Aikasarja-analyysi, 10 op

Arto Luoma
Tilastotiede
Informaatiotieteiden yksikkö
33014 TAMPEREEN YLIOPISTO

Syksy 2013

Sisältö

Kirjallisuutta

Pääasialliset lähteet

- Brockwell, Davis: Introduction to Time Series and Forecasting
- Tsay: Analysis of Financial Time Series, Third Edition

Muuta kirjallisuutta

- Cowpertwait, Metcalfe: Introductory Time Series with R (Use R!)
- Pfaff: Analysis of Integrated and Cointegrated Time Series with R (Use R!)
- Brockwell, Davis: Time Series: Theory and Methods (lyh. TSTM).
- Shumway, Stoffer: Time Series Analysis and Its Applications: With R Examples
- Cryer, Chan: Time Series Analysis With Applications in R

Luku 1

Johdanto

1.1 Yleistä

Aikasarja syntyy, kun jotain suuretta mitataan peräkkäisinä ajankohtina, esim. seurattaessa talouden kehitystä, teollisuuden prosessia tai vaikkapa sään kehitystä. Aikasarjojen matemaattinen malli on stokastinen prosessi $\{X_t,\ t \in T\}$, missä indeksijoukko T voi olla diskreetti (esim. $\mathbb{Z} = \{0, 1, 2, ...\}$) tai jatkuva (esim. $[0, \infty)$). Tällä kurssilla rajoitutaan diskreetteihin aikasarjoihin.

Yleensä aikasarjoista on saatavilla yksi realisaatio tietyllä aikavälillä, esim. $x_1, x_2, ..., x_n$. Aikasarjaa analysoimalla pyritään löytämään stokastinen prosessi, joka voisi tuottaa kyseisen sarjan. Tämä voisi auttaa 1) ymmärtämään jotain tutkittavan ilmiön luonteesta ja 2) ennustamaan sarjan tulevia arvoja. Aikasarja-analyysiä sovelletaan myös 3) prosessien kontrolloinnissa esim. teollisuudessa, kun säädettävissä oleva prosessi tuottaa jonon havaintoarvoja. 4) Yhden sarjan vaihtelua voidaan pyrkiä selittämään muiden sarjojen vaihtelulla, jolloin sivutaan samoja ongelmia kuin regressioanalyysissä.

1.2 Stationaarinen prosessi

Yleensä aikasarja pyritään erilaisilla muunnoksilla saattamaan sellaiseen muotoon, että sen voi katsoa olevan realisaatio stationaarisesta prosessista. Stationaarista prosessia voidaan yleensä mallintaa tilastollisesti ns. ARMA-prosessin avulla. Alkuperäinen aikasarja voidaan sitten kuvata käyttäen sta-

tionaarista aikasarjaa ja niitä muunnoksia, joilla siitä saadaan alkuperäinen sarja.

Prosessin $\{X_t\}$ sanotaan olevan vahvasti stationaarinen (strictly stationary), jos peräkkäisten havaintojen yhteisjakauma ei muutu, kun siirrytään ajassa eteenpäin. Toisin sanoen prosessi on vahvasti stationaarinen, jos satunnaisvektorilla $(X_{1+h}, X_{2+h}, ..., X_{n+h})$ on sama jakauma kuin vektorilla $(X_1, X_2, ..., X_n)$, missä h ja n ovat mitä tahansa positiivisia kokonaislukuja.

Kun puhutaan stationaarisuudesta, tarkoitetaan kuitenkin yleensä ns. heikkoa stationaarisuutta. Prosessin $\{X_t\}$ sanotaan olevan heikosti stationaarinen t. stationaarinen laajassa mielessä (weakly stationary, stationary in the wide sense), jos prosessin odotusarvo eikä peräkkäisten havaintojen kovarianssimatriisi muutu siirryttäessä ajassa eteenpäin. Tällöin oletetaan, että on äärellinen kaikissa aikapisteissä t, jotta odotusarvo ja kovarianssit olisivat määriteltyjä. Huomaa, että vahvasta stationaarisuudesta seuraa heikko stationaarisuus, mikäli on äärellinen, mutta heikosta stationaarisuudesta ei välttämättä seuraa vahva stationaarisuus.

Määritellään odotusarvofunktio $\mu_X(t) = \mathsf{E}(X_t)$ ja autokovarianssifunktio $\gamma_X(r, s) = \mathsf{Cov}(X_r, X_s)$ kaikille kokonaisluvuille r ja s. Näiden funktioiden avulla määriteltynä prosessi on (heikosti) stationaarinen, jos odotusarvofunktio $\mu_X(t)$ on ajasta t riippumaton vakio ja kovarianssifunktio $\gamma_X(t + h, t)$ ei riipu ajasta t millään kokonaisluvulla h. Jos kyseessä on stationaarinen prosessi, kovarianssifunktio voidaan ilmaista yhden argumentin avulla: $\gamma_X(t + h, t) = \gamma_X(h, 0) = \gamma_X(h)$. Lisäksi stationaariselle prosessille voidaan määritellä autokorrelaatiofunktio $\rho_X(h) = \gamma_X(h)/\gamma_X(0) = \mathsf{Cor}(X_{t+h}, X_t)$.

Esim. 1. Olkoon $\{X_t,\ t = 1, 2, ..\}$ jono riippumattomia ja samoin jakautuneita satunnaismuuttujia, joille $\mathsf{E}(X_t) = \mu$ ja $\mathsf{Var}(X_t) = \sigma^2$. Tarkastellaan satunnaisprosessia $\{Y_t,\ t = 1, 2, ...\}$, missä $Y_0 = 0$ ja $Y_t = X_1 + X_2 + ... + X_t$, kun $t \geq 1$. Prosessia kutsutaan satunnaiskävelyksi. Kyseessä on symmetrinen satunnaiskävely, jos $\mu = 0$ ja yksinkertainen satunnaiskävely, jos X_t voi saada arvokseen joko 1 tai -1. Prosessin odotusarvofunktio on $\mu_Y(t) = \mathsf{E}(X_1 + X_2 + ... + X_t) = t\mu$ ja varianssi $\mathsf{Var}(Y_t) = \mathsf{Var}(X_1) + \mathsf{Var}(X_2) + ... + \mathsf{Var}(X_t) = t\sigma^2$. Koska varianssi kasvaa, kun t kasvaa, prosessi ei ole stationaarinen vaikka μ olisi 0. Alkuperäinen sarja $\{X_t\}$ saadaan sarjasta $\{Y_t\}$ differoimalla: $X_t = Y_t - Y_{t-1}$.

Esim. 2. *IID-kohina.* Olkoon $\{X_t\}$ jono riippumattomia ja samoin jakautuneita satunnaismuuttujia, joiden odotusarvo on 0. Tällöin jonoa sanotaan IID-kohinaksi (IID=independent identically distributed). Prosessi on ehkä

yksinkertaisin esimerkki vahvasta stationaarisuudesta. Koska satunnaismuuttujat ovat riippumattomia ja samoin jakautuneita,

$$\begin{aligned}
&\Pr(X_1 \leq x_1, X_2 \leq x_2, .., X_n \leq x_n) \\
=&\Pr(X_1 \leq x_1) \Pr(X_2 \leq x_2) ... \Pr(X_n \leq x_n) \\
=&\Pr(X_{1+h} \leq x_1) \Pr(X_{2+h} \leq x_2) ... \Pr(X_{n+h} \leq x_n) \\
=&\Pr(X_{1+h} \leq x_1, X_{2+h} \leq x_2, ..., X_{n+h} \leq x_n),
\end{aligned}$$

josta nähdään, että peräkkäisten havaintojen yhteisjakauma pysyy samana ajassa siirryttäessä. Merkitään IID-kohinaa $\{X_t\} \sim \text{IID}(0, \sigma^2)$.

Esim. 3. *Valkoinen kohina.* Olkoon $\{X_t\}$ jono korreloimattomia satunnaismuuttujia, joilla on odotusarvo 0 ja varianssi σ^2. Silloin jonoa sanotaan valkoiseksi kohinaksi ja merkitään $\{X_t\} \sim \text{WN}(0, \sigma^2)$. Jono on esimerkki (heikosti) stationaarisesta prosessista sillä määritelmän perusteella $\mu_X(t) = \mathsf{E}X_t = 0$ ja

$$\gamma_X(t+h, t) = \begin{cases} \sigma^2, & \text{kun } h = 0, \\ 0, & \text{kun } h \neq 0. \end{cases}$$

Odotusarvo- ja autokovarianssifunktiot eivät siis riipu t:stä. IID-kohina on myös valkoista kohinaa, jos prosessin varianssi on äärellinen. Valkoinen kohina ei sen sijaan ole välttämättä IID-kohinaa.

Esim. 4. *MA(1)-prosessi.* Oletetaan, että jono $\{Z_t,\ t = 0, 1, 2, ...\}$ on valkoista kohinaa. Määritellään $X_t = Z_t + \theta Z_{t-1}$, missä θ on reaaliluku. Tällöin $\mu_X(t) = 0$ ja $\mathsf{E}X_t^2 = \sigma^2(1+\theta^2) < \infty$. Lisäksi

$$\gamma_X(t+h, t) = \begin{cases} \sigma^2(1+\theta^2), & \text{kun } h = 0, \\ \sigma^2\theta, & \text{kun } h \pm 1, \\ 0, & \text{kun } |h| > 1. \end{cases}$$

Koska $\mu_X(t)$ ja $\gamma_X(t+h, t)$ eivät riipu t:stä, kyseessä on stationaarinen prosessi. Autokorrelaatiofunktio prosessille $\{X_t\}$ on

$$\rho_X(t+h, t) = \begin{cases} 1, & \text{kun } h = 0, \\ \theta/(1+\theta^2), & \text{kun } h \pm 1, \\ 0, & \text{kun } |h| > 1. \end{cases}$$

Esim 5. *AR(1)-prosessi.* Oletetaan että $\{X_t\}$ on stationaarinen aikasarja, joka toteuttaa yhtälöt

$$X_t = \phi X_{t-1} + Z_t, t = 0, \pm 1, ..., \tag{1.1}$$

missä $\{Z_t\} \sim \text{WN}\ (0, \sigma^2)$, $|\phi| < 1$ ja Z_t on korreloimaton satunnaismuuttujien X_s kanssa, kun $s < t$. Tällöin sanotaan, että $\{X_t\}$ noudattaa autoregressiivistä prosessia viiveellä 1 (eli AR(1)-prosessia). Ottamalla odotusarvo yhtälön (1) eri puolista saadaan prosessin odotusarvoksi $\mathsf{E}X_t = 0$. Autokovarianssifunktion määrittämiseksi kerrotaan yhtälö puolittain X_{t-h}:lla, missä $h > 0$, ja otetaan odotusarvo puolittain:

$$\begin{aligned} & \mathsf{E}(X_t X_{t-h}) = \phi \mathsf{E}(X_{t-1} X_{t-h}) + \mathsf{E}(Z_t X_{t-h}) \\ \Leftrightarrow \quad & \mathsf{Cov}(X_t, X_{t-h}) = \phi \mathsf{Cov}(X_{t-1}, X_{t-h}) + 0 \\ \Leftrightarrow \quad & \gamma_X(h) = \phi \gamma_X(h-1). \end{aligned}$$

Rekursiivisesti voidaan päätellä, että $\gamma_X(h) = \phi^h \gamma_X(0)$. Koska $\gamma_X(h) = \mathsf{Cov}(X_{t+h}, X_t) = \mathsf{Cov}(X_t, X_{t+h}) = \gamma_X(-h)$, positiivisille ja negatiivisille viiveille h soveltuva autokovarianssin kaava on $\gamma_X(h) = \phi^{|h|} \gamma_X(0)$. Autokorrelaatio puolestaan on $\rho_X(h) = \gamma_X(h)/\gamma_X(0) = \phi^{|h|}, h = 0, \pm 1, ...$ Voidaksemme määrittää, mikä on $\gamma_X(0)$, todetaan ensin että

$$\begin{aligned} \mathsf{Cov}(X_t, Z_t) &= \mathsf{Cov}(\phi X_{t-1} + Z_t, Z_t) = \phi \mathsf{Cov}(X_{t-1}, Z_t) + \mathsf{Cov}(Z_t, Z_t) \\ &= \mathsf{Cov}(Z_t, Z_t) = \sigma^2, \end{aligned}$$

sillä oletuksen mukaan $\mathsf{Cov}(X_{t-1}, Z_t) = 0$. Täten

$$\gamma_X(0) = \mathsf{Cov}(X_t, X_t) = \mathsf{Cov}(X_t, \phi X_{t-1} + Z_t) = \phi \gamma_X(1) + \sigma^2 = \phi^2 \gamma_X(0) + \sigma^2,$$

josta saadaan ratkaistua, että $\gamma_X(0) = \sigma^2/(1 - \phi^2)$.

1.3 Otosautokorrelaatiofunktio

Yleensä aikasarjan autokovarianssi ja -korrelaatiofunktioita ei tunneta, vaan ne joudutaan estimoimaan otoksen perusteella. Oletetaan, että meillä on otos $x_1, x_2, ..., x_n$ jostain aikasarjan realisaatiosta. Tällöin määritellään *otoskeskiarvo*

$$\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t,$$

otosautokovarianssifunktio

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}), \quad -n < h < n,$$

ja *otosautokorrelaatiofunktio*

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n.$$

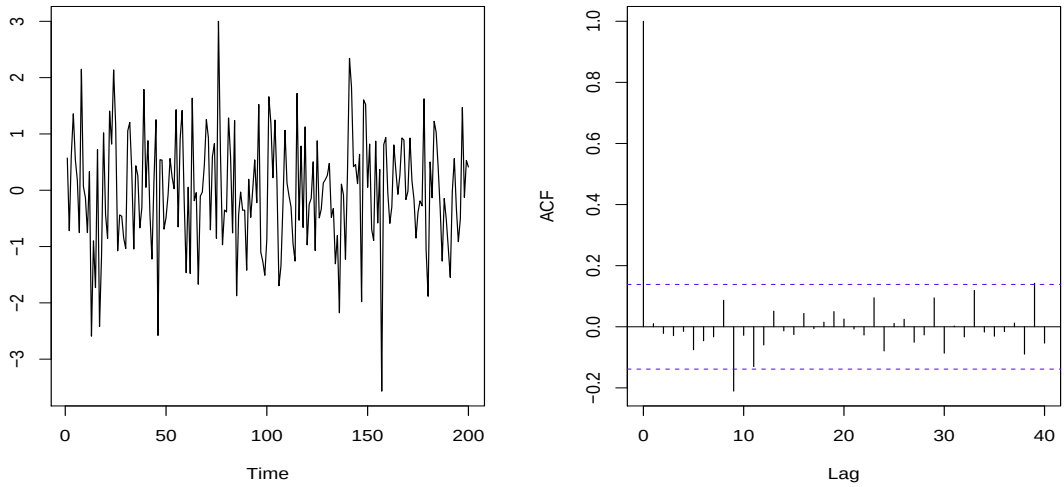
Huomaa, että kyseessä ei ole tavallinen havaintopareista $(x_t, x_{t+|h|})$ laskettu otoskovarianssi. Annettu määritelmä kuitenkin takaa sen otoskovarianssi- ja otoskorrelaatiomatriisi ovat ei-negatiivisesti definiittejä. Teoreettiset kovarianssi- ja korrelaatiomatriisit ovat aina ei-negatiivisesti definiittejä. Lisäksi ei-negatiivisesti definiittisyys takaa eräiden matriisihajotelmien olemassaolon.

Otosautokorrelaatiofunktiosta voidaan tehdä päätelmiä aikasarjan generoineen satunnaisprosessin autokorrelaatiofunktiosta, mikä voi auttaa sopivan mallin löytämisessä. Valkoisen kohinan tapauksessa $\rho(h) = 0$, kun $h \neq 0$. Tällöin voisi olettaa, että valkoisen kohinan prosessin tuottaman aikasarjan otosautokorrelaatiofunktio olisi lähellä nollaa, kun $h \neq 0$. Itse asiassa voidaan osoittaa (TSTM, s. 222), että IID-kohinan tapauksessa, kun prosessin varianssi on äärellinen, otosautokorrelaatiot $\hat{\rho}(h),\ h > 0$, ovat riippumattomia ja noudattavat normaalijakaumaa $N(0, 1/n)$, kun n on suuri. Täten noin 95% otosautokorrelaatioista pitäisi osua rajojen $\pm 1.96/\sqrt{n}$ sisälle, jos kysessä on IID-kohinan prosessi.

Kuviossa 1.1 on simuloitu 200 arvoa IID N(0,1) -kohinaa. Kuvion b-osassa on piirretty otosautokorrelaatiofunktio simuloidun aineiston perusteella. Nähdään että kaikki arvot sijoittuvat rajojen $\pm 1.96/\sqrt{n}$ sisälle. Kuviot on tuotettu R-käskyillä

```
whitenoise<-rnorm(200)
plot(ts(whitenoise),main="",ylab="")
acf(whitenoise,main="",lag.max=40)
```

Tutkitaan seuraavaksi hypoteesia, että logaritmoitu osakekurssi olisi satunnaiskävelyprosessi. Kuviossa 1.2 näkyy Nokian osakekurssi ajalta 4.9.2012 – 3.9.2013 sekä muunnettu sarja, joka on saatu logaritmoimalla ja differoimalla alkuperäinen sarja. Viimeisen päivän kurssinousu on seuraus ilmoituksesta Nokian matkapuhelinliiketoiminnan myymisestä Microsoftille. Autokorrelaatiofunktion perusteella muunnettu sarja näyttää korreloimattomalta prosessilta, niin kuin pitäisikin. Koska lisäksi sarjan varianssi näyttää pysyvän likimain vakiona ja sen odotusarvo ei eroa tilastollisesti merkittävästi nollasta, sitä voidaan pitää valkoisen kohinan prosessina. Hajonnan σ estimaatti on 0.03821.

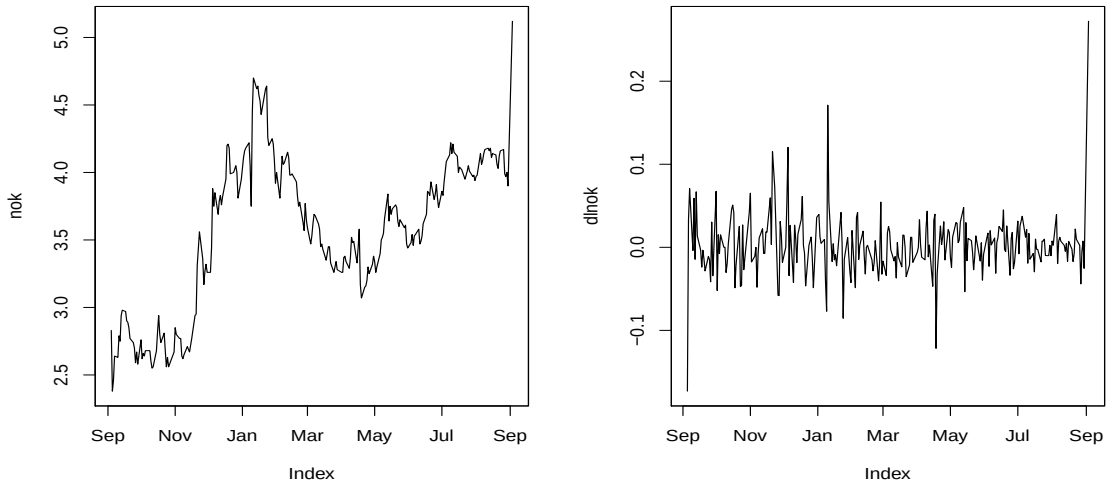


(a) Aikasarja. (b) Otosautokorrelaatiofunktio.

Kuvio 1.1: Valkoisen kohinan simulointia.

```
library(tseries)
nok <- get.hist.quote(instrument="nok",start="2012-9-1", end="2013-9-3",
        quote="Close")
dlnok <- diff(log(nok))

x <- coredata(dlnok)
acf(x,main="")

summary(lm(dlnok~  1))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.002381   0.002422   0.983    0.326

Residual standard error: 0.03821 on 248 degrees of freedom
```

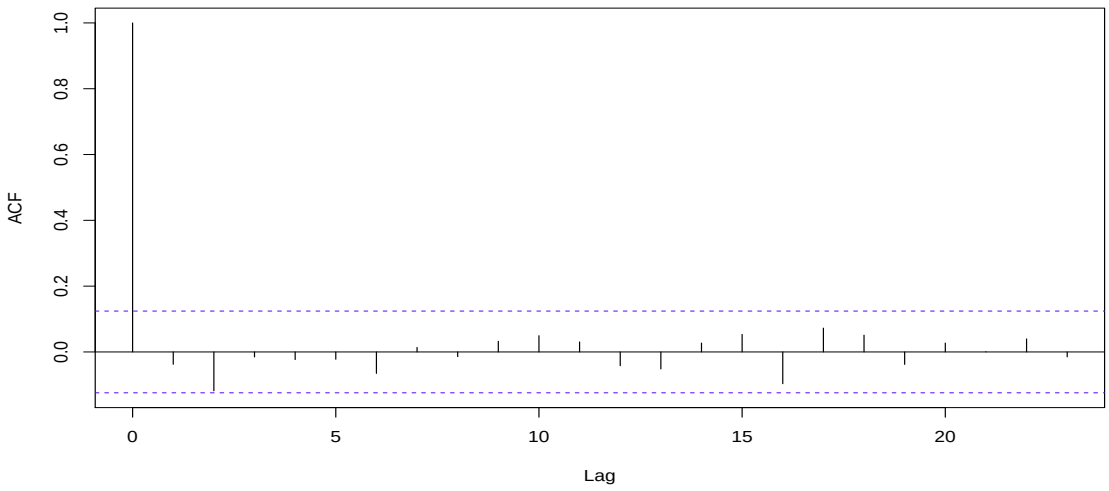
Kuviossa 1.4 on piirrettynä Lake Huron -järven pinnan tasoa osoittava kuvaaja. Selvästikin aikasarja ei ole stationaarinen, sillä siinä esiintyy



(a) Aikasarja

(b) Aikasarja logaritmoituna ja differoituna

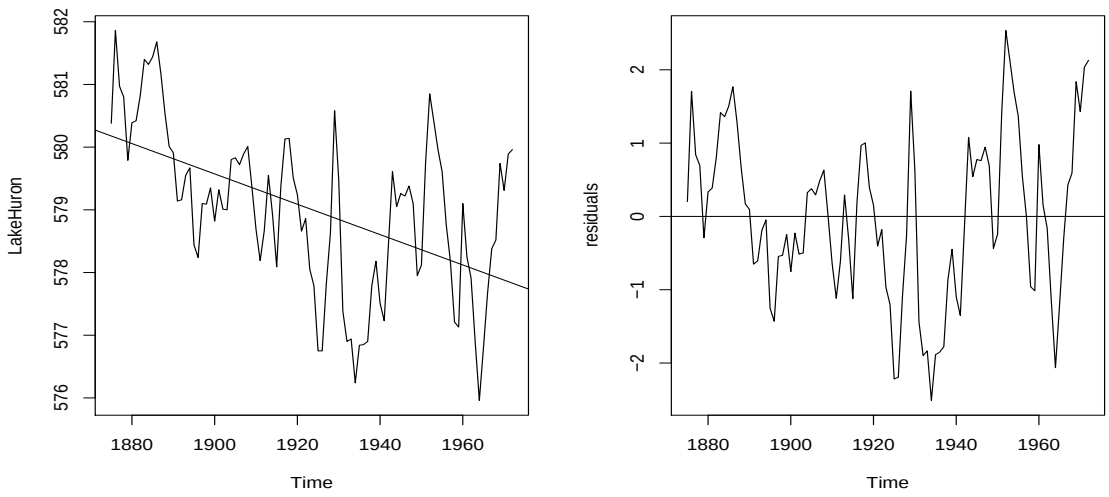
Kuvio 1.2: Nokian osakekurssi 4.9.2012 – 3.9.2013



Kuvio 1.3: Nokian osaketuottosarjan (1.2 b) otosautokorrelaatiofunktio

laskeva trendi. Kun estimoidaan lineaarinen trendi tavallisella pienimmän neliösumman menetelmällä (OLS), havaitaan jäännössarjassa positiivista autokorrelaatiota pienillä viiveillä. Positiivinen autokorrelaatio näkyy myös tarkasteltaessa otosautokorrelaatiofunktiota. Koska otosautokorrelaatiofunktio vähenee suurin piirtein geometrisesti, voidaan ehdottaa jäännössarjan generoivaksi prosessiksi AR(1)-prosessia parametrina $\phi \approx 0.8$. Teoreettinen autokorrelaatiofunktiohan AR(1)-prosessille on $\rho(h) = \phi^{|h|}$.

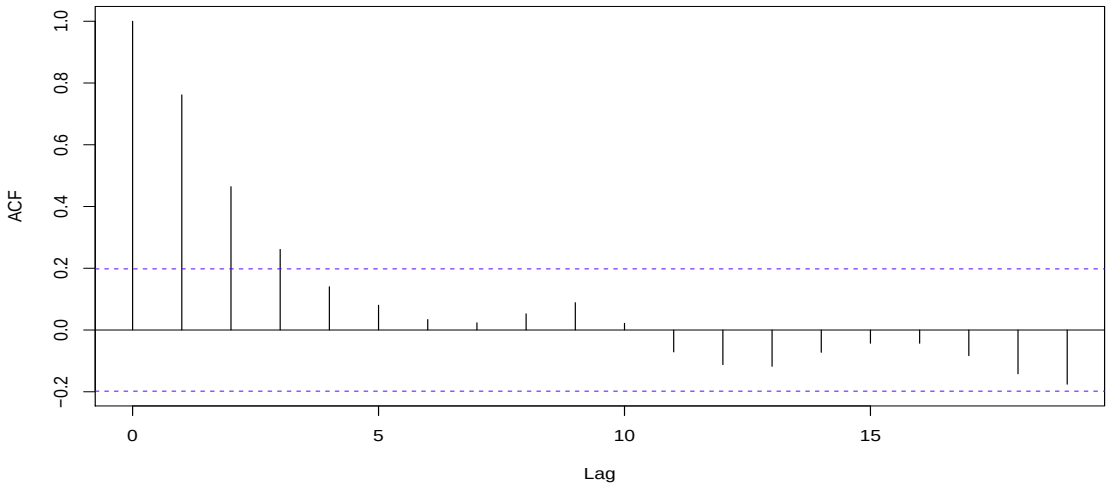
```
plot(LakeHuron)
a <- lm(LakeHuron~  time(LakeHuron))
abline(a)

resid <- ts(residuals(a),start=start(LakeHuron),
            frequency=frequency(LakeHuron))
plot(resid,ylab="residuals")
abline(0,0)

acf(residuals(a),main="")
```



(a) Aikasarja

(b) Jäännössarja

Kuvio 1.4: Lake Huron -järven pinta



Kuvio 1.5: Lake Huron -järven pinnan jäännössarjan otosautokorrelaatiofunktio

Otosautokorrelaatio voidaan laskea vaikka sarja ei olisikaan stationaarinen, jolloin siitä voidaan mahdollisesti havaita ei-stationaarisuus. Esimerkiksi, jos sarjassa on lineaarinen trendi, otosautokorrelaatiofunktio pysyy pitkään positiivisena, kun h kasvaa. Tämä johtuu siitä, että sarjan alkupää on keskiarvon toisella puolella ja loppupää toisella puolella, jolloin otosautokovarianssin lausekkeessa tekijät $x_t - \bar{x}$ ja $x_{t+|h|} - \bar{x}$ ovat samanmerkkisiä eivätkä niiden tulot kumoa toisiaan. Symmetrinen satunnaiskävely ei ole stationaarinen prosessi, mutta se muistuttaa AR(1)-prosessia, kun ϕ on lähellä ykköstä. Tässä tapauksessa $\hat{\rho}(h)$ vähenee hitaasti ykkösestä, kun $|h|$ kasvaa. Lisäksi jos sarjassa on säännöllistä jaksollista vaihtelua, myös otosautokorrelaatiofunktiossa on havaittavissa samanlaista vaihtelua samalla jaksolla.

1.4 Klassinen hajotelma

Klassisessa kausihajotelmamallissa (the classical decomposition model) oletetaan, että prosessi voidaan jakaa kolmeen komponenttiin, nimittäin hitaasti vaihtelevaan trendiin, kausikomponenttiin ja stationaarisen kohinan sarjaan.

Hajotelma voidaan esittää muodossa

$$X_t = m_t + s_t + Y_t,$$

missä m_t edustaa trendiä, s_t kausikomponenttia ja Y_t stationaarista kohinaa. Mallissa oletetaan, että $\mathsf{E}(Y_t) = 0$, $s_t = s_{t+d}$ ja $\sum_{j=1}^{d} s_j = 0$, missä d on kausien lukumäärä. Trendi m_t voidaan erottaa sarjasta käyttämällä liukuvaa keskiarvoa, joka on erikoistapaus ns. lineaarisesta suotimesta. Trendin estimaatiksi saadaan

$$\hat{m}_t = (x_{t-q} + x_{t-q+1} + ... + x_{t+q})/d, \quad q < t \leq n - q,$$

kun kausien lukumäärä on pariton $d = 2q + 1$ ja

$$\hat{m}_t = (0.5x_{t-q} + x_{t-q+1} + ... + x_{t+q-1} + 0.5x_{t+q})/d, \quad q < t \leq n - q,$$

kun kausien lukumäärä on parillinen $d = 2q$. Tällä suotimella saadaan kausikomponenttien vaikutus häviämään ja kohinan vaikutus hyvin pieneksi.

Kausikomponentin estimoimiseksi lasketaan kullekin kaudelle k, $k = 1, ..., d$ keskiarvo w_k poikkeamista $x_{k+jd} - \hat{m}_{k+jd}$, $q < k + jd \leq n - q$. Koska näiden kausikeskiarvojen summa ei välttämättä ole 0, muodostetaan kausitermille s_k estimaatti

$$\hat{s}_k = w_k - \frac{1}{d}\sum_{i=1}^{d} w_i, \ k = 1, ..., d.$$

Kausipuhdistettu (deseasonalised) sarja saadaan vähentämällä alkuperäisestä sarjasta kausikomponentti:

$$d_t = x_t - \hat{s}_t, \ t = 1, ..., n.$$

Lopuksi kausipuhdistetusta sarjasta voidaan erottaa trendi esimerkiksi sovittamalla polynomifunktio pienimmän neliösumman menetelmällä. Trendille siis muodostetaan uusi estimaatti $\hat{m}_t$. Kohinasarjan Y_t estimaatti on tällöin jäännössarja

$$\hat{Y}_t = x_t - \hat{m}_t - \hat{s}_t.$$

Lukuun ottamatta viimeistä vaihetta, deterministisen trendifunktion sovitusta, klassinen hajotelma voidaan tehdä R-funktiolla "decompose".

1.5 Kausivaihtelun mallintaminen käyttäen lineaarista regressiota

Yksi tapa mallintaa kausivaihtelua voidaan mallintaa on ns. harmoninen regressio. Siinä aikasarjaa selitetään eri vaiheessa olevilla ja eri taajuuksia edustavilla sinifunktioilla. Jos yhden jakson pituus (kausien lkm) on d, kausikomponentti s_t voidaan esittää muodossa

$$s_t = a_0 + \sum_{j=1}^{k} \left(a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)\right),$$

missä taajuudet λ_j ovat lausekkeen $2\pi/d$ kokonaisia monikertoja. Tuntemattomat kertoimet a_0 ja $a_j,\ b_j,\ j = 1, ..., k$, voidaan estimoida aineistosta tavallisella pienimmän neliösumman menetelmällä. Systemaattisempi menetelmä esitetään spektrianalyysin yhteydessä.

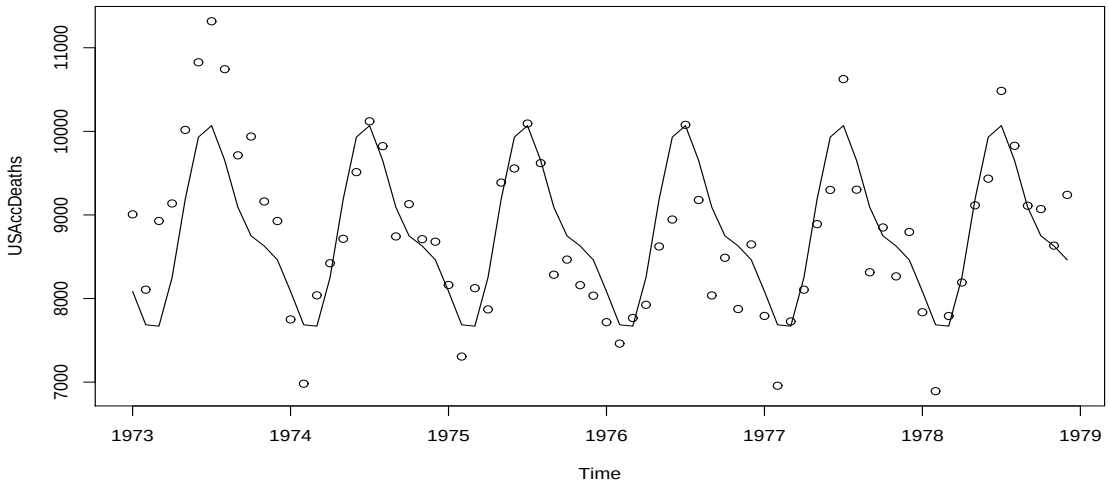
Kuviossa 1.6 on USAccDeaths-aineistoon sovellettu harmonista regressiota. Siinä on käytetty ainoastaan kahta matalataajuisinta sinifunktiota, jolloin $k = 2,\ \lambda_1 = 2\pi/d, \lambda_2 = 4\pi/d$. Kuvio voidaan tuottaa R-käskyillä

```
t <- time(USAccDeaths)
plot(USAccDeaths,lty=0,type="b")
a <- lm(USAccDeaths~  sin(2*pi*t)+cos(2*pi*t)+sin(4*pi*t)+cos(4*pi*t))
lines(ts(fitted(a),start=1973,frequency=12))
```

Jos trendiä mallinnetaan parametrien suhteen lineaarisella funktiolla, kuten polynomilla, kausikomponentti ja trendi voidaan estimoida yleistetyllä pienimmän neliösumman menetelmällä (GLS). Tavallista PNS-menetelmää (OLS) käytetään, jos voidaan olettaa, että jäännössarja on valkoisen kohinan prosessi. Malli, jossa trendi on toisen asteen polynomi ja johon sisältyy kausikomponentti, voidaan esittää muodossa

$$X_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \gamma_1 u_{t1} + ... + \gamma_{d-1} u_{t,d-1} + Y_t,$$

missä $u_{tj},\ j = 1, ..., d-1$, on osoitinmuuttuja, joka saa arvon 1, jos ajanhetkellä t on menossa kausi j, ja arvon 0 muuten. Jos mallissa olisi mukana kauden d indikaattorimuuttuja, parametrit eivät olisi estimoituvia. Kausikomponentti voidaan laskea kaavalla $s_j = \gamma_j - (\gamma_1 + ... + \gamma_{d-1})/d$, kun $j = 1, ..., d-1$, ja $s_d = -(\gamma_1 + ... + \gamma_{d-1})/d$.



