E-kursuse "Aegridade Analüüs"
materjalid

Aine maht 6 EAP

Raul Kangro (Tartu Ülikool), 2012

Sisukord

Sissejuhatus

Nii firmade kui ka tavainimeste elus mängivad suurt rolli ajas toimuvad ning teatud juhuslikkuse komponenti sisaldavad sündmused, mille efekt on sageli väljendatav erinevatele ajahetkedele vastavate numbrite jadadena ehk aegridadena. Näitena võib tuua sissetulekud ja väljaminekud, toodete läbimüügi maht, päevaste ja kuiste sademete hulk jms. Sellistes valdkondades juhuslikkus toob kaasa riske, mille haldamiseks on väga tähtis osata juhuslikkuse iseloomu kindlaks teha, mineviku andmete põhjal võimalikult täpseid prognoose leida ning mõningatel juhtudel ka ebasoovitavate tendentside ilmnemisel õigeaegselt sekkuda. Kõike seda võimaldab aegridade teooria.

Ajalooliselt on praktikute poolt kasutusele võetud mitmeid **meetodeid** aegridadega seotud ülesannete (nt. trendi leidmine, tulevikuväärtuste prognoosimine jms) lahendamiseks. Meetodi all mõistame siin kursuses arvutuseeskirja, mille rakendamine peaks andma soovitud tulemuse. Meetodid tuginevad enamasti nn tervel mõistusel ja intuitsioonil ning neid võib rakendada suvalisele ajas järjestatud andmete kogumile, kuid lahtiseks jääb küsimus tulemuste tegelikkusele vastavuse ja usaldusväärsuse osas.

Selleks, et olla (piisavalt) kindel selles, et arvutatud tulemused kajastavad reaalsust ning on kasutatavad ka tuleviku prognoosimisel, tuleb lähtuda aegrea matemaatilistest **mudelitest**. Mudel on matemaatiline kirjeldus selle kohta, kuidas juhuslikkus mõjutab aegreale vastavate andmete tekkimist. Mudelist lähtuvalt on võimalik kontrollida selle sobivust konkreetse aegrea kirjeldamiseks ning tuletada teoreetiliselt põhjendatud arvutuseeskirjad vaadeldavale mudelile vastava aegrea erinevate komponentide leidmiseks ning tuleviku prognoosimiseks koos konkreetsete usalduspiiridega leitavate hinnangute jaoks. Aegridade teooria seisneb mudelite kirjeldamises ja nendele vastavate arvutuseeskirjade ning veahinnangute tuletamises.

Teooria rakendamine koosneb mitmetest etappidest, milleks on

1. Sobiva matemaatilise mudeli valik. Nagu me kursuse jooksul näeme, on võimalike mudelite hulk väga lai ning äärmiselt tähtis on leida võimalikult lihtne mudel, mis võimaldaks tegelikkust adekvaatselt kirjeldada.

2. Leitud mudeli kalibreerimine olemasolevate andmetega ning saadud konkreetse mudeli kirjeldusvõime kontroll. Kui selgub, et kirjeldusvõime on liiga madal, siis tuleb minna tagasi mudeli valiku juurde.

3. Kalibreeritud mudeli kasutamine tuleviku ennustamiseks, ennustuste veapiiride kindlakstegemine, vajadusel sobivate juhtimismehhanismide valik soovitud

tulemusest tekkinud kõrvalekallete vähendamiseks.

Kõiki neid küsimusi (välja arvatud juhtimismehhanismide valik) vaadeldakse käesoleva kursuse raames. Samas tuleb silmas pidada, et tegemist on sissejuhatava kursusega aegridade teooriast ning küllalt palju olulisi mudeleid ning tehnilisi vahendeid jääb selle kursuse raames käsitlemata. Aegridade teooria aktuaalsusest annab aga tunnistust näiteks see fakt, et 2003. aasta Nobeli majanduspreemia anti välja just aegridade teooria alaste tööde põhjal (R.F. Engle)

Peatükk 1

Tähistused ja mõisted. Aegrea komponendid

Mitteformaalselt on aegrida mingi ajas muutuva ja juhuslikest teguritest sõltuva suuruse erinevatele järjestatud ajavahemikele vastavate väärtuste kogum. Selleks võib olla näiteks teatud ajavahemike tagant mõõdetud konkreetse inimese kaal, aktsiahind, firma aastane kasum, kindlustusfirmale laekuvate kahjunõuete kogusumma päevade kaupa vms. Kui mõõtmised toimuvad pidevalt, siis on tegemist pideva ajaga aegreaga, vastasel korral öeldakse, et aegrida on diskreetse ajaga. Käesolevas kursuses käsitleme ainult selliseid diskreetseid aegridasid, kus väärtused vastavad võrdsete ajavahemike tagant tehtud mõõtmistele. Olgu selle ajavahemiku pikkus h, seega eeldame, et huvipakkuva suuruse Z väärtusi mõõdetakse ajamomentidel $\tau_i = \tau_0 + i\,h$, kus $i \in \mathrm{N}$ või $i \in \mathbf{Z}$. Selleks, et hoida tähistusi võimalikult lihtsana ning olla kooskõlas aegridade alase kirjanduse tavadega, tähistame ajamomendile τ_t vastavat juhuslikku suurust Z kujul Z_t ning selle teadaolevat väärtust kujul z_t, kus t on täisarvuline (või naturaalarvuline) indeks.

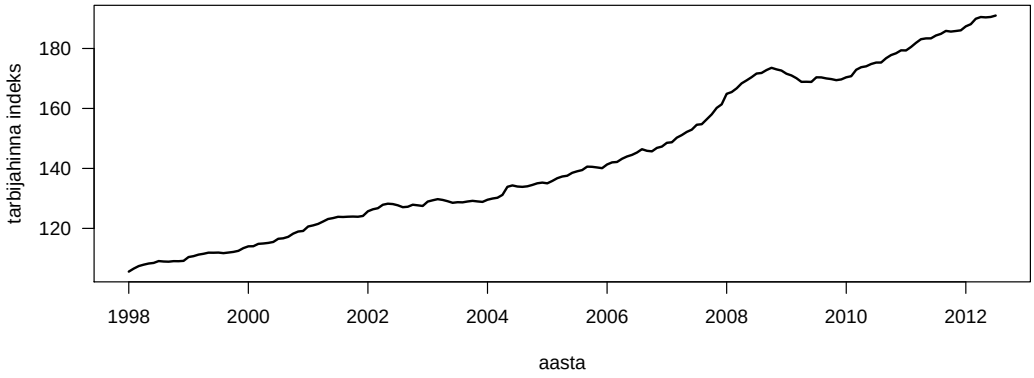
1.1 Aegrea komponendid

Aegreaga kirjeldatud juhusliku suuruse muutumisel ajas võib olla mitmeid erinevaid põhjuseid:

- Ümbritseva (majandus)keskkonna, firma juhtimiskultuuri vms tegurite pikaajaline mõju, mida nimetatakse **trendiks**.
- Kellaajast, nädalapäevast, kalendrikuust vms sõltuvad kindla perioodiga muutused. Kui perioodiks on aasta, siis nimetatakse selliseid muutusi **sesoonseteks muutusteks**.
- Jooksva aasta kalendrist sõltuvad muutused. Osade vaadeldavate suuruste väärtused sõltuvad näiteks töö- või kalendripäevade arvust kuus või kvartalis.
- Ebaregulaarsed, lühikeste ajavahemike järel toimuvad muutused.

Majandusest rääkides eristatakse sageli veel pikaajalist kindla suunaga muutumist ning nn majandustsüklist sõltuvaid ebaregulaarse pikkusega küllaltki pikaajalisi tõuse ja langusi, kuid andmete põhjal on neid muutumise tüüpe praktiliselt võimatu eristada. Aegrea osadeks jaotamisel nimetatakse seetõttu sageli suhteliselt aeglaselt toimuvat muutumist trend-tsükliks (inglise keeles *trend-cycle*).

Vaatlema näidetena kahte kuude kaupa defineeritud Eesti Statistikaameti veebilehelt allalaaditud aegrida - tarbijahinna indeksit ning majutatud turistide arvu. Tarbijahinna indeks on kujutatud joonisel 1.1. Joonise põhjal paistab, et tarbija-



Joonis 1.1: Tarbijahinna indeks 01.1998-07.2012(Statistikaameti andmed [1])

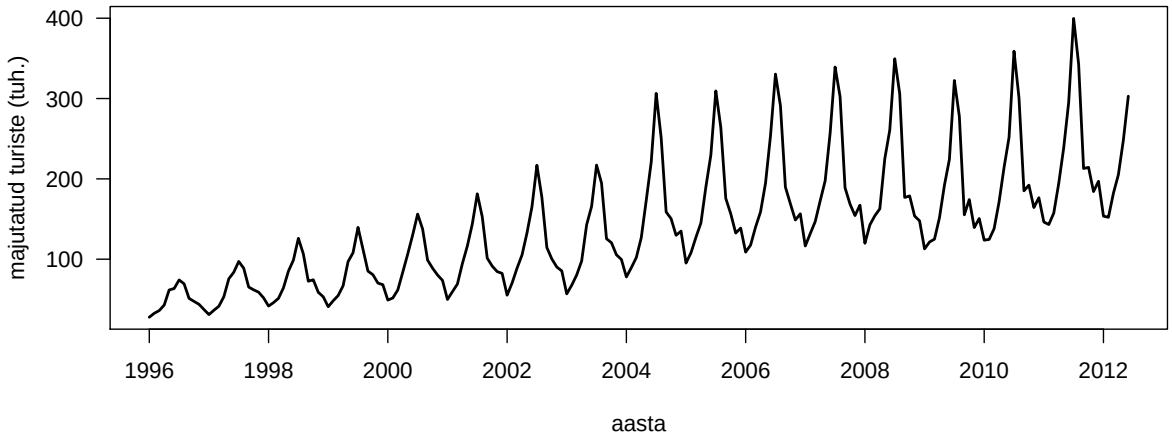
hinna indeksil on selgelt kasvav trend ning silmaga nähtavaid sesoonseid muutuseid ei paista. Samas aga majutatud turistide arvu aegreal paistab olema nii kasvav trend kui ka selge sesoonne komponent (vt. joonis 1.2. Oluline on aga mõista, et eelnevalt toodud kirjeldused trendi ja sesoonse (perioodilise) komponendi kohta ei ole matemaatilised definitsioonid ning erinevad inimesed võivad neid mõsta erinevalt. Selleks, et vastavate komponentide olemasolus korrektselt veenduda ning etteantud aegrida osadeks jaotada, tuleb kõigepealt täpsustada, mida täpselt mõeldakse. Sõltuvalt valitud definitsioonidest võib saada vägagi erinevaid tulemusi. Järgnevalt vaatleme põgusalt mõningaid võimalusi, kuidas aegrida jaotada erinevat tüüpi osadeks.

1.1.1 Trendi leidmise meetodid

Trendi puhul eristatakse nn globaalset, ajas muutumatu iseloomuga trendi ja lokaalset trendi, mis võib ajas pikkamööda muutuda.

Globaalse trendi eraldamine

Mõnikord on otstarbekas eeldada, et vaadeldava juhusliku suuruse pikaajalist käitumist ajas iseloomustab mingi küllalt lihtsal kujul olev funktsioon (lineaarne, ruut-



Joonis 1.2: Majutatud turistide arv kuude lõikes 01.1996-07.2012 (Statistikaameti andmed [2])

funktsioon, trigonomeetriline funktsioon, eksponentfunktsioon), mille ümber toimub võnkumine ebaregulaarsete häirituste ning perioodilise mõjutegurite tõttu. Vaatleme lihtsuse mõttes ainult juhtu, kus perioodilist komponenti ei ole; sel juhul tehakse sageli oletus, et aegrea andmed on kujul

$$z_t = f(\beta, t) + v_t,$$

kus f on mingi teadaolev parameetritest $\beta = (\beta_1, \ldots, \beta_p)'$ sõltuv funktsioon ning v_t on juhuslik kõrvalekalle. Sageli kasutatavaks meetodiks parameetrite β leidmiseks on vähimruutude meetod, mille korral leitakse β avaldise

$$\sum_t (f(\beta, t) - z_t)^2,$$

minimiseerimise teel. Summeerimine toimub siinjuures üle kõikide teadaolevate andmete. Sageli loetakse heaks suvalist funktsiooni f, mille korral saavutatakse vaadeldava summa piisavalt väike väärtus ning kasutatakse seda funktsiooni (trendi) tuleviku ennustamiseks. Selline lähenemine on aga sageli põhjendamatu, sest eriti aegridade puhul ei pruugi mineviku andmetega hästi sobiv funktsioon tuleviku ennustamiseks üldse sobida. Selleks, et veendunud olla vaadeldava meetodi sobivuses konkreetse andmestiku jaoks, tuleb lähtuda aegrea mudelist, mille korral vastav meetod annab mõistliku tulemuse. Selliseks mudeliks on

$$Z_t = f(\beta, t) + \varepsilon_t,$$

kus ε_t on sõltumatud sama jaotusega juhuslikud suurused (tegelikult piisab ka mittekorreleeritusest ja konstantsest dispersioonist). Juhul, kui funktsioon f sõltub kordajatest β lineaarselt, nimetatakse sellist lähenemist statistikas ka lineaarseks

regressiooniks; mittelineaarse sõltuvuse korral on tegemist mittelineaarse regressiooniga. Seega võime lugeda tulemusi usaldatavateks siis, kui pärast parameetrite leidmist järgi jäävad vead võib lugeda sõltumatute juhuslike suuruste väärtustele vastavaks; aegridade puhul juhtub seda harva.