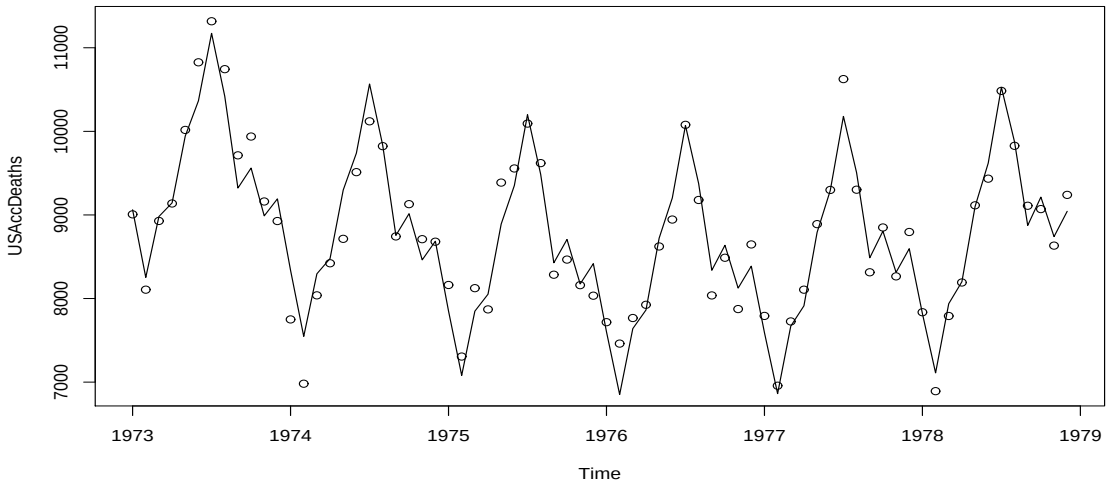
Kuvio 1.6: Harmonisen regression sovittaminen USAccDeaths-aineistoon

Alla on malli estimoitu OLS-menetelmällä käyttäen funktiota lm ja USAccDeaths-aineistoa. Osoitinmuuttujien matriisi $\mathbf{U}$ on muodostettu käyttämällä Kroneckerin tuloa $\mathbf{1}_6 \otimes \mathbf{I}_{12}$, missä $\mathbf{1}_6$ on 6-ulotteinen ykkösvektori ja $\mathbf{I}_{12}$ on 12×12-identiteettimatriisi. Tulomatriisissa on 6 identiteettimatriisia alakkain. Sovituksen tulos on nähtävissä kuviossa 1.7.

```
t <- time(USAccDeaths)
U <- kronecker(rep(1,6),diag(12))
a <- lm(USAccDeaths~  t+I(t^2)+U[,-12])
plot(USAccDeaths,lty=0,type="b")
lines(ts(fitted(a),start=1973,frequency=12))
```

1.6 Muita signaalinhajoituksen menetelmiä

Edellä kuvatuissa menetelmissä oletetaan, että kausikomponentti pysyy vakiona vuodesta toiseen. Kehittyneemmissä menetelmissä sen sallitaan vaihdella. Näissä menetelmissä on usein myös esipuhdistusosio, joka etsii aikasarjasta poikkeavat havainnot, luokittelee ne sekä poistaa niiden vaikutuksen. Esipuhdistusvaiheessa voidaan ottaa huomioon myös kauppa- tai



Kuvio 1.7: Neliöllisen trendin ja kausikomponentin sovittaminen USAccDeaths-aineistoon

työpäivien ja lomapäivien vaikutus. Varsinaisen komponentteihin jaon jälkeen jäännössarjan puhtaus voidaan tarkistaa diagnostisin testein.

Muista menetelmistä, jotka hajoittavat aikasarjan komponentteihin, tunnetuimpia on Yhdysvaltain tilastokeskuksen Census Bureaun kehittämä Census II. Menetelmän uusin variaatio on X-12-REGARIMA. Menetelmän esipuhdistusosio on REGARIMA. Ohjelmiston toinen osa X-12 jakaa aikasarjan komponentteihin ja se perustuu eri pituisten liukuvan keskiarvon suodinten peräkkäiseen soveltamiseen. Jotta liukuvan keskiarvon suotimia voitaisiin soveltaa sarjan molemmissa päissä, sarjaa ennustetaan eteen- ja taaksepäin sopivalla ARIMA-mallilla.

Toinen tunnettu menetelmä on Espanjan pankin kehittämä TRAMO/SEATS. Siinä TRAMO huolehtii esipuhdistuksesta ja SEATS komponentteihin jaosta. Menetelmää sanotaan mallipohjaiseksi, sillä se perustuu aikasarjan mallintamiseen ARIMA-prosessilla. (SEATS = Signal Extraction in ARIMA Time Series).

Kolmas tunnettu menetelmä on STL (A Seasonal-Trend decomposition based on Loess). Menetelmässä sekä trendi- että kausikomponentti tasoitetaan ns. lössimenetelmällä (loess). Lössimenetelmä perustuu paikalliseen reg-

ressioon (local regression), joka sovittaa aikasarjaan tasaisen käyrän. Lössimenetelmä poikkeaa paikallisesta regressiosta siinä, että siinä poikkeavien havaintojen vaikutusta pyritään vähentämään.

STL-menetelmä sisältyy R-ohjelmaan. Tasoitusta voidaan säätää parametreilla s.window ja t.window. Parametri s.window säätää kausi-ikkunan leveyttä. Mitä suuremmaksi parametrin arvo valitaan, sitä hitaammin kausikomponentti muuttuu. Jos parametrin arvoksi valitaan "per", tai "periodic", kausikomponentti ei vaihtele vuodesta toiseen. Parametri t.window määrää trendi-ikkunan leveyden. Mitä suurempi parametri on, sitä tasaisempi estimoitu trendikäyrä on ja sitä hitaammin se reagoi sarjassa tapahtuviin muutoksiin.

TRAMO/SEATS -ohjelmaa on esitellään esim. kirjoissa Maravall, A. (1995): Unobserved Components in Economic Time Series. The Handbook of Applied Econometrics, s.12-72, ja Maravall, Gomez (2001): Seasonal Adjustment and Signal Extraction in Economic Time Series. A Course in Time Series Analysis. X-12-REGARIMA ja STL -menetelmiä kuvataan esim. kirjassa Makridakis, Wheelwright, Hyndman: Forecasting: Methods and applications.

1.7 Lineaariset suotimet

Aiemmin esiteltiin liukuva keskiarvo esimerkkinä ns. lineaarisesta suotimesta. Yleisemmin voidaan määritellä lineaarinen suodin

$$m_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}.$$

Liukuvan keskiarvon tapauksessa $a_j = 1/(2q+1)$, kun $-q \leq j \leq q$, ja $a_j = 0$ muulloin. Liukuva keskiarvo on tyypillinen alipäästösuodin, joka suodattaa aikasarjasta pois nopeasti vaihtelevat komponentit (korkeat taajuudet). Liukuva keskiarvo ei vaikuta lineaariseen trendiin mitenkään, sillä sovellettaessa liukuvaa keskiarvoa sarjaan $X_t = a + bt + Y_t$ saadaan

$$\sum_{j=-q}^{q} \frac{1}{2q+1}\left(a + b(t-j) + Y_{t-j}\right) = a + bt + \sum_{j=-q}^{q} \frac{1}{2q+1} Y_{t-j}.$$

Yleensä mitä suurempi q valitaan, sen suurempi tasoitus saadaan aikaan. Jos q valitaan liian suureksi, trendin estimaatista saattaa tulla huono, ellei

trendi ole lineaarinen. Voidaan myös suunnitella suotimia, jotka säilyttävät useampiasteiset polynomifunktiot koskemattomina.

Liukuvan keskiarvon suotimia voi olla useampia peräkkäin. Esimerkiksi 3×5 MA -suodin toimii niin, että ensin lasketaan 3 havainnon liukuva keskiarvo ja tämän jälkeen 5 havainnon keskiarvo.

R-ohjelman stats-pakettiin funktio filter, jolla voidaan tehdä lineaarisia suodatuksia. Suotimet voivat olla tyypiltään konvoluutiosuotimia tai rekursiivisia suotimia. Seuraavassa on annettu esimerkkejä erityyppisistä suotimista ja siitä, miten suodatukset voidaan toteuttaa.

Yksipuolinen konvoluutiosuodin:

$$y_t = a_0 x_t + a_1 x_{t-1} + ... + a_p x_{t-p}$$

R-toteutus:

a <- c(a_0,a_1,...,a_p) y <- filter(x,a,sides=1)

Kaksipuoleinen konvoluutiosuodin:

$$y_t = a_{-p} x_{t+p} + a_{-p+1} x_{t+p-1} + ... + a_{p-1} x_{t-p+1} + a_p x_{t-p}$$

a <- c(a_{-p},a_{-p+1},...,a_{p-1},a_p) y <- filter(x,a,sides=2)

Rekursiivinen suodin:

$$y_t = x_t + a_1 y_{t-1} + ... + a_p y_{t-p}$$

a <- c(a_1,...,a_p) y <- filter(x, a, method="recursive")

Alapuolella on esitetty R-kielinen funktio maverage, joka laskee liukuvan keskiarvon. Funktiolle annetaan argumenttina suodatettava sarja ja q, missä q ilmaisee, montako edeltävää ja seuraavaa havaintoa sisällytetään keskiarvoon. Jotta liukuva keskiarvo voitaisiin laskea myös sarjan alussa ja lopussa, sarjan alkuun lisätään q kertaa ensimmäinen havainto x_1 ja loppuun q kertaa viimeinen havainto x_n.

```
maverage<-function(tseries,q)
{
  n <- length(tseries)
  y <- c(rep(tseries[1],q),tseries,rep(tseries[n],q))
  filt <- rep(1,2*q+1)/(2*q+1)
  y <- filter(y,filt,sides=2)
  y[(q+1):(q+n)]
}
```

Trendi voidaan estimoida käyttäen myös ns. eksponentiaalista tasoitusta. Se määritellään rekursiivisesti seuraavasti:

$$\hat{m}_t = aX_t + (1-a)\hat{m}_{t-1}, \quad t = 2, ..., n,$$

ja

$$\hat{m}_1 = X_1,$$

missä a on reaaliluku väliltä (0,1). Kun a on lähellä ykköstä saadaan pieni tasoitus ja kun se on lähellä 0:aa, tasoitus on suuri. Suodin voidaan ilmaista myös muodossa

$$\hat{m}_t = \sum_{j=0}^{t-2} a(1-a)^j X_{t-j} + (1-a)^{t-1} X_1,$$

kun $t \geq 2$, jolloin nähdään, että kyseessä on painotettu liukuva keskiarvo. Koska painokertoimet vähenevät eksponentiaalisesti, puhutaan eksponentiaalisesta tasoituksesta.

Funktiolla expsmooth voidaan tehdä eksponentiaalista tasoitusta.

```
expsmooth <- function(tseries, a=0.5)
{
  tseries[1] <- tseries[1]/a
  a*filter(tseries,1-a,method="recursive")
}
```

1.8 Aikasarjojen saattaminen stationaariseksi differoimalla

Edellä on esitetty menetelmiä aikasarjan trendin ja mahdollisen kausikomponentin estimoimiseksi. Kun alkuperäisestä sarjasta vähennetään trendi

ja kausikomponentti, tuloksena voidaan saada stationaarinen jäännössarja. Toinen menetelmä sarjan saattamiseksi stationaariseksi on differointi. Kun suoritetaan differointi viiveellä 1, kustakin aikasarjan havainnosta vähennetään edellinen havainto. Yhdellä askelella viivästettyä havaintoa merkitään operaattorilla B ja differointia viiveellä yksi merkitään operaattorilla ∇. Jos alkuperäinen sarja on $\{X_t\}$, viiveellä 1 differoitu sarja on $\{Y_t\}$, missä

$$Y_t = X_t - X_{t-1} = X_t - \mathsf{B}X_t = (1 - \mathsf{B})X_t = \nabla X_t. \tag{1.2}$$

Differointi viiveellä 1 hävittää sarjasta lineaarisen trendikomponentin. Jos sarja voidaan esittää muodossa $X_t = a + bt + Z_t$, missä $\mathsf{E}(Z_t) = 0$, differoitu sarja on

$$X_t - X_{t-1} = (a + bt + Z_t) - [a + b(t-1) + Z_{t-1}] = b + Z_t - Z_{t-1},$$

jonka odotusarvofunktio on vakio b. Jos trendi on p-asteinen polynomifunktio, differoituun sarjaan jää astetta $p-1$ oleva trendi, (ks. harjoitustehtävä).

Differointia voidaan käyttää myös muun kuin trendistä johtuvan epästationaarisuuden poistamiseen. Esimerkissä 1, esiteltiin satunnaiskävelyn prosessi, joka on epästationaarinen, koska sen varianssi kasvaa lineaarisesti ajan suhteen. Differoimalla prosessi saatiin IID-prosessi, joka on stationaarinen.

Differointi voidaan myös toistaa. Soveltamalla differointia toisen kerran kaavassa (1.2) määriteltyyn prosessiin Y_t saamme

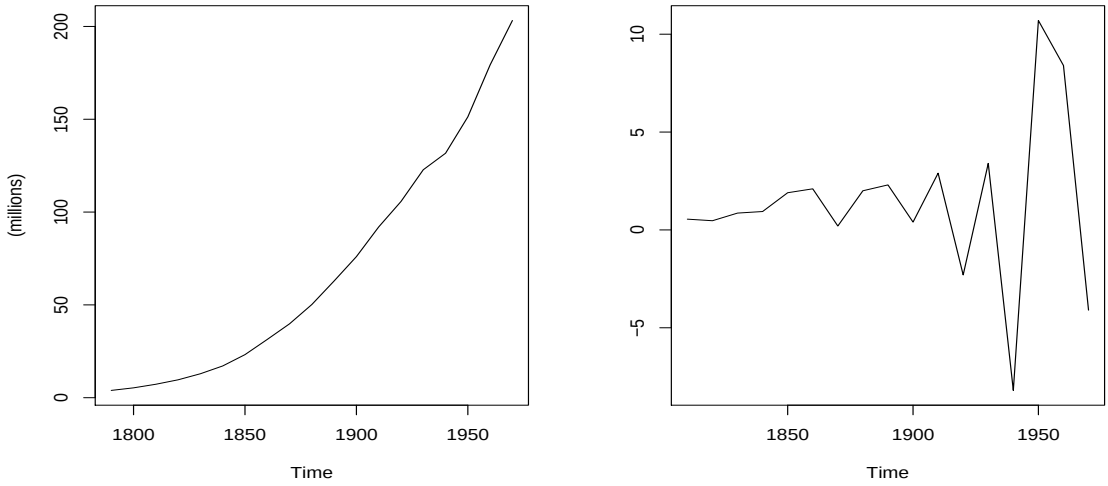
$$Y_t - Y_{t-1} = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2}. \tag{1.3}$$

Viiveoperaattorin B lausekkeet käyttäytyvät kuin polynomit. Määritellään, että potenssi B^d tarkoittaa sarjan viivästämistä d aikayksikköä: $\mathsf{B}^d X_t = X_{t-d}$, ja ∇^k tarkoittaa sarjan differointia k kertaa. Kaavassa (1.3) esitetty toisen kertaluvun differenssi olisi voitu laskea suoraan operaattorien avulla seuraavasti:

$$\nabla^2 X_t = (1 - \mathsf{B})^2 X_t = (1 - 2\mathsf{B} + \mathsf{B}^2)X_t = X_t - 2X_{t-1} + X_{t-2}.$$

Jos sarjan trendi on polynomifunktio astetta p, differoimalla sarja p kertaa saadaan trendi poistettua. Voidaan osoittaa, että jos polynomifunktio on $m(t) = c_0 + c_1 t + ... + c_p t^p$, niin $\nabla^p m(t) = p! c_p$. Yleensä trendin poistamiseen riittää yksi tai kaksi differointia viiveellä 1. Tämä johtuu siitä, että ainakin paikallisesti trendiä voidaan usein mallintaa suoralla tai matalaasteisella polynomilla.

Kuvioon 1.8 on piirretty USA:n populaatiota kuvaava aikasarja ja kaksi kertaa differoitu aikasarja. Havaitaan, ettei kaksi kertaa differoidussa sarjassa ole havaittavissa trendiä. Sarja ei kuitenkaan ole stationaarinen, sillä varianssi kasvaa selvästi. Ongelma voidaan poistaa logaritmoimalla alkuperäinen sarja. Jos alkuperäisessä sarjassa varianssi kasvaa samalla kun trendi kasvaa, ongelmasta voidaan päästä logaritmoimalla tai tekemällä Box-Cox -muunnos.



(a) Aikasarja

(b) Kahdesti differoitu sarja

Kuvio 1.8: USA:n väestö 10 vuoden välein

Kuviot on tulostettu käskyillä

```
plot(uspop,ylab="(millions)")
plot(diff(uspop,1,2),ylab="")
```

Funktion diff argumentti 1 tarkoittaa differointia viiveellä yksi ja 2 sitä, että differointi toistetaan kaksi kertaa.

Differoimalla voidaan päästä myös eroon kausikomponentista. Jos sarja voidaan esittää muodossa $X_t = s_t + Y_t$, missä s_t on kausivaihtelutermi, jonka jakson pituus on d, niin differoimalla viiveellä d saadaan

$$\nabla_d X_t := (1 - \mathsf{B}^d)X_t = X_t - X_{t-d} = (s_t + Y_t) - (s_{t-d} + Y_{t-d}) = Y_t - Y_{t-d},$$

sillä $s_t = s_{t-d}$. (Älä sekoita operaattoria ∇_d , joka kuvaa differoimista viiveellä d edellä määriteltyyn operaattoriin ∇^d , joka kuvaa differoimista d kertaa viiveellä 1). Jos jäännössarja $Y_t - Y_{t-d}$ sisältää vielä trendiä, se voidaan poistaa toistamalla differointia viiveellä 1.

1.9 Jäännössarjan 'valkoisuuden' testaaminen

Edellä on pyritty poistamaan sarjasta trendi ja kausivaihtelu joko suoraan vähentämällä sarjasta kyseiset komponentit tai differoimalla. Tavoitteena on ollut saada stationaarinen jäännössarja. Jos jäännössarja muistuttaa valkoista kohinaa, voidaan sarjan mallintaminen lopettaa. Tällöin ennusteet voidaan laatia trendin ja kausivaihtelun perusteella. Jos sen sijaan jäännöstermeillä on korrelaatiota, korrelaatiorakennetta voidaan mallintaa ja sitä voidaan käyttää hyväksi ennustamisessa. Siksi on tärkeätä testata, vastaako jäännössarja valkoisen kohinan prosessia.

Kappaleessa 1.3 todettiin, että jos aikasarja noudattaa IID(0,σ^2)-prosessia, otosautokorrelaatio-kertoimet ovat likimain $N(0, 1/n)$-jakautuneita ja riippumattomia. Tämä merkitsee sitä, että noin 95% otosautokorrelaatiokertoimista sijoittuu rajojen $\pm 1.96/\sqrt{n}$ sisälle. Jos autokorrelaatiokertoimet on laskettu viiveeseen 40 asti ja enemmän kuin 4 on rajojen ulkopuolella, riskitasolla 5 % voidaan hylätä hypoteesi, että kyseessä on valkoisen kohinan prosessi. Myös jos jokin otosautokorrelaatio-kertoimista on huomattavasti rajojen ulkopuolella, hypoteesi voidaan hylätä.

Kaikki lasketut otosautokorrelaatiokertoimet voidaan sisällyttää ns. *Portmanteau*-testiin. Jos hypoteesi siitä, että aikasarja noudattaa IID(0,σ^2)-prosessia, pitää paikkansa, testisuure

$$Q = n \sum_{j=1}^{h} \hat{\rho}^2(j)$$

on likimain $\chi^2(h)$-jakautunut. Suuret Q:n arvot kertovat siitä, että autokorrelaatiokertoimet ovat itseisarvoltaan liian suuria, jotta kyseessä olisi IID(0,σ^2)-prosessi. Testiä sanotaan Box-Pierce-testiksi.

Ljung ja Box ovat ehdottaneet testisuuretta

$$Q_{LB} = n(n+2) \sum_{j=1}^{h} \hat{\rho}^2(j)/(n-j)$$

joka noudattaa vieläkin tarkemmin $\chi^2(h)$-jakaumaa. McLeod ja Li ovat kehittäneet Portmanteau-testin, jonka avulla voidaan testata, onko jäännössarjassa epälineaarista riippuvuutta. Testisuure on Q_{LB} sovellettuna aikasarjaan, joka on saatu korottamalla alkuperäinen sarja toiseen potenssiin, ja sen jakauma on $\chi^2(h)$. Testi sopinee parhaiten tilanteeseen, jossa vaihtoehtoisena hypoteesina IID-prosessille on ns. ARCH-prosessi.

Peräkkäisten havaintojen riippuvuuden tutkimiseen voidaan käyttää myös ns. *käännepistetestiä* (turning point test). Jos $y_1, y_2, ..., y_n$ on jono havaintoja, sanotaan, että ajanhetkellä i on käännepiste, jos $y_{i-1} < y_i$ ja $y_i > y_{i+1}$ tai $y_{i-1} > y_i$ ja $y_i < y_{i+1}$. Olkoon T käännepisteiden lukumäärä IID- sarjassa, jonka pituus on n. Koska kääntymäpisteen todennäköisyys hetkellä i on $2/3$,

$$\mu_T = \mathsf{E}(T) = 2(n-2)/3.$$

Voidaan myös osoittaa, että

$$\sigma_T^2 = \mathsf{Var}(T) = (16n - 29)/90.$$

Jos $T - \mu_T$ on paljon nollaa suurempi, voidaan päätellä, että sarja vaihtaa suuntaansa nopeammin kuin voisi odottaa IID-prosessilta, ja jos $T - \mu_T$ on paljon nollaa pienempi, se on merkki positiivisesta autokorrelaatiosta. Kun n on suuri ja aikasarja noudattaa IID-prosessia, likimain

$$T \sim N(\mu_T, \sigma_T^2),$$

mitä voidaan käyttää hyväksi testaamisessa.

Lineaarisen trendin toteamisessa on hyödyllinen ns. järjestystesti (rank test). Merkitään P:llä sellaisten parien (y_i, y_j) määrää, että $y_j > y_i$, kun $j > i$ ja $i = 1, 2, ..., n-1$. Kaikestaan pareja (y_i, y_j), joilla $j > i$, on $n(n-1)/2$. Jos aikasarja on IID-prosessista, todennäköisyys, että $y_j > y_i$, on $1/2$, joten

$$\mu_P = \mathsf{E}(P) = (1/4)n(n-1).$$

Voidaan myös osoittaa, että

$$\sigma_P^2 = \mathsf{Var}(P) = n(n-1)(2n+5)/72.$$

Suurilla arvoilla n likimain

$$P \sim N(\mu_P, \sigma_P^2).$$

Jos jäännössarjan valkoisuuden lisäksi halutaan tutkia sen havaintojen normaalijakautuneisuutta, voidaan käyttää kvantiili-kvantiili-kuviota (qq-plot). Olkoon $X_1, X_2, ..., X_n$ satunnaisotos jakaumasta $N(0,1)$ ja $Y_1, Y_2, ..., Y_n$ jakaumasta $N(\mu, \sigma^2)$. Merkitään otosten järjestystunnuslukuja $X_{(i)}$ ja $Y_{(i)}$, $i = 1, 2, .., n$, jolloin $X_{(1)} < X_{(2)} < ... < X_{(n)}$ ja $Y_{(1)} < Y_{(2)} < ... < Y_{(n)}$. Tällöin

$$\mathsf{E}(Y_{(j)}) = \mu + \sigma m_j,$$

missä $m_j = \mathsf{E}(X_{(j)}),\ j = 1, 2, ..., n$. Tästä seuraa, että piirrettäessä parien $(m_1, Y_{(1)}), ..., (m_n, Y_{(n)})$ muodostama pisteparvi (qq-kuvio) pisteiden tulisi sijaita suurin piirtein suoralla. Jos aikasarjasta piirretyssä kuviossa havainnot eivät ole suurin piirtein suoralla, se on merkki siitä, etteivät havainnot noudata normaalijakaumaa. Odotusarvoja m_j voidaan approksimoida kaavalla $\Phi^{-1}[(j - 0.5)/n]$.

Alla olevilla funktioilla voi tehdä kääntymäpiste- ja järjestystestit. Ljung-Box-testin ja sen muunnelmat voi tehdä kirjastoon 'stats' sisältyvällä funktiolla Box.test. Normaalijakaumaa vastaavan qq-kuvion voi piirtää käskyllä qqnorm(x) ja siihen liittyvän viivan käskyllä qqline(x), missä x on tutkittava jäännössarja. Lisäksi normaalisuutta voidaan testata Jarque-Bera-testillä, joka sisältyy kirjastoon 'tseries'. Testi voidaan toteuttaa käskyllä jarque.bera.test(x) ja se perustuu vinouden ja huipukkuuden laskemiseen.

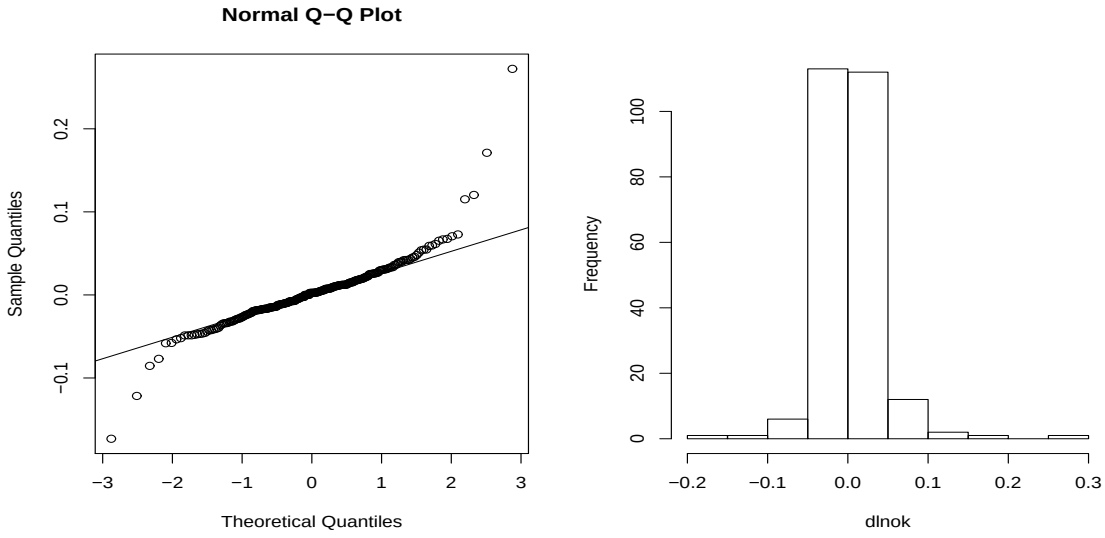
```
turning.point.test <- function(x)
{
DNAME <- deparse(substitute(x))
n<-length(x)
METHOD <- "Turning point test"
X<-embed(x,3)
STATISTIC<-sum((X[,2] > X[,1] & X[,2] > X[,3]) |
    (X[,2] < X[,1] & X[,2] < X[,3]))
mu <- 2*(n-2)/3
sigma2 <- (16*n-29)/90
PVAL<-2*(1-pnorm(abs(STATISTIC -  mu) / sqrt(sigma2)))
PARAMETER <- c(mu,sigma2)
names(STATISTIC) <- "normal"
names(PARAMETER) <- c("mu", "sigma2")
structure(list(statistic = STATISTIC, parameter = PARAMETER,
          p.value = PVAL, method = METHOD, data.name = DNAME),
        class = "htest")
}
```

```
rank.test <- function(x)
{
DNAME <- deparse(substitute(x))
n<-length(x)
METHOD <- "Rank test"
STATISTIC<-sum(outer(x,x,"<")[outer(1:n,1:n,"<")])
mu <- n*(n-1)/4
sigma2 <- n*(n-1)*(2*n+5)/72
PVAL<-2*(1-pnorm(abs(STATISTIC-mu) / sqrt(sigma2)))
PARAMETER <- c(mu,sigma2)
names(STATISTIC) <- "normal"
names(PARAMETER) <- c("mu", "sigma2")
structure(list(statistic = STATISTIC, parameter = PARAMETER,
        p.value = PVAL, method = METHOD, data.name = DNAME),
        class = "htest")}
```

Alla on tehty esitellyt testit sarjalle 'dlnok' (ks. 1.3). Testien perusteella nollahypoteesi siitä, että kyseessä on valkoisen kohinan prosessi, säilyy. Kuitenkin Jarque-Pera-testi hylkää hypoteesin jäännösten normaalisuudesta. Tämä voidaan todeta myös qq-kuviosta 1.9 a. Sen perusteella aineiston suuret havainnot ovat liian suuria verrattuna normaalijakaumaan, joten jäännösten jakauma on oikealle vino. Myös pienimmät havainnot ovat 'liian pieniä', joten jakauma on paksuhäntäinen molempiin suuntiin. Samat asiat voidaan todeta histogrammista 1.9 b.

```
Box.test(dlnok,lag=40,"Ljung")

X-squared = 29.1238, df = 40, p-value = 0.8982


Box.test(dlnok^2,lag=40,"Ljung")

X-squared = 4.3286, df = 40, p-value = 1


turning.point.test(dlnok)

normal = 158, mu = 164.667, sigma2 = 43.944, p-value = 0.3146
```

```
rank.test(dlnok)

normal = 15486, mu = 15438.0, sigma2 = 431406.3, p-value = 0.9417


jarque.bera.test(dlnok)

X-squared = 1677.061, df = 2, p-value < 2.2e-16

#QQ-kuvio ja histogrammi
qqnorm(dlnok)
qqline(coredata(dlnok))
hist(coredata(dlnok),main="",xlab="dlnok")
```



(a) QQ-kuvio (b) Histogrammi

Kuvio 1.9: Nokian osakekurssin tuottojakauma

Luku 2

Stationaariset prosessit

2.1 Yleistä ennustamisesta

Oletetaan, että tunnetaan aikasarjan $\{X_t, t = 0, \pm 1, ...\}$ realisaatiosta havainnot $x_1, x_2, ..., x_n$ ja halutaan ennustaa h askelta eteenpäin havaintoa x_{n+h}. Ajatellaan, että ennustevirheestä e aiheutuvaa tappiota voidaan mitata tappiofunktiolla $C(e)$. Tällöin on luonnollista hakea sellaista havainnoista $x_1, x_2, ..., x_n$ riippuvaa ennustefunktiota $f(x_1, x_2, ..., x_n)$, että $\mathsf{E}\ C(e) = \mathsf{E}\ C[X_{n+h} - f(X_1, ..., X_n)]$ minimoituu. Tappiofunktio $C(e)$ saa arvon 0, kun ennustevirhe $e = 0$, ja $C(e)$ kasvaa, kun $|e|$ kasvaa. Kuitenkaan $C(e)$ ei ole välttämättä nollan suhteen symmetrinen funktio.

Yleisimmin käytetty tappiofunktio on muotoa $C(e) = ae^2$, missä a on jokin positiivinen reaaliluku. Tällöin optimaalinen ennustaminen merkitsee keskineliövirheen $\mathsf{E}[X_{n+h} - f(X_1, ..., X_n)]^2$ minimoimista. Voidaan osoittaa, että minimoiva ratkaisu on $f(X_1, ..., X_n) = \mathsf{E}(X_{n+h}|X_1, ..., X_n)$, (vrt. harj. teht.). Ehdollisen odotusarvon laskemiseksi on tunnettava satunnaismuuttujien $X_1, ..., X_n, X_{n+h}$ yhteisjakauma.

Esim. 1. Oletetaan, että satunnaismuuttujat X_1 ja X_2 noudattavat binormaalijakaumaa (kaksiulotteista normaalijakaumaa). Merkitään odotusarvoja $\mathsf{E}(X_i) = \mu_i$, $i = 1, 2$, variansseja $\mathsf{Var}(X_i) = \sigma_i^2$ ja korrelaatiota $\rho = \mathsf{Cor}(X_1, X_2)$. Tällöin satunnaismuuttujan X_2 jakauma ehdolla X_1 on myös normaalijakauma odotusarvolla

$$\mathsf{E}(X_2|X_1) = \mu_2 + \rho\sigma_2\sigma_1^{-1}(X_1 - \mu_1)$$

ja varianssilla

$$\mathsf{Var}(X_2|X_1) = \sigma_2^2(1-\rho^2).$$

Edellisessä esimerkissä ehdollinen odotusarvo $E(X_2|X_1)$ on lineaarinen eli se on muotoa $aX_1 + b$. Yleisemmin voidaan osoittaa, että $\mathsf{E}(X_{n+h}|X_1, ..., X_n)$ on lineaarinen moniulotteisen normaalijakauman tapauksessa ja voidaan siis esittää muodossa

$$\mathsf{E}(X_{n+h}|X_1, ..., X_n) = a_0 + a_1X_1 + ... + a_nX_n,$$

missä $a_0, ..., a_n$ ovat vakiota. Jos havaintojen yhteisjakauma ei ole normaalinen (ts. kyseessä ei ole gaussinen prosessi), optimaalinen ennustin ei välttämättä ole lineaarinen. Tällä kurssilla kuitenkin rajoitutaan tarkastelemaan lineaarisia ennustimia ja optimaalisuuden kriteerinä käytetään keskineliövirhettä. Tämä merkitsee sitä, että prosessien määrittelyssä rajoitutaan ensimmäisen ja toisen asteen ominaisuuksiin, ts. odotusarvoon ja kovarianssifunktioon.

2.2 Autokovarianssifunktion ominaisuuksia

Ennen kuin siirrymme stationaaristen prosessien ennustamiseen, on hyvä tutkia hiukan lähemmin näiden prosessien ominaisuuksia. Stationaarisen prosessin autokovarianssifunktiolla on seuraavat ominaisuudet:

1. $\gamma(0) \geq 0$
2. $|\gamma(h)| \leq \gamma(0)$ kaikilla h
3. $\gamma(h)$ on parillinen funktio, ts. $\gamma(h) = \gamma(-h)$.

Ominaisuus (1) seuraa siitä, että varianssi on aina ei-negatiivinen ja ominaisuus (2) Cauchy-Schwarz-epäyhtälöstä. Ominaisuus (3) johtuu siitä, että $\gamma(h) = \mathsf{Cov}(X_{t+h}, X_t) = \mathsf{Cov}(X_t, X_{t+h}) = \gamma(-h)$.

Lisäksi autokovarianssifunktio on ei-negatiivisesti definiitti. Määritellään, että kokonaisluvuille määritelty, reaaliarvoinen funktio κ on *ei-negatiivisesti definiitti*, jos

$$\sum_{i,j=1}^{n} a_i\kappa(i-j)a_j \geq 0 \tag{2.1}$$

kaikille positiivisille kokonaisluvuille n ja vektoreille $\mathbf{a} = (a_1, a_2, ..., a_n)'$ missä komponentit a_i ovat reaalilukuja. Huomaa, että ehto 2.1 voidaan esittää matriisimuodossa $\mathbf{a}'\mathbf{K}\mathbf{a} \geq 0$, missä $\mathbf{K}$ on matriisi $[\kappa(i-j)]_{i,j=1}^n$.

Lause 1. Reaaliarvoinen, kokonaisluvuille määritelty funktio on stationaarisen aikasarjan autokovarianssifunktio jos ja vain jos se on ei-negatiivisesti definiitti.

Todistus. Oletetaan, että $\gamma(h)$ on stationaarisen aikasarjan $\{X_t\}$ autokovarianssifunktio. Olkoon n mikä tahansa positiivinen kokonaisluku ja $\mathbf{a} = (a_1, a_2, ..., a_n)'$ mikä tahansa reaaliluvuista koostuva vektori, jonka pituus on n. Tällöin

$$\mathsf{Var}(a_1X_1 + a_2X_2 + ... + a_nX_n) = \sum_{i,j=1}^{n} \mathsf{Cov}(a_iX_i, a_jX_j) = \sum_{i,j=1}^{n} a_i\gamma(i-j)a_j \geq 0,$$

sillä varianssi on aina ei-negatiivinen. Nähdään siis, että $\gamma(h)$ on ei-negatiivisesti definiitti.

Todistus toiseen suuntaan on vaativampi. Ks. esim. TSTM, Theorem 1.5.1.

Esim. 2. Osoitetaan, että funktio

$$\kappa(h) = \begin{cases} 1, & \text{kun } h = 0, \\ \rho, & \text{kun } h = \pm 1, \\ 0, & \text{muulloin.} \end{cases}$$

on ei-negatiivisesti definiitti, kun $|\rho| \leq 0.5$. Huomataan, että $\kappa(h)$ on MA(1)-prosessin autokovarianssifunktio, kun asetetaan $\sigma^2(1+\theta^2) = 1$ ja $\sigma^2\theta = \rho$. Ratkaisuksi saadaan

$$\begin{cases} \theta = \frac{1 \pm \sqrt{1-4\rho^2}}{2\rho} \\ \sigma^2 = \frac{1}{\theta^2} \end{cases}$$

kun $|\rho| \leq 0.5$. Ratkaisua ei ole olemassa, jos $|\rho| > 0.5$.

2.3 Lineaariset prosessit

Tärkeän ryhmän stationaarisista prosesesseista muodostavat ns. *lineaariset prosessit*, joiden mallintaminen ja ennustaminen ARMA-malleilla muodostaa

kurssin ydinosan. Aikasarjaa $\{X_t\}$ sanotaan lineaariseksi prosessiksi, jos se voidaan esittää muodossa

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \tag{2.2}$$

kaikilla arvoilla t, missä $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ ja $\{\psi_j\}$ on jono vakioita, jotka toteuttavat ehdon $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Käyttämällä siirto-operaattoria B, kaava (2.2) voidaan esittää tiiviimmässä muodossa

$$X_t = \psi(\mathsf{B}) Z_t,$$

missä $\psi(\mathsf{B}) = \sum_{j=-\infty}^{\infty} \psi_j \mathsf{B}^j$. Lineaarista prosessia sanotaan *liukuvan keskiarvon prosessiksi*, jos $\psi_j = 0$, kun $j < 0$. Tällöin prosessi voidaan esittää muodossa

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

Huomautus. Ehto $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ takaa sarjan (2.2) suppenemisen (todennäköisyydellä 1). Cauhcy-Schwarzin epäyhtälön perusteella $\mathsf{E}|Z_t| \leq \sigma$, sillä $[\mathsf{E}(|Z_t \cdot 1|]^2 \leq \mathsf{E}|Z|^2 \mathsf{E}(1) = \sigma^2$. Itseisarvosarjan odotusarvo on äärellinen:

$$\mathsf{E} \sum_{j=-\infty}^{\infty} |\psi_j||Z_{t-j}| = \sum_{j=-\infty}^{\infty} |\psi_j| \, \mathsf{E}|Z_{t-j}| \leq \sigma \sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

Koska itseisarvosarjan odotusarvo on äärellinen, itseisarvosarja on äärellinen todennäköisyydellä 1, josta seuraa, että myös alkuperäinen sarja suppenee todennäköisyydellä 1.

Operaattorin $\psi(\mathsf{B})$ voidaan ajatella olevan lineaarinen suodin, jonka syötteenä on sarja $\{Z_t\}$ ja ulostulona sarja $\{X_t\}$. Seuraavaksi osoitetaan, että mikä tahansa stationaarinen sarja antaa ulostulona stationaarisen sarjan, kun siihen sovelletaan suodinta $\psi(\mathsf{B})$.

Lause 2. Olkoon $\{Y_t\}$ stationaarinen aikasarja, jonka odotusarvo on 0 ja kovarianssifunktio γ_Y. Jos $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, niin aikasarja

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} = \psi(B) Y_t \tag{2.3}$$

on stationaarinen odotusarvona 0 ja autokovarianssifunktiona

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h+k-j). \tag{2.4}$$

Siinä erikoistapauksessa, että $\{Y_t\}$ on valkoisen kohinan prosessi WN $(0, \sigma^2)$,

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \sigma^2. \tag{2.5}$$

Todistus. Samoin kuin edellisessä huomautuksessa voidaan perustella, että sarja $\{X_t\}$ suppenee (hajonnan σ tilalla on $\sqrt{\gamma(0)}$). Koska $\mathsf{E}(Y_t) = 0$,

$$\mathsf{E}(X_t) = \mathsf{E}\left(\sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} \right) = \sum_{j=-\infty}^{\infty} \psi_j \mathsf{E}(Y_{t-j}) = 0$$

ja

$$\begin{aligned} \mathsf{E}(X_{t+h} X_t) =& \mathsf{E}\left[\left(\sum_{j=-\infty}^{\infty} \psi_j Y_{t+h-j} \right) \left(\sum_{k=-\infty}^{\infty} \psi_k Y_{t-k} \right) \right] \\ =& \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \mathsf{E}(Y_{t+h-j} Y_{t-k}) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h+k-j), \end{aligned}$$

josta voidaan päätellä aikasarjan $\{X_t\}$ olevan stationaarinen kovarianssifunktiolla (2.4). (Se, että odotusarvon ja summauksen järjestystä voidaan vaihtaa, perustuu oletukseen $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.) Jos $\{Y_t\}$ on valkoisen kohinan prosessi, $\gamma_Y(h+k-j) = \sigma^2$, kun $k = j-h$, ja 0 muuten, mistä seuraa (2.5).

Huomautus. Suotimia, joilla on itseisesti summautuvat kertoimet, voidaan soveltaa peräkkäin stationaariseen aikasarjaan ja tuloksena saadaan aina stationaarinen aikasarja. Lopputuloksen kannalta ei ole merkitystä sillä, missä järjestyksessä suotimia sovelletaan. Jos suotimia $\alpha(\mathsf{B}) = \sum_{j=-\infty}^{\infty} \alpha_j \mathsf{B}^j$ ja $\beta(\mathsf{B}) = \sum_{j=-\infty}^{\infty} \beta_j \mathsf{B}^j$ sovelletaan peräkkäin sarjaan $\{Y_t\}$, tuloksena saadaan sarja

$$\alpha(\mathsf{B})\beta(\mathsf{B})Y_t = \beta(\mathsf{B})\alpha(\mathsf{B})Y_t = \psi(\mathsf{B})Y_t,$$

missä $\psi(\mathsf{B}) = \alpha(\mathsf{B})\beta(\mathsf{B})$. Lopputulokseen päästään siis myös soveltamalla sarjaan $\{Y_t\}$ yhtä suodinta $\psi(\mathsf{B})$, joka saadaan kertomalla keskenään suotimet $\alpha(\mathsf{B})$ ja $\beta(\mathsf{B})$, jotka ovat siirto-operaattorin B potenssisarjoja.

Esim 1. *MA(q)-prosessi.* Esimerkki liukuvan keskiarvon prosessista on MA(q)-prosessi, joka voidaan määritellä yhtälöllä

$$X_t = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q},$$

missä $Z_t \sim \text{WN}\ (0, \sigma^2)$. Kaavan (2.5) erikoistapauksena saadaan MA(q)-prosessin kovarianssifunktio

$$\gamma_X(h) = \begin{cases} \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|}, & \text{kun } 0 \leq |h| \leq q, \\ 0, & \text{muuten,} \end{cases}$$

missä $\theta_0 = 1$. Havaitaan, että autokovarianssifunktio häviää, kun $|h| > q$. Tällaista prosessia, jonka autokovarianssifunktio $= 0$, kun $|h| > q$, sanotaan q-korreloituneeksi. Voidaan yleisesti osoittaa, että q-korreloitunut stationaarinen prosessi voidaan esittää MA(q)-prosessina (TSTM, Section 3.2).

Esim 2. *AR(1)-prosessi.* Ensisilmäyksellä AR(1)-prosessi ei näytä lineaariselta prosessilta, sillä se määritellään stationaariseksi prosessiksi, joka toteuttaa differenssiyhtälön

$$X_t - \phi X_{t-1} = Z_t, \tag{2.6}$$

missä $Z_t \sim \text{WN}\ (0, \sigma^2)$. Prosessi voidaan kuitenkin esittää liukuvan keskiarvon prosessina, kun $|\phi| < 1$. Differenssiyhtälö (2.6) voidaan esittää muodossa

$$\phi(\mathsf{B}) X_t = Z_t, \tag{2.7}$$

missä $\phi(\mathsf{B}) = 1 - \phi\mathsf{B}$. Tarkastellaan seuraavaksi suodinta $\psi(\mathsf{B}) = 1/\phi(\mathsf{B})$, joka voidaan geometrisen sarjan summakaavaa käyttäen esittää muodossa

$$\psi(\mathsf{B}) = \frac{1}{1 - \phi\mathsf{B}} = 1 + \phi\mathsf{B} + \phi^2\mathsf{B}^2 + ...$$

Suotimella on itseisesti summautuvat kertoimet, sillä $1 + |\phi| + |\phi|^2 + ... = 1/(1 - |\phi|) < \infty$. Soveltamalla suodinta $\psi(\mathsf{B})$ yhtälön (2.7) molempiin puoliin saadaan $X_t = \psi(\mathsf{B}) Z_t$, eli

$$X_t = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + ...,$$

josta nähdään, että $\{X_t\}$ on liukuvan keskiarvon prosessi.

Huomautus. Jokainen stationaarinen aikasarja voidaan esittää liukuvan keskiarvon prosessin ja deterministisen prosessin summana. Prosessi on deterministinen, jos prosessin havainnot määräytyvät täysin edeltävien havaintojen avulla. Summaesitystä kutsutaan *Woldin hajotelmaksi.* On erittäin harvinaista, että esim. taloudellisissa aikasarjoissa esiintyisi deterministinen komponentti. Tällä kurssilla ei käsitellä enempää Woldin hajotelmaa.

2.4 Stationaarisen aikasarjan odotuarvon ja kovarianssifunktion estimoiminen

Stationaarisen prosessin $\{X_t\}$ odotusarvon momenttiestimaattori on otoskeskiarvo

$$\bar{X}_n = (X_1 + X_2 + ... + X_n)/n.$$

Se on harhaton, sillä $\mathsf{E}(\bar{X}_n) = \mu$. Otoskeskiarvon keskineliövirhe on

$$\begin{aligned}\mathsf{E}(\bar{X}_n - \mu)^2 =& \mathsf{Var}(\bar{X}_n) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathsf{Cov}(X_i, X_j) \\ =& n^{-2} \sum_{i-j=-n}^{n} (n - |i-j|)\gamma(i-j) \\ =& n^{-1} \sum_{h=-n}^{n} \left(1 - \frac{|h|}{n}\right) \gamma(h).\end{aligned}$$

Keskineliövirhettä voidaan estimoida lausekkeella $n^{-1} \sum_{h=-a}^{a} \left(1 - \frac{|h|}{n}\right) \hat{\gamma}(h)$, jos $\hat{\gamma}(h) \approx 0$, kun $|h| > a$.

Huomaa, että $\hat{\gamma}(h)$ ei ole luotettava $\gamma(h)$:n estimaattori, jos $h > n/4$, missä n on aikasarjan havaintojen lukumäärä. Keskineliövirheen lauseketta voidaan käyttää hyväksi, kun määritetään luottamusväli μ:lle. Jos $\{X_t\}$ on lineaarinen tai ARMA-prosessi, otoskeskiarvo on likimain normaalisti jakautunut. Tällöin 95% luottamusväli μ:lle on $(\bar{X}_n - 1.96\sqrt{\mathsf{Var}(\bar{X}_n)}, \bar{X}_n + 1.96\sqrt{\mathsf{Var}(\bar{X}_n)})$.

Otosautokovarianssi- ja otosautokorrelaatiomatriisi ovat ei-negatiivisesti definiittejä, samoin kuin niiden teoreettiset vastineet (ks. todistus Brockwell&Davis s.58). Itse asiassa ne ovat positiivisesti definiittejä, elleivät kaikki sarjan arvot ole yhtä suuria. Lineaaristen ja erityisesti ARMA-prosessien tapauksessa otosautokorrelaatiokerrointen vektori noudattaa likimain moniulotteista normaalijakaumaa:

$$(\hat{\rho}(1), ..., \hat{\rho}(k))' \sim N((\rho(1), ..., \rho(k))', n^{-1}\mathbf{W}),$$

missä matriisin $\mathbf{W}$ elementit saadaan *Bartlettin kaavasta*

$$\begin{aligned}w_{ij} = \sum_{k=1}^{\infty} &\{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \\ &\times \{\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\}.\end{aligned}$$

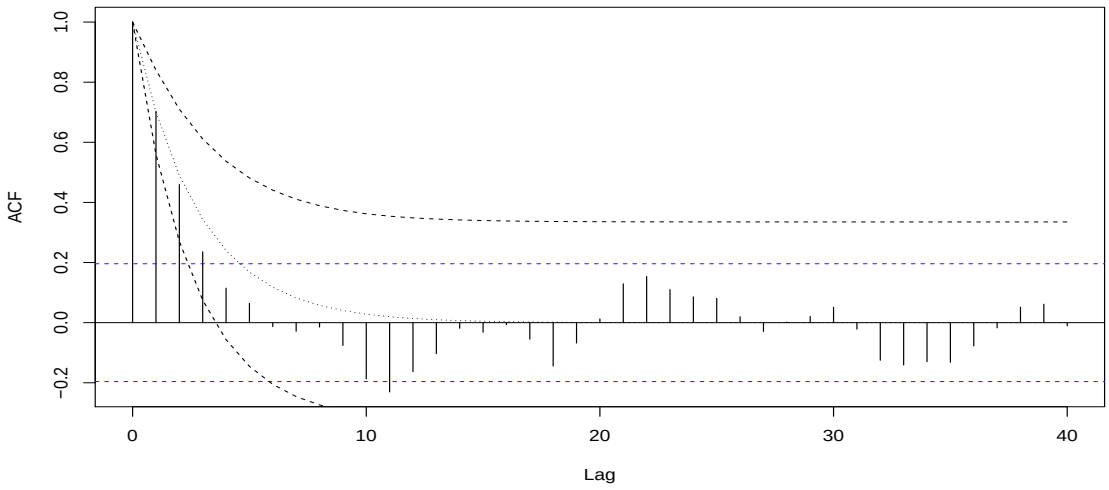
Esim 1. AR(1)-prosessin

$$X_t = \phi X_{t-1} + Z_t, \quad \{Z_t\} \sim \text{WN } (0, \sigma^2),$$

tapauksessa, missä $|\phi| < 1$, autokorrelaatiokerroin viiveellä h on $\rho(h) = \phi^{|h|}$. Otosautokorrelaatio-kertoimen varianssi on w_{ii}/n, missä Bartlettin kaavan perusteella voidaan johtaa, että

$$w_{ii} = (1 - \phi^{2i})(1 + \phi^2)(1 - \phi^2)^{-1} - 2i\phi^{2i}.$$

Kuviossa 2.1 on piirretty otosautokorrelaatiofunktio, joka on saatu simuloidusta 100 havainnon pituisesta AR(1)-prosessista, jossa $\phi = 0.7$. Lisäksi on piirretty teoreettinen autokorrelaatiofunktio ja otosautokorrelaatioille 95% todennäköisyysväli. Havaitaan, että väli suurilla viiveillä on suurempi, kuin IID-kohinaa vastaava todennäköisyysväli, joka myös on piirretty kuvioon. Havaitaan myös, että otosautokorrelaatiofunktio ikään kuin 'lainehtii' eli peräkkäisillä arvoilla on positiivista autokorrelaatiota, vaikka teoreettinen autokorrelaatiofunktio vähenee tasaisesti.



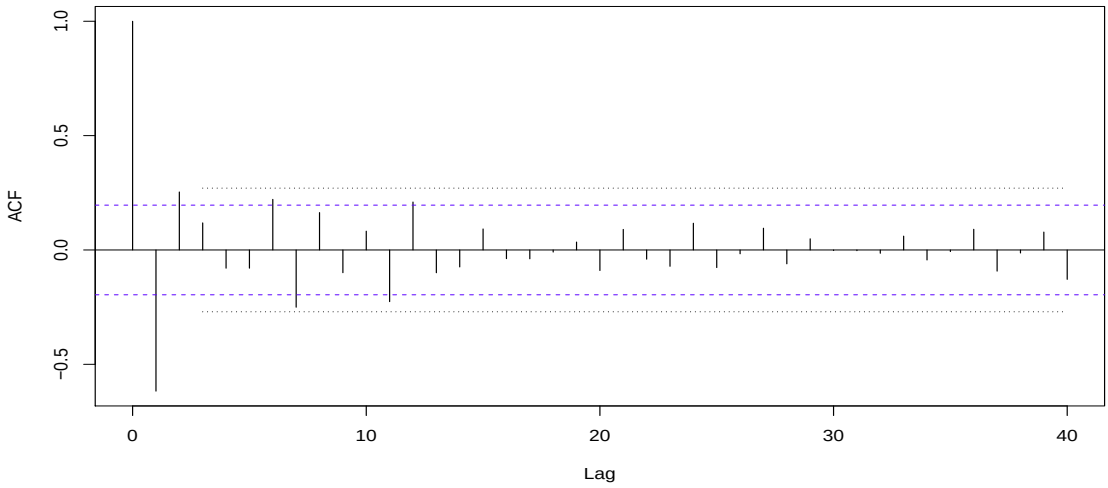
Kuvio 2.1: AR(1)-prosessin otosautokorrelaatiofunktio todennäköisyysväleineen

```
n <- 100
phi <- 0.7
u <- arima.sim(list(ar=0.7),n)
x <- 0:40
y <- phi^x
z <- (1-phi^(2*x))*(1+phi^2)/(1-phi^2)-2*x*phi^(2*x)
acf(u,40,main="")
lines(x,y,lty=3)
lines(x,y+1.96*(z/n)^0.5,lty=2)
lines(x,y-1.96*(z/n)^0.5,lty=2)
```

AR(p) -prosessien identifiointi eli viivepituuden p määritys onnistuu parhaiten osittaisautokorrelaatiofunktion avulla, joka esitellään myöhemmin.

Esim. 2. *MA(q)-prosessin autokorrelaatiofunktio.* Kappaleessa 2.3 todettiin, että MA(q)-prosessi on q-korreloitunut, eli autokorrelaatiofunktio $\rho(h)$ häviää (eli on 0), kun $|h| > q$. Tämän vuoksi MA(q) prosessia voidaan yrittää identifioida tarkastelemalla sen otosautokorrelaatiofunktiota. Bartlettin kaavan avulla on helppo osoittaa, että q-korreloituneen prosessin tapauksessa $Var(\hat{\rho}(i) \approx [1 + 2\rho^2(1) + ... + 2\rho^2(q)]/n$, kun $i > q$. Kuviossa 2.2 on otosautokorrelaatiofunktio aikasarjalle, joka on saatu simuloimalla MA(2)-prosessia $X_t = Z_t - 0.6Z_{t-1} + 0.8Z_{t-2}$. Kuvioon on piirretty 95% todennäköisyysrajat $\pm 1.96n^{-1/2}(1 + 2\rho^2(1) + 2\rho^2(2))^{1/2}$, sekä tavanomaiset rajat $\pm 1.96n^{-1/2}$. Yleensä käytetään tiukempia tavanomaisia rajoja, sillä korrelaatiokertoimet $\rho(i)$ eivät ole tavallisesti tunnettuja. Tässä tapauksessa $\rho(1) = (\theta_1 + \theta_1\theta_2)/(1 + \theta_1^2 + \theta_2^2) = (-0.6 - 0.6 * 0.8)/(1 + 0.36 + 0.64) = -0.54$ ja $\rho(2) = \theta_2/(1 + \theta_1^2 + \theta_2^2) = 0.8/(1 + 0.36 + 0.64) = 0.4$.

```
u<-arima.sim(list(ma=c(-0.6,0.8)),100)
r1<-(-0.6-0.6*0.8)/(1+0.36+0.64)
r2<-0.8/(1+0.36+0.64)
upper<-1.96*sqrt(1+2*r1^2+2*r2^2)/10
lower <- -upper
acf(u,40,main="")
lines(c(3,40),c(upper,upper),lty=3)
lines(c(3,40),c(lower,lower),lty=3)
```



Kuvio 2.2: MA(2)-prosessin otosautokorrelaatiofunktio todennäköisyysväleineen

2.5 Stationaaristen aikasarjojen ennustaminen

Seuraavaksi johdetaan lineaarinen ennustin stationaarisen aikasarjan $\{X_t\}$ havainnolle X_{n+h}, kun käytettävissä on havinnot $X_1, X_2, ..., X_n$. Odotusarvo μ ja autokovarianssifunktio $\gamma(h)$ oletetaan tunnetuiksi. Merkitään keskineliövirheen mielessä optimaalista ennustinta

$$\mathsf{P_n}X_{n+h} = a_0 + a_1X_n + a_2X_{n-1} + .. + a_nX_1.$$

Kertoimien a_i määrittämiseksi minimoidaan lauseke

$$\mathsf{E}(X_{n+h} - a_0 - a_1X_n - a_2X_{n-1} - ... - a_nX_1)^2.$$

Derivoimalla lauseke kertoimien a_i suhteen ja asettamalla derivaatat nolliksi saadaan yhtälöt

$$\mathsf{E}(X_{n+h} - a_0 - a_1X_n - a_2X_{n-1} - ... - a_nX_1) = 0 \tag{2.8}$$

$$\mathsf{E}(X_{n+h} - a_0 - a_1X_n - a_2X_{n-1} - ... - a_nX_1)X_{n+1-i} = 0, \quad i = 1, 2, ..., n \tag{2.9}$$

Ensimmäisestä yhtälöstä voidaan päätellä, että ennustevirheen odotusarvo on 0 eli optimaalinen ennustin on harhaton. Yhtälöstä saadaan ratkaistua

$$a_0 = \mu(1 - a_1 - a_2 - ... - a_n).$$

Yhtälöt (2.9) voidaan kirjoittaa muotoon

$$\mathsf{Cov}(X_{n+h} - a_0 - a_1 X_n - a_2 X_{n-1} - ... - a_n X_1, X_{n+1-i}) = 0, \quad i = 1, 2, ..., n$$

josta nähdään, että ennustevirhe on korreloimaton selittävien muuttujien $X_n, X_{n-1}, ..., X_1$ kanssa. Yhtälöt voidaan edelleen esittää muodossa

$$a_1\gamma(i-1) + a_2\gamma(i-2) + ... + a_n\gamma(i-n) = \gamma(h+i-1), \quad i = 1, 2, ..., n$$

tai matriisimuodossa

$$\begin{pmatrix} \gamma(0) & \gamma(1) & ... & \gamma(n-1) \\ \gamma(1) & \gamma(0) & ... & \gamma(n-2) \\ ... & ... & ... & .... \\ \gamma(n-1) & \gamma(n-2) & ... & \gamma(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ ... \\ a_n \end{pmatrix} = \begin{pmatrix} \gamma(h) \\ \gamma(h+1) \\ ... \\ \gamma(h+n-1) \end{pmatrix}.$$

Merkitään matriisiyhtälö lyhyesti

$$\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}_n(h). \tag{2.10}$$

Optimaalinen ennustin on siis

$$\mathsf{P_n} X_{n+h} = \mu + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu)$$

missä kertoimet a_i määräytyvät yhtälön (2.10) perusteella.

Koska kaavan (2.8) perusteella ennustevirheen odotusarvo on 0, ennustevirheen keskineliövirhe on

$$\mathsf{E}(X_{n+h} - \mathsf{P_n}X_{n+h})^2 = \mathsf{E}\left[X_{n+h} - \mu - \sum_{i=1}^{n} a_i(X_{n+1-i} - \mu)\right]^2$$

$$= \mathsf{E}(X_{n+h} - \mu)^2 - 2\sum_{i=1}^{n} a_i(X_{n+1-i} - \mu)(X_{n+h} - \mu)$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathsf{E}(X_{n+1-i} - \mu)(X_{n+1-j} - \mu)$$

$$= \gamma(0) - 2\sum_{i=1}^{n} a_i\gamma(h+i-1) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(i-j)$$

$$= \gamma(0) - 2\mathbf{a}'\boldsymbol{\gamma}_n(h) + \mathbf{a}'\Gamma_n\mathbf{a}$$

$$= \gamma(0) - \mathbf{a}'\boldsymbol{\gamma}_n(h).$$

Viimeisessä välivaiheessa on käytetty hyväksi kaavaa (2.10).

2.6 Ennusteoperaattorin $\mathsf{P_n}$ ominaisuuksia

Lineaarisella ennustimella $\mathsf{P_n}Y$ tarkoitetaan satunnaismuuttujien $1, X_n, X_{n-1}, ..., X_1$ lineaarikombi-naatiota $a_0 + a_1X_n + ... + a_nX_1$, joka minimoi keskineliövirheen

$$\mathsf{E}(Y - a_0 - a_1X_n - ... - a_nX_1)^2.$$

Voidaan osoittaa, että riittävät ja välttämättömät ehdot kertoimille $a_0, a_1, ..., a_n$ ovat (vrt. kaavat (2.8) ja (2.9))

$$\mathsf{E}(Y - a_0 - a_1X_n - a_2X_{n-1} - ... - a_nX_1) = 0 \tag{2.11}$$

$$\mathsf{E}(Y - a_0 - a_1X_n - a_2X_{n-1} - ... - a_nX_1)X_{n+1-i} = 0, \quad i = 1, 2, ..., n \tag{2.12}$$

Ensimmäinen ehto merkitsee sitä, että ennustin on harhaton ja toinen, että ennustevirhe on korreloimaton ennustavien muuttujien kanssa.

Ennusteoperaattorilla $\mathsf{P_n}$ on seuraavat ominaisuudet, jotka on hyvä tuntea:

1. $\mathsf{P_n}(\beta_0 + \sum_{j=1}^{k} \beta_j Y_j) = \beta_0 + \sum_{j=1}^{k} \beta_j \mathsf{P_n} Y_j$ (lineaarisuus)
2. $\mathsf{P_n} X_m = X_m$, jos $m \leq n$,
3. $$\mathsf{P_n P_m} Y = \begin{cases} \mathsf{P_n} Y, & \text{jos } m > n \\ \mathsf{P_m} Y, & \text{jos } m \leq n. \end{cases}$$

Todistus:

1. Osoitetaan, että lauseke $\beta_0 + \sum_{j=1}^{k} \beta_j \mathsf{P_n} Y_j$ toteuttaa ehdon (2.12), missä $Y = \beta_0 + \sum_{j=1}^{k} \beta_j Y_j$. (Vastaavasti voidaan tarkistaa ehto (2.11)). Merkitään Y_j:n optimaalista ennustinta $\mathsf{P_n} Y_j = a_{j0} + a_{j1} X_n + a_{j2} X_{n-1} + ... + a_{jn} X_1$, jolloin $\mathsf{E}(Y_j - a_{j0} - a_{j1} X_n - a_{j2} X_{n-1} - ... - a_{jn} X_1) X_{n+1-i} = 0,\ i = 1, 2, .., n.$
   Nyt
   $$\begin{aligned}
   &\mathsf{E}\left\{ \left[Y - (\beta_0 + \sum_{j=1}^{k} \beta_j \mathsf{P_n} Y_j) \right] X_{n+1-j} \right\} \\
   &= \mathsf{E}\left\{ \left\{ \beta_0 + \sum_{j=1}^{k} \beta_j Y_j) - \left[\beta_0 + \sum_{j=1}^{k} \beta_j (a_{j0} + a_{j1} X_n + ... + a_{jn} X_1) \right] \right\} X_{n+1-i} \right\} \\
   &= \mathsf{E}\left[\sum_{j=1}^{k} \beta_j (Y_j - a_{j0} - a_{j1} X_n - ... - a_{jn} X_1) X_{n+1-j} \right] \\
   &= \sum_{j=1}^{k} \beta_j \mathsf{E}\left[(Y_j - a_{j0} - a_{j1} X_n - ... - a_{jn} X_1) X_{n+1-j} \right] = 0, \quad i = 1, 2, ..., n.
   \end{aligned}$$
2. Jos $m \leq n$, on ilmiselvää, että $\mathsf{P_n} X_m = X_m$, sillä ennustettava muuttuja X_m sisältyy ennustaviin muuttujiin $X_n, X_{n-1}, ..., X_1$.
3. Merkitään $\mathsf{P_m} Y = a_{m0} + a_{m1} X_m + ... + a_{mm} X_1$ ja $\mathsf{P_n} Y = a_{n0} + a_{n1} X_n + ... + a_{nn} X_1$. Oletetaan aluksi, että $m \leq n$. Tällöin $\mathsf{P_n P_m} Y = \mathsf{P_n}(a_{m0} + a_{m1} X_m + ... + a_{mm} X_1) = a_{m0} + a_{m1} \mathsf{P_n} X_m + ... + a_{mm} \mathsf{P_n} X_1 = a_{m0} + a_{m1} X_m + ... + a_{mm} X_1 = \mathsf{P_m} Y$ lineaarisuuden ja kohdan 2) perusteella.
   Kun $m > n$, on siis todistettava, että $\mathsf{P_n} Y$ toteuttaa ehdot $\mathsf{E}(\mathsf{P_m} Y - \mathsf{P_n} Y) = 0$ ja $\mathsf{E}(\mathsf{P_m} Y - \mathsf{P_n} Y) X_{n+1-i} = 0$. Ensimmäinen ehto toteutuu, koska sekä $\mathsf{P_m} Y$ että $\mathsf{P_n} Y$ ovat harhattomia ennustimia Y:lle. Jälkimmäinen ehto toteutuu, koska
   $$\begin{aligned}
   &\mathsf{E}[(\mathsf{P_m} Y - \mathsf{P_n} Y) X_{n+1-i}] \\
   &= \mathsf{E}[(a_{m0} + a_{m1} X_m + ... + a_{mm} X_1 - a_{n0} - a_{n1} X_n - ... - a_{nn}) X_{n+1-i}] \\
   &= \mathsf{E}[(a_{m0} + a_{m1} X_m + ... + a_{mm} X_1 - Y) X_{n+1-i}] \\
   &\quad + \mathsf{E}[(Y - a_{n0} - a_{n1} X_n - ... - a_{nn}) X_{n+1-i}] = 0.
   \end{aligned}$$

Luku 3

ARMA-mallit

3.1 ARMA(p,q)-prosessit

Tähän mennessä on käsitelty autoregressiivisiä AR(p)- ja liukuvan keskiarvon MA(q)- malleja. On myös mahdollista yhdistää nämä mallit. Määritellään, että $\{X_t\}$ on ARMA(p,q)-prosessi, jos $\{X_t\}$ on stationaarinen ja kaikkina ajanhetkinä t

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}, \qquad (3.1)$$

missä $Z_t \sim$ WN $(0, \sigma^2)$. Yksinkertaisuuden vuoksi oletetaan lisäksi, ettei polynomeilla $1 - \phi_1 z - ... - \phi_p z^p$ ja $1 + \theta_1 z + ... + \theta_q z^q$ ei ole yhteisiä tekijöitä.

Voidaan osoittaa että yhtälöllä (3.1) on yksikäsitteinen stationaarinen ratkaisu, jos kompleksitasossa määritellyllä polynomilla $1 - \phi_1 z - ... - \phi_p z^p$ ei ole nollakohtia yksikköympyrällä $\{z \in \mathbb{C} : |z| = 1\}$. Yleensä kuitenkin rajoitutaan *kausaalisiin* ja *invertoituviin* ARMA-prosesseihin, sillä nämä prosessit riittävät stationaaristen ARMA-prosessien kovarianssirakenteen mallintamiseen.

Sanotaan, että ARMA-prosessi $\{X_t\}$ on kausaalinen, jos se voidaan esittää äärettömänä liukuvan keskiarvon prosessina

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + ...,$$

missä $\{Z_t\} \sim$ WN $(0, \sigma^2)$ ja $\sum_{j=1}^{n} |\psi_j| < \infty$. Ehto on yhtäpitävä sen kanssa, että polynomilla $1 - \phi_1 z - ... - \phi_p z^p$ ei ole nollakohtia kompleksitason yksikkökiekossa $\{z \in \mathbb{C} : |z| \leq 1\}$.