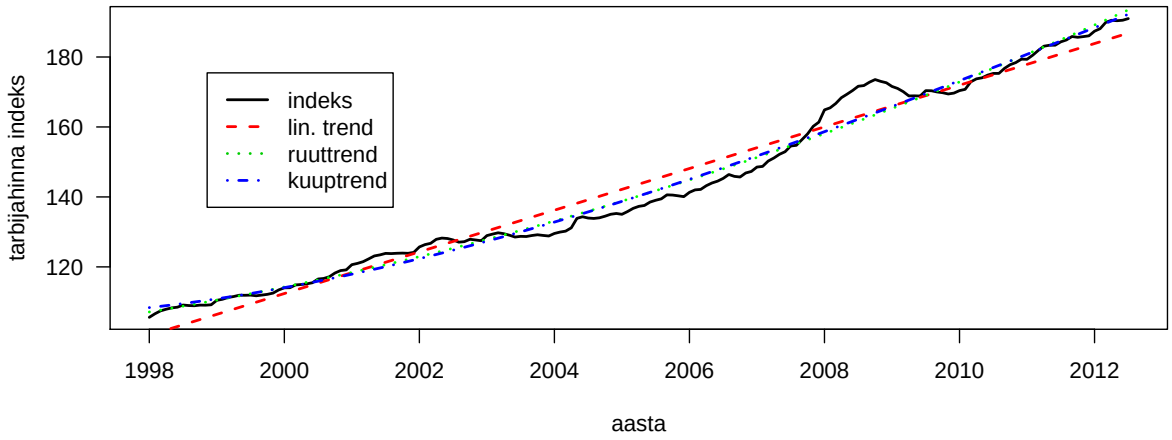
Vaatleme näitena lineaarse, ruut- ja kuupfunktsiooni sobitamist eelnevalt vaadeldud tarbijahinna indeksi andmetele. Näiteks lineaarse trendi sobitamise korral on funktsiooni f kujuks

$$f(\beta, t) = \beta_1 + \beta_2\, t$$

ning vähimruutude meetodil saame parimaks lineaarseks lähendiks

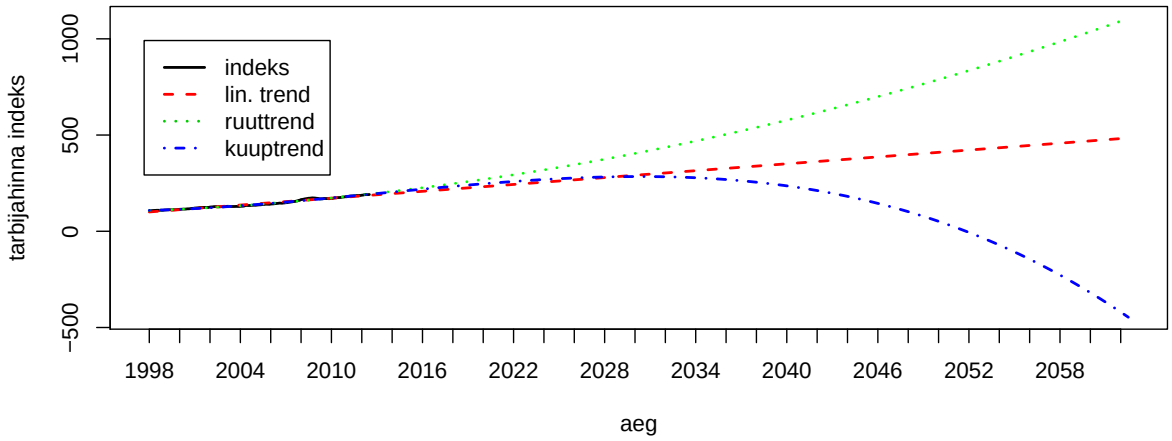
$$f(\hat{\beta}, t) = 99.97 + 0,4963\, t,$$

kus t on väljendatud kuudes alates 1998-nda aasta algusest (st 1998 jaanuar vastab väärtusele $t = 1$). Joonisel 1.3 näeme tarbijahinna indeksi väärtuseid koos lineaar-



Joonis 1.3: Tarbijahinna indeksi globaalse trendi hinnangud

sele, ruut -ja kuuptrendile vastavate kõveratega. Kasutades nende kõverate sobitamiseks mingit statistikatarkvara, võib kogenematul statistika rakendajal jääda mulje, et nad kõik on väga head tarbijahinna indeksi käitumise kirjeldamiseks (kõikidel juhtudel on kõik mudeli kordajad olulised väga madala olulisusnivoo korral ja determinatsioonikordajad on vägagi lähedased ühele), kusjuures ruut- ja kuuptrendid on vägagi sarnased, nii et nende vahel on raske valida. Samas on aga selge, et üldine majanduskeskkond on muutuv ning seetõttu on globaalse trendi olemasolu vägagi kaheldav; konkreetselt ei rahulda vaadeldud juhtudel andmed regressioonanalüüsi eelduseid (jääkide sõltumatus!) ja seetõttu tarkvara poolt väljastatud head sobivusnäitajad ei oma mõtet. Ka kaine mõistus peaks manitsema ettevaatusele: globaalse trendi eeldamine võib pikemas perspektiivis tähendada küllalt omapäraste tulemuste aktsepteerimist.



Joonis 1.4: Tarbijahinna indeksi globaalsete trendide tulevikuprognoosid

Lisades eelmisele graafikule trendikõverate poolt ennustatavad käitumised järgmiseks 50ks aastaks (vt. joonis 1.4), ennustab ruuttrend meile järjest kiirenevat tarbijahinna indeksi kasvu (nii et rahakotid peavad vägagi mahukaks muutuma), kuuptrendi uskumine aga tähendab, et 50 aasta pärast saab iga poeskäija lisaks kaubale ka vaevatasuks kopsaka rahasumma. Kokkuvõtteks: globaalse trendi olemasolu eeldus on praktilises andmeanalüüsis väga harva õigustatud ning lihtsa regressiooni abil sobitatud trendikõverate kasutamisel tuleviku ennustamiseks tuleb olla väga ettevaatlik. Konkreetseid meetodeid selle kindlakstegemiseks, et vaadeldud trendimudelid ei sobi käesoleval juhul tuleviku ennustamiseks, vaatleme hilisemates alapunktides.

Lokaalse trendi eraldamine

Kuna globaalse trendi olemasolu on väga harva põhjendatav, siis mõistetakse trendikõvera all enamasti aegrea suhteliselt aeglaselt muutuvat, "siledat" osa. Kahjuks ei ole aga olemas üldiselt aktsepteeritavat lokaalse trendi definitsiooni, mistõttu ei ole tegemist matemaatilise mõistega ning seetõttu on trendist rääkides vaja alati täpsustada, mida konkreetsel juhul selle all mõistetakse.

Lokaalse trendi leidmiseks tuleb aegreast eemaldada juhuslikest häiritustest tekitatud müra, seda protsessi nimetatakse *silumiseks* või *filtreerimiseks*. Väga sageli seisneb silumine uue aegrea tekitamises nn silutud keskmise leidmise abil.

Definitsioon 1 *Rea* (z_t) *teisendust kujul*

$$y_t = \sum_{i=-q}^{r} w_i z_{t-i}, \tag{1.1}$$

kus $w_i \geq 0$ *ja* $\sum_{i=-q}^{r} w_i = 1$, *nimetatakse* **libiseva keskmise** *leidmiseks. Juhul kui* $q = r$ *ja* $w_{-i} = w_i,\ i \leq q$, *nimetatakse sellist teisendust* **sümmeetriliseks libise-**

vaks keskmiseks *ning kui kõik kaalud w_i on võrdsed, on tegemist lihtsa libiseva keskmisega.*

Mõningad näited:

- Lihtne sümmeetriline libisev keskmine:

$$y_t = \frac{1}{2q+1} \sum_{i=-q}^{q} z_{t-i}.$$

  Sobib ka juhul, kui andmetes on paarituarvulise perioodiga $2q+1$ perioodiline komponent.

- Paarisarvulise perioodiga perioodilise komponendi olemasolu korral kasutatav lihtsa sümmeetrilise keskmistamise modifikatsioon:

$$y_t = \frac{1}{2q} \left(\frac{1}{2}(z_{-q} + z_q) + \sum_{i=-q+1}^{q-1} z_{t-i} \right), \tag{1.2}$$

  kus perioodi pikkuseks on $2q$.

- Eksponentsiaalne silumine:

$$y_t = \alpha \sum_{i=0}^{\infty} (1-\alpha)^i z_{t-i},$$

  kus $\alpha \in (0,\, 1)$ on mingi positiivne number. Praktilistes arvutustes kasutatakse eksponentsilumise omadust

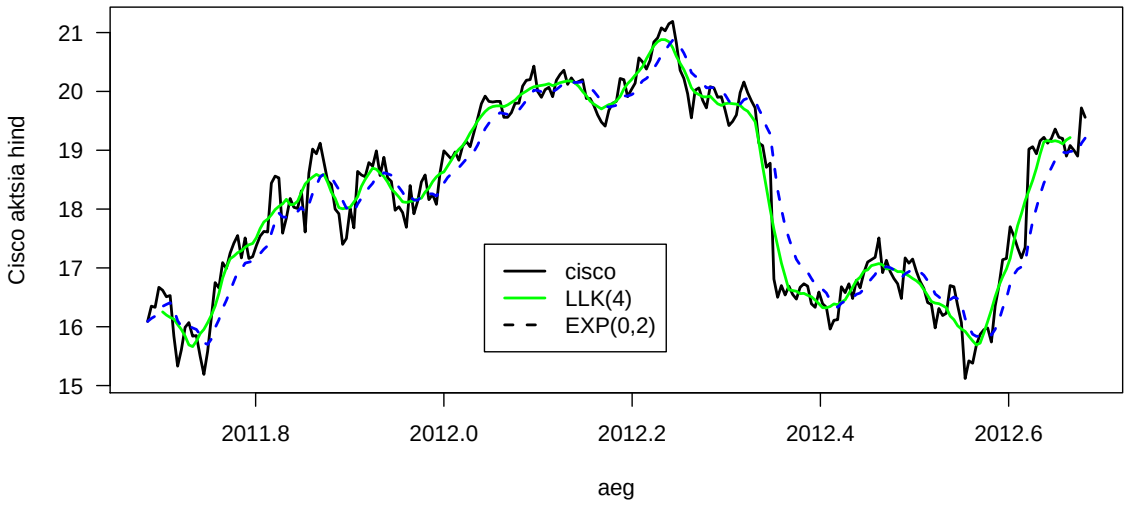
$$y_t = \alpha z_t + (1-\alpha) y_{t-1},$$

  mis võimaldab lihtsalt siluda lõplikku aegrida. Kui α läheneb ühele, siis silumist praktiliselt ei toimu ning mida väiksem on α, seda tugevam on silumine.

Joonisel 1.5 on toodud näited Cisco aktsia hinna silumisel saadud kõveratest.

Igasugune silumine peaks vähendama müra (eriti kuna müra kohta eeldame, et see on keskmiselt null). Samuti võib argumenteerida, et perioodilise komponendi puudumisel peaks silutud rida olema lähedane trendile, kuna müra on vähenenud ja aeglaselt muutuva trendi korral on selle väärtuste keskmine (vähemalt juhul, kui keskmist arvutatakse üle suhteliselt lühikese perioodi) lähedane tema hetkeväärtusele. Ideaalne oleks aga juht, kus vähemalt lihtsamate trendide korral saaksime müra puudumisel trendikõvera täpselt leida. Osutub, et see on võimalik.

Harjutus 1 *Näidata, et kui aegrea väärtused on antud lineaarse funktsiooni $f(t) = a + b\,t$ poolt (st $z_t = f(t)$), siis sümmeetrilise libiseva keskmise kasutamisel kehtib $y_t = f(t)$.*



Joonis 1.5: Lihtsa 9-tööpäevase libiseva keskmise ning eksponentsiaalse silumise ($\alpha = 0,2$) abil teisendatud Cisco aktsia hind

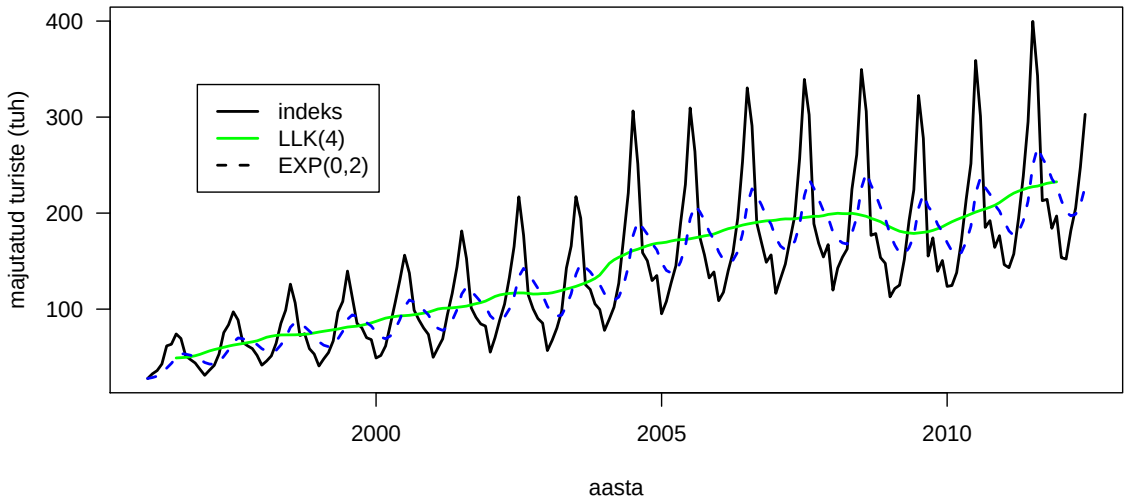
Sümmeetrilist keskmistamist ei ole aga alati võimalik rakendada. Näiteks aegrea lõpuosas puuduvad meil vajalikud tulevikuväärtused ning seetõttu on tuleviku prognoosimisel võimalik kasutada ainult ühepoolseid keskmisi, näiteks eksponentsiaalset keskmistamist. Sel juhul aga ei pruugi keskmistamisel leitud trendikõver isegi müra puudumisel langeda kokku õige trendiga.