Vastaavasti ARMA-prosessin $\{X_t\}$ sanotaan oleva invertoituva, jos Z_t voidaan esittää havaintojen $X_t, X_{t-1}, ...$ avulla:

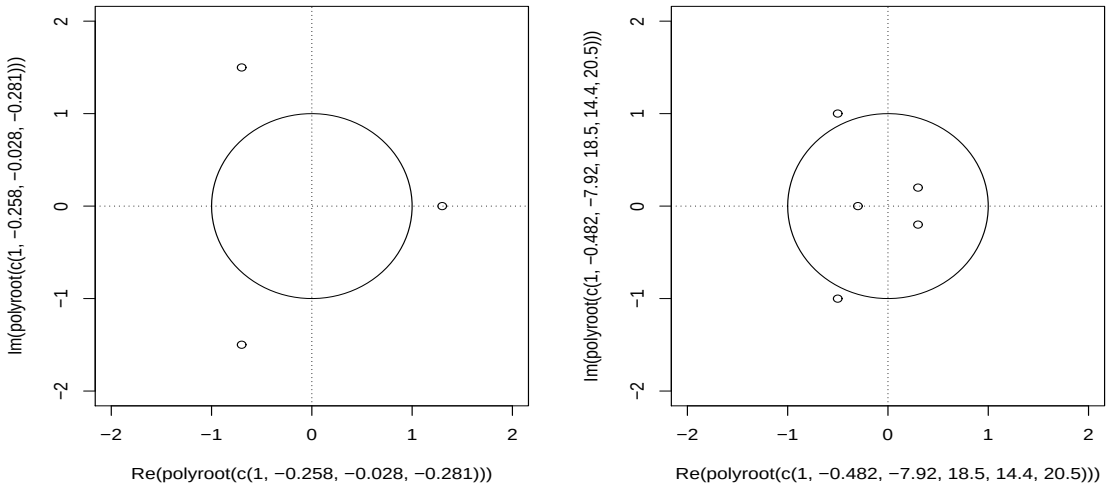
$$Z_t = X_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + ...,$$

missä $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ ja $\sum_{j=1}^{n} |\pi_j| < \infty$. Ehto on yhtäpitävä sen kanssa, ettei polynomilla $1 + \theta_1 z - ... + \theta_q z^q$ ole nollakohtia yksikkökiekossa $\{z \in \mathbb{C} : |z| \leq 1\}$. (Yhtäpitävät ehdot ovat seurausta potenssisarjojen perusominaisuuksista, vrt. TSTM, Section 3.1.)

Esim. 1. Tarkastellaan ARMA-mallia

$$(1 - 0.258\mathsf{B} - 0.028\mathsf{B}^2 - 0.281\mathsf{B}^3)X_t = (1 - 0.4821\mathsf{B} - 7.92\mathsf{B}^2 + 18.5\mathsf{B}^3 + 14.4\mathsf{B}^4 + 20.5\mathsf{B}^5)Z_t.$$

Kuvioon 3.1 on piirretty polynomien $\phi(z) = 1 - 0.258z - 0.028z^2 - 0.281z^3$ ja $\theta(z) = 1 - 0.4821z - 7.92z^2 + 18.5z^3 + 14.4z^4 + 20.5z^5$ nollakohdat kompleksitasossa. Koska polynomin $\phi(z)$ nollakohdat ovat yksikköympyrän ulkopuolella, prosessi on kausaalinen. Koska osa polynomin $\theta(z)$ nollakohdista on yksikköympyrän sisäpuolella, prosessi ei ole invertoituva.



Kuvio 3.1: AR- ja MA-polynomien nollakohdat kompleksitasossa

Oikeanpuoleinen kuvio on piirretty käyttäen R-käskyjä

```
plot(polyroot(c(1,-0.482,-7.92,18.5,14.4,20.5)),
    xlim=c(-2,2),ylim=c(-2,2))
lines(exp(1i*seq(-pi,pi,len=100)))
abline(v=0,lty=3)
abline(h=0,lty=3)
```

Funktiolla polyroot saadaan funktion juuret. Funktiolla abs saataisiin suoraan selville juurten itseisarvot, esim.

```
abs(polyroot(c(1,-0.482,-7.92,18.5,14.4,20.5)))
```

antaa tuloksen

```
[1] 0.3603599 0.2998975 0.3603599 1.1191807 1.1191807
```

josta nähdään että kolme juurta on yksikköympyrän sisäpuolella ja kaksi niukasti ulkopuolella.

3.2 ARMA-prosessin muuntaminen liukuvan keskiarvon prosessiksi

Oletetaan, että ARMA-prosessi

$$(1 - \phi_1 \mathsf{B} - ... - \phi_p \mathsf{B}^p) X_t = (1 + \theta_1 \mathsf{B} + ... + \theta_q \mathsf{B}^q) Z_t, \tag{3.2}$$

missä $Z_t \sim \text{WN}\ (0, \sigma^2)$, on kausaalinen. Tällöin se voidaan esittää liukuvan keskiarvon prosessina

$$X_t = (1 + \psi_1 \mathsf{B} + \psi_2 \mathsf{B}^2 + ...) Z_t. \tag{3.3}$$

Sijoittamalla X_t:n lauseke (3.3) yhtälöön (3.2) saadaan yhtälö

$$(1 - \phi_1 \mathsf{B} - ... - \phi_p \mathsf{B}^p)(1 + \psi_1 \mathsf{B} + \psi_2 \mathsf{B}^2 + ...) Z_t = (1 + \theta_1 \mathsf{B} + ... + \theta_q \mathsf{B}^q) Z_t,$$

jonka avulla voidaan ratkaista tuntemattomat kertoimet. Merkitsemällä eri B:n potensseja vastaavat kertoimet yhtä suuriksi saadaan yhtälöt

$$\begin{aligned}
\theta_1 &= \psi_1 - \phi_1 \\
\theta_2 &= \psi_2 - \psi_1 \phi_1 - \phi_2 \\
\theta_3 &= \psi_3 - \psi_2 \phi_1 - \psi_1 \phi_2 - \phi_3 \\
&... \\
\end{aligned}$$

Esim. 1. Tarkastellaan ARMA-prosessia

$$(1 - 0.8\mathsf{B} + 0.2\mathsf{B}^2)X_t = (1 - 0.3\mathsf{B})Z_t.$$

Polynomilla $\phi(z) = 1 - 0.8z + 0.2z^2$ on nollakohdat $2 \pm i$, jotka sijaitsevat yksikköympyrän ulkopuolella. ARMA-prosessi on siis kausaalinen ja voidaan esittää liukuvan keskiarvon prosessina (3.3).

Tuntemattomat kertoimet ψ_j voidaan ratkaista yhtälön

$$(1 - 0.8\mathsf{B} + 0.2\mathsf{B}^2)(1 + \psi_1\mathsf{B} + \psi_2\mathsf{B}^2 + ...) = 1 - 0.3\mathsf{B}$$

perusteella. Saamme B:n eri potenssien kertoimiksi

$$\begin{aligned}
\mathsf{B}: &\quad \psi_1 - 0.8 = -0.3 \\
\mathsf{B}^2: &\quad \psi_2 - 0.8\psi_1 + 0.2 = 0 \\
\mathsf{B}^3: &\quad \psi_3 - 0.8\psi_2 + 0.2\psi_1 = 0 \\
&\quad ...
\end{aligned}$$

mistä saadaan ratkaistua

$$\begin{aligned}
\psi_1 &= -0.3 + 0.8 = 0.5 \\
\psi_2 &= -0.2 + 0.8\psi_1 = -0.2 + 0.8 \cdot 0.5 = 0.2 \\
\psi_3 &= 0.8\psi_2 - 0.2\psi_1 = 0.8 \cdot 0.5 - 0.2 \cdot 0.8 = 0.06 \\
&\quad ...
\end{aligned}$$

Prosessi voidaan siis esittää sarjana

$$X_t = Z_t + 0.5Z_{t-1} + 0.2Z_{t-2} + 0.06Z_{t-3} + ...$$

3.3 ARMA-prosessin autokovarianssifunktion määrittäminen

Yksi tapa määrittää ARMA-prosessin autokovarianssifunktio on kertoa yhtälö

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}$$

puolittain satunnaismuuttujalla X_{t-k} ja ottaa odotusarvot puolittain. Koska X_{t-k} on riippumaton Z_{t-h}:n kanssa, kun $h < k$, saadaan

$$\begin{aligned} &\mathsf{E}(X_{t-k}X_t) - \phi_1\mathsf{E}(X_{t-k}X_{t-1}) - ... - \phi_p\mathsf{E}(X_{t-k}X_{t-p}) \\ &= \theta_k\mathsf{E}(X_{t-k}Z_{t-k}) + ... + \theta_q\mathsf{E}(X_{t-k}Z_{t-q}), \end{aligned} \tag{3.4}$$

kun $0 \leq k \leq q$. (Merkitään, että $\theta_0 = 1$). Käyttämällä hyväksi esitysmuotoa

$$X_{t-k} = Z_{t-k} + \psi_1 Z_{t-k-1} + \psi_2 Z_{t-k-2} + ...$$

yhtälö (3.4) saadaan muotoon

$$\gamma(k) - \phi_1\gamma(k-1) - ... - \phi_p\gamma(k-p) = (\theta_k + \theta_{k+1}\psi_1 + \theta_{k+2}\psi_2 + ... + \theta_q\psi_{q-k})\sigma^2. \tag{3.5}$$

Kun $k > q$, on voimassa yhtälö

$$\gamma(k) - \phi_1\gamma(k-1) - ... - \phi_p\gamma(k-p) = 0. \tag{3.6}$$

Käyttämällä $p+1$ ensimmäistä yhtälöä yhtälöistä (3.5) ja (3.6) voidaan ratkaista autokovarianssit $\gamma(0), \gamma(1), \gamma(2), ..., \gamma(p)$. Tämän jälkeen rekursiivisesti voidaan ratkaista loput autokovarianssit yhtälöstä (3.6).

Esim 1 (jatkoa). Jatketaan kappaleen 3.2 esimerkin käsittelyä, jossa tarkasteltiin prosessia

$$X_t - 0.8X_{t-1} + 0.2X_{t-2} = Z_t - 0.3Z_{t-1},$$

jolla oli esitysmuoto

$$X_t = Z_t + 0.5Z_{t-1} + 0.2Z_{t-2} + 0.06Z_{t-3} + ...$$

Autokovarianssit $\gamma(0)$, $\gamma(1)$ ja $\gamma(2)$ voidaan ratkaista lineaarisesta yhtälöryhmästä

$$\begin{aligned} \gamma(0) - 0.8\gamma(1) + 0.2\gamma(2) &= (1 - 0.3 \cdot 0.5)\sigma^2, \\ \gamma(1) - 0.8\gamma(0) + 0.2\gamma(1) &= -0.3\sigma^2, \\ \gamma(2) - 0.8\gamma(1) + 0.2\gamma(0) &= 0. \end{aligned}$$

Loput autokovarianssit voidaan ratkaista rekursiivisesti yhtälön

$$\gamma(k) - 0.8\gamma(k-1) + 0.2\gamma(k-2) = 0$$

avulla. Yleinen ratkaisu voitaisiin määrittää käyttämällä hyväksi differenssiyhtälöiden ratkaisukaavoja, mutta niihin ei puututa tässä.

3.4 Osittaisautokorrelaatiofunktio

Stationaarisen aikasarjan $\{X_t\}$ *osittaisautokorrelaatiofunktio* $\alpha(h)$ määritellään yhtälöillä

$$\begin{aligned}\alpha(0) &= 1,\\ \alpha(h) &= \phi_{hh}, \quad h \geq 1,\end{aligned}$$

missä ϕ_{hh} on vektorin $\boldsymbol{\phi}_h = \Gamma_h^{-1}\boldsymbol{\gamma}_h$ viimeinen komponentti, $\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$ ja $\boldsymbol{\gamma}_h = [\gamma(1), \gamma(2), ..., \gamma(h)]'$. Vastavasti havainnoille $x_1, x_2, ..., x_n$ voidaan määritellä otos-osittaisautokorrelaatiofunktio yhtälöillä

$$\begin{aligned}\hat{\alpha}(0) &= 1,\\ \hat{\alpha}(h) &= \hat{\phi}_{hh}, \quad h \geq 1,\end{aligned}$$

missä $\hat{\phi}_{hh}$ on vektorin $\hat{\boldsymbol{\phi}}_h = \hat{\Gamma}_h^{-1}\hat{\boldsymbol{\gamma}}_h$ viimeinen komponentti.

Esim. 1. AR(p)-prosessin osittaisautokorrelaatiofunktio. Kausaalinen AR(p)-prosessi määritellään yhtälöllä

$$X_t - \phi_1 X_{t-1} - .... - \phi_p X_{t-p} = Z_t,$$

missä $\{Z_t\} \sim \text{WN}\ (0, \sigma^2)$ ja $\mathsf{Cov}(X_s, Z_t) = 0$, kun $s < t$. Paras lineaarinen ennustin havainnolle X_{h+1} havaintojen $X_1, ..., X_h$ avulla on

$$\hat{X}_{h+1} = \phi_1 X_h + \phi_2 X_{h-1} + ... + \phi_p X_{h+1-p},$$

kun $h > p$ (todistus harjoitustehtävä). Toisaalta havaintojen $X_h, X_{h-1}, ..., X_1$ kertoimet $a_1, a_2, ..., a_h$ saadaan yhtälöstä $\boldsymbol{\gamma}_h = \Gamma_h \mathbf{a}_h$, vrt. kaava (2.10). Nyt siis $a_i = \phi_i$, kun $i \leq p$, ja $a_i = 0$, kun $i > p$. Nähdään siis, että $\alpha(p) = a_p = \phi_p$ ja $\alpha(h) = a_h = 0$, kun $h > p$.

Siis AR(p)-prosessin tapauksessa osittaiskorrelaatiokerroin viiveellä h on nolla, kun $h > p$. Voisi odottaa että otososittaisautokorrelaatiokerroin olisi likimain nolla, kun $h > p$. Voidaankin osoittaa, että AR(p)-prosessin tapauksessa osittaisautokorrelaatiot $\hat{\alpha}(h)$ ovat likimain riippumattomia ja $N(0, 1/n)$-jakautuneita, kun $h > p$ ja n on havaintojen lukumäärä (ks. TSTM, Section 8.10). Jos siis viiveestä $p + 1$ lähtien otososittaisautokorrelaatio asettuu rajojen $\pm 1.96/\sqrt{n}$ sisäpuolelle, voidaan alustavasti ehdottaa AR(p)-mallia.

Huomautus. Osittaisautokorrelaatio voidaan määritellä myös korrelaationa

$$\alpha(h) = \mathsf{Cor}(X_{h+1} - \mathsf{P}(X_{h+1}|X_2, X_3, .., X_h), X_1 - \mathsf{P}(X_1|X_2, ...X_h)),$$

missä $\mathsf{P}(X_s|X_2, X_3, .., X_h)$ tarkoittaa havainnon X_s parasta lineaarista ennustinta havaintojen $X_2, X_3, .., X_h$ avulla. Määritelmä on yhtäpitävä aiemmin annetun määritelmän kanssa (vrt. TSTM, p.171). Intuitiivisesti voidaan ajatella, että laskettaessa osittaisautokorrelaatio on havaintojen X_1 ja X_{h+1} korrelaatiosta poistettu välissä olevien havaintojen $X_2, ..., X_h$ vaikutus.

3.5 Periodogrammi ja ARMA-prosessien jaksollisuus

Reaalilukuvektorille $\mathbf{x} = (x_0, x_1, ..., x_{n-1})'$ voidaan määritellä diskreetti Fourier-muunnos $\mathbf{a} = (a_0, a_1, ..., a_{n-1})'$, missä

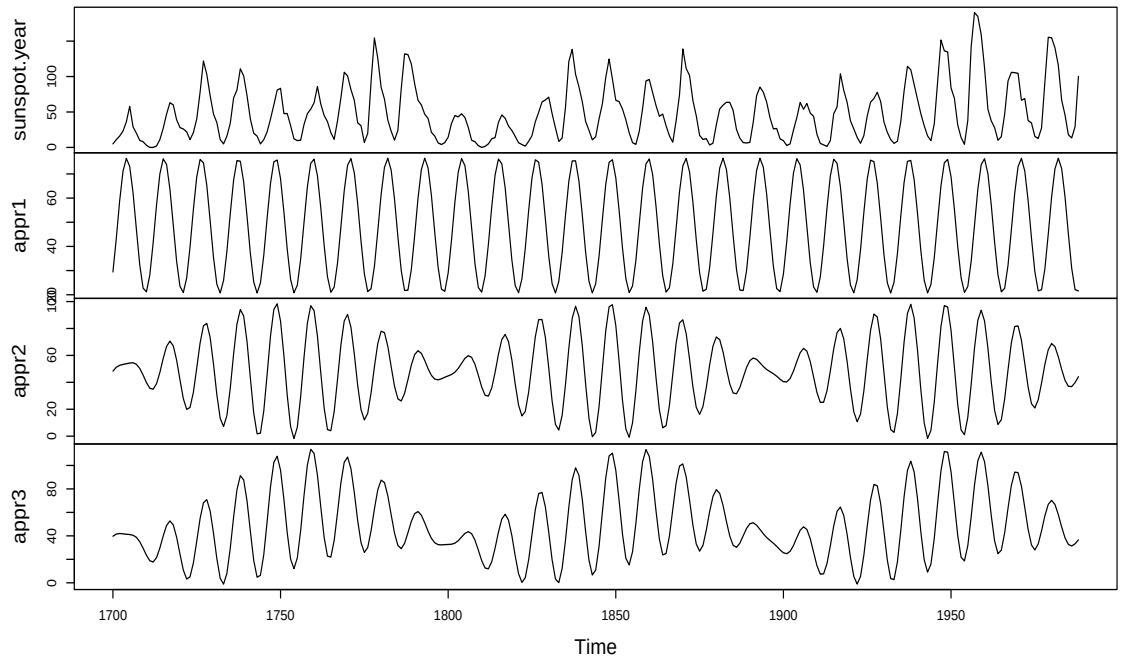
$$a_k = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t e^{-it\omega_k} = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \left[\cos(\omega_k t) - i \sin(\omega_k t)\right]$$

ja luvut $\omega_k = 2\pi k/n$ ovat ns. *Fourierin taajuuksia.* Kertoimet a_k ovat kompleksilukuja. Kun tunnetaan Fourier-muunnos, alkuperäinen lukujono $\mathbf{x}$ voidaan palauttaa kaavalla

$$x_t = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} a_k e^{i\omega_k} = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} a_k \left[\cos(\omega_k t) + i \sin(\omega_k t)\right].$$

Jokainen n:n pituinen lukujono $\mathbf{x}$ voidaan siis esittää ω_k-taajuuksisten siniaaltojen lineaarikombinaationa. Fourierin kertoimen a_k itseisarvo $|a_k|$ kertoo vastaavan siniaallon amplitudin eli heilahteluvälin. Fourierin taajuus $\omega_{n-k} = 2\pi - 2\pi k/n$ vastaa negatiivista taajuutta $-\omega_k$ ja on helppo osoittaa, että $a_{n-k} = \bar{a}_k$, missä viiva yläpuolella tarkoittaa liittolukua. Siksi $|a_{n-k}| = |a_k|$. Fourierin kerroin a_0 on reaaliluku ja on lukujonon $\mathbf{x}$ keskiarvo kerrottuna vakiolla $\sqrt{n}$.

Kuviossa 3.2 on piirretty aikasarja sunspot.year sekä erilaisia approksimaatioita käyttäen amplitudiltaan suurimpia siniaaltoja.



Kuvio 3.2: Sunspot-sarjan Fourier-approksimaatioita

```
f <- fft(sunspot.year)
f0 <- rep(0,289)
f0[order(Mod(f))[287:289]] <- f[order(Mod(f))[287:289]]
appr1<-Re(fft(f0,inv=T)/289)
f0[order(Mod(f))[285:289]] <- f[order(Mod(f))[285:289]]
appr2<-Re(fft(f0,inv=T)/289)
f0[order(Mod(f))[283:289]] <- f[order(Mod(f))[283:289]]
appr3<-Re(fft(f0,inv=T)/289)

plot(cbind(sunspot.year,appr1,appr2,appr3),main="")
```

Koska R-funktio fft ei suorita vakiolla $1/\sqrt{n}$ kertomista, täytyy käänteismuunnosta muodostettaessa jakaa luvulla $n = 289$. Kertoimen a_0 jälkeen suurin itseisarvo on kertoimella a_{26}, joka vastaa taajuutta $\omega_{26} = 2\pi(26/289)$. Tätä taajuutta vastaa jakso $2\pi/\omega_{26} = 11.11538$, joka on lähinnä au-

ringonpilkkujen tunnettua jaksoa 11.

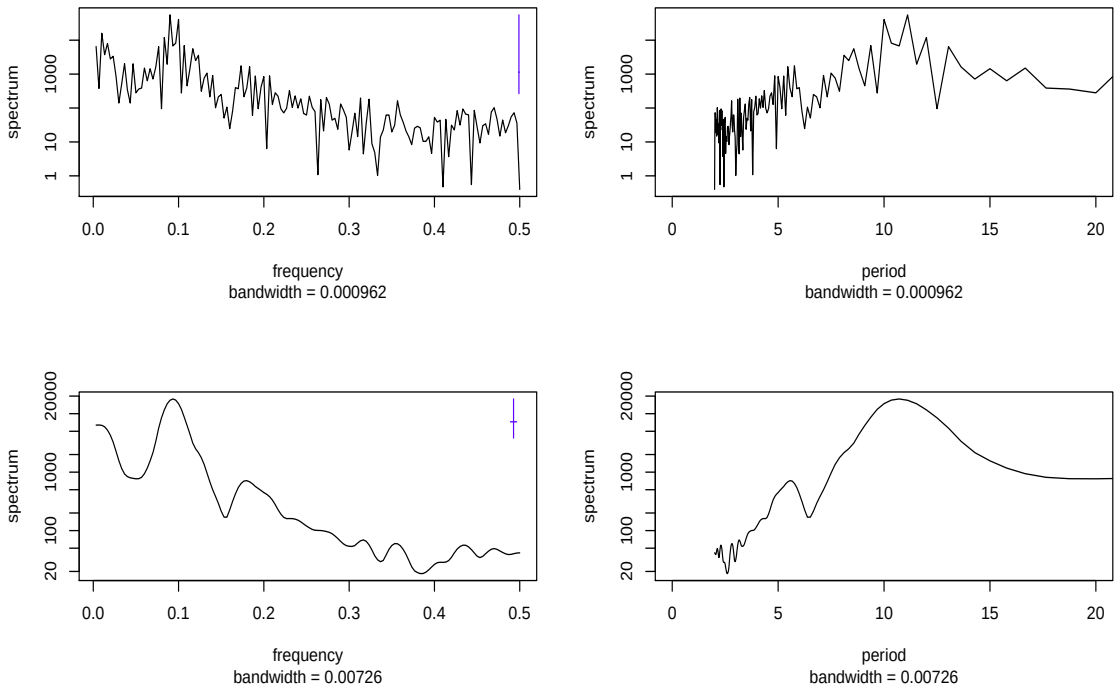
Aikasarjoja analysoitaessa käytetään yleisesti *periodogrammia*, joka on jaksollinen reaaliluvuille määritelty funktio

$$I_n(\lambda) = \frac{1}{n}\left|\sum_{t=0}^{n-1} x_t e^{-it\lambda}\right|^2$$

Kun $\lambda = \omega_k$, missä ω_k on yksi Fourierin taajuuksista, niin $I_n(\omega_k) = |a_k|^2$. Periodogrammi voidaan siis laskea Fourierin muunnoksella. Yleensä periodogrammi piirretään välille $[0, \pi]$, sillä periodogrammin jakso on 2π ja $I_n(\lambda) = I_n(-\lambda)$. Periodogrammi ilmaisee aikasarjan $\mathbf{x}$ varianssin jakautumisen eri taajuuskomponentteihin, sillä $\hat{\gamma}(0) = \frac{1}{n}\sum_{k=1}^{n-1} I(\omega_k)$. Keskiarvon neliö toteuttaa yhtälön $\bar{x}^2 = \frac{1}{n}I(0)$.

Yleensä periodogrammia tasoitetaan peräkkäisillä liukuvan keskiarvon suotimilla, jotta jakauma tulisi paremmin näkyviin. Tasoitusta tarvitaan, sillä raakaperiodogrammi ei lähesty teoreettista spektritiheysfunktiota, vaikka otoskoko $n \to \infty$ (ks. Brockwell&Davis, s.122). Mitä useampia termejä liukuviin keskiarvoihin otetaan, sitä pienempi on periodogrammin varianssi ja sitä tasaisempi kuvio. Kuitenkin suotimen pituuden kasvattaminen lisää periodogrammin harhaa. Kuvioon 3.3 on piirretty auringonpilkkuaineiston periodogrammi sekä periodogrammi, jota on tasoitettu 5 ja 7 termin liukuvilla keskiarvoilla. Huomaa, että R skaalaa periodogrammin välille [0,0.5] välin $[0, \pi]$ sijasta. Oikeanpuoleiset periodogrammit on piirretty käyttäen asteikkona jaksoa taajuuden sijasta. Periodogrammeissa on huippu taajuudella $1/11$, joka vastaa jaksoa 11.

```
par(mfrow=c(2,2))
a <- spectrum(sunspot.year,main="")
a$freq <- 1/a$freq
plot(a,xlab="period",xlim=c(0,20),main="")

a <- spectrum(sunspot.year,spans=c(5,7),main="")
a$freq <- 1/a$freq
plot(a,xlab="period",xlim=c(0,20),main="")
```



Kuvio 3.3: Sunspot-sarjan periodogrammikuvioita

Koska aikasarjan jaksollisuus näkyy myös autokorrelaatiofunktiosta, ei liene yllättävää, että periodogrammi voidaan laskea otosautokovarianssifunktiosta: $I_n(\omega_k) = \sum_{|h|<n} \hat{\gamma}(h)e^{-ih\omega_k}$. Periodogrammin teoreettinen vastine on *spektritiheysfunktio*, joka voidaan määritellä stationaarisille ja nollakeskiselle prosesseille $\{X_t\}$ kaavalla

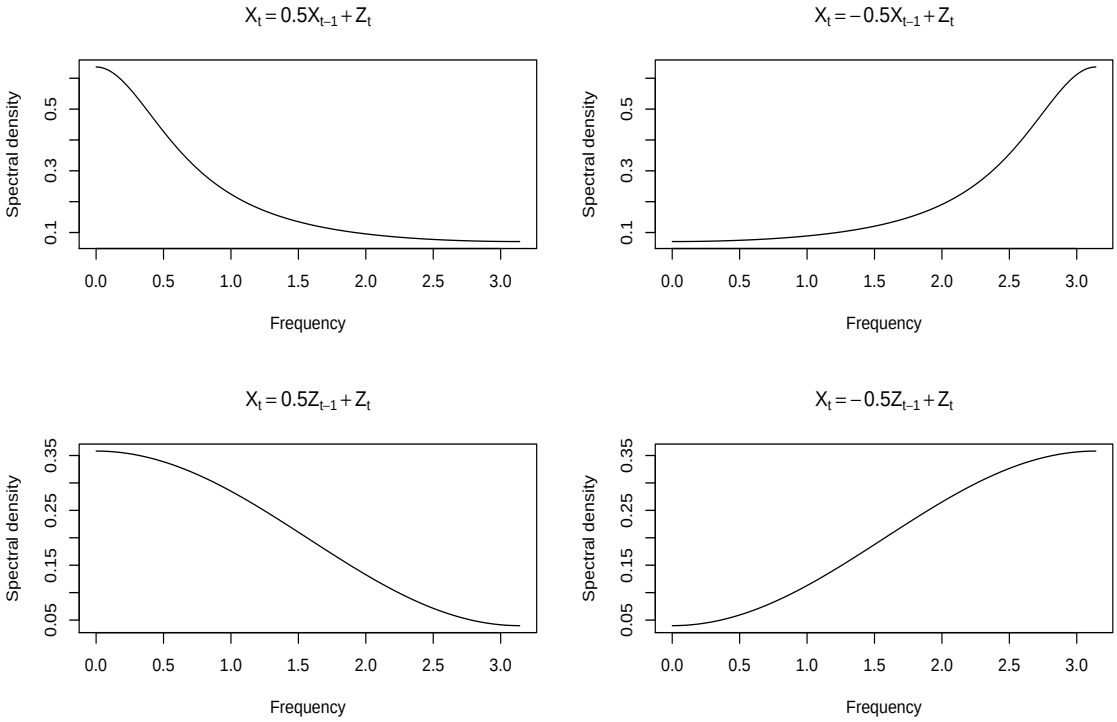
$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-ih\lambda}\gamma(h),$$

kun autokovarianssifunktio $\gamma(h)$ toteuttaa ehdon $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$.

Voidaan osoittaa (Brockwell&Davis, s.130), että ARMA(p,q)-prosessin $\{X_t\}$, joka toteuttaa differenssiyhtälön $\phi(\mathsf{B})X_t = \theta(\mathsf{B})Z_t$, spektritiheysfunktio on

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\theta(e^{-i\lambda})|^2}{|\phi(e^{-i\lambda})|^2}.$$

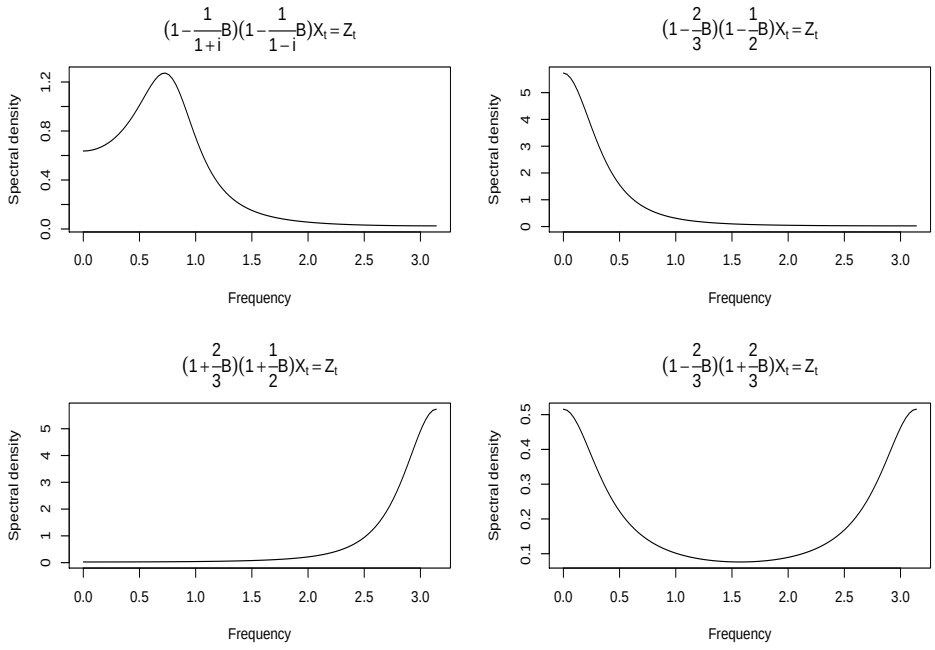
Kuviossa 3.4 on piirretty AR(1) ja MA(1) -prosessien spektritiheysfunktioita. Havaitaan, että AR(1)-prosessin tapauksessa taajuusjakauma on keskittynyt pienemmälle alueelle. Kun ϕ tai θ on negatiivinen, spetritiheysfunktiossa on huippu taajuudella π, joka vastaa jaksoa 2.



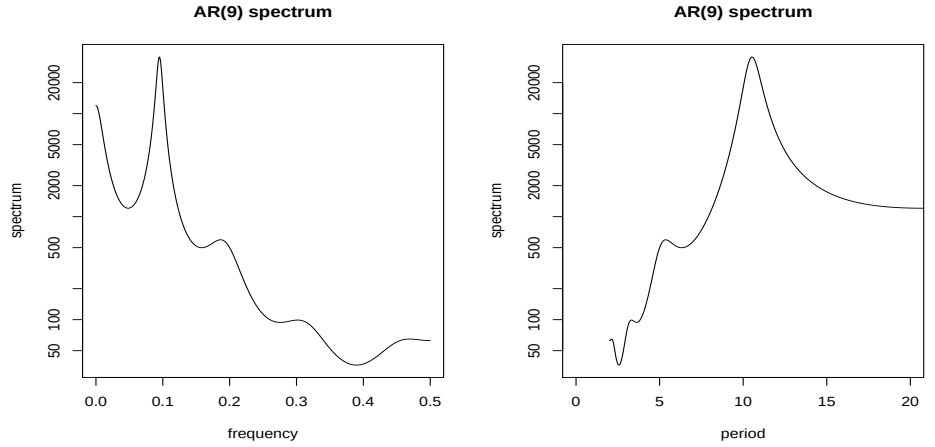
Kuvio 3.4: AR(1) ja MA(1)-prosessien spektritiheysfunktioita

Kuviossa 3.5 on piirretty AR(2)-funktion spektrejä eri tapauksissa. Ensimmäisessä tapauksessa, kun polynomilla on kaksi kompleksijuurta, spektrillä on huippu taajuudella $\pi/4$. AR-prosessissa siis esiintyy periodista vaihtelua, jonka jakso on 8 ($= 2\pi/(\pi/4)$). Huomaa kuitenkin, että ARMA-prosessien tapauksessa ei ole kyse aidosta jaksollisuudesta, sillä 'aalloissa' tapahtuu vaihesiirtymää.

Periodogrammi voidaan tasoittaa myös niin, että estimoidaan sopiva AR-malli ja piirretään sitä vastaava spektritiheysfunktio. Lopputulos voi olla kuitenkin harhaanjohtava, jos malli on identifioitu väärin. Kuviossa 3.6 on piirretty sunspot.year-sarjalle periodogrammi tällä menetelmällä.



Kuvio 3.5: AR(2)-prosessien spektritiheysfunktioita



Kuvio 3.6: Periodogrammi AR-menetelmällä sunspot.year-aineistolle

```
a<-spectrum(sunspot.year,method="ar",main="AR(9) spectrum")
a$freq<-1/a$freq; plot(a,xlim=c(0,20),xlab="period",main="AR(9) spectrum")
```

Luku 4

Mallintaminen ja ennustaminen ARMA-prosesseilla

4.1 Alustava estimointi Yule-Walker-yhtälöillä

Yleensä parhaana estimointimenetelmänä pidetään suurimman uskottavuuden estimointia, sillä suurimman uskottavuuden estimaateilla on hyvät suuren otoksen ominaisuudet. Lisäksi monissa erikoistapauksissa voidaan osoittaa, että SU-estimaattorilla on muihin estimaattoreihin verrattuna pieni keskineliövirhe, vaikka se ei yleensä olekaan harhaton. Aikasarjamallien tapauksessa SU-estimointi on numeerisesti raskas toimenpide, joten alustavissa tarkasteluissa käytetään usein muita estimointimenetelmiä.