Harjutus 2 *Olgu aegrea väärtused antud lineaarse funktsiooni $f(t) = a + b\,t$ poolt (st $z_t = f(t)$). Näidata, et sel juhul eksponentsiaalsel keskmistamisel saadav funktsioon on samuti lineaarne, leida selle kordajad. (Näpunäide: tekkiva lõpmatu summa leidmisel on võimalik kasutada geomeetrilise jaotusega juhusliku suuruse keskväärtuse valemit)*

Mittelineaarsete trendikõverate olemasolul ei anna ka sümmeetriline keskmistamine täpset tulemust, kuid on küllalt lihtne näidata, et juhul, kui andmeid on mõõdetud väikese intervalliga (ehk, ekvivalentselt, kui trendikõver muutub piisavalt aeglaselt), on sümmeetrilise keskmistamise tulemus müra puudumisel vähemalt piisavalt väikese ajaintervalli korral väga lähedane tegelikule trendikõverale. Samas aga järgneva harjutuse tulemus näitab, et juhul, kui andmeid on mõõdetud väikese intervalliga (ehk, ekvivalentselt, kui nn. trendikõver muutub piisavalt aeglaselt), on sümmeetrilise keskmistamise tulemus müra puudumisel vähemalt piisavalt väikese ajaintervalli korral väga lähedane tegelikule trendikõverale.

Harjutus 3 *(*) (lisapunktide saamiseks esitamise tähtaeg 27.09.2012) Olgu antud aegrida $z_t = f(\tau_t)$, kus $\tau_t = t\,h$, $t \in \mathbf{Z}$ ja h on fikseeritud ajaintervall. Olgu teada, et $|f''(\tau)| \leq c\ \forall \tau \in R$ mingi konstandi c korral. Näidata, et sel juhul lõpliku arvu nullist*

erinevate kaaludega sümmeetrilise libiseva keskmise kasutamisel kehtib hinnang $|y_t - f(\tau_t)| \le c_1\, h^2$, *kus* $y_t, t \in \mathbb{Z}$ *on keskmistamisel saadud rida* c_1 *sõltub ainult kaaludest ja konstandist* c.



Joonis 1.6: Paarisarvulisele perioodile vastava lihtsa sümmeetrilise keskmistamise modifikatsiooni abil silutud majutatud turistide arv ajas

Kui perioodilist komponenti sisaldava (sesoonse) rea trendi soovitakse silumise teel eraldada, siis peab silmas pidama, et keskmistamisel on sel juhul kaks eesmärki - juhuslike häirituste eemaldamine ning perioodiliste muutuste eemaldamine. Selleks on võimalik kasutada perioodi pikkusega kooskõlas oleva silumisaknaga keskmistamist, kuna üle terve perioodi summeerimisel peaks perioodiliste muutuste summa null olema. Näitena vaatleme majutatud turistide arvu lihtsat silumist. Kuna periood on antud juhul paarisarvuline (12 kuud), siis kasutame valemit 1.2. Tulemus on toodud joonisel 1.6. Nagu näha, eemaldab antud juhul lihtne keskmistamine aegreast perioodilised võnkumised ning tulemust võime lugeda trendikõveraks samal ajal, kui eksponentsiaalse keskmistamise korral jäävad silutud aegritta perioodilised muutused alles.

Näitame ka matemaatiliselt, et perioodilist komponenti sisaldava rea silumine eelmainitud tüüpi keskmistamise korral aitab küllalt hästi trendi eraldada. Selleks näitame, et kui andmed on kujul

$$z_t = a + b\,t + g(t),$$

kus g on perioodiga $2q$ funktsioon (st $g(t+2q) = g(t)\ \forall t$), siis valemiga (1.2) saadud silutud rida langeb kokku trendiga. Selleks, et lahutus trendiks ja perioodiliseks osaks oleks üheselt määratud, nõuame täiendavalt, et $\sum_{i=1}^{2q} g(t-i) = 0\ \forall t$. Seega

on meie eesmärgiks näidata, et $y_t = a + b\,t$. Arvutame:

$$\begin{aligned}
y_t &= \frac{1}{2q}\left(\frac{1}{2}(z_{-q}+z_q) + \sum_{i=-q+1}^{q-1} z_{t-i}\right)\\
&= \frac{1}{2q}\left(\frac{1}{2}(a+b\,(t+q)+g(t+q)) + \sum_{i=-q+1}^{q-1}(a+b\,(t-i)+g(t-i))\right.\\
&\qquad\left. +\frac{1}{2}(a+b\,(t-q)+g(t-q))\right)\\
&= \frac{1}{2q}\left(\frac{a+b\,t}{2} + \sum_{i=-q+1}^{q-1}(a+b\,t) + \frac{a+b\,t}{2}\right)\\
&\quad + \frac{1}{2q}\left(\frac{b\,q}{2} - \sum_{i=-q+1}^{q-1}(b\,i) - \frac{b\,q}{2}\right)\\
&\quad + \frac{1}{2q}\left(\frac{g(t+q)}{2} + \sum_{i=-q+1}^{q-1} g(t-i) + \frac{g(t-q)}{2}\right)\\
&= a + b\,t + 0 + \quad \frac{1}{2q}\left(\frac{g(t+q)}{2} + \sum_{i=-q+1}^{q-1} g(t-i) + \frac{g(t-q)}{2}\right).
\end{aligned}$$

Nüüd kasutame g perioodilisust:

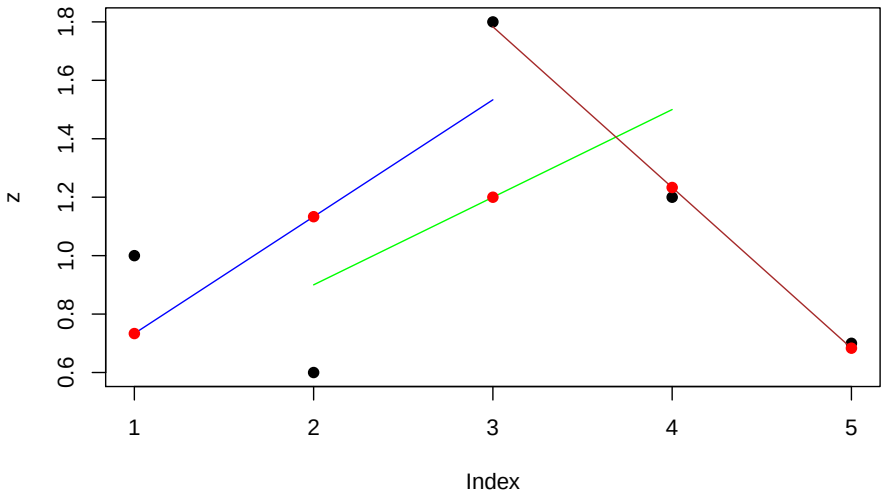
$$\begin{aligned}
&\frac{g(t+q)}{2} + \sum_{i=-q+1}^{q-1} g(t-i) + \frac{g(t-q)}{2}\\
&\qquad = \frac{g(t-q)}{2} + \sum_{i=-q+1}^{0} g(t-i-2\,q) + \sum_{i=1}^{q-1} g(t-i) + \frac{g(t-q)}{2}\\
&\qquad = \sum_{i=q+1}^{2\,q} g(t-i) + \sum_{i=1}^{q-1} g(t-i) + g(t-q)\\
&\qquad = \sum_{i=1}^{2\,q} g(t-i) = 0.
\end{aligned}$$

Sellega oleme näidanud, et $y_t = a + b\,t$.

Üheks küllaltki populaarseks meetodiks aegridade silumisel on ka nn *loess* või *lowess* (inglise keeles *locally weighted scatterplot smoothing*) meetod, mis silutud kõvera väärtuse leidmiseks mingis punktis sobitab kaalutud vähimruutude meetodil madala astme polünoomi läbi antud punktile lähedastele ajamomentidele vastavate aegrea väärtuste ning arvutades silutud kõvera väärtuse selle polünoomi abil. Idee tutvustamiseks vaatleme juhtu, kus meil on antud aegrida

$$z = (1, 0.6, 1.8, 1.2, 0.7).$$

Kõige lihtsamal kujul lokaalse regressiooni puhul tuleb otsustada polünoomi aste ning kasutatavate naabrite arv. Vaatleme lineaarset polünoomi (ehk sirget) ning



Joonis 1.7: Lokaalse regressiooni abil aegrea silumine

kasutame silutud rea leidmisel kolme lähimat väärtust (st jooksvale ajahetkele vastavat väärtust ja veel kahele lähimale ajahetkele vastavaid vaatluseid). Silumise protsess on kujutatud joonisel 1.7, kus esialgse aegrea väärtused on kujutatud mustade punktidena. Silutud rea esimese väärtuse leidmiseks leiame vähimruutude meetodil sirge, mis lähendab esimest kolme vaatlust võimalikult hästi (joonisel sinine sirge). Selle sirge väärtus ajal $t = 1$ ongi silutud rea esimeseks väärtuseks (joonisel punane punkt). Silutud rea teise väärtuse leidmine toimub sama sirge abil, kuna kasutusele tulevad samad z väärtused. Kolmanda punkti leidmisel tuleb sobitada sirge läbi teise, kolmanda ja neljanda z väärtuse (joonisel roheline) ning viimased kaks väärtust leitakse läbi viimase kolme z väärtuse sobitatud sirge abil (joonisel pruun).

Praktilisel kasutamisel antakse polünoomi sobitamisel igale kasutatavale z väärtusele veel kaal sõltuvalt selle kaugusest arvutatavast väärtusest. Täpsemalt võib nende meetodite kohta lugeda näiteks Wikipedia artiklist [3].

1.1.2 Dekompositsioonimeetodid. Sesoonne kohandamine.

Aegridade käitumisest arusaamine ja nende tõlgendamine on majanduses väga suure tähtsusega, seetõttu on loodud mitmeid meetodeid ja töövahendeid, mis võimaldavad neid osadeks lahutada. Ennem mõningate enim tunnustatud vahendite tutvustamist aga selgitame kasutatavaid mõisteid.

Aegrea osadeks lahutamisel tuleb kõigepealt otsustada, mismoodi need osad tervikus sisalduvad. Valdavalt vaadeldakse kahte juhtu: **aditiivne dekompositsioon**, mille

korral eeldatakse, et vaadeldav juhuslik suurus avaldub kujul

$$Z_t = T_t + S_t + I_t,$$

kus T on trend (või trend-tsükkel), S on perioodiline (sesoonne) komponent ja I on ebaregulaarne (juhuslik) komponent ehk müra; ning **multiplikatiivne dekompositsioon**, mille korral eeldatakse käitumist

$$Z_t = T_t S_t I_t.$$

Mõlemal juhul eeldatakse, et T sisaldab kogu informatsiooni keskmise taseme kohta ning S ja I kirjeldavad kõikumist keskmise ümber, st aditiivse lahutuse korral on S ja I keskmiselt nullid ning multiplikatiivsel juhul väljendavad nad suhet keskmisse väärtusesse (st on ise keskmiselt võrdsed ühega). Multiplikatiivset juhtu on võimalik taandada aditiivsele juhule esialgse aegrea logaritmimise teel. Loomulikult on võimalikud ka vahepealsed variandid, kus esineb nii liitmist kui ka korrutamist.