Johdetaan seuraavaksi ns. Yule-Walker-yhtälöt AR(p)-mallin parametrien estimoimiseksi. Kerrotaan nollakeskisen ja kausaalisen AR(p)-prosessin $\{X_t\}$ määrittelevä yhtälö

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + Z_t$$

puolittain viivästetyllä havainnolla X_{t-k} ja otetaan odotusarvo puolittain, jolloin saadaan yhtälöt

$$\gamma(0) = \phi_1\gamma(1) + \phi_2\gamma(2) + ... + \phi_p\gamma(p) + \sigma^2,$$
$$\gamma(k) = \phi_1\gamma(k-1) + \phi_2\gamma(k-1) + ... + \phi_p\gamma(k-p), \quad k = 1, 2, ..., p.$$

Yhtälöt voidaan esittää matriisimuodossa

$$\Gamma_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p, \tag{4.1}$$
$$\sigma^2 = \gamma(0) - \boldsymbol{\phi}'\boldsymbol{\gamma}_p, \tag{4.2}$$

missä Γ_p on autokovarianssimatriisi $\Gamma_p = [\gamma(i-j)]_{i,j=1}^p$ ja $\boldsymbol{\gamma}_p = (\gamma(1), ..., \gamma(p))'$. Korvaamalla tuntemattomat kovarianssit otoksen perusteella estimoiduilla saadaan Yule-Walker -estimaattorit

$$\hat{\boldsymbol{\phi}} = \hat{\Gamma}_p^{-1}\hat{\boldsymbol{\gamma}}_p \tag{4.3}$$

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}'\hat{\boldsymbol{\gamma}}_p, \tag{4.4}$$

parametreille ϕ_j ja σ^2. (Matriisi $\hat{\Gamma}_p$ on kääntyvä, kun $\hat{\gamma}(0) > 0$; ks. Brockwell&Davis, Section 2.4.2.). Voidaan osoittaa, että saatu ratkaisu on kausaalinen (ks. TSTM, Problem 8.3). Vertaamalla ratkaisua otososittaisautokorrelaatiokertoimen määrittelevään kaavaan kappaleessa 3.4 havaitaan, että $\hat{\alpha}(p) = \hat{\phi}_p$, missä $\hat{\phi}_p$ on parametrin ϕ_p estimaattori, kun estimoidaan AR(p)-malli Yule-Walker-yhtälöillä.

Yule-Walker -estimointi perustuu momenttimenetelmään, jossa teoreettiset ja estimoidut momentit asetetaan yhtä suuriksi. Yleensä momenttiestimaattoreilla on paljon suurempi varianssi kuin vastaavilla SU-estimaattoreilla. Kuitenkin voidaan osoittaa, että suuren otoksen tapauksessa Yule-Walker estimaattorilla $\hat{\boldsymbol{\phi}}$ on sama jakauma kuin SU-estimaattorilla. Voidaan osoitaa, että suuren otoksen tapauksessa likimain $\hat{\boldsymbol{\phi}} \sim N(\boldsymbol{\phi}, \sigma^2\Gamma_p^{-1}/n)$. Yule-Walker-ratkaisu voidaan esittää myös kaavoilla

$$\hat{\boldsymbol{\phi}} = \hat{\mathbf{R}}_p^{-1}\hat{\boldsymbol{\rho}}_p \tag{4.5}$$

$$\hat{\sigma}^2 = \hat{\gamma}(0)[1 - \hat{\boldsymbol{\rho}}'\hat{\mathbf{R}}_p^{-1}\hat{\boldsymbol{\rho}}_p], \tag{4.6}$$

missä $\hat{\mathbf{R}}_p = \hat{\Gamma}_p/\hat{\gamma}(0)$ on otosautokorrelaatiomatriisi ja $\hat{\boldsymbol{\rho}}_p = (\hat{\rho}(1), ..., \hat{\rho}(p))' = \hat{\boldsymbol{\gamma}}_p/\hat{\gamma}(0)$. Lisäksi keskistetyn aikasarjan $\{X_t\}$ otosautokovarianssimatriisi voidaan esittää matriisilausekkeena

$$\hat{\Gamma}_p = \mathbf{X}'\mathbf{X}/n \tag{4.7}$$

ja estimaattori $\hat{\boldsymbol{\phi}}$ muodossa

$$\hat{\boldsymbol{\phi}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{4.8}$$

missä

$$\mathbf{X} = \begin{pmatrix} x_1 & 0 & ... & 0 & 0 \\ x_2 & x_1 & ... & 0 & 0 \\ x_3 & x_2 & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & x_n & x_{n-1} \\ 0 & 0 & ... & 0 & x_n \end{pmatrix} \quad \text{ja} \quad \mathbf{y} = \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Yule-Walker-ratkaisu siis vastaa tietynlaisen regressiomallin estimointia. Ratkaisu riippuu ainoastaan korrelaatioista ja kovariansseista, joihin aikasarjan keskiarvolla ei ole siihen vaikutusta. Aikasarja voidaan aina keskistää vähentämällä siitä keskiarvo. Lausekkeesta (4.7) voidaan havaita, että otos-autokovarianssimatriisi on aina ei-negatiivisesti definiitti.

4.2 Alustava estimointi Hannan-Rissanen-algoritmilla

Hannan-Rissanen -algoritmia voidaan käyttää alustavaan ARMA(p,q)-prosessin estimointiin. Algoritmin idea on siinä, että invertoituvan ARMA-prosessin tapauksessa virheprosessilla $\{Z_t\} \sim \text{WN}\ (0, \sigma^2)$ on esitysmuoto

$$Z_t = X_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + ...$$

Estimoidaan AR(m)-malli, missä m on suurehko luku ($m > \max(p, q)$) ja oletetaan, että kertoimet π_j eivät ole merkittäviä, kun $j > m$. Luku m voidaan valita niin, että se minimoi Akaiken informaatiokriteerin. R-ohjelman funktio `ar` tekee tämän automaattisesti. Estimoinnista jäävää residuaalisarjaa voidaan sitten käyttää valkoisen kohinan $\{Z_t\}$ estimaattina.

Tämän jälkeen estimoidaan ARMA(p,q)-mallin parametrit regressiomallin

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \theta_1 \hat{Z}_{t-1} + ... + \theta_q \hat{Z}_{t-q} + Z_t$$

avulla. Parametrivektorin $\boldsymbol{\beta} = (\phi_1, ...\phi_p, \theta_1, ..., \theta_q)'$ pienimmän neliösumman estimaattori (OLSE) voidaan antaa matriisimuodossa

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

missä $\mathbf{y} = (X_{m+1+q}, ..., X_n)'$ ja $\mathbf{X}$ on $(n-m-q) \times (p+q)$-matriisi

$$\mathbf{X} = \begin{pmatrix} X_{m+q} & X_{m+q-1} & ... & X_{m+q+1-p} & \hat{Z}_{m+q} & \hat{Z}_{m+q-1} & ... & \hat{Z}_{m+1} \\ X_{m+q+1} & X_{m+q} & ... & X_{m+q+2-p} & \hat{Z}_{m+q+1} & \hat{Z}_{m+q} & ... & \hat{Z}_{m+2} \\ \vdots & \vdots & ... & \vdots & \vdots & \vdots & ... & \vdots \\ X_{n-1} & X_{n-2} & ... & X_{n-p} & \hat{Z}_{n-1} & \hat{Z}_{n-2} & ... & \hat{Z}_{n-q} \end{pmatrix}.$$

Hannan-Rissanen -algoritmiin sisältyy vielä kolmaskin vaihe, mutta sitä ei tässä esitetä. Akaiken informaatiokriteerin laskeminen alustavien estimaattien avulla on epävarmaa ARMA(p,q)-mallin tapauksessa ($q > 0$), joten mallien vertailu AIC-kriteerin perusteella kannattaa tehdä SU-estimoinnin jälkeen.

Alla on annettu funktio hanris, jonka avulla voidaan algoritmia soveltaa.

```
hanris <- function(x,phi,theta){
# Hannan-Rissanen -algoritmi ilman 3. askelta
# Käyttö:
# hanris(series,1:3,1:5) kun sovitetaan sarjaan "series" arma(3,5)-mallia

   p <- max(phi)
   q <- max(theta)
   n <- length(x)
   y <- ar(x)
   if (y$order < max(p,q)+1)
      y <- ar(x,F,max(p,q)+1)
   m <- y$order
   X <- embed(x,p)
   Z <- embed(y$resid,q)
   lm(x[(m+q+1):n]~  cbind(X[(m+q-p+1):(n-p),phi],Z[(m+1):(n-q),theta])) }
```

Esim. malli

$$X_t = \delta + \phi_1 X_{t-1} + \phi_5 X_{t-5} + \theta_2 Z_{t-2} + \theta_3 Z_{t-3} + Z_t, \quad Z_t \sim \text{WN}\ (0, \sigma^2)$$

voidaan estimoida "treering-aineistolle käskyllä

`tulos <- hanris(treering,c(1,5),c(2,3))`.

Enemmän tuloksia saa käskyllä

`summary(tulos)`.

4.3 Suurimman uskottavuuden estimointi ARMA-mallille

Oletetaan, että $\{X_t\}$ on 0-keskinen gaussinen aikasarja, jonka autokovarianssifunktio on $\gamma(i,j) = \mathsf{E}(X_i X_j)$. (Gaussisuus tarkoittaa sitä, että kaikilla $n = 1, 2, ...$ prosessin havainnoista muodostetut n-ulotteiset jakaumat noudattavat n-ulotteista normaalijakaumaa). Olkoon aikasarjan havaintovektori $\mathbf{X}_n = (X_1, ..., X_n)'$ ja olkoon kovarianssimatriisi $\Gamma_n = \mathsf{E}(\mathbf{X}_n \mathbf{X}_n')$ on epäsingulaarinen. Tällöin uskottavuusfunktio voidaan esittää muodossa

$$L(\Gamma_n) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \exp\left(-\frac{1}{2} \mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n \right). \tag{4.9}$$

Matriisien Γ_n ja Γ_n^{-1} muodostaminen on kuitenkin numeerisesti raskas operaatio. Käyttämällä hyväksi ehdollisia jakaumia uskottavuusfunktio voidaan esittää muodossa

$$L(\Gamma_n) = \prod_{j=1}^{n} f_j(X_j | X_{j-1}, ..., X_1) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\nu_{j-1}}} \exp\left(-\frac{1}{2} (X_j - \hat{X}_j)^2 / \nu_{j-1} \right),$$

missä $\hat{X}_1 = 0$, $\hat{X}_j = \mathsf{E}(X_j | X_1, ..., X_{j-1}) = \mathsf{P}_{j-1} X_j$, ovat yhden askelen ennusteita ja $\nu_0 = \mathsf{Var}(X_1)$, $\nu_{j-1} = \mathsf{Var}(X_j | X_1, ..., X_{j-1})$, $j \geq 2$ ovat yhden askelen ennustevirheiden eli innovaatioiden variansseja.

Kun $\{X_t\}$ on ARMA-prosessi, joka toteuttaa yhtälön $\phi(\mathsf{B}) X_t = \theta(\mathsf{B}) Z_t$, $Z_t \sim \mathrm{WN}\ (0, \sigma^2)$, ennustevirheen varianssi voidaan esittää muodossa $\nu_j = \sigma^2 r_n$, missä r_n riippuu parametreista $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)$ ja $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)$ muttei parametrista σ^2. Voidaan osoittaa (TSTM, Problem 5.6) että $r_n \to 1$, kun $n \to \infty$, jos prosessi on invertoituva. Uskottavuusfunktio voidaan siis kirjoittaa myös muodossa

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \cdots r_{n-1}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (X_j - \hat{X}_j)^2 / r_{j-1} \right). \tag{4.10}$$

Yhden askelen ennusteet $\hat{X}_j$ ja niiden varianssit ν_j voidaan tehokkaasti laskea käyttäen joko *innovaatioalgoritmia* tai *Kalman rekursioita*, joita on kuvattu mm. Brockwellin ja Davisin kirjassa.

Koska lausekkeessa (4.10) termit r_j ja innovaatiot $X_j - \hat{X}_j$ eivät riipu parametrista σ^2, logaritmoitu uskottavuusfunktio voidaan maksimoida de-

rivoimalla se aluksi parametrin σ^2 suhteen ja asettamalla derivaatta nollaksi (harjoitustehtävä). Sijoittamalla saatu σ^2:n ratkaisu uskottavuusfunktioon (4.10), saadaan suurimman uskottavuuden estimaattorit

$$\hat{\sigma}^2 = n^{-1} S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}), \tag{4.11}$$

missä

$$S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n} (X_j - \hat{X}_j)^2 / r_{j-1} \tag{4.12}$$

ja $\hat{\boldsymbol{\phi}}$, $\hat{\boldsymbol{\theta}}$ ovat parametrien $\boldsymbol{\phi}$, $\boldsymbol{\theta}$ arvot, jotka maksimoivat lausekkeen

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}) = -\frac{n}{2} \ln(S(\boldsymbol{\phi}, \boldsymbol{\theta})) - \frac{1}{2} \sum_{j=1}^{n} \ln r_{j-1}. \tag{4.13}$$

Suuren otoksen tapauksessa suurimman uskottavuuden estimaattori $\hat{\boldsymbol{\beta}}$ parametrivektorille noudattaa likimain normaalijakaumaa $N(\boldsymbol{\beta}, -\mathbf{H}^{-1}(\boldsymbol{\beta}))$, missä $\mathbf{H}$ on Hessin matriisi $[\partial^2 l(\boldsymbol{\beta})/\partial\beta_i\beta_j]_{i,j=1}^n$. Tämän asymptootisen jakauman avulla voidaan määrittää luottamusvälejä ja -alueita tuntemattomille parametreille. Tietokoneohjelma, joka hakee suurimman uskottavuuden ratkaisun, määrittää myös matriisin $\mathbf{H}$ numeerisesti pisteessä $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ ja antaa siihen perustuvat keskivirheet parametreille.

Estimaattorin $\hat{\boldsymbol{\beta}}$ asymptoottiselle jakaumalle olemassa yleinen kaava ARMA-mallien tapauksessa (ks. TSTM, Section 8.8). AR(p)-mallin tapauksessa likimain $\hat{\boldsymbol{\phi}} \sim N(\boldsymbol{\phi}, \sigma^2 \Gamma_p^{-1}/n)$.

4.4 Mallinvalinta informaatiokriteerien avulla

ARMA-mallien hyvyyttä voidaan vertailla käyttäen ns. Akaiken informaatiokriteeriä. Akaiken informaatiokriteeri on estimoitu Kullback-Leibler-etäisyys oikean mallin ja ehdokkaana olevan mallin välillä. Oletetaan, että sekä oikea malli että ehdokkaana olevat mallit perustuvat moniulotteiseen normaalijakaumaan. Valitaan sellainen malli, jonka Akaiken kriteeri on pienin. Akaiken kriteeri voidaan ilmoittaa uskottavuusfunktion $L(\boldsymbol{\beta}, \sigma^2)$ ja parametrien lukumäärän avulla, joka ARMA(p,q)-mallin tapauksessa on $p + q + 1$. Akaiken kriteerillä (AIC) on taipumus valita malli, jossa on liian paljon parametreja. Kriteerille on kehitetty myös korjattu versio (AICC)

josta harha on poistettu. Lisäksi käytetään bayesiläistä informaatiokriteeriä (BIC), joka rankaisee enemmän parametrien lukumäärästä kuin AIC tai AICC. Kriteerit AIC ja AICC saadaan laskettua kaavoista

$$\text{AICC} = -2\ln L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + 2(p+q+1)n/(n-p-q-2) \tag{4.14}$$

$$\text{AIC} = -2\ln L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + 2(p+q+1) \tag{4.15}$$

missä $\hat{\boldsymbol{\beta}}$ ja $\hat{\sigma}$ ovat SU-estimaatteja.

R-ohjelman funktio ar etsii parhaan AR-mallin käyttäen AIC-kriteeriä. Seuraavassa esitetään funktio arc, joka perustuu funktioon ar, mutta käyttää AIC-kriteerin sijasta AICC-kriteeriä.

```
arc<-function(x,aic=TRUE,order.max=NULL){
  y <- ar(x,aic,order.max)
  p <- 0:y$order.max
  n <- y$n.used
  aicc <- y$aic+2*(p+1)*(p+2)/(n-p-2)
  order <- min(p[aicc==min(aicc)])
  if(aic && order > 0)
    y <- ar(x,F,order)
  if(aic && order ==0)
  {
    y$order <- 0
    y$ar <- numeric(0)
    y$var.pred <- var(x)
    y$resid <- x-mean(x)
  }
  y$aic <- aicc-min(aicc)
  y
}
```

4.5 Mallin sopivuuden tarkistaminen ja ennustaminen

Yleensä pitäisi pyrkiä valitsemaan sellainen malli, joka on mahdollisimman yksinkertainen ja sisältää vähän parametreja. Toisaalta mallilla pitäisi olla riittävä yhteensopivuus aineiston kanssa. Parametrien määrän kasvaessa yhteensopivuus kasvaa, mutta toisaalta parametrien estimoinnin tarkkuus

huononee. Jos parametrien estimaatit ovat huonoja, ennustaminen käy epäluotettavaksi. Sopivaa ARMA-mallia valitessa voidaan käyttää informaatiokriteerejä (AIC, AICC, BIC) jotka ottavat huomioon mallin yhteensopivuuden aineiston kanssa mutta myös rankaisevat liioista parametreista. Malli tulisi valita niiden mallien joukosta, jotka antavat pienimmän informaatiokriteerin arvon. Pienellä erolla AICC-kriteerin arvossa (n. 2) ei ole oleellista merkitystä, kun halutaan valita yksinkertaisin malli.

Ennen mallin hyväksymistä tulisi tarkastella jäännössarjaa. ARMA-mallin tapauksessa jäännökset lasketaan kaavalla

$$\hat{W}_t = (X_t - \hat{X}_t(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}))/(r_{t-1}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})^{1/2}$$

missä on käytetty samaa merkintää kuin kappaleessa 4.3. Jäännössarja on estimaatti valkoisen kohinan sarjalle, joten sillä pitäisi olla samat ominaisuudet kuin valkoisella kohinalla, jos malli on hyvä. Jäännössarjan tutkiminen kannattaa aloittaa sen kuvaajan piirtämisellä, jolloin siitä voidaan havaita mahdollinen trendi, sykliset komponentit ja varianssin vaihtelu. Jäännössarjan hyvyyttä voidaan myös testata käyttämällä kappaleessa 1.9 kuvattuja tapoja. Lisäksi voidaan tutkia, mikä on informaatiokriteerin mielessä paras AR(p)-malli jäännössarjalle (esim. ar-funktion avulla R:ssä). Tällöin pitäisi päätyä AR(0)-malliin (WN $(0, \sigma^2)$, mikä osoittaa, ettei jäännössarjassa ole jäljellä autokorrelaatiorakennetta.

Parhaan mallin löytymisen jälkeen voidaan tehdä ennuste, kun tuntemattomat ARMA-mallin parametrit on korvattu estimoiduilla. Tällöin on hyvä ottaa huomioon, että todellinen ennustevirhe voi olla suurempi kuin teoreettinen ennustevirhe, mikä aiheutuu parametrien estimointivirheestä. Brockwellin ja Davisin kirjassa on tarkasteltu yhden askelen ennustevirhettä AR(1)-mallin tapauksessa. Teoreettinen keskineliövirhe on σ^2 ja todellinen likimäärin $\sigma^2(1 + 1/n)$. Havaitaan, että näiden ero pienenee, kun otoskoko n kasvaa.

4.6 ARMA-prosessin asymptoottinen ennustaminen

Oletetaan, että ARMA(p,q)-prosessi $\{X_t\}$, joka toteuttaa differenssiyhtälön

$$\phi(\mathsf{B})X_t = \theta(\mathsf{B})Z_t, Z_t \sim \text{WN } (0, \sigma^2),$$

on kausaalinen ja invertoituva. Tällöin satunnaismuuttuja X_t voidaan esittää äärettömänä liukuvan keskiarvon prosessina

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + ... \tag{4.16}$$

ja äärettömänä autoregressiivisenä prosessina

$$X_t = Z_t - \pi_1 X_{t-1} - \pi_2 X_{t-2} - ... \tag{4.17}$$

Olkoon havainnon $\tilde{\mathsf{P}}_{\mathsf{n}} X_{n+h} = \mathsf{P}(X_{n+h}|X_n, X_{n-1}, ...)$ paras lineaarinen ennustin havaintojen $X_n, X_{n-1}, ...$ avulla. Siis teoriassa oletetaan, että tunnetaan sarja äärettömään menneisyyteen asti. Soveltamalla ennusteoperaattoria yhtälön (4.17) molempiin puoliin saadaan palautuskaava ennusteiden laskemiseksi:

$$\tilde{\mathsf{P}}_{\mathsf{n}} X_{n+h} = -\sum_{j=1}^{\infty} \pi_j \tilde{\mathsf{P}}_{\mathsf{n}} X_{n+h-j}. \tag{4.18}$$

Yhden askelen ennuste $\tilde{\mathsf{P}}_{\mathsf{n}} X_{n+1}$ voidaan laskea, sillä yksinkertaisesti $\tilde{\mathsf{P}}_{\mathsf{n}} X_{n+1-j} = X_{n+1-j}$, kun $j = 1, 2, ...$, sillä ajanhetkellä $t = n$ havainnot X_{n+1-j} jo tunnetaan. Tämän jälkeen voidaan laskea kahden askelen ennuste jne. Käytännössä sarjasta $\{X_t\}$ tunnetaan vain havainnot $X_n, X_{n-1}, ..., X_1$, jonka jälkeen sarja joudutaan katkaisemaan. Ennuste on kuitenkin hyvä, jos kertoimet π_j lähestyvät nopeasti nollaa, kun $j \to \infty$. Jos $\{X_t\}$ on AR(p)-prosessi, ennuste on itse asiassa täysin oikea, sillä $\pi_j = 0$, kun $j > p$.

Ennustevirhe voidaan laskea käyttämällä esitysmuotoa (4.16). Koska invertoituvuuden ja kausaalisuuden vuoksi havainnot $X_t, X_{t-1}, ...$ voidaan ilmoittaa havaintojen $Z_t, Z_{t-1}, ...$ avulla ja päinvastoin, saadaan

$$\begin{aligned}
\tilde{\mathsf{P}}_{\mathsf{n}} X_{n+h} &= \mathsf{P}(X_{n+h}|X_n, X_{n-1}, ...) \\
&= \mathsf{P}(Z_{n+h} + \psi_1 Z_{n+h-1} + ...|Z_n, Z_{n-1}, ...) \\
&= \mathsf{P}(Z_{n+h}|Z_n, Z_{n-1}, ...) + \psi_1 \mathsf{P}(Z_{n+h-1}|Z_n, Z_{n-1}, ...) + ... \\
&= \psi_h Z_n + \psi_{h+1} Z_{n-1} + ...,
\end{aligned}$$

joten h askelen ennustevirhe on

$$e_h = X_{n+h} - \tilde{\mathsf{P}}_{\mathsf{n}} X_{n+h} = Z_{n+h} + \psi_1 Z_{n+h-1} + ... + \psi_{h-1} Z_{n+1}$$

ja ennustevirheen varianssi

$$\tilde{\sigma}_h^2 = \sigma^2 (1 + \psi_1^2 + ... + \psi_{h-1}^2). \tag{4.19}$$

Luku 5

Epästationaaristen aikasarjojen mallinnus

5.1 Aikasarjan alustava tarkastelu

Jotta aikasarjan autokorrelaatiorakennetta voitaisiin mallintaa, se on ensin muunnettava stationaariseksi. Jos sarjassa on havaittavissa varianssin kasvua, samalla kun sen odotusarvo kasvaa, tarpeen voi olla tehdä sille alustava muunnos varianssin tasaamiseksi, kuten logaritmointi tai Box-Cox-muunnos. Tämän jälkeen sarjassa saattaa olla jäljellä trendi ja kausivaihtelukomponentti. Luvussa 1 esitettiin kaksi vaihtoehtoista menetelmää näiden poistamiseksi. Ensimmäinen tapa on vähentää sarjasta estimoitu trendifunktio ja/tai kausikomponentti. Toinen tapa on differointi, johon seuraavassa kappaleessa esiteltävä ARIMA-mallintaminen perustuu.

Alustavassa muunnoksessa tarpeellinen Box-Cox-muunnos määritellään yhtälöillä

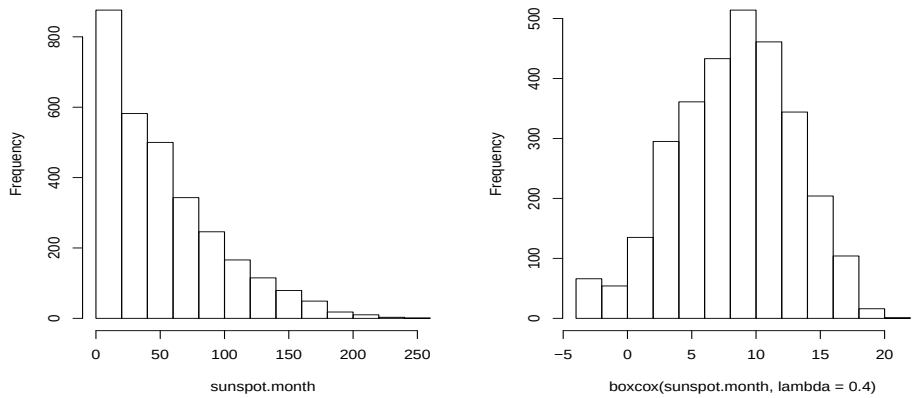
$$f_\lambda(U_t) = \begin{cases} \lambda^{-1}(U_t^\lambda - 1), & U_t \geq 0,\ \lambda > 0, \\ \ln U_t, & U_t > 0,\ \lambda = 0. \end{cases}$$

Käytännössä yleensä riittävät muunnokset, joissa $\lambda = 0$ tai $\lambda = 0.5$. Jos sarja $\{U_t\}$ menee hiukan nollan alapuolelle, asian voi korjata lisäämällä sarjaan pieni positiivinen vakio ennen Box-Cox-muunnoksen tekemistä.

Seuraavalla R-funktiolla voidaan Box-Cox -muunnos:

```
boxcox<-function(series,lambda){if(lambda>0) (series^lambda-1)/lambda
  else log(series)}
```

Kuviossa 5.1 on piirretty histogrammi aineistosta sunspot.month. Silmämääräisesti katsottuna auringonpilkkujen määrä näyttää eksponentiaalisesti jakautuneelta. Oikealla puolella on aineisto sen jälkeen, kun siihen on sovellettu Box-Cox-muunnosta parametrilla $\lambda = 0.4$, ja se näyttää nyt paremmin normaalijakautuneelta (alinta luokkaa lukuun ottamatta).



Kuvio 5.1: Histogrammi sunspot.month-sarjalle ennen ja jälkeen Box-Cox-muunnoksen

5.2 ARIMA-malli

Aikasarjojen epästationaarisuudella voi olla monta syytä. Aikasarja on epästationaarinen jos siihen sisältyy trendi tai kausikomponentteja. Trendin ja kausikomponenttien poistamisenkin jälkeen aikasarja on epästationaarinen, jos sitä voidaan kuvata ARMA-yhtälöllä (3.1), jossa autoregressiivisellä polynomilla $\phi(z)$ yksikköjuuria eli juuria, jotka sijaitsevat yksikköympyrällä kompleksitasossa. Mikäli aikasarja voidaan kuvata kausaalisella ARMA(p,q)-mallilla sen jälkeen, kun sitä on differoitu d kertaa, alkuperäistä sarjaa vastaa ns. ARIMA(p,d,q)-malli.

Määritellään, että $\{X_t\}$ on ARIMA(p,d,q)-prosessi, jos sitä kuvaa differenssiyhtälö

$$\phi(\mathsf{B})(1-\mathsf{B})^d X_t = \theta(\mathsf{B})Z_t, \quad Z_t \sim \text{WN}\ (0,\sigma^2), \tag{5.1}$$

missä $\phi(z)$ on p-asteinen polynomi, jolla ei ole juuria, kun $|z| \leq 1$, ja $\theta(z)$ on q-asteinen polynomi. Jos aikasarjan odotusarvo ei ole 0 vaan μ, kun sitä on differoitu d kertaa, sitä voidaan kuvata yhtälöllä

$$\phi(\mathsf{B})(\nabla^d X_t - \mu) = \theta(\mathsf{B})Z_t, \quad Z_t \sim \text{WN } (0, \sigma^2), \tag{5.2}$$

tai yhtälöllä

$$\phi(\mathsf{B})\nabla^d X_t = \delta + \theta(\mathsf{B})Z_t, \quad Z_t \sim \text{WN } (0, \sigma^2), \tag{5.3}$$

missä $\delta = \phi(\mathsf{B})\mu = \mu(1 - \phi_1 - \phi_2 - ... - \phi_p)$.

On huomattava, että yhtälöt (5.1), (5.2) ja (5.3) eivät määrittele alkuperäisen sarjan $\{X_t\}$ odotusarvoa, kun $d > 0$, sillä differoitaessa se häviää. On selvää, että ARIMA-mallin estimointi voidaan palauttaa ARMA-mallin estimointiin. Parametrivektorit $\boldsymbol{\phi}$, $\boldsymbol{\theta}$ ja varianssi σ^2 voidaan estimoida differoidusta aikasarjasta $\nabla^d X_t$ käyttäen luvussa 4 kuvattuja menetelmiä sen jälkeen, kun se on tarpeen vaatiessa keskistetty. Myös ennustaminen voidaan palauttaa ARMA-prosessien ennustamiseen käyttäen myöhemmin kuvattavaa menetelmää.

5.3 Yksikköjuuren olemassaolon testaaminen

Epästationaarista aikasarjaa ei voida aina muuntaa stationaariseksi vähentämällä siitä deterministinen trendi tai kausikomponentti. Jos aikasarjaa voidaan kuvata ARMA-yhtälöllä (3.1), se on epästationaarinen, jos autoregressiivisellä polynomilla $\phi(z)$ on nollakohtia yksikköympyrällä. Jos polynomilla $\phi(z)$ on nolla kohdassa $z = 1$, ns. yksikköjuuri, tämä ilmenee niin, että otosautokorrelaatiofunktio vähenee hitaasti ykkösestä. Itse aikasarja näyttää satunnaiskävelytyyppiseltä, jolloin sen varianssi kasvaa ajan myötä eikä sarjalla näytä olevan kiinteää keskiarvoa.

Yksikköjuuren olemassaoloa voidaan testata laajennetulla Dickey-Fuller-testillä (augmented Dickey-Fuller test). Tarkastellaan havaintoja $X_1, X_2, ..., X_n$ AR(1)-prosessista

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + Z_t, \quad \{Z_t\} \sim \text{WN } (0, \sigma^2), \tag{5.4}$$

missä $|\phi_1| < 1$ ja $\mu = \mathsf{E}(X_t)$. Yhtälö (5.4) voidaan myös esittää muodossa

$$\nabla X_t = X_t - X_{t-1} = \phi_0^* + \phi_1^* X_{t-1} + Z_t, \quad Z_t \sim \text{WN } (0, \sigma^2), \tag{5.5}$$

missä $\phi_0^* = \mu(1-\phi_1)$ ja $\phi_1^* = \phi_1 - 1$. Olkoon $\hat{\phi}_1^*$ tavallinen pienimmän neliösumman estimaattori (OLSE) parametrille ϕ_1^*, kun differenssiä ∇X_t selitetään viivästetyllä havainnolla X_{t-1} ja vakiolla 1. Estimaattorin estimoitu keskivirhe on

$$\mathrm{SE}(\hat{\phi}_1^*) = S/\left(\sum_{t=2}^{n}(X_{t-1}-\bar{X})^2\right)^{1/2},$$

missä $S^2 = \sum_{t=2}^{n}(\nabla X_t - \hat{\phi}_0^* - \hat{\phi}_1^* X_{t-1})^2/(n-3)$ ja $\bar{X}$ on havaintojen $X_1, X_2, ..., X_n$ otoskeskiarvo. Testisuureen

$$\hat{\tau}_\mu = \hat{\phi}_1^*/\mathrm{SE}(\hat{\phi}_1^*) \tag{5.6}$$

asymptoottisen jakauman 0.01-, 0.05- ja 0.1-kvantiilit ovat -3.43, -2.86 ja -2.57, jos $\phi_1 = 1$. Testisuuretta voidaan siis käyttää hypoteesin $H_0: \ \phi_1 = 1$ testaamiseen hypoteesia $H_1: \ \phi_1 < 1$ vastaan. Nollahypoteesi hylätään esim. riskitasolla 0.05, jos $\hat{\tau}_\mu < -2.86$. Jos nollahypoteesi hylätään, se merkitsee sitä, ettei sarjaa $\{X_t\}$ tarvitse differoida stationaarisuuden saavuttamiseksi. Huomaa, että testisuureen kriittiset arvot poikkeavat vastaavista tavallisen t-testisuureen arvoista.

Edellä kuvattua testiä voidaan käyttää myös AR(p)-prosessin

$$X_t - \mu = \phi_1(X_{t-1}-\mu) + ... + \phi_p(X_{t-p}-\mu) + Z_t, \quad \{Z_t\} \sim \mathrm{WN}\ (0, \sigma^2),$$

tapauksessa. Malli voidaan kirjoittaa muodossa (ks. harjoitustehtävä)

$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + \phi_2^* \nabla X_{t-1} + ... + \phi_p^* \nabla X_{t-p+1} + Z_t \tag{5.7}$$

missä $\phi_0^* = \mu(1-\phi_1-...-\phi_p)$, $\phi_1^* = \sum_{i=1}^{p}\phi_i - 1$ ja $\phi_j^* = -\sum_{i=j}^{p}\phi_i, \ j = 2, ..., p$. Jos autoregressiivisellä polynomilla $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$ on nolla kohdassa $z = 1$, niin $\phi_1^* = 0$ ja differoitu sarja $\{\nabla X_t\}$ on AR(p-1)-prosessi. Hypoteesi, että polynomilla $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$ on nollakohta $z = 1$ on siis yhtäpitävää hypoteesin $H_0: \ \phi_1^* = 0$ kanssa. Tätä hypoteesia voidaan testata samalla tavoin kuin AR(1)-mallin tapauksessa. Estimoidaan käyttäen tavallista pienimmän neliösumman regressiota selittämällä differenssiä ∇X_t muuttujilla $1, X_{t-1}, \nabla X_{t-1}, ..., \nabla X_{t-p+1}$. Kun n on suuri ja hypoteesi H_0 tosi, suhde (5.6) noudattaa samaa jakaumaa kuin AR(1)-mallin tapauksessa.