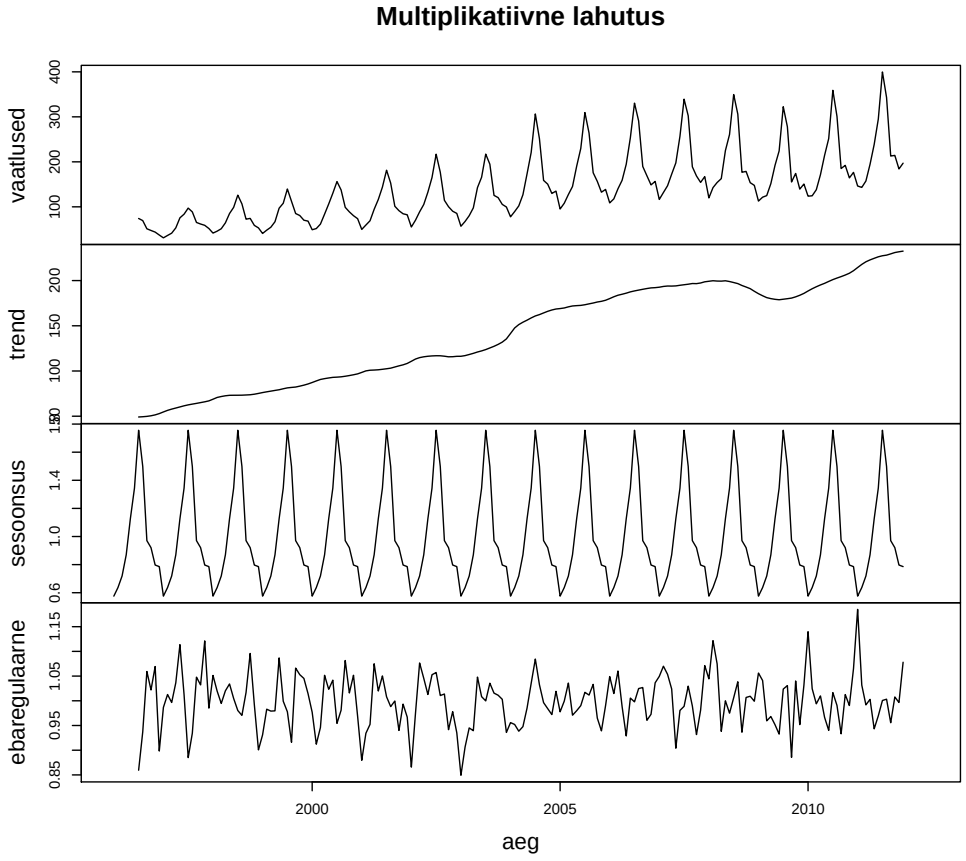
Sageli pakub suurt huvi eriti sesoonse komponendi eemaldamine, mida nimetatakse *sesoonseks kohandamiseks* (*seasonal adjustment*). Sesoonne kohandamine võimaldab paremini võrrelda aegrea järjestikuseid väärtuseid (näiteks uurida, kas majandus on tõusuteel, kui teise kvartali tulemus on parem esimese kvartali tulemusest).

Aegrea nn. klassikaline osadeks lahutamine toimub nii:

- kõigepealt leitakse trend sobiva sümmeetrilise libiseva keskmise abil,
- seejärel eemaldatakse reast trend (lahutades aditiivse rea korral ja jagades multiplikatiivse rea puhul),
- seejärel leitakse perioodiline komponent perioodile vastavaid alamridasid keskmistades,
- leitud perioodiline komponent normeeritakse nii, et aditiivse rea puhul oleks summa üle perioodi null ja multiplikatiivse rea korral oleks keskmine üle perioodi 1,
- Viimasena eraldatakse ebaregulaarne komponent esialgsest reast trendi ja perioodilise osa eemaldamise teel.

Majutatud turistide andmestiku korral saadud tulemused on kujutatud joonisel 1.8 Võimalik on vaadeldud protsessi ka itereerida, et trendi ja sesoonsust paremini eraldada, nimelt võib pärast esimest sesoonse osa leidmist omakorda selle esialgsest reast eemaldada ja seejärel leida uuesti trend (võib-olla teistsuguse libiseva keskmise kasutamisega kui varem), seejärel jällegi eemaldada esialgsest reast trend ja leida uuesti sesoonsus jne.

Eelnev nn klassikaline lahutus omab mitmeid puuduseid: ei ole selge, kas ja millisel määral saab leitud osasid tuleviku prognoosimisel kasutada ning samuti on eeldatud, et sesoonsuse iseloom ajas ei muutu. Seetõttu on välja töötatud mitmesuguseid täiendavate võimalustega osadeks jaotamise vahendeid, kuid lisavõimaluste

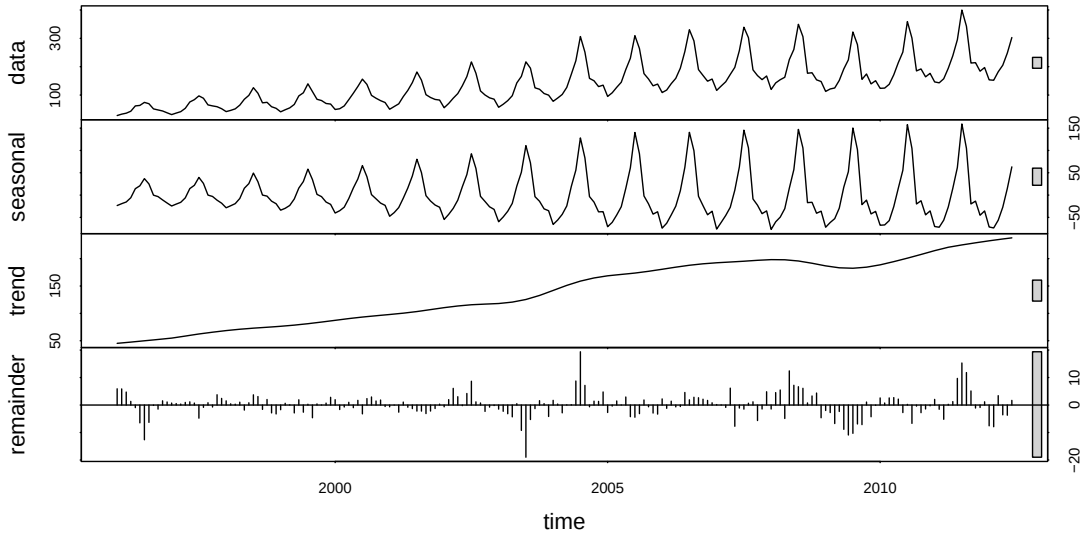


Joonis 1.8: Turistide rea osadeks lahutus multiplikatiivse esituse eeldusel

olemasolu muudab veelgi olulisemaks arusaama, et aegrea osade mõisted ei ole matemaatiliselt defineeritud ning nendest rääkides tuleb alati kokku leppida selles, mida konkreetselt silmas peetakse.

Mainime tuntud meetoditest kahte. Esiteks, STL (*Seasonal-Trend decomposition based on LOESS*) võimaldab jaotada aegrida aditiivseteks komponentideks, on kasutatav ka puuduvate väärtuste korral ning on realiseeritud näiteks tarkvarapaketis R. Teiseks, Ameerika Ühendriikide statistikaameti (*U. S. Census Bureau*) poolt kasutatavad dekompositsiooni ja sesoonse kohandamise meetodid on realiseeritud nende poolt hallatavas ning tasuta allalaaditavas tarkvarapaketis X-13ARIMA-SEATS, mis võimaldab aegreast kõrvaldada erindeid, võtta arvesse töö- ja puhkepäevade efekte, jaotada aegrida nii aditiivseteks kui ka multiplikatiivseteks osadeks ning arvutada ka mitmesugustel meetoditel põhinevaid prognoose ja sooritada diagnostilisi teste. Varasemale versioonile X-12-ARIMA vastav protseduur on olemas ka tarkvarapaketis SAS; tasuta versioonid mitmesuguste operatsioonisüsteemide jaoks on saadaval internetis aadressil [5].