Testaaminen onnistuu kätevästi kirjastoon urca sisältyvällä funktiolla ur.df. Esim. LakeHuron-aineiston tapauksessa nähdään, ettei aineistoa

tarvitse differoida. Tulostuksen testisuure tau2 on kaavan (5.6) $\hat{\tau}_\mu$. Testisuuretta phi1 käytetään yhteishypoteesin $\phi_0^* = \phi_1^* = 0$ testaamiseen (ks. Dickey, D.,Fuller, W.A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root, *Econometrica* 49, 1057–72.

```
summary(ur.df(LakeHuron, type = "drift", lags = 1, selectlags = "Fixed"))

###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71212 -0.42575  0.00349  0.43147  1.69043

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.94994   32.06259   3.897 0.000183 ***
z.lag.1      -0.21584    0.05538  -3.898 0.000183 ***
z.diff.lag    0.23757    0.09714   2.446 0.016337 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6846 on 93 degrees of freedom
Multiple R-squared: 0.156,Adjusted R-squared: 0.1379
F-statistic: 8.596 on 2 and 93 DF,  p-value: 0.0003754


Value of test-statistic is: -3.8977 7.6333

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.51 -2.89 -2.58
phi1  6.70  4.71  3.86
```

Mallia (5.7) voidaan yleistää sisällyttämällä niihen lineaarinen aikatrendi:

$$\nabla X_t = \phi_0^* + \beta t + \phi_1^* X_{t-1} + \phi_2^* \nabla X_{t-1} + ... + \phi_p^* \nabla X_{t-p+1} + Z_t \qquad (5.8)$$

Vaihtoehtoisena hypoteesina on nyt, että prosessi on lineaarisen trendin ja stationaarisen vaihtelun summa. Nollahypoteesina on edelleen, että $\phi 1* = 0$, jolloin $\{X_t\}$ on ARIMA(p-1,1,0)-prosessi mahdollisesti lisättynä neliöllisellä trendillä, jos β poikkeaa nollasta. Testattaessa nollahypoteesia estimoidaan (5.8) OLS-menetelmällä. Nyt kuitenkin parametriin liittyvän t-testisuureen τ_τ jakauma poikkeaa aiemmin käytetyn $\tau\mu$:n jakaumasta. Testaus onnistuu jälleen funktiolla ar.df, kun asetetaan type="trend". Kiinteän viivepituuden lisäksi sopiva viivepituus voidaan valita AIC- ja BIC-kriteereiden avulla.

Kirjastoon urca sisältyy myös Phillips-Perron testi (ur.pp) ja Kwiatkowski-Phillips-Schmidt-Shin (KPSS) -testi (ur.kpss), joiden avulla stationaarisuutta voidaan testata. PP-testi on luenteeltaan epäparametrinen ja sallii yleisemmän aikasarjarakenteen kuin ARMA-prosessit. KPSS-testi eroaa ADF- ja PP-testeistä siinä, että stationaarisuus tai trendistationaarisuus on nollahypoteesi eikä vaihtoehtoinen hypoteesi. Alla on testattu GNP-sarjaa, joka sisältyy USeconomic-aineistoon. Ainoastaan siinä tapauksessa, että käytetään KPSS-testiä, ja asetetaan parametrin lshort arvoksi FALSE, saadaan se tulos, että GNP on trendistationaarinen.

```
data(USeconomic)

summary(ur.kpss(GNP, type="tau", lags="long"))
Test is of type: tau with 12 lags.

Value of test-statistic is: 0.0831

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.119 0.146  0.176 0.216
```

```
summary(ur.kpss(GNP, type="tau", lags="short"))
Test is of type: tau with 4 lags.

Value of test-statistic is: 0.134

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.119 0.146  0.176 0.216

summary(ur.df(GNP, type = "trend", selectlags = "AIC"))

Value of test-statistic is: -2.1857 10.4565 3.3009

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.99 -3.43 -3.13
phi2  6.22  4.75  4.07
phi3  8.43  6.49  5.47

summary(ur.pp(GNP,  model =  "trend", lags = "short"))

Value of test-statistic, type: Z-alpha  is: -11.5388
```

On myös mahdollista testata yksikköjuuren olemassaoloa ARMA-mallin polynomissa $\theta(z)$. Yksikköjuuren olemassaolo kertoo tässä tapauksessa sitä, että sarjaa on ylidifferoitu. Tässä tapauksessa juuren olemassaololla ei kuitenkaan ole vaikutusta stationaarisuuteen, invertoituvuuteen kylläkin. Brockwellin ja Davisin kirjassa on esitetty testi MA(1)-mallin tapauksessa.

5.4 ARIMA-prosessin ennustaminen

Oletetaan, että meillä on ARIMA(p,d,q)-prosessista $\{X_t\}$ havainnot $X_{1-d}, X_{2-d}, ..., X_{n-1}, X_n$. Differoimalla prosessia $\{X_t\}$ d kertaa saadaan ARMA(p,q)-prosessi $\{Y_t\}$, josta on havainnot $Y_1, Y_2, ..., Y_n$. Jotta ennustaminen olisi yksinkertaista, oletetaan, että havainnot $X_{1-d}, X_{2-d}, .., X_0$ ovat korreloimattomia differenssien $Y_1, Y_2, ...$ kanssa. Sarja Y_t voidaan laskea kaavalla

$$Y_t = (1 - \mathsf{B})^d X_t = X_t + \sum_{j=1}^{d} \binom{d}{j} (-1)^j X_{t-j}, \ t = 1, 2, ..., \qquad (5.9)$$

ja sarja $\{X_t\}$ voidaan palauttaa kaavalla

$$X_t = Y_t - \sum_{j=1}^{d} \binom{d}{j} (-1)^j X_{t-j}, \ t = 1, 2, ..., \tag{5.10}$$

kun tunnetaan sarja $\{Y_t\}$ ja sarjan $\{X_t\}$ alkupään havainnot $X_{1-d}, X_{2-d}, .., X_0$. Merkitään ennusteoperaattorin $\mathsf{P_n}$ avulla keskineliövirheen mielessä parasta lineaarinen ennustetta, joka perustuu havaintoihin $X_{1-d}, X_{2-d}, ..., X_{n-1}, X_n$ tai vaihtoehtoisesti havaintoihin $X_{1-d}, X_{2-d}, ..., X_0, Y_1, Y_2, ..., Y_n$. Ennuste $\mathsf{P_n}Y_{n+h}$ perustuu havaintoihin $Y_1, Y_2, ..., Y_n$, sillä havainnot $X_{1-d}, X_{2-d}, ..., X_0$ oletettiin korreloimattomiksi havaintojen Y_i, $i = 1, 2, ...$ kanssa. (Harjoitustehtävässä osoitetaan, että X_0 ei vaikuta havainnon Y_{n+1} ennustamiseen ARIMA(p,1,q)-mallin tapauksessa.)

Soveltamalla ennusteoperaattoria $\mathsf{P_n}$ yhtälön (5.10) molempiin puoliin, kun $t = n+h$, saadaan rekursiivinen kaava h askelen ennusteiden laskemiseksi:

$$\mathsf{P_n}X_{n+h} = \mathsf{P_n}Y_{n+h} - \sum_{j=1}^{d} \binom{d}{j} (-1)^j \mathsf{P_n}X_{n+h-j}. \tag{5.11}$$

Kun $h = 1$, $\mathsf{P_n}X_{n+h-j} = X_{n+h-j}$, sillä havainnot $X_n, X_{n-1}, ..., X_{1-d}$ tunnetaan hetkellä $t = n$. Kun $h = 2$, voidaan käyttää hyväksi aiemmin laskettua ennustetta $\mathsf{P_n}X_{n+1}$ jne. Ennusteet $\mathsf{P_n}Y_{n+h}$ voidaan laskea käyttämällä hyväksi mitä tahansa ARMA(p,q)-prosessin ennustamiseen soveltuvaa menetelmää.

Ennuste ja ennustevirhe voidaan laskea käyttäen kappaleessa 4.6 kuvattua asymptoottista menetelmää, jos oletetaan, että differoitu prosessi $\{Y_t\}$, $Y_t = \nabla^d X_t$, on invertoituva ja kausaalinen. Prosessille $\{Y_t\}$ voidaan laskea ennuste kaavan (4.18) avulla (kun X_t:n tilalla on Y_t). Tämän jälkeen voidaan laskea ennusteet prosessille $\{X_t\}$ käyttäen kaavaa (5.11). Ennustevirhe lasketaan niin, että esitetään X_t päättymättömänä sarjana

$$X_t = (1 + \psi_1 \mathsf{B} + \psi_2 \mathsf{B}^2 + ...)Z_t, \tag{5.12}$$

jolloin ennustevirheen varianssi saadaan kaavasta (4.19). Kertoimet ψ_j saadaan määritettyä sijoittamalla lauseke (5.12) X_t:n paikalle yhtälöön

$$(1 - \phi_1 \mathsf{B} - ... - \phi_p \mathsf{B}^p)(1 - \mathsf{B})^d X_t = \theta(\mathsf{B})Z_t, \tag{5.13}$$

ja merkitsemällä potenssien B^j kertoimet yhtä suuriksi yhtälön molemmin puolin, kun $j \geq 1$. (Harjoitustehtävänä on esimerkki edellä mainittujen kaavojen soveltamisesta).

Kuviossa 5.2 on esitetty ennuste GNP-sarjalle kahdella tavalla ja 95% ennustevälit normaalijakaumaoletuksella. (Huomaa, että GNP on jo itsessään logaritmoitu aikasarja). Ensimmäisessä tavassa oletetaan, että GNP on trendistationaarinen ja sitä voidaan kuvata mallilla

$$\text{GNP} = a_0 + a_1 t + Y_t, \quad Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + Z_t,\ Z_t \sim \text{WN}\ (0, \sigma^2).$$

Toisessa tavassa oletetaan, että GNP on differenssistationaarinen ja sille sopii malli

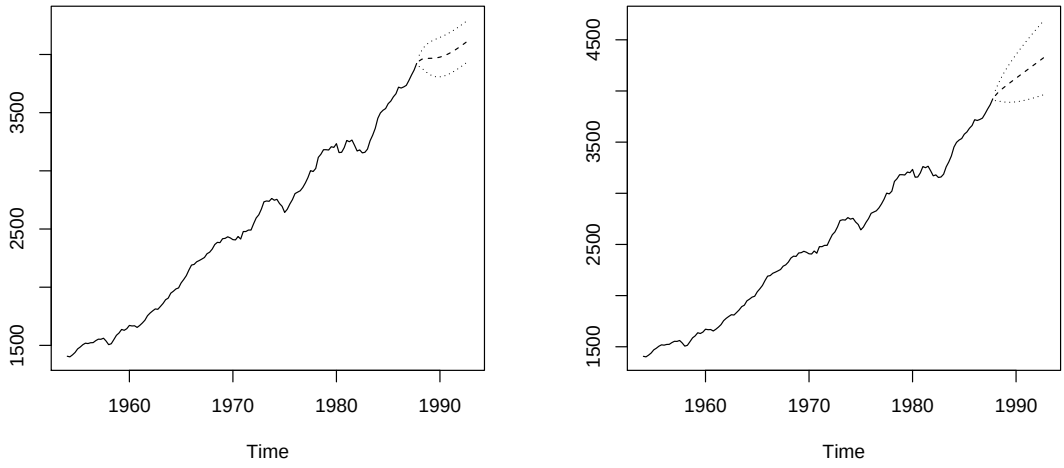
$$\nabla \text{GNP}_t = \phi_0 + \phi_1 \nabla \text{GNP}_{t-1} + \phi_2 \nabla \text{GNP}_{t-2} + Z_t, \quad Z_t \sim \text{WN}\ (0, \sigma^2).$$

```
library(tseries)
data(USeconomic)

a <- arima(GNP,c(3,0,0),xreg=1:136)
pr <- predict(a,n.ahead=20,newxreg=137:156)
ts.plot(GNP,pr$pred,pr$pred-2*pr$se,pr$pred+2*pr$se,
         gpars=list(lty=c(1,2,3,3)))

a <- arima(GNP,c(2,1,0),xreg=1:136)
pr <- predict(a,n.ahead=20,newxreg=137:156)
ts.plot(GNP,pr$pred,pr$pred-2*pr$se,pr$pred+2*pr$se,
           gpars=list(lty=c(1,2,3,3)))
```

Huomaa, että ennusteväleissä ei oteta huomioon parametrien estimoinnin epätarkkuudesta aiheutuvaa virhettä muuten kuin sillä tavoin että ennustejakauman hajonta kerrotaan 2:lla normaalijakaumaa vastaavan 1.96 sijasta! Ennustevälit ovat siis hiukan liian kapeat vaikka oletettaisiinkin, että malli on oikea. Ennustettaessa pitemmälle estimointivirheen vaikutus korostuu.

Jos oletetaan, että kyseessä on ns. paikallisen lineaarisen trendin prosessi, se vastaa korrelaatiorakenteeltaan ARIMA(0,2,2)-prosessia ja siinä 2 kertaa differoidun sarjan odotusarvo on nolla. Tällaisen prosessin ennustaminen vastaa Holt-Winters-menetelmällä laadittavaa ennustetta.

```
a <- arima(GNP,c(0,2,2))
pr <- predict(a,n.ahead=20)
ts.plot(GNP,pr$pred,pr$pred-2*pr$se,pr$pred+2*pr$se,
      gpars=list(lty=c(1,2,3,3)))
```



(a) Trendistationaarinen malli.

(b) Differenssistationaarinen malli.

Kuvio 5.2: Logaritmoitu GNP-sarja ja kaksi ennustetta.

5.5 SARIMA-malli (Kerrannainen kausivaihtelumalli)

Oletetaan, että aikasarja $\{X_t\}$, joka kuvaa kausiaineistoa, on saatettu stationaariseksi differoimalla se d kertaa viiveellä 1 ja D kertaa viiveellä s, missä s on kausien lukumäärä. Tällöin sarjaa $\{Y_t\}$, $Y_t = \nabla^d \nabla_s^D X_t$, voidaan kuvata SARIMA$(p,d,q) \times (P,D,Q)_s$ -mallilla

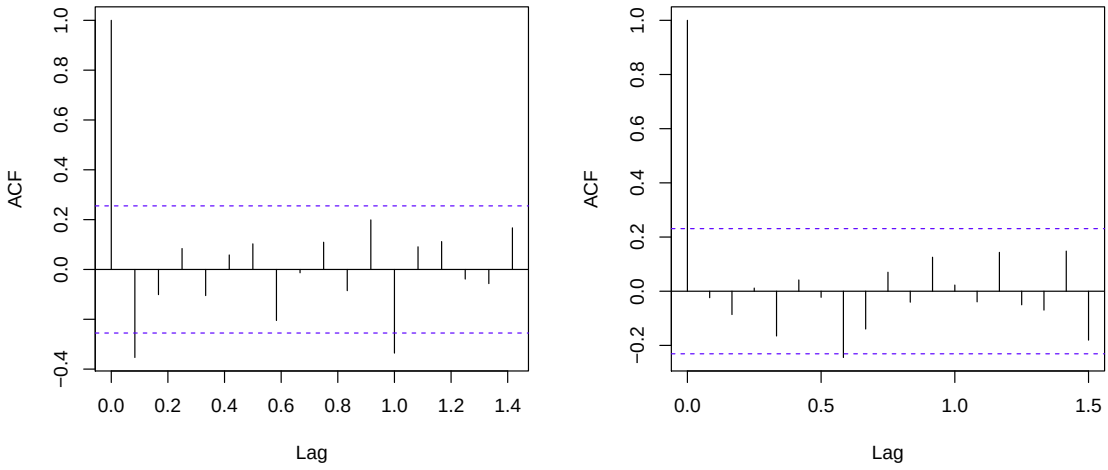
$$\phi(B)\Phi(\mathsf{B}^s)(Y_t - \mu) = \theta(\mathsf{B})\Theta(\mathsf{B}^s)Z_t, \quad Z_t \sim \text{WN } (0, \sigma^2) \tag{5.14}$$

missä $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$, $\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q$, $\Phi(z) = 1 - \Phi_1 z - ... - \Phi_P z^P$, $\Theta(z) = 1 + \Theta_1 z + ... + \Theta_Q z^Q$ ja $\mu = \mathsf{E}(Y_t)$. Polynomien $\phi(\mathsf{B})$ ja $\theta(\mathsf{B})$ avulla mallinnetaan peräkkäisten havaintojen välistä autokorrelaatiorakennetta ja polynomien $\Phi(\mathsf{B}^s)$ ja $\Theta(\mathsf{B}^s)$ avulla peräkkäisten jaksojen välistä korrelaatiorakennetta.

Esim 1. Aineistossa USAccDeaths esiintyy neliöllinen trendi ja kausivaihtelua viiveellä 12, jotka saadaan poistettua differoimalla aineisto viiveillä 1 ja 12. Kuviossa 5.3a on esitetty differoidun aineiston autokorrelaatiofunktio. Havaitaan negatiiviset piikit viiveillä 1 ja 12, jotka kertovat ylidiffe-

roinnista. Piikkien poistamiseksi voidaan sekä perättäisten kuukausien että vuosien välistä autokorrelaatiota mallintaa MA(1)-prosessilla, sillä autokorrelaatio katoaa kummassakin tapauksessa viiveen 1 jälkeen. Päädytään siis SARIMA$(0,1,1)\times(0,1,1)_{12}$-malliin aineistolle USAccDeaths. Kuviossa 5.3b on piirretty kuva jäännössarjan autokorrelaatiofunktio mallin sovituksen jälkeen. R:ssä mallin estimointi, kun oletetaan, että $\mu = 0$, toteutetaan käskyllä

```
a <- arima(USAccDeaths,c(0,1,1),seasonal=list(order=c(0,1,1)))
```



(a) Differoitu viiveillä 1 ja 12. (b) Jäännössarja.

Kuvio 5.3: Käsitellyn US accidental deaths -sarjan ACF

Malli SARIMA$(0,1,1)\times(0,1,1)_{12}$ aikasarjalle $\{X_t\}$ voidaan siis esittää muodossa

$$(1-\mathsf{B})(1-\mathsf{B}^{12})X_t-\mu=(1+\theta_1\mathsf{B})(1+\Theta_1\mathsf{B}^{12})Z_t,\quad Z_t\sim \text{WN}\ (0,\sigma^2),\ (5.15)$$

tai kun kerrotaan auki oikeanpuoleinen tulolauseke, muodossa

$$\nabla\nabla^{12}X_t-\mu=(1+\theta_1\mathsf{B}+\Theta_1\mathsf{B}^{12}+\theta_1\Theta_1\mathsf{B}^{13})Z_t,\quad Z_t\sim \text{WN}\ (0,\sigma^2),$$

josta nähdään, että sarja $\nabla\nabla_{12}X_t$ noudattaa tietynlaista MA(13)-prosessia. Vastaavasti yleinen SARIMA$(p,d,q)\times(P,D,Q)_s$ on erikoistapaus ARMA$(p+sP, q+sQ)$-mallista prosessille $\nabla^d\nabla_s^D X_t$.

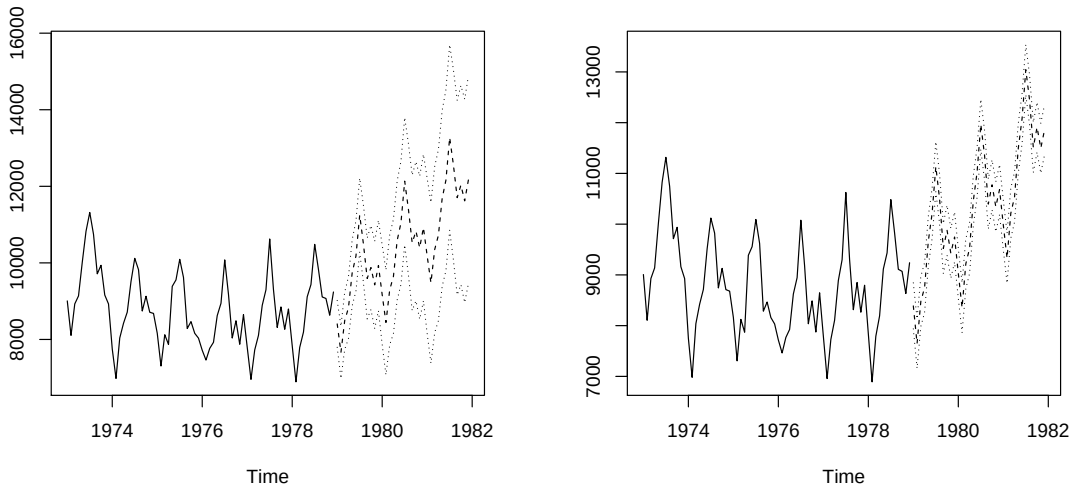
Koska SARIMA- prosessit voidaan palauttaa ARMA-prosessiksi differoidulle sarjalle, niiden ennustaminen ja ennustevälien muodostaminen voidaan tehdä käyttäen samoja periaatteita kuin ARIMA-prosessien tapauksessa. Alapuolella on laadittu kolmen vuoden ennuste ja ennusteväli USAccDeaths-sarjalle käyttäen kahta erilaista mallia. Ensimmäinen malli on (5.15) ja jälkimmäisessä kvadraattiseen trendiin ja kausikomponenttiin on lisätty AR(1)-prosessi:

$$\begin{aligned} X_t &= a_0 + a_1 t + a_2 t^2 + \gamma_1 u_{t1} + ... + \gamma_{11} u_{t,11} + Y_t \\ Y_t &= \phi Y_{t-1} + Z_t,\ Z_t \sim \text{WN}\ (0, \sigma^2). \end{aligned} \tag{5.16}$$

Kummassakin ennusteessa näkyy selvästi neliöllinen trendi (kuvio 5.4). Ennusteeseen on kuitenkin syytä suhtautua varauksella, sillä trendin jatkuminen tulevaisuudessa on epävarmaa. Ennustevälissä ei ole otettu huomioon parametrien estimaattoreiden varianssia. Mallissa (5.16) ennusteväli on hämmästyttävän kapea, mikä johtuu siitä, että toisen asteen trendifunktio istuu poikkeuksellisen hyvin alkuperäiseen aikasarjaan.

```
t <- 1:72
tnew <- 73:108

a <- arima(USAccDeaths,c(0,1,1), xreg= cbind(t,t^2),
            seasonal=list(order=c(0,1,1)))
pr <- predict(a,n.ahead=36,newxreg=cbind(tnew,tnew^2))
ts.plot(USAccDeaths,pr$pred,pr$pred+2*pr$se,pr$pred-2*pr$se,
          gpars=list(lty=c(1,2,3,3)))

U <- kronecker(rep(1,6),diag(12))
Unew <- kronecker(rep(1,3),diag(12))
a <- arima(USAccDeaths,c(1,0,0),xreg=cbind(t,t^2,U[,-12]))
pr <- predict(a,n.ahead=36,newxreg=cbind(tnew,tnew^2,Unew[,-12]))
ts.plot(USAccDeaths,pr$pred,pr$pred+2*pr$se,pr$pred-2*pr$se,
         gpars=list(lty=c(1,2,3,3)))
```

Parempi ennuste saataneen, jos oletetaan, että aikasarjassa ei ole neliöllistä trendiä vaan paikallinen lineaarinen trendi. Tämä vastaa SARIMA-mallia, jossa oletetaan viivepituuksilla 1 ja 12 differoidun sarjan odotusarvoksi 0. R tekee tämän oletuksen automaattisesti, joten ennusteen laatiminen on helppoa. Ennuste vastaa likimäärin Holt-Winters-menetelmällä laadittavaa ennustetta.



(a) Differenssistationaarinen malli.

(b) Determ. kausikomp. + trendi.

Kuvio 5.4: US accidental deaths -sarja ennusteineen.

```
a <- arima(USAccDeaths,c(0,1,1),seasonal=list(order=c(0,1,1))
pr <- predict(a,n.ahead=36)
ts.plot(USAccDeaths,pr$pred,pr$pred+2*pr$se,pr$pred-2*pr$se,
        gpars=list(lty=c(1,2,3,3)))
```

5.6 Regressioanalyysi, kun virhetermi noudattaa ARMA-prosessia

Tarkastellaan yleistä regressiomallia havainnoille

$$Y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + W_t, \quad t = 1, 2, ..., n, \tag{5.17}$$

missä havaintoja $Y_1, Y_2, ..., Y_n$ selitetään *ei-satunnaisilla* muuttujilla $x_{t1}, ..., x_{tk}$ ja virhetermillä W_t. Matriisimuodossa yhtälö (5.17) on

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{W},$$

missä $\mathbf{Y} = (Y_1, ..., Y_n)$ on selitettävien havaintojen vektori, $\boldsymbol{\beta} = (\beta_1, ...\beta_k)'$ on parametrivektori, matriisin X rivit ovat regressiovektoreita $\mathbf{x}_t = (x_{t1}, x_{t2}, ..., x_{tk})'$, $t = 1, .., n$, ja $\mathbf{W} = (W_1, ..., W_n)'$ on virhetermien vektori.

Erikoistapauksena voidaan ajatella trendin ja kausikomponenttien estimointia, kun aikasarjalle $\{Y_t\}$ halutaan tehdä klassinen hajotelma. Jos kyseessä on kuukausiaineisto, jossa kuukaudet ovat kausia, ja käytetään neliöllistä trendiä, regressiomalli voidaan esittää muodossa

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \gamma_1 u_{t1} + ... + \gamma_{11} u_{t,11} + W_t,$$

missä osoitinmuuttuja u_{tj} saa arvon 1, kun on kyseessä j:s kuukausi ja arvon 0 muuten.

Yleensä regressioanalyysissä oletetaan, että virhetermit W_t ovat korreloimattomia ja niillä on sama varianssi, toisin sanoen $W_t \sim \text{WN}\ (0, \sigma^2)$. Tällöin voidaan osoittaa, että tavallinen pienimmän neliösumman estimaattori (OLSE)

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Y} \tag{5.18}$$

on paras lineaarinen harhaton estimaattori (BLUE), missä 'paras' tarkoittaa sitä, että $\mathsf{Var}(\mathbf{c}'\hat{\boldsymbol{\beta}})$ minimoituu kaikilla vektoreilla $\mathbf{c} \in \mathbb{R}^k$. Yleisemmässä tapauksessa $\mathsf{Cov}(\mathbf{W})$ ei ole $\sigma^2 I_n$ vaan Γ_n. Tällöin BLUE-estimaatori on yleistetty pienimmän neliösumman estimaattori

$$\hat{\boldsymbol{\beta}}_{GLS} = (X'\Gamma_n^{-1}X)^{-1}X'\Gamma_n^{-1}\mathbf{Y} \tag{5.19}$$

missä oletetaan, että $\det(\Gamma_n) > 0$. Kaavan (5.19) käyttö tietysti edellyttää, että Γ_n tunnetaan. Jos oletetaan, että virhetermi $\{W_t\}$ noudattaa gaussista ARMA(p,q)-prosessia, parametrivektorille $\boldsymbol{\beta}$ ja matriisin Γ_n sisältämille parametreille saadaan suurimman uskottavuuden ratkaisu maksimoimalla SU-funktiota

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2; \mathbf{Y}) = (2\pi)^{-n/2}(\det \Gamma_n)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{Y} - X\boldsymbol{\beta})'\Gamma_n^{-1}(\mathbf{Y} - X\boldsymbol{\beta})\right). \tag{5.20}$$

Maksimointi voidaan toteuttaa iteroimalla niin, että ensin muodostetaan $\boldsymbol{\beta}$:lle alustava estimaatti kaavan (5.18) avulla, jolloin saadaan muodostettua keskistetty havaintovektori $\mathbf{Y}^* = \mathbf{Y} - X\boldsymbol{\beta}$. Tämän jälkeen lauseke (5.20) maksimoidaan numeerisesti parametrien $\boldsymbol{\phi}$, $\boldsymbol{\theta}$ ja σ^2 suhteen käyttämällä hyväksi uskottavuusfunktion innovaatioesitystä (4.10). Saadaan estimaatti matriisille Γ_n, minkä jälkeen voidaan voidaan estimoida $\boldsymbol{\beta}$ BLUE-estimaattorilla (5.19). Palataan jälleen lausekkeen (5.20) maksimointiin ja jatketaan silmukkaa, kunnes vektorin $\boldsymbol{\beta}$ arvo ei enää muutu.

Luku 6

Ehdollisen heteroskedastisuuden mallit

Monissa finanssiaikasarjoissa esiintyy hajonnan eli volatiliteetin vaihtelua, mitä mallinnetaan ehdollisen heteroskedastisuuden malleilla. Tyypillisimmin volatiliteetin vaihtelu esiintyy osakkeiden tuottosarjoissa. Yleensä tarkastellaan ns. *log-tuottoja*

$$r_t = \ln \frac{P_t}{P_{t-1}} = p_t - p_{t-1},$$

missä P_t on osakkeen hinta ajanhetkellä t ja $p_t = \ln(P_t)$.

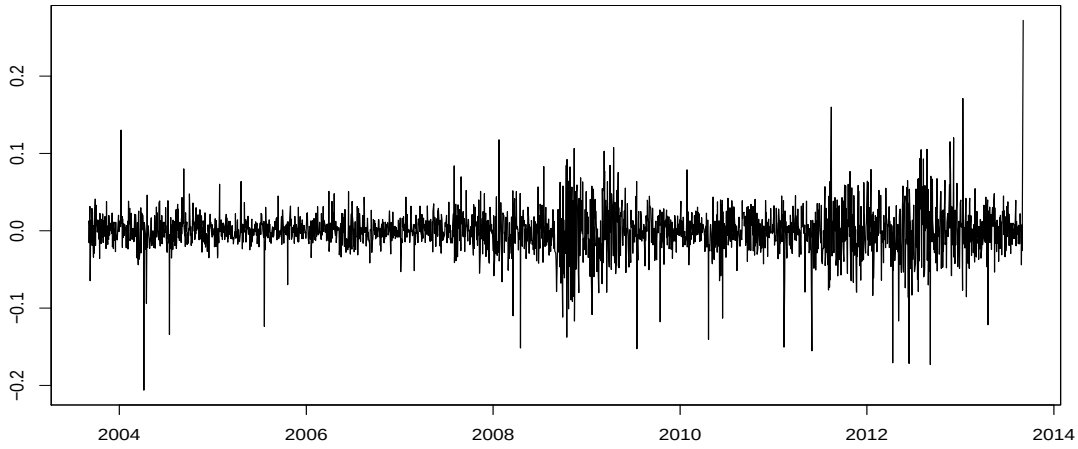
Tuottosarjojen volatiliteetin mallintaminen on erityisen kiinnostuksen kohteena mm. seuraavista syistä: 1) Optioiden hinnoittelussa käytettävä Black-Scholes-kaava perustuu tuottosarjan volatiliteettiin. 2) Volatiliteetin avulla voidaan laskea riskienhallinnassa käytettävä *value at risk* (VaR) -luku. 3) Volatiliteetillä on tärkeä osa sijoitussalkun optimoinnissa pyrittäessä maksimoimaan tuoton odotusarvoa ja minimoimaan sen varianssia. 4) Volatiliteetin huomioon ottaminen parantaa aikasarjan parametrien estimointitarkkuutta ja sitä kautta myös ennustetarkkuutta. 5) Volatiliteettiindeksistä (VIX) on tullut rahoitusväline. (Tsay, luku 3).

Volatiliteetti itsessään ei ole havaittava suure. Sen sijaan sitä voidaan pyrkiä estimoimaan monin tavoin. Sen lisäksi, että se voidaan estimoida suoraan historiallisen aikasarjan avulla, se voidaan epäsuorasti ratkaista optioiden hinnoittelukaavasta käyttämällä havaittuja optioiden hintoja. Tällöin puhutaan ns. *implisiittisestä volatiliteetistä* (implied volatility). Myös VIX-indeksi perustuu implisiittiseen volatiliteettiin.

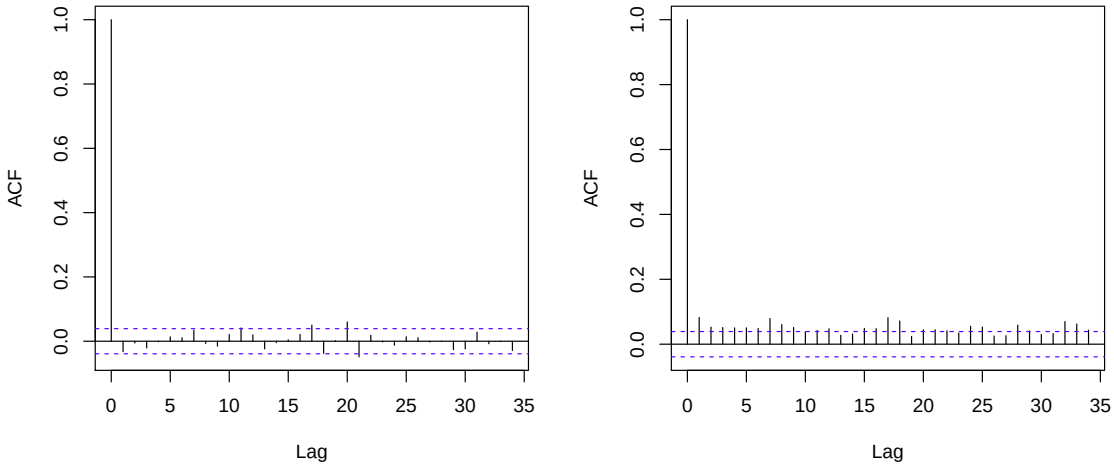
6.1 Ehdollisen heteroskedastisuuden luonnehdintaa

Volatiliteetin vaihtelulla tarkoitetaan varianssin ryvästymistä, joka voidaan havaita monien finanssiaikasarjan kuvaajasta. Ilmiötä voidaan mallintaa mm. *autoregressiivisillä ehdollisen heteroskedastisuuden* (ARCH) malleilla. Siksi puhutaan ns. *ARCH-efektistä*. Aikasarjojen peräkkäisten havaintojen välillä ei välttämättä ole autokorrelaatiota tai se voi olla hyvin heikkoa. Sen sijaan voidaan havaita havaintojen neliöiden tai itseisarvojen välinen autokorrelaatio joko ACF- ja PACF -funktioista tai erilaisten testien avulla.

Kuviossa 1.2b näytettiin Nokian osakkeen tuottosarja ajalta 4.9.2012 – 3.9.2013. Piirtämällä kuvaaja pitemmältä aikavaliltä 4.9.2003 – 3.9.2013 ARCH-efekti tulee selvästi näkyviin (kuvio 6.1). Kuviossa 6.2 on piirretty alkuperäisen ja neliöidyn tuottosarjan ACF. Vaikka havainnot ovat korreloimattomia, samaa ei voi sanoa neliöidyistä havainnoista. Myös Ljung-Box-testit osoittavat saman ilmiön.



Kuvio 6.1: Nokian osaketuottosarja 4.9.2003 – 3.9.201



(a) Havaintojen ACF (b) Neliöityjen havaintojen ACF.

Kuvio 6.2: Nokian osaketuottosarjan ACF

```
1> Box.test(x,lag=40,type="Ljung")

Box-Ljung test

X-squared = 55.5122, df = 40, p-value = 0.05232

1> Box.test(x^2,lag=40,type="Ljung")

Box-Ljung test

X-squared = 238.0425, df = 40, p-value < 2.2e-16
```

6.2 ARCH-malli

Tähän mennessä käsitellyt stationaariset aikasarjat ovat olleet kausaalisia ARMA-prosesseja, joilla on liukuvan keskiarvon esitysmuoto

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + ..., \quad Z_t \sim \text{WN } (0, \sigma^2).$$

Aikasarja $\{X_t\}$ siis on lineaarinen kombinaatio valkoisen kohinan prosessista $\{Z_t\}$. Yleisemmissä epälineaarisissa malleissa lähtökohtana on IID(0,1)-prosessi $\{Z_t\}$ (havainnot Z_t ovat samoin jakautuneita ja riippumattomia satunnaismuuttujia odotusarvolla 0 ja varianssilla 1) ja X_t voidaan ilmoittaa sarjan $\{Z_t\}$ havaintojen epälineaarisena funktiona. Kausaalisuus tässä yleisemmässä yhteydessä tarkoittaa sitä, että X_t voidaan ilmoittaa havaintojen $Z_s, s \leq t$, funktiona. Kausaalisuudesta seuraa, että Z_t on riippumaton havaintojen $X_{t-1}, X_{t-2}, ...$ kanssa.

Yleisimmin käytettyjä epälineaarisia malleja on ARCH-prosessi (Autoregressive Conditional Heteroscedasticity), jonka avulla voidaan mallintaa aikasarjan varianssia tai hajontaa (volatiliteettia). Sanotaan, että $\{X_t\}$ noudattaa ARCH(q)-prosessia, jos

$$X_t = \sigma_t Z_t, \text{ missä } Z_t \sim IID(0,1), \text{ ja} \tag{6.1}$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \alpha_2 X_{t-2}^2 + ... + \alpha_q X_{t-q}^2. \tag{6.2}$$

Jotta σ_t^2 olisi aina positiivinen, oletetaan lisäksi, että $\alpha_0 > 0$ ja $\alpha_j \geq 0,\ j = 1, \ldots, q$.

Esim. 1. *ARCH(1)-prosessi.* Kun $\{X_t\}$ on kausaalinen ja $\alpha_1 < 1$, voidaan osoittaa (ks. Brockwell&Davis) että sillä on epälineaarinen esitysmuoto

$$X_t = Z_t \sqrt{\alpha_0 \left(1 + \sum_{j=1}^{\infty} \alpha_1^j Z_{t-1}^2 ... Z_{t-j}^2\right)},$$

jonka perusteella voidaan päätellä, että $\{X_t\}$ on vahvasti stationaarinen. Lisäksi voidaan laskea, että $\mathsf{E}X_t^2 = \alpha_0/(1-\alpha_1) < \infty$, joten se on myös heikosti stationaarinen.

Eräissä sovelluksissa oletetaan myös korkeampien momenttien olemassaolo sarjalle $\{X_t\}$. Tästä seuraa lisärajoitus α_1:lle. Oletetaan seuraavassa tarkastelussa, että prosessi $\{Z_t\}$ on normaalinen, jolloin $\mathsf{E}Z_t^4 = 3$. Merkitään hetkeen t mennessä kertynyttä informaatiota $\mathcal{F}_t$, ts. havaintojen $X_t, X_{t-1}, ...$ antamaa informaatiota. Tällöin

$$\mathsf{E}(X_t^4|\mathcal{F}_{t-1}) = \mathsf{E}(\sigma_t^4 Z_t^4|\mathcal{F}_{t-1}) = \sigma_t^4 \mathsf{E}(Z_t^4|\mathcal{F}_{t-1}) = \sigma_t^4 \mathsf{E}(Z_t^4) = 3\sigma_t^4,$$

joten

$$\mathsf{E}X_t^4 = \mathsf{E}[\mathsf{E}(X_t^4|\mathcal{F}_{t-1})] = 3\mathsf{E}(\alpha_0 + \alpha_1 X_{t-1}^2)^2 = 3\mathsf{E}(\alpha_0^2 + 2\alpha_0\alpha_1 X_{t-1}^2 + \alpha_1^2 X_{t-1}^4).$$

Jos X_t on neljänteen momenttiin asti stationaarinen, ja merkitään $m_4 = \mathsf{E}X_t^4$, niin

$$\begin{aligned} m_4 &= 3[\alpha_0^2 + 2\alpha_0\alpha_1\mathsf{Var}(X_t) + \alpha_1^2 m_4] \\ &= 3\alpha_0^2\left(1 + 2\frac{\alpha_1}{1-\alpha_1}\right) + 3\alpha_1^2 m_4, \end{aligned}$$

mistä saadaan ratkaistua

$$m_4 = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

Jotta ratkaisu olisi positiivinen, on asetettava lisäehto $\alpha_1^2 < \frac{1}{3}$. Huipukkuus (kurtositeetti) on

$$\frac{\mathsf{E}X_t^4}{[\mathsf{Var}(X_t)]^2} = 3\frac{1-\alpha_1^2}{1-3\alpha_1^2} > 3.$$

Tämä on suurempi kuin normaalijakauman huipukkuus 3, joten ARCH(1)-malli selittää osittain tuottosarjoissa havaittavan huipukkuuden.

Jos ARCH(q)-prosessilla $\{X_t\}$ on äärellinen varianssi σ^2, voidaan helposti helposti osoittaa, että se on valkoisen kohinan prosessi (vaikka se ei ole IID-prosesi).

Ensinnäkin prosessin odotusarvo on 0, sillä

$$\mathsf{E}(X_t) = \mathsf{E}(\mathsf{E}(X_t|\mathcal{F}_{t-1})) = \mathsf{E}(\sigma_t\mathsf{E}(Z_t|\mathcal{F}_{t-1})) = \mathsf{E}(\sigma_t\mathsf{E}(Z_t)) = 0.$$

Toinen yhtäsuuruus seuraa siitä, että että σ_t voidaan palauttaa satunnaismuuttujiin $X_{t-1}, X_{t-2}, ..., X_{t-q}$. Kolmas yhtäsuuruus seuraa siitä, että kausaalisuuden perusteella Z_t on riippumaton informaatiosta $\mathcal{F}_{t-1}$.

Toiseksi $\sigma^2 = \mathsf{E}X_t^2 = \mathsf{E}(\sigma_t^2 Z_t^2) = \mathsf{E}(\sigma_t^2)\mathsf{E}(Z_t^2) = \mathsf{E}\sigma_t^2$, missä kolmas yhtäsuuruus perustuu siihen että σ_t^2 ja Z_t ovat riippumattomia. Riippumattomuus perustuu siihen, että σ_t^2 voidaan palauttaa satunnaismuuttujiin $X_{t-1}, X_{t-2}, ..., X_{t-q}$, jotka ovat kausaalisuuden perusteella riippumattomia Z_t:n kanssa. Ottamalla odotusarvo puolittain yhtälöstä (6.2) saadaan

$$\sigma^2 = \alpha_0 + \alpha_1\sigma^2 + ... + \alpha_q\sigma^2,$$

josta saadaan ratkaistua $\sigma^2 = \alpha_0/(1 - \alpha_1 - ... - \alpha_q)$. Havainnot X_{t+h} ja X_t ovat korreloimattomia, sillä

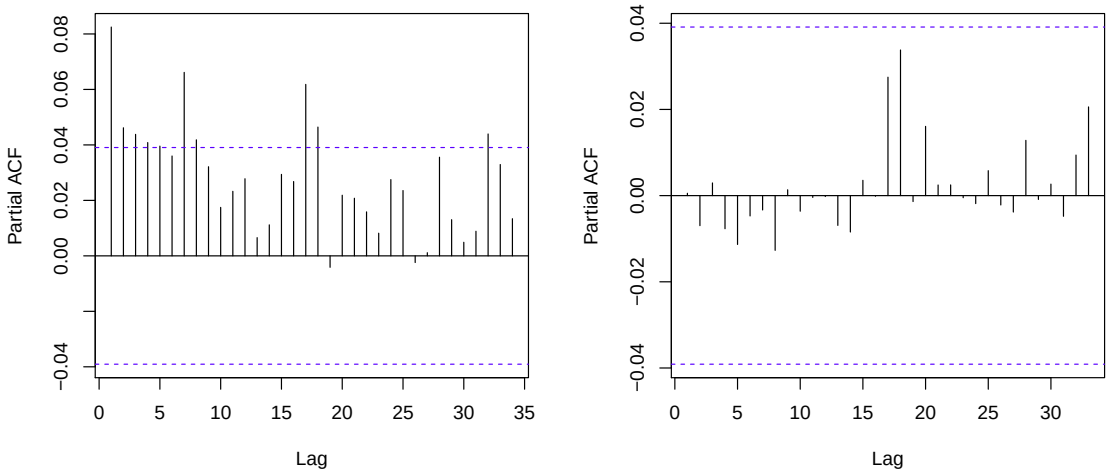
$$\mathsf{E}(X_{t+h}X_t) = \mathsf{E}(\mathsf{E}(X_{t+h}X_t|\mathcal{F}_{t+h-1})) = \mathsf{E}(X_t\sigma_{t+h}\mathsf{E}(Z_{t+h}|\mathcal{F}_{t+h-1})) = 0.$$

Lisäämällä yhtälöön (6.2) puolittain $X_t^2 - \sigma_t^2$ saadaan

$$X_t^2 = \alpha_0 + W_t + \alpha_1 X_{t-1}^2 + ... + \alpha_q X_{t-q}^2,$$

missä $W_t = X_t^2 - \sigma_t^2$. ARCH(q)-prosessi voidaan siis tulkita neliöityjen havaintojen AR(q)-prosessiksi, missä virheprosessina on $\{W_t\}$. Prosessi $\{W_t\}$ on valkoisen kohinan prosessi, mikäli parametrit $\alpha_1, ..., \alpha_q$ on rajoitettu siten, että $\mathsf{E}X_t^4 < \infty$ (harjoitustehtävä). Tämän vuoksi sopiva viivepituus q voidaan pyrkiä määrittämään tarkastelemalla havaintojen X_t^2 otososittaisautokorrelaatiofunktiota. Tämä menetelmä ei kuitenkaan ole välttämättä kovin tehokas.

Kuviossa 6.3a nähdään neliöityjen Nokia-tuottojen PACF. Osittaisautokorrelaatio näyttää häviävän viiveen 8 jälkeen, joten voisimme sovittaa ARCH(8)-mallia. Myös viiveillä 17 ja 18 näkyy merkittäviä piikkejä, mutta nämä voivat olla 'sattumaa'. Kuvion oikeassa osassa on piirretty ARCH-mallin sovittamisesta saatujen neliöityjen jäännösten PACF, ja näyttää siltä, että autokorrelaatio on hävinnyt. Sovittamisessa käytettiin tseries-paketin funktiota garch.

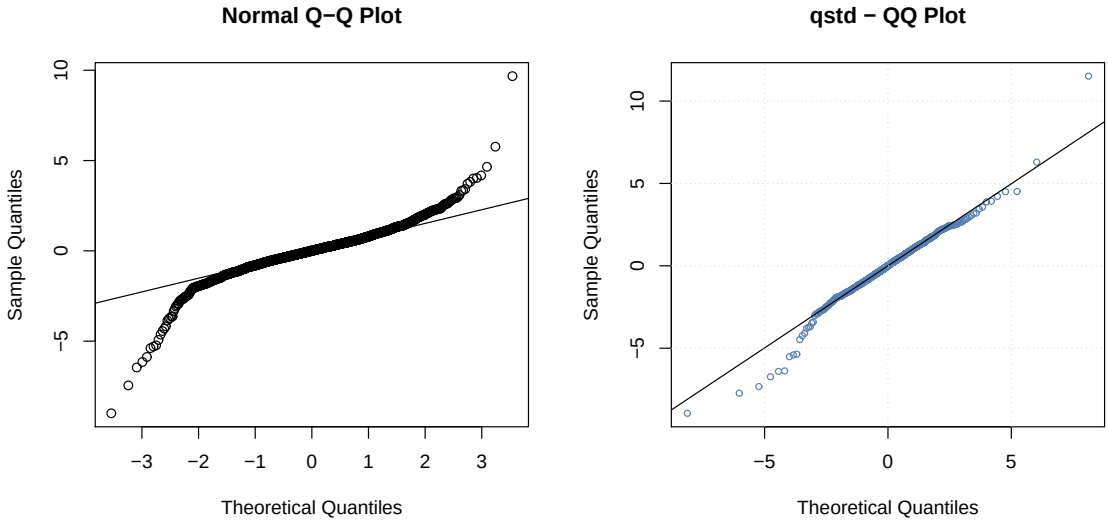


(a) Neliöityjen havaintojen PACF

(b) Neliöityjen jäännösten PACF.

Kuvio 6.3: Nokian osaketuottosarjaan liittyviä osittaisautokorrelaatiokuvioita. Jäännökset on saatu sovittamalla ARCH(8)-malli.

Kuviossa 6.4a on piirretty jäännösten QQ-kuvio. Sen perusteella ARCH-malli ei yksin selitä havaittua huipukkuutta. Teemme toisen sovituksen käyttämällä fGarch-pakettiin sisältyvää funktiota garchFit, jolloin innovaatioprosessin $\{Z_t\}$ jakaumaksi voidaan asettaa Studentin t-jakauma. Oikeanpuoleisessa QQ-kuviossa jäännösten jakaumaa verrataan estimoituun t-jakaumaan, ja nyt sopivuus näyttää paremmalta.



(a) Normaalijakauman QQ-kuvio

(b) Studentin t-jakauman QQ-kuvio.

Kuvio 6.4: Nokian osaketuottosarjaan sovitetun ARCH(8)-mallin jäännösten QQ-kuvioita

Studentin t-jakauman muotoparametrin ν estimaatiksi saadaan 3.783, mikä tarkoittaa sitä, että jakaumalla ei ole 4. momenttia. Tällöin on kyseenalaista piirtää ACF tai PACF neliöidyille jäännöksille, koska niiden teoreettista vastinetta ei ole olemassa. Tosin muotoparametri ei ole merkitsevästi poikkea 4:stä.

```
pacf(x^2,main="")
library(garch)
a <- garch(x,order=c(0,8),trace=FALSE)
pacf(a$res[-(1:8)]^2,main="")
qqnorm(a$res)
qqline(a$res)
```

```
library(fGarch)
a2 <- garchFit(~garch(8,0),data=x,trace=FALSE,include.mean=FALSE,
                cond.dist="std")
plot(a2)
summary(a2)
```

```
        Estimate  Std. Error  t value Pr(>|t|)
omega  2.058e-04   2.703e-05    7.613 2.69e-14 ***
alpha1 2.226e-01   5.313e-02    4.189 2.80e-05 ***
alpha2 9.029e-02   3.908e-02    2.310  0.02086 *
alpha3 6.424e-02   3.668e-02    1.751  0.07987 .
alpha4 7.380e-02   3.904e-02    1.890  0.05870 .
alpha5 1.606e-01   5.379e-02    2.985  0.00284 **
alpha6 6.240e-02   3.198e-02    1.951  0.05101 .
alpha7 9.101e-02   3.704e-02    2.457  0.01401 *
alpha8 1.632e-01   5.046e-02    3.234  0.00122 **
shape  3.783e+00   2.884e-01   13.117  < 2e-16 ***
```

6.3 GARCH-malli

Usein ARCH-efektiä ei pystytä mallintamaan vähäparametrisella ARCH-mallilla. Tehokkaampi malli tässä suhteessa on yleistetty versio ARCH-prosessista, GARCH (Generalized ARCH). Prosessin $\{X_t\}$ sanotaan noudattava GARCH(p,q)-prosessia, jos yhtälö (6.2) on korvattu yhtälöllä

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + ... + \alpha_q X_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + ... + \beta_p \sigma_{t-p}^2. \qquad (6.3)$$

Jotta σ_t^2 olisi aina positiivinen, oletetaan, että $\alpha_0 > 0,\ \alpha_1 \geq 0, ..., \alpha_q \geq 0,\ \beta_1 \geq 0, ..., \beta_p \geq 0$.

Lisäämällä yhtälöön puolittain $W_t = X_t^2 - \sigma_t^2$ ja järjestelemällä termejä yhtälö voidaan esittää muodossa

$$X_t^2 = \alpha_0 + (\alpha_1 + \beta_1) X_{t-1}^2 + ... + (\alpha_r + \beta_r) X_{t-r}^2 + W_t - \beta_1 W_{t-1} - ... - \beta_p W_{t-p}, \quad (6.4)$$

missä $r = \max(p, q)$ ja $\alpha_j = 0$, kun $j > q$, ja $\beta_j = 0$, kun $j > p$. Nähdään siis, että GARCH(p,q)-prosessi on ARMA(r,p)-prosessi neliöidylle sarjalle. GARCH-prosessilla voidaan siis selittää neliöidyssä sarjassa esiintyvää autokorrelaatiota, mikäli neliöidyllä sarjalla on äärellinen varianssi.

Jos $(\alpha_1 + \beta_1) + ... + (\alpha_r + \beta_r) < 1$, GARCH-prosessi on stationaarinen ja sen ei-ehdollinen varianssi on $\sigma^2 = \alpha_0/(1 - (\alpha_1 + \beta_1) - ... - (\alpha_r + \beta_r))$ (harjoitustehtävä).

Kun oletetaan, että $Z_t \sim N(0,1)$, suurimman uskottavuuden funktio GARCH-prosessille on

$$L(\alpha, \beta) = \prod_{t=q+1}^{n} f_t(x_t|x_{t-1}, ..., x_{t-q}, \sigma^2_{t-1}, ..., \sigma^2_{t-p}) = \prod_{t=q+1}^{n} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{-\frac{1}{2}\frac{x_t^2}{\sigma_t^2}\right\},$$

missä on ehdollistettu havaintojen $X_1, ..., X_q$ ja tuntemattomien ehdollisten varianssien $\sigma_1^2, ..., \sigma_p^2$ suhteen. Arvot $\sigma_1^2, ..., \sigma_p^2$ voidaan korvata sarjan $\{X_t\}$ ensimmäisten havaintojen varianssilla. Maksimoilla uskottavuusfunktiota numeerisesti parametrien suhteen saadaan niille SU-estimaatit.

Melkein aina käytännön mallinnuksessa riittävät matala-asteiset mallit GARCH(1,1), GARCH(1,2) ja GARCH(2,1). Useimmiten selvitään jo GARCH(1,1)-mallilla.

Tarkastellaan seuraavaksi varianssin ennustamista GARCH(1,1)-mallin avulla. Oletetaan, että X_t ja σ_t^2 ovat tunnettuja ajanhetkellä h, ja merkitään l askelen ennustetta $\sigma_h^2(l)$. Yhden aika-askelen ennuste on

$$\sigma_h^2(1) = \sigma_{h+1}^2 = \alpha_0 + \alpha_1 X_h^2 + \beta_1 \sigma_h^2.$$

Käyttämällä hyväksi yhtälöä $X_t^2 = \sigma_t^2 Z_t^2$, yhtälö 6.3 voidaan esittää muodossa

$$\sigma_{t+1}^2 = \alpha_0 + (\alpha_1 + \beta_1)\sigma_t^2 + \alpha_1 \sigma_t^2 (Z_t^2 - 1),$$

kun $p = q = 1$. Kun $t = h + 1$, yhtälöstä tulee

$$\sigma_{h+2}^2 = \alpha_0 + (\alpha_1 + \beta_1)\sigma_{h+1}^2 + \alpha_1 \sigma_{h+1}^2 (Z_{h+1}^2 - 1).$$

Koska $\mathsf{E}(Z_{h+1}^2 - 1|\mathcal{F}_h) = 0$, kahden askelen ennuste toteuttaa yhtälön

$$\sigma_h^2(2) = \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(1).$$

Yleisesti

$$\sigma_h^2(l) = \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(l-1), \quad l > 1. \tag{6.5}$$

Tämän rekursion perusteella voidaan johtaa (harjoitustehtävä) yleinen kaava

$$\sigma_h^2(l) = \frac{\alpha_0[1 - (\alpha_1 + \beta_1)^{l-1}]}{1 - \alpha_1 - \beta_1} + (\alpha_1 + \beta_1)^{l-1}\sigma_h^2(1).$$

Sovitetaan lopuksi aiemmin tässä luvussa analysoituun Nokian tuottosarjaan GARCH(1,1)-malli. Nyt onnistutaan selittämään ehdollinen heteroskedastisuus vähemmillä parametreilla. AIC on nyt -4.67, kun se ARCH(8)-mallin tapauksessa oli -4.64. Huomaa, että fGarch 'standardoi' informaatiokriteerit jakamalla ne havaintojen lukumäärällä.

```
b <- garchFit(~garch(1,1),data=x,trace=FALSE,include.mean=FALSE,cond.dist="std")
summary(b)
```

```
        Estimate  Std. Error  t value Pr(>|t|)
omega  2.426e-06   1.165e-06    2.082   0.0374 *
alpha1 3.742e-02   6.441e-03    5.810 6.25e-09 ***
beta1  9.613e-01   6.123e-03  156.989  < 2e-16 ***
shape  3.919e+00   2.932e-01   13.364  < 2e-16 ***

Information Criterion Statistics:
      AIC       BIC       SIC      HQIC
-4.665666 -4.656403 -4.665671 -4.662304
```

6.4 Integroitunut GARCH-malli

Jos AR-polynomilla GARCH-mallin esitysmuodossa 6.3 on yksikköjuuri, kyseessä on IGARCH-malli. Tällöin ehdollisella varianssilla $\{\sigma_t^2\}$ on satunnaiskävelyä muistuttava käyttäytyminen. Innovaatioiden $W_t = X_t^2 - \sigma_t^2$ vaikutus sarjan tulevien havaintojen varianssiin jää pysyväksi. Prosessi $\{\sigma_t^2\}$ ei ole heikosti stationaarinen, koska sillä ei ole kahta ensimmäistä momenttia. Sen sijaan se saattaa olla vahvasti stationaarinen.

IGARCH(1,1)-prosessi on muotoa

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + (1 - \beta_1) X_{t-1}^2,$$

missä $0 < \beta_1 < 1$ ja $\{Z_t\} \sim \text{IID}(0, 1)$. Tälle mallille saadaan yhtälöstä 6.3 l askelen ennusteeksi

$$\sigma_h^2(l) = \sigma_h^2(1) + (l - 1)\alpha_0.$$

Siis ehdollisen varianssin ennuste kasvaa lineaarisesti kulmakertoimella α_0.

Kun $\alpha_0 = 0$, saadaan mielenkiintoinen erikoistapaus. Tällöin

$$\begin{aligned}\sigma_t^2 &= (1-\beta_1)X_{t-1}^2 + \beta_1\sigma_{t-1}^2 \\ &= (1-\beta_1)X_{t-1}^2 + \beta_1[(1-\beta_1)X_{t-2}^2 + \beta_1\sigma_{t-2}^2] \\ &= .... \\ &= (1-\beta_1)(X_{t-1}^2 + \beta_1 X_{t-2}^2 + \beta_1^2 X_{t-3}^3 + ...),\end{aligned}$$

joten $\{\sigma_t^2\}$ saadaan eksponentiaalisella tasoituksella prosessista $\{X_t\}$, kun tasoitusparametrina on β_1.

Kun Nokian tuottosarja estimoitiin GARCH(1,1)-mallilla saatiin ratkaisu, joka on hyvin lähellä IGARCH-prosessia, sillä $\hat{\alpha}_1 + \hat{\beta}_1 = 0.99872$ on hyvin lähellä ykköstä. Parametrin α_0 ('omega') estimaatti ei ole kovin merkitsevä, joten eksponentiaalisen tasoituksen malli saattaisi olla hyvä approksimaatio kyseiselle aikasarjalle.

6.5 GARCH-M-malli

Rahoitusteoriassa arvopaperin tuoton odotusarvo ja varianssi ovat usein sidoksissa toisiinsa. Yksi tapa mallintaa tätä ilmiötä ovat GARCH in mean (GARCH-M) -mallit. Yksinkertainen GARCH(1,1)-M-malli voidaan esittää muodossa

$$\begin{aligned}&Y_t = \mu + c\sigma_t^2 + X_t, \quad X_t = \sigma_t Z_t, \\ &\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1\sigma_{t-1}^2. \end{aligned} \tag{6.6}$$

Tässä c on ns. riskilisäparametri (risk premium parameter). Jos c on positiivinen, se tarkoittaa sitä, että tuotto-odotus kasvaa riskin lisääntyessä. Muita mahdollisia määrittelyitä riskilisälle ovat $Y_t = \mu + c\sigma_t + X_t$ ja $Y_t = \mu + c\ln(\sigma_t^2) + X_t$.

GARCH-M-mallin määrittelystä 6.6 seuraa, että prosessi $\{Y_t\}$ on autokorreloitunut. Tämä aiheutuu prosessin $\{\sigma_t^2\}$ autokorreloituneisuudesta. GARCH-M on siis yksi mahdollinen selitys, jos tuottosarjassa havaitaan autokorrelaatiota.

Luku 7

Moniulotteiset aikasarjat

Usein on järkevää tarkastella aikasarjoja niin, että ne että ne muodostavat yhdessä moniulotteisen prosessin. Tällöin voidaan ottaa huomioon havaintojen autokorrelaation lisäksi eri sarjojen havaintojen välinen korrelaatio, ns. ristikorrelaatio. Esim. haluttaessa muodostaa hyvin hajautettu sijoitussalkku, joka maksimoi tuotto-odotuksen ja minimoi varianssin, on välttämätöntä tarkastella arvopaperien tuottosarjoja yhdessä ottaen huomioon niiden välinen korrelaatio. Monet makrotaloudelliset muuttujat ovat sidoksissa toisiinsa tavalla, jota on mielekästä tutkia moniulotteisen aikasarja-analyysin keinoin.

Monet yksiulotteisen aikasarja-analyysin menetelmät ja mallit yleistyvät melko suoraviivaisesti moniulotteiseen tapaukseen. Kuitenkin uusia ongelmia ilmenee ja aivan uudentyyppisiä riippuvuussuhteita on mielekästä määritellä. Seuraavassa kappaleessa yleistetään eräitä yksiulotteisen aikasarja-analyysin määritelmiä.

7.1 Heikko stationaarisuus ja ristikorrelaatiofunktio

Tarkastellaan m aikasarjaa $\{X_{ti}\, t = 0, \pm 1, \pm 2, ...\}$, $i = 1, ..., m$, missä $\mathsf{E}X_{ti}^2 < \infty$ kaikilla t ja i. Jos kaikki satunnaismuuttujien X_{tj} äärellisulotteiset jakaumat olisivat multinormaalisia, kaikki aikasarjojen jakaumaominaisuudet voitaisiin palauttaa odotusarvoihin

$$\mu_{ti} = \mathsf{E}X_{ti}$$

ja kovariansseihin

$$\gamma_{ij}(t+h,t) = \mathsf{E}[(X_{t+h,i} - \mu_{ti})(X_{tj} - \mu_{tj})].$$

Vaikka multinormaalisuus ei olisikaan voimassa, nämä toisen asteen ominaisuudet tarjoavat hyvän lähtökohdan aikasarjojen riippuvuuden analysoinnille.

Vektorimerkintää käyttäen moniulotteinen aikasarja voidaan esittää muodossa

$$\mathbf{X}_t = \begin{bmatrix} X_{t1} \\ \vdots \\ X_{tm} \end{bmatrix}, \quad t = 0, \pm 1, ...$$

Moniulotteisen aikasarjan $\{\mathbf{X}_t\}$ toisen asteen ominaisuudet määräytyvät odotusarvovektorien

$$\boldsymbol{\mu}_t = \mathsf{E}\mathbf{X}_t = \begin{bmatrix} \mu_{t1} \\ \vdots \\ \mu_{tm} \end{bmatrix}$$

ja kovarianssimatriisien

$$\Gamma(t+h,t) = \begin{bmatrix} \gamma_{11}(t+h,t) & \cdots & \gamma_{1m}(t+h,t) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(t+h,t) & \cdots & \gamma_{mm}(t+h,t) \end{bmatrix}$$

avulla. Kovarianssimatriisit voidaan myös lyhyesti määritellä $\Gamma(t+h,t) = \mathsf{E}[(\mathbf{X}_{t+h} - \boldsymbol{\mu}_{t+h})(\mathbf{X}_t - \boldsymbol{\mu}_t)']$.

Vastaavasti kuin yksiulotteisessa tapauksessa voidaan määritellä, että $\{\mathbf{X}_t\}$ on *(heikosti) stationaarinen*, jos $\boldsymbol{\mu}(t)$ on riippumaton ajasta t ja $\Gamma(t+h,t$ on riippumaton ajasta t kaikilla viiveillä h. Stationaariselle aikasarjalle voimme käyttää lyhennettyjä merkintöjä

$$\boldsymbol{\mu} = \mathsf{E}\mathbf{X}_t = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix}$$

ja

$$\Gamma(h) = \mathsf{E}[(\mathbf{X}_{t+h} - \boldsymbol{\mu})(\mathbf{X}_t - \boldsymbol{\mu})'] = \begin{bmatrix} \gamma_{11}(h) & \cdots & \gamma_{1m}(h) \\ \vdots & \ddots & \vdots \\ \gamma_{m1}(h) & \cdots & \gamma_{mm}(h) \end{bmatrix}.$$

Huomaa, että matriisin $\Gamma(h)$ diagonaalilta saadaan yksiulotteisten prosessien $\{X_{ti}\}$ autokovarianssifunktiot $\gamma_{ii}(h),\ i = 1, ..., m$. Diagonaalin ulkopuolista elementtiä $\gamma_{ij}(h),\ i \neq j$ kutsutaan sarjojen $\{X_{ti}\}$ ja $\{X_{tj}\}$ väliseksi *ristikovarianssifunktioksi*. Tässä on tärkeää huomata, että yleensä $\gamma_{ij}(h)$ ei ole sama kuin $\gamma_{ji}(h)$. Korrelaatiomatriisifunktio määritellään

$$R(h) = \begin{bmatrix} \rho_{11}(h) & \cdots & \rho_{1m}(h) \\ \vdots & \ddots & \vdots \\ \rho_{m1}(h) & \cdots & \rho_{mm}(h) \end{bmatrix},$$

missä $\rho_{ij}(h) = \gamma_{ij}(h)/[\gamma_{ii}(0)\gamma_{jj}(0)]^{1/2}$. Funktio $R(h)$ on kovarianssifunktio standardoidulle sarjalle, joka saadaan vähentämällä $\boldsymbol{\mu}$ sarjasta $\{\mathbf{X}_t\}$ ja jakamalla kukin komponenttisarja keskihajonnallaan.

Stationaarisen prosessin ristikovarianssifunktiolla on seuraava symmetriaominaisuus: $\gamma_{ij}(h) = \gamma_{ji}(-h)$. Tämä seuraa suoraan määritelmästä ja stationaarisuudesta: $\gamma_{ij}(h) = \mathsf{Cov}(X_{t+h,i}, X_{tj}) = \mathsf{Cov}(X_{tj}, X_{t+h,i}) = \mathsf{Cov}(X_{t-h,j}, X_{ti}) = \gamma_{ji}(-h)$. Matriisimuodossa tämä yhteys on: $\Gamma(h) = \Gamma'(-h)$. Vastaava ominaisuus on tietysti myös voimassa ristikorrelaatiofunktolla ja korrelaatiomatriisifunktiolla.

Ristikorrelaatiofunktiota $\rho_{ij}(h)$ voidaan käyttää sarjojen i ja j välisen lineaarisen riippuvuuden tutkimiseen. Riippuvuus voi ilmetä mm. seuraavilla tavoilla:

- Sarjoilla $\{X_{ti}\}$ ja $\{X_{tj}\}$ on *samanaikasta riippuvuutta*, jos $\rho_{ij}(0) \neq 0$.
- Jos $\rho_{ij}(h) = 0$ kaikkilla h, sarjat ovat lineaarisesti riippumattomat (korreloimattomat).
- Jos $\rho_{ij}(h) \neq 0$ jollakin $h > 0$ mutta $\rho_{ij}(h) = 0$ kaikilla $h < 0$, sarja $\{X_{ti}\}$ *seuraa (lags)* sarjaa $\{X_{tj}\}$.
- Jos $\rho_{ij}(h) \neq 0$ jollakin $h < 0$ mutta $\rho_{ij}(h) = 0$ kaikilla $h > 0$, sarja $\{X_{ti}\}$ *johtaa (leads)* sarjaa $\{X_{tj}\}$.
- Jos $\rho_{ij}(h) \neq 0$ jollakin $h < 0$ ja jollakin $h > 0$, sarjojen välillä on *takaisinkytkentä (feedback relationship)*.

Stationaarisen, moniulotteisen aikasarjan odotusarvovektori voidaan estimoida harhattomasti keskiarvovektorilla

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}_t.$$

Kovarianssimatriisifunktio voidaan puolestaan estimoida kaavoilla

$$\hat{\Gamma}(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-h} (\mathbf{X}_{t+h} - \bar{\mathbf{X}}_n)(\mathbf{X}_t - \bar{\mathbf{X}}_n)' & \text{kun } 0 \leq h \leq n-1, \\ \hat{\Gamma}'(-h) & \text{kun } -n+1 \leq h < 0. \end{cases}$$

Ristikorrelaatiokertoimen estimaatit ovat

$$\hat{\rho}_{ij}(h) = \hat{\gamma}_{ij}(h)(\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0))^{-1/2},$$

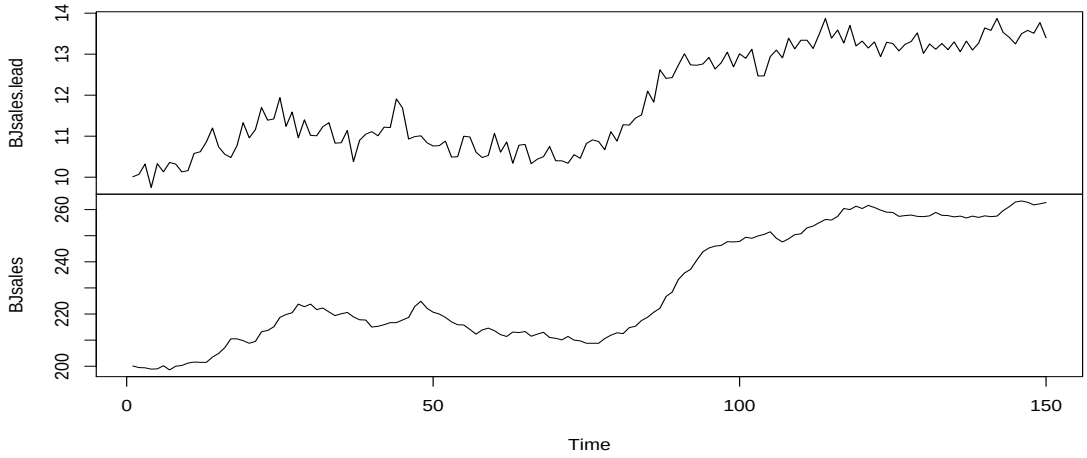
missä $\hat{\gamma}_{ij}(h)$ tarkoittaa matriisin $\hat{\Gamma}(h)$ rivin i sarakkeen j komponenttia.

Otosristikorrelaatiokertoimille $\hat{\rho}_{ij}$ pätee seuraava asymptoottinen tulos: Jos sarjat $\{X_{ti}\}$ ja ovat $\{X_{tj}\}$ riippumattomat ja ainakin toinen sarjoista on iid-kohinaa, suurella otoskoolla likimain $\hat{\rho}_{ij} \sim N(0, 1/n)$. Tämän tuloksen perusteella ristikorrelaation merkitsevyyden tutkimisessa voidaan käyttää samoja rajoja kuin tutkittaessa, onko yksittäinen sarja valkoista kohinaa. Laajemmin asymptoottisia tuloksia on esitelty mm. Brockwellin ja Davisin kirjassa.

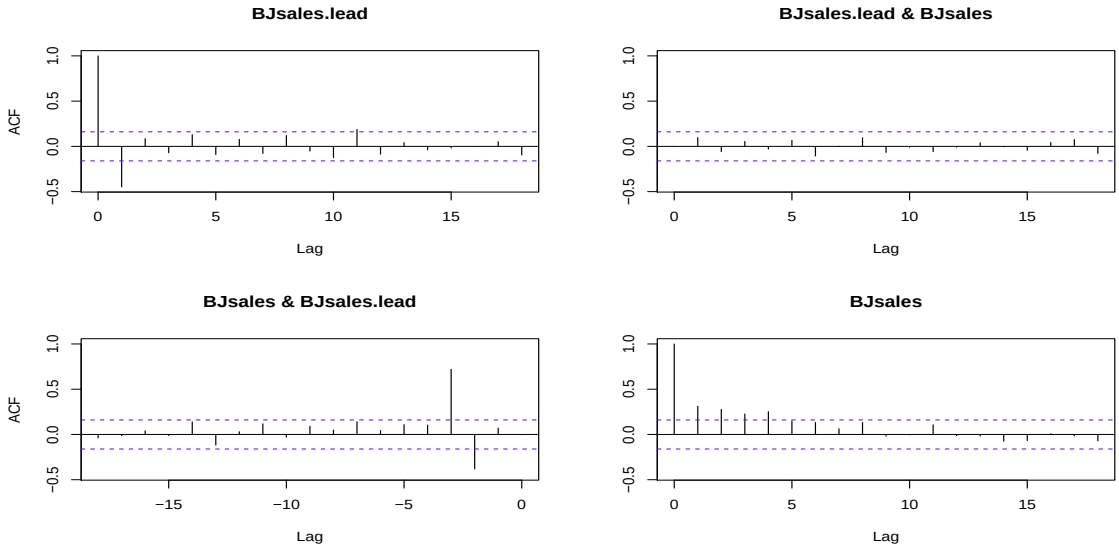
Esim. 1 *Myyntiaineisto ja johtava indeksi.* Seuraava esimerkkiaineisto esiintyy Boxin ja Jenkinsin kirjassa (1976) Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco, p. 537. Kuten kuviosta 7.1 nähdään, sarjat BJsales.lead ja BJsales eivät ole stationaarisia. Voidaksemme tutkia niiden riippuvuutta ristikorrelaatiofunktion avulla, tarkastelemme differoituja sarjoja $\{D_{t1}\}$ ja $\{D_{t2}\}$.

Kuviossa 7.2 näemme sarjojen autokorrelaatiofunktiot ja ristikorrelaatiofunktion. Ristikorrelaatiofunktion perusteella näyttäisi siltä, että sarjoilla ei ole samanaikaista korrelaatiota mutta BJsales.lead ennakoi (johtaa) sarjaa BJsales viiveillä 2 ja 3, sillä $\hat{\rho}_{12}(h) \neq 0$, kun $h = -2$ tai -3. Tulkinnoissa on kuitenkin syytä olla varovainen, sillä kummatkin sarjat ovat autokorreloituneita. Parempi olisi ensin mallintaa differoitujen sarjojen autokorrelaatiorakenne ARMA-mallilla ja tarkastella jäännössarjojen ristikorrelaatiofunktiota.

```
X <- cbind(BJsales.lead,BJsales)
plot(X,main="")
D <- diff(X)
acf(D)
```



Kuvio 7.1: BJsales.lead ja BJsales



Kuvio 7.2: Sarjojen BJsales.lead ja BJsales differenssisarjojen autokorrelaatiofunktiot ja ristikorrelaatiofunktio

7.2 Valkoinen kohina ja lineaarinen prosessi

Analogisesti yksiulotteisten prosessien kanssa määrittelemme, että m-ulotteinen aikasarja $\{\mathbf{Z}_t\}$ on *valkoista kohinaa odotusarvolla* $\mathbf{0}$ *ja kovarianssimatriisilla* Σ , merk. $\{\mathbf{Z}_t\} \sim \text{WN}\ (\mathbf{0}, \Sigma)$, jos $\{\mathbf{Z}_t\}$ on stationaarinen odotusarvolla $\mathbf{0}$ ja kovarianssifunkitiolla

$$\Gamma(h) = \begin{cases} \Sigma, & \text{kun } h = 0, \\ 0, & \text{muuten.} \end{cases}$$

Sarja $\{\mathbf{Z}_t\}$ on *riippumatonta kohinaa odotusarvolla* $\mathbf{0}$ *ja kovarianssimatriisilla* Σ , merk. $\{\mathbf{Z}_t\} \sim \text{IID}(\mathbf{0}, \Sigma)$, jos satunnaisvektorit $\{\mathbf{Z}_t\}$ ovat riippumattomia ja samoin jakautuneita odotusarvolla $\mathbf{0}$ ja kovarianssimatriisilla Σ.

Prosessin 'valkoisuutta' voidaan tutkia auto- ja ristikorrelaatiofunktioiden lisäksi portmanteau-testeillä. Yksiulotteiselle Ljung-Box-testille ovat kehittäneet yleistyksiä moniulotteiseen tapaukseen mm. Hoskin (1980,1981) ja Li&McLeod (1981). Testien nollahypoteesi on $R(1) = R(2) = ... = R(h) = 0$ ja vaihtoehtoinen hypoteesi $R(i) \neq 0$ jollain $i = 1, ..., h$. Hoskinin testisuure on muotoa

$$Q_m(h) = n^2 \sum_{j=1}^{h} \frac{1}{n-j} \text{tr}[\hat{\Gamma}(j)'\hat{\Gamma}(0)^{-1}\hat{\Gamma}(j)\hat{\Gamma}(0)^{-1}],$$

missä n on havaintojen lkm ja m aikasarjan dimensio. Nollahypoteesin ja tiettyjen säännöllisyysehtojen vallitessa $Q_m(h)$ noudattaa suurella otoskoolla likimain χ^2-jakaumaa m^2h vapausasteella.

Aikasarja $\{\mathbf{X}_t\}$ on *lineaarinen prosessi*, jos se voidaan esittää muodossa

$$\mathbf{X}_t = \sum_{j=-\infty}^{\infty} C_j \mathbf{Z}_{t-j}, \quad \{\mathbf{Z}_t\} \sim \text{WN}\ (\mathbf{0}, \Sigma),$$

missä $\{C_j\}$ on jono $m \times m$-matriiseja, joiden komponentit ovat itseisesti summautuvia. Tällöin voidaan osoittaa (harjoitustehtävä), että prosessi on stationaarinen odotusarvolla $\mathbf{0}$ ja kovarianssifunktiolla

$$\Gamma(h) = \sum_{j=-\infty}^{\infty} C_{j+h} \Sigma C_j'.$$

7.3 Vektori-autoregressiiviset (VAR) mallit

Moniulotteiselle prosessille voidaan määritellä ARMA-malli (ks. Brockwell&Davis tai Tsay). Rajoitumme kuitenkin seuraavassa tarkastelemaan vektoriarvoista AR(p)-mallia, jota kuvaa yhtälö

$$\mathbf{X}_t = \boldsymbol{\phi}_0 + \Phi_1 \mathbf{X}_{t-1} + ... + \Phi_p \mathbf{X}_{t-p} + \mathbf{Z}_t, \quad \mathbf{Z}_t \sim \text{WN } (\mathbf{0}, \Sigma). \tag{7.1}$$

Käyttäen viiveoperaattoria B tämä voidaan esittää edelleen muodossa

$$(I - \Phi_1 \mathsf{B} - ... - \Phi_p \mathsf{B}) \mathbf{X}_t = \boldsymbol{\phi}_0 + \mathbf{Z}_t,$$

missä I on identiteettimatriisi, tai vielä lyhyemmin

$$\Phi(\mathsf{B}) \mathbf{X}_t = \boldsymbol{\phi}_0 + \mathbf{Z}_t,$$

missä $\Phi(z) = I - \Phi_1 z - ... - \Phi_p z^p$ on matriisi-arvoinen polynomi.

Kun oletetaan prosessin $\{\mathbf{X}_t\}$ stationaarisuus, ottamalla yhtälöstä (7.1) puolittain odotusarvo voidaan ratkaista odotusarvovektori

$$\boldsymbol{\mu} = \mathsf{E}(\mathbf{X}_t) = (I - \Phi_1 - .... - \Phi_p)^{-1} \boldsymbol{\phi}_0 = [\Phi(1)]^{-1} \boldsymbol{\phi}.$$

Prosessi (7.1) on *kausaalinen*, jos se voidaan esittää MA(∞)-prosessina

$$\mathbf{X}_t - \boldsymbol{\mu} = \sum_{j=0}^{\infty} \Psi_j \mathbf{Z}_{t-j} \quad \text{kaikilla } t,$$

missä $\{\Psi_j\}$ on jono matriiseja itseisesti summautuvilla komponenteilla. Tämän kanssa yhtäpitävää on (ks. todistus TSTM), että determinantilla $\det \Phi(z)$ ei ole nollakohtia kompleksitason yksikkökiekossa $\{z \in \mathbb{C} : |z| \leq 1\}$.

Kertoimet Ψ_j voidaan ratkaista rekursiivisesti yhtälöistä

$$\Psi_j = \sum_{k=1}^{\min(p,j)} \Phi_k \Psi_{j-k}, \ j = 1, 2, ...,$$

missä $\Psi_0 = I$ ja $\Psi_j = 0$, kun $j < 0$. Esim. VAR(1)-prosessin tapauksessa

$$\begin{aligned}
\Psi_0 &= I, \\
\Psi_1 &= \Phi_1 \Psi_0 = \Phi_1, \\
\Psi_2 &= \Phi_1 \Psi_1 = \Phi_1^2, \\
&\vdots \\
\Psi_j &= \Phi_1 \Psi_{j-1} = \Phi^j, \ j \geq 3.
\end{aligned}$$

Keskineliövirheen mielessä paras yhden askelen ennuste VAR(p)-prosessille on

$$\mathsf{P_n}\mathbf{X}_{n+1} = \boldsymbol{\phi}_0 + \Phi_1\mathbf{X}_n + ... + \Phi_p\mathbf{X}_{n-p+1},$$

kun $n \geq p$. Ennustevirhe on $\mathbf{Z}_{n+1}$, jonka kovarianssimatriisi on Σ. Useamman askelen ennusteet saadaan rekursiivisesti kaavasta

$$\mathsf{P_n}\mathbf{X}_{n+h} = \boldsymbol{\phi}_0 + \Phi_1\mathsf{P_n}\mathbf{X}_{n+h-1} + ... + \Phi_p\mathsf{P_n}\mathbf{X}_{n+h-p},$$

missä $\mathsf{P_n}\mathbf{X}_{n+h-j} = \mathbf{X}_{n+h-j}$, kun $j \geq h$. Ennustevirhe h askelen ennusteessa on

$$\mathbf{X}_{n+h} - \mathsf{P_n}\mathbf{X}_{n+h} = \sum_{j=0}^{h-1} \Psi_j\mathbf{Z}_{n+h-j},$$

jonka kovarianssimatriisi on $\sum_{j=0}^{h-1} \Psi_j\Sigma\Psi_j'$.

VAR-mallin estimointi voidaan tehdä suurimman uskottavuuden menetelmällä, mutta se on moniulotteisessa tapauksessa numeerisesti huomattavasti vaativampi kuin yksiulotteisessa tapauksessa. Yksityiskohtia on esitetty Brocwellin ja Davisin kirjassa. Joka tapauksessa on syytä käyttää alkuarvoina alustavilla estimointimenetelmillä saatavia estimaatteja. Alustavina menetelminä voidaan käyttää joko moniulotteista yleistystä Yule-Walker-menetelmästä tai ehdollista uskottavuusmenetelmää, jossa uskottavuusfunktio lasketaan ehdollisena ensimmäisten havaintovektorien $\mathbf{X}_1, ...\mathbf{X}_p$ suhteen. Tämä ehdollinen uskottavuusmenetelmä on yhtäpitävä OLS-menetelmän kanssa, kun käytetään normaalista uskottavuusfunktiota.

Ehdollinen uskottavuusfunktio on muotoa

$$L(\boldsymbol{\phi}, \boldsymbol{\Phi}, \Sigma) = (2\pi)^{(n-p)m/2}(\det\Sigma)^{-(n-p)/2} \exp\left[-\frac{1}{2}\sum_{t=p+1}^{n}(\mathbf{x}_t - \hat{\mathbf{x}}_t)'\Sigma^{-1}(\mathbf{x}_t - \hat{\mathbf{x}}_t)\right],$$

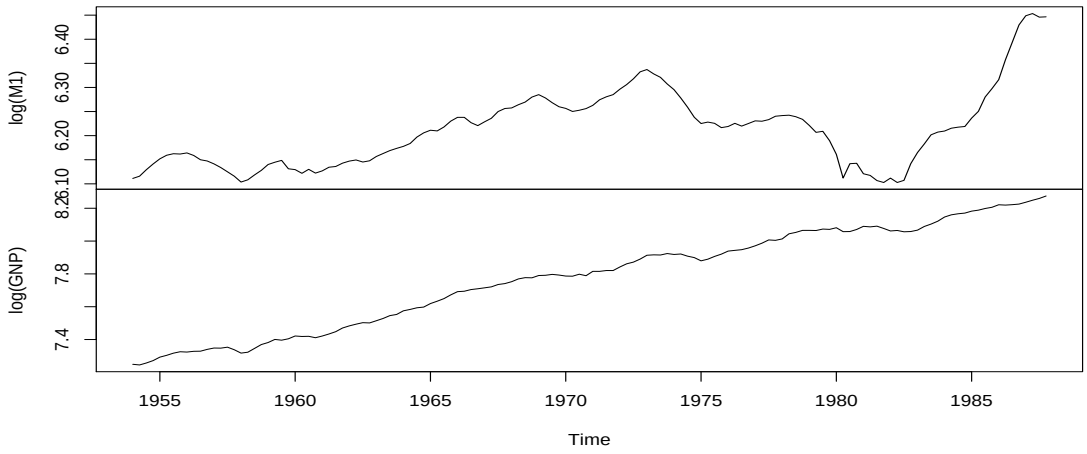
missä $\boldsymbol{\Phi} = (\Phi_1, ..., \Phi_p)$ ja $\hat{\mathbf{x}}_t = \boldsymbol{\phi}_0 + \sum_{j=1}^{p} \Phi_j\mathbf{x}_{t-j}$ ovat yhden askelen ennusteita. Estimoinnissa voidaan hyödyntää regressioyhtälöä

$$\begin{bmatrix} \mathbf{x}_{p+1}' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{x}_p' & \mathbf{x}_{p-1}' & \dots & \mathbf{x}_1' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{x}_{n-1}' & \mathbf{x}_{n-2}' & \dots & \mathbf{x}_{n-p}' \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}_0' \\ \Phi_1' \\ \Phi_2' \\ \vdots \\ \Phi_p' \end{bmatrix} + \begin{bmatrix} \mathbf{z}_{p+1}' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix}$$

tai lyhyesti $X = UB + Z$, josta saadaan OLS-estimaattori $\hat{B} = (U'U)^{-1}U'X$. Suurimman uskottavuuden estimaattori Σ:lle on $\hat{\Sigma} = \frac{1}{n-p}\hat{Z}'\hat{Z}$, missä $\hat{Z} = X - U\hat{B}$. Harhaton estimaattori on puolestaan $S = \frac{1}{n-2p-1}\hat{Z}'\hat{Z}$

Sopivan viivepituuden valinnassa voidaan käyttää informaatiokriteereitä AIC, AICC ja BIC. Jos käytetään ehdollista uskottavuusfunktiota, on syytä tehdä korjaus selittävien havaintojen määrän $n - p$ suhteen, jotta tulokset olisivat vertailukelpoisia.

Esim. 2 *USeconomic-sarja.* Pakettiin tseries sisältyvässä aineistossa USeconomic on 4-ulotteinen aikasarja, joka sisältää USA:n taloutta kuvaavia muuttujia vuosilta 1954—1987. Rajoitumme tässä tarkastelemaan muuttujien log(M1) ja log(GNP) muodostamaa kaksiulotteista aikasarjaa. Kuten kuviosta 7.3 havaitaan, aikasarjat ovat epästationaarisia. Tarkastelemme jatkossa differoitujen aikasarjojen muodostamaa kaksiulotteista aikasarjaa.



Kuvio 7.3: USeconomic-aineiston aikasarjoja

R:n funktio ar määrittää parhaaksi viivepituudeksi 1. Malli siis voidaan esittää muodossa $\mathbf{Y}_t = \boldsymbol{\phi} + \Phi\mathbf{Y}_{t-1} + \mathbf{Z}_t$ tai

$$Y_{t1} = \phi_1 + \phi_{11}Y_{t-1,1} + \phi_{12}Y_{t-1,2} + Z_{t1},$$
$$Y_{t2} = \phi_2 + \phi_{21}Y_{t-1,1} + \phi_{22}Y_{t-1,2} + Z_{t2}.$$

Jotta voisimme vielä helpommin tutkia kerrointen merkitsevyyttä, estimoimme mallin myös OLS-menetelmällä. Tulosten perusteella näyttää siltä, että kerroin ϕ_{12} ei ole merkisevä. Sarjan $\{Y_{t1}\}$ (∇log(M1)) aikaisempi havainto siis ennakoi sarjan $\{Y_{t2}\}$ (∇log(GNP)) nykyistä havaintoa muttei päinvastoin. Tällaisessa tilanteessa puhutaan *Granger-kausaalisuudesta*. Toisin sanoen M1 Granger-vaikuttaa (Granger-causes) muuttujaan GNP muttei päinvastoin. Asia voidaan myös ilmaista sanomalla, että muuttuja M1 on *eksogeeninen* muuttujan GNP suhteen. Saman päätelmän voisi tehdä tutkimalla ristikorrelaatiofunktiota.

```
library(tseries)
X <- USeconomic[,1:2]
Y <- diff(X)
#Yule-Walker solution
(a <- ar(Y))
$ar
, , 1
         log(M1) log(GNP)
log(M1)   0.6046  -0.1413
log(GNP)  0.2647   0.1887

#OLS-solution
n <- dim(Y)[1]
summary(lm(Y[-1,]~Y[-n,]))

Response log(M1) :

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.002032   0.001112   1.827    0.070 .
Y[-n, ]log(M1)   0.605185   0.075740   7.990 6.09e-13 ***
Y[-n, ]log(GNP) -0.142638   0.093215  -1.530    0.128
---

Response log(GNP) :

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0055944  0.0009962   5.616 1.12e-07 ***
Y[-n, ]log(M1)  0.2644613  0.0678459   3.898 0.000154 ***
Y[-n, ]log(GNP) 0.1893458  0.0834998   2.268 0.024990 *
```

Voimme testata mallin riittävyyden Hoskinin testin avulla.

```
library(portes)
Hosking(a)
Lags Statistic  df   p-value
    5  18.84034  16 0.2770294
   10  35.33721  36 0.4999223
   15  50.23717  56 0.6917513
   20  70.91884  76 0.6432960
   25  89.18418  96 0.6756652
   30 104.08878 116 0.7782603
```

Lopuksi voimme tulostaa alkuperäisen sarjan ja viiden vuoden ennusteen seuraavasti:

```
pr <- predict(a,n.ahead=20)
Y2 <- rbind.zoo(Y,pr$pred)
init <- ts(rbind(X[1,]),start=start(X),frequency=frequency(X))
X2 <- cumsum(rbind.zoo(init,Y2))
plot(X2)
```

7.4 Yksikköjuuri-epästationaarisuus ja yhteisintegroituneisuus

Olemme nähneet, että epästationaarisen yksiulotteisen aikasarjan voi usein muuntaa stationaariseksi soveltamalla siihen differenssioperaattoria $\nabla = 1 - \mathsf{B}$ yhden tai useamman kerran. Jos $\{\nabla^d X_t\}$ on stationaarinen jollain positiivisella kokonaisluvulla d mutta $\{\nabla^{d-1} X_t\}$ on epästationaarinen, sanomme, että sarjan $\{X_t\}$ *integroitumisaste on* d (integrated of order d), merk. $\{X_t\} \sim I(d)$. Määrittelemme differenssioperaattorin m-ulotteiselle aikasarjalle $\{\mathbf{X}_t\}$ niin, että jokainen sarjan komponentti differoidaan: $\nabla \mathbf{X}_t = (\nabla X_{t1}, ..., \nabla X_{tm})'$.

Moniulotteisessa tapauksessa voimme määritellä integroitumisasteen samalla tavalla. Sanomme, että $\{\mathbf{X}_t\} \sim I(d)$, jos $\{\nabla^d \mathbf{X}_t\}$ on stationaarinen, mutta $\{\nabla^{d-1} \mathbf{X}_t\}$ ei. Kuitenkin moniulotteisessa tapauksessa integroitumisasteilla on eräänlaisia välimuotoja. Sanomme, että $I(d)$-prosessi $\{\mathbf{X}_t\}$ on yhteisintegroitunut yhteisintegroitumisvektorilla $\boldsymbol{\beta}$, jos $\boldsymbol{\beta}$ on sellainen m-vektori, että prosessin $\{\boldsymbol{\beta}' \mathbf{X}_t\}$ integroitumisaste on pienempi kuin d.

7.5 Yhteisintegroituneet VAR-mallit

Tarkastellaan seuraavassa VAR-prosessia, joka saattaa sisältää aikatrendin:

$$\mathbf{X}_t = \boldsymbol{\delta}_t + \Phi_1 \mathbf{X}_{t-1} + ... + \Phi_p \mathbf{X}_{t-p} + \mathbf{Z}_t, \tag{7.2}$$

missä $\mathbf{Z}_t \sim \text{WN}(\mathbf{0}, \Sigma)$ ja $\boldsymbol{\delta}_t = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 t$. Tässä AR-polynomi on $\Phi(z) = I - \Phi_1 z - ... - \Phi_p z^p$. Jos determinantin $|\Phi(z)|$ nollakohdat ovat yksikköympyrän ulkopuolella, prosessi on (trendi)stationaarinen ja kausaalinen. Jos $|\Phi(1)| = 0$, prosessi on yksikköjuuri-epästationaarinen. Oletamme seuraavassa tarkastelussa yksinkertaisuuden vuoksi, että prosessin $\{\mathbf{X}_t\}$ integroitumisaste on 0 tai 1.

Malli (7.2) voidaan esittää ns. *virheenkorjausmallina* (error correction model, ECM) seuraavasti:

$$\nabla \mathbf{X}_t = \boldsymbol{\delta}_t + \Pi \mathbf{X}_{t-1} + \Phi_1^* \nabla \mathbf{X}_{t-1} + \ldots + \Phi_{p-1}^* \nabla \mathbf{X}_{t-p+1} + \mathbf{Z}_t, \tag{7.3}$$

missä

$$\begin{aligned} \Phi_j^* &= -\sum_{i=j+1}^{p} \Phi_i, \quad j = 1, ..., p-1, \\ \Pi &= -\Phi(1) = \Phi_1 + \Phi_2 + ... + \Phi_p - I. \end{aligned}$$

Koska $\{\mathbf{X}_t\}$ on enintään I(1)-prosessi, $\{\nabla \mathbf{X}_t\}$ on I(0)-prosessi.

Voimme erottaa eri tapauksia sen mukaan, mikä matriisin Π aste on. Yleisesti, jos aste on r, on olemassa hajotelma $\Pi = \boldsymbol{\alpha}\boldsymbol{\beta}'$, missä $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$ ovat täysiasteisia $m \times r$-matriiseja.

1. Jos $\text{rank}(\Pi) = 0$, niin $\Pi = 0$ ja $\{X_t\}$ ei ole yhteisintegroitunut. Yhtälön (7.3) ECM yksinkertaistuu muotoon

   $$\nabla \mathbf{X}_t = \boldsymbol{\delta}_t + \Phi_1^* \nabla \mathbf{X}_{t-1} + \ldots + \Phi_{p-1}^* \nabla \mathbf{X}_{t-p+1} + \mathbf{Z}_t,$$

   joten $\{\nabla X_t\}$ noudattaa VAR$(p-1)$-prosessia deterministisellä trendillä.

2. Jos $\text{rank}(\Pi) = m$, niin $|\Phi(1)| \neq 0$ eikä AR-polynomin determinantilla ole yksikkö-juuria. Siis $\{\mathbf{X}_t\} \sim I(0)$. Tällöin ECM-mallista ei ole hyötyä ja voidaan tutkia suoraan AR(p)-mallia (7.2).

3. Kun $0 < \text{rank}(\Pi) = r < m$, niin $\{\mathbf{X}_t\}$ on yhteisintegroitunut r yhteisintegroitumisvektorilla; yhteisintegroituvat relaatiot ovat $\mathbf{W}_t = \boldsymbol{\beta}' \mathbf{X}_t$.

Sen mukaan, miten deterministinen termi $\boldsymbol{\delta}_t = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 t$ määritellään, voidaan erotella seuraavat trendityypit:

1. Jos $\boldsymbol{\delta}_t = \mathbf{0}$, kaikki komponentit $\{X_{ti}\}$ ovat I(1)-prosesseja ilman trendiä ja stationaarisen sarjan $\mathbf{W}_t = \boldsymbol{\beta}'\mathbf{X}_t$ odotusarvo on $\mathbf{0}$.

2. Jos $\boldsymbol{\delta}_t = \boldsymbol{\alpha}\mathbf{c}_0$, missä $\mathbf{c}_0$ on r-ulotteinen, nollasta poikkeava vakiovektori, ECM-malli tulee muotoon

$$\nabla\mathbf{X}_t = \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{X}_{t-1} + \mathbf{c}_0) + \Phi_1^*\nabla\mathbf{X}_{t-1} + \ldots + \Phi_{p-1}^*\nabla\mathbf{X}_{t-p+1} + \mathbf{Z}_t,$$

   jolloin komponentit $\{X_{ti}\}$ ovat I(1)-prosesseja ilman trendiä mutta sarjan $\{\mathbf{W}_t\}$ odotusarvo on $-\mathbf{c}_0$.

3. Jos $\boldsymbol{\delta}_t$ on vakiovektori, se voidaan esittää muodossa $\boldsymbol{\delta}_0 = \boldsymbol{\alpha}\mathbf{c}_0 + \boldsymbol{\alpha}_\perp\mathbf{d}_0$, missä $\boldsymbol{\alpha}_\perp$ on sellainen $m \times (m-r)$-matriisi, että $\boldsymbol{\alpha}'_\perp\boldsymbol{\alpha} = 0$. Tällöin ECM voidaan esittää

$$\nabla\mathbf{X}_t = \boldsymbol{\alpha}_\perp\mathbf{d}_0 + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{X}_{t-1} + \mathbf{c}_0) + \Phi_1^*\nabla\mathbf{X}_{t-1} + \ldots + \Phi_{p-1}^*\nabla\mathbf{X}_{t-p+1} + \mathbf{Z}_t,$$

   joten komponenttisarjoilla $\{X_{ti}\}$ on lineaarinen trendi ja sarjalla $\{\mathbf{W}_t\}$ odotusarvo $-\mathbf{c}_0$.

4. Jos $\boldsymbol{\delta}_t = \boldsymbol{\delta}_0 + \boldsymbol{\alpha}\mathbf{c}_1 t$, ECM on muotoa

$$\nabla\mathbf{X}_t = \boldsymbol{\alpha}_\perp\mathbf{d}_0 + \boldsymbol{\alpha}(\boldsymbol{\beta}'\mathbf{X}_{t-1} + \mathbf{c}_0 + \mathbf{c}_1 t) + \Phi_1^*\nabla\mathbf{X}_{t-1} + \ldots + \Phi_{p-1}^*\nabla\mathbf{X}_{t-p+1} + \mathbf{Z}_t,$$

   joten komponenttisarjoilla $\{X_{ti}\}$ on lineaarinen trendi ja komponenttisarjat $\{W_{ti}\}$ ovat trendistationaarisia.

5. Jos $\boldsymbol{\delta}_t = \boldsymbol{\delta}_0 + \boldsymbol{\delta}_1 t$, komponenteilla $\{X_{ti}\}$ on neliöllinen trendi ja komponentit $\{W_{ti}\}$ ovat trendistationaarisia.

ECM-malli voidaan estimoida suurimman uskottavuuden menetelmällä. Menetelmä on kuitenkin jossain määrin monimutkainen eikä sitä esitellä tässä (ks. Tsay tai Hamilton: Time Series Analysis tai Søren Johansen: Likelihood-based inference in cointegrated vector auto-regressive models).

Søren Johansen (1988) on esittänyt seuraavat testityypit yhteisintegroituneiden relaatioiden määrän testaamiseksi.

- trace-testi: $H_0 : \text{rank}(\Pi) = r$ vs. $H_a : \text{rank}(\Pi) > r$
- suurimman ominaisarvon testi (maximum eigenvalue test): $H_0 : \text{rank}(\Pi) = r$ vs. $H_a : \text{rank}(\Pi) = r + 1$

Ks. testisuureiden lausekkeet esim. Tsayn kirjasta. Kummassakin tapauksessa kyseessä on uskottavuussuhdetesti. Testisuureet eivät kuitenkaan noudata nollahypoteesin tapauksessa asymptoottisesti tavanomaisia χ^2-jakaumia. Sen sijaan jakaumien kriittisiä arvoja on taulukoitu simulointien perusteella.

Esim. 4 . *Myyntiaineisto ja johtava indeksi (jatkoa).* Kuvion 7.1 perusteella voisi päätellä, että sarjojen BJsales ja BJsales.lead välillä voisi vallita likimainen lineaarinen riippuvuus. Testaamme seuraavaksi yhteisintegroituvan relaation olemassaoloa käyttäen R-pakettiin urca sisältyvää funktiota ca.jo. Esimerkin testityyppinä on 'eigen' eli suurimman ominaisarvon testi. Mallin deterministisen osan määrittely on 'const', joka vastaa tapausta $\boldsymbol{\delta}_t = \boldsymbol{\alpha}\mathbf{c}_0$. (Määrittely 'trend' vastaa tapausta $\boldsymbol{\delta}_t = \boldsymbol{\delta}_0 + \boldsymbol{\alpha}\mathbf{c}_1 t$. Tapauksessa 'none' ilmeisesti R-koodissa on väärä taulukko). Määrittely $K = 3$ tarkoittaa sitä, että mallin AR-muodon (7.2) asteeksi valitaan on 3. Määrittely spec='transitory' vastaa ECM-mallin määrittelyä (7.3).

Testien perusteella hypoteesi $r = 0$ hylätään mutta ei hypoteesia $r = 1$. Tämä tarkoittaa sitä, että prosessilla on yksi yhteisintegroituva vektori. Jos myös hypoteesi $r = 1$ hylättäisiin, tämä tarkoittaisi sitä, että sarja olisi stationaarinen, mikä ei vaikuta kovin uskottavalta kuvion 7.1 perusteella. Estimointitulosten perusteella $\hat{\boldsymbol{\alpha}} = (-0.0546, 4.72)'$ ja $\hat{\boldsymbol{\beta}} = (1, -0.0550)'$. Yhteisintegroituva relaatio on muotoa $BJsales.lead = 0.055 \cdot BJsales - 0.951 + \tilde{W}_t$, missä $\{\tilde{W}_t\}$ on 0-keskinen stationaarinen prosessi.

Tulokseen on kuitenkin syytä suhtautua varauksella, sillä BJSales-sarjan jäännössarjassa esiintyy autokorrelaatiota (kuvio 7.4). Testit perustuvat oletukseen jäännösten riippumattomuudesta. Jos AR-muodon viivepituudeksi asetetaan $K = 8$, jäännösten autokorreloituneisuus häviää. Samalla kuitenkin testin $H_0 :$ rank$(\Pi) = 0$ tulos muuttuu ei-merkitseväksi eikä yhteisintegroituvan relaation olemassaoloa voida osoittaa.

```
library(urca)
X <- cbind(BJsales.lead,BJsales)
a <- ca.jo(X,type="eigen",ecdet="const",K=3,spec="transitory")
```

```
summary(a); resplot(a)

######################
# Johansen-Procedure #
######################

Test type: maximal eigenvalue statistic (lambda max) , without linear
trend and constant in cointegration

Eigenvalues (lambda):
[1] 9.468938e-01 1.911559e-02 6.002439e-17

Values of teststatistic and critical values of test:

           test 10pct  5pct  1pct
r <= 1 |   2.84  7.52  9.24 12.97
r = 0  | 431.51 13.75 15.67 20.20

Eigenvectors, normalised to first column:
(These are the cointegration relations)

               BJsales.lead.l1  BJsales.l1   constant
BJsales.lead.l1      1.00000000  1.00000000  1.0000000
BJsales.l1          -0.05504036 -0.02891266 -0.2672937
constant             0.95111756 -5.99894266 46.6758372

Weights W:
(This is the loading matrix)

              BJsales.lead.l1   BJsales.l1      constant
BJsales.lead.d     -0.05458765 -0.038047745 -3.641671e-17
BJsales.d           4.71969207 -0.001594596  4.322637e-15
```

Kuvio 7.4: BJsales-sarjan jäännössarja ECM-mallin sovittamisen jälkeen