Näitena aegrea osadeks jaotamisel saadavatest tulemustes on joonisel 1.9. STL kasutamise korral saab valida mitmesuguseid parameetreid trendi ja perioodilise osa leidmise konkretiseerimiseks. Näiteks joonisel toodud juhul on lubatud perioodilisel



Joonis 1.9: STL meetodil saadud majutatud turistide aegrea komponendid

komponendil ajas küllalt kiiresti muutuda ning seetõttu on tulemus oluliselt erinev sellest, mille saaksime muutumatut perioodilist komponenti eeldades.

Käesoleva kursuse põhirõhk on siiski tuleviku prognoosimisel, mistõttu me rohkem aegrea osadeks jaotamise küsimustel ei peatu.

Peatükk 2

Keskmistamisel põhinevad prognoosimeetodid. Prognoosimudeli headuse mõõdikud

2.1 Silumisel põhinevad lihtsamad prognoosivõtted perioodilist komponenti mittesisaldavate ridade jaoks

Olgu meil antud teatud kogus aegrea väärtusi $z_t,\ t = 1, 2, \ldots, T$. Sageli pakub huvi tulevikuväärtuste võimalikult täpne ennustamine, kusjuures on selge, et me saame selleks kasutada ainult teadaolevaid väärtuseid. Tähistame kujul $\hat{z}_{t_1|t}$ prognoosi aja t_1 jaoks, mis on saadud, kasutades teadaolevaid väärtuseid ajani kuni ajani t ning kui $t_1 = t + 1$, siis kasutame prognoosi jaoks lihtsustatud tähist $\hat{z}_{t_1}$.

2.1.1 Ilma trendita aegrea prognoosimine

Kui aegrea väärtused tunduvad käituvat täiesti juhuslikult või kui trendist põhjustatud muutlikuse osa on väga väike, siis on tuleviku jaoks küllaltki mõistlikuks ennustuseks eelnevate vaatluste keskmine (kuna sõltumatute sama jaotusega juhuslike suuruste korral on parimaks ennustuseks keskväärtus). Kui on alust arvata, et vaatlused on täiesti juhuslikud (st vastavad sõltumatute sama jaotusega juhuslike suuruste väärtustele), siis võib kasutada kõigi teadaolevate vaatluste keskmist:

$$\hat{z}_{t+p|t} = \frac{1}{t}\sum_{i=1}^{t} z_i.$$

Harjutus 4 *Hinnata juhul, kui Z_t on sõltumatud, sama jaotusega ning dispersiooniga σ^2 juhuslikud suurused, tõenäosust, et eelneva valemiga ennustatud tulemus erineb Z_{t+1} tegelikust väärtusest rohkem kui ε, st leida hinnang tõenäosusele*

$$P(|Z_{t+1} - \frac{1}{t}\sum_{i=1}^{t} Z_i| \geq \varepsilon).$$

Kui aga keskmine on ajas siiski muutuv, on parem kasutada selliseid keskmise valemeid, mis kõige värskematele vaatlustele annavad suurema kaalu, näiteks lihtsat libisevat keskmist

$$\hat{z}_{t+p|t} = \frac{1}{q}\sum_{i=0}^{q-1} z_{t-i}$$

või siis eksponentsiaalset keskmistamist

$$\hat{z}_{t+p|t} = \hat{z}_{t+1} = \alpha z_t + (1-\alpha)\hat{z}_t,$$

kusjuures tavaliselt võetakse sellisel juhul $\hat{z}_1 = z_1$

2.1.2 Trendiga aegrea prognoosimine. Holti meetod

Trendiga aegridade korral on vägagi loomulik nõuda, et prognoosimudel annaks täpse ennustuse vähemalt sellistel juhtudel, kui aegrida vastab täpselt lineaarsele funktsioonile. Libiseva keskmise kasutamisel tekib aga prognoosiviga: kui $z_t = a+b\,t$, siis libiseva keskmisega arvutatud prognoosi korral saame (arvestades, et kaalud summeeruvad üheks)

$$\begin{aligned}\hat{z}_{t+1} = y_t &= \sum_{i=0}^{q-1} w_i z_{t-i} = \sum_{i=0}^{q-1} w_i(a + b\,t - b\,i)\\ &= a + b\,t - b\sum_{i=0}^{q-1} w_i\,i.\end{aligned}$$

Vea parandamise üheks võimaluseks on parandada prognoosi kahekordse keskmistamise (st kasutades prognoosimisel ka keskmistatud suuruste y_t keskmist) abil. Nimelt kui me arvutame eelneval juhul

$$\bar{y}_t = \sum_{i=0}^{q-1} w_i y_{t-i} = a + b\,t - 2\,b\sum_{i=0}^{q-1} w_i\,i,$$

siis saame avaldada

$$a + b\cdot t = 2y_t - \bar{y}_t,\ b = \frac{y_t - \bar{y}_t}{\sum_{i=0}^{q-1} w_i\,i}$$

ning seega saame soovitud omadusega prognoosimeetodi kujul

$$\hat{z}_{t+p|t} = 2y_t - \bar{y}_t + \frac{y_t - \bar{y}_t}{\sum_{i=0}^{q-1} w_i\,i}p.$$

Harjutus 5 *Lihtsustage eelnevat avaldist lihtsa libiseva keskmise* $y_t = \frac{1}{q}\sum_{i=0}^{q-1} z_{t-i}$ *korral.*

Üheks küllalt populaarseks meetodiks trendiga aegridade prognoosimiseks on **Holti meetod**, mis tugineb eksponentsiaalsel keskmistamisel. Holti meetodi korral prognoosid vastavad lineaarsele funktsioonile:

$$\hat{z}_{t+p|t} = a_t + b_t\,p,$$

kus a_t arvustamisel kasutatakse z_t väärtust ning eelnevate andmete põhjal tehtud prognoosi:

$$a_t = \alpha\, z_t + (1-\alpha)\hat{z}_t = \alpha\, z_t + (1-\alpha)(a_{t-1} + b_{t-1})$$

ning b_t arvutamisel kasutatakse selle eelmise väärtuse ning a muutuse keskmist:

$$b_t = \beta\,(a_t - a_{t-1}) + (1-\beta)b_{t-1}.$$

Meetodi kasutamiseks tuleb valida või andmete põhjal hinnata väärtused a_1, b_1, α, β. Sageli valitakse $a_1 = z_1,\ b_1 = 0$ või siis kasutada nendeks teatud arvu esimeste z väärtuste jaoks leitud lineaarse regressioonikõvera vastavaid väärtusi. Parameetrite α ja β valikul võib kasutada näiteks mingi prognoosivea mõõdiku minimiseerimist, näiteks võib minimiseerida nende parameetrite järgi ennustusvigade ruutude summat

$$\sum_{t=2}^{n} (\hat{z}_t - z_t)^2.$$

Harjutus 6 *Näidata, et Holti meetod on sobivalt valitud a_1 ja b_1 korral täpne juhul, kui aegrida vastab lineaarsele funktsioonile.*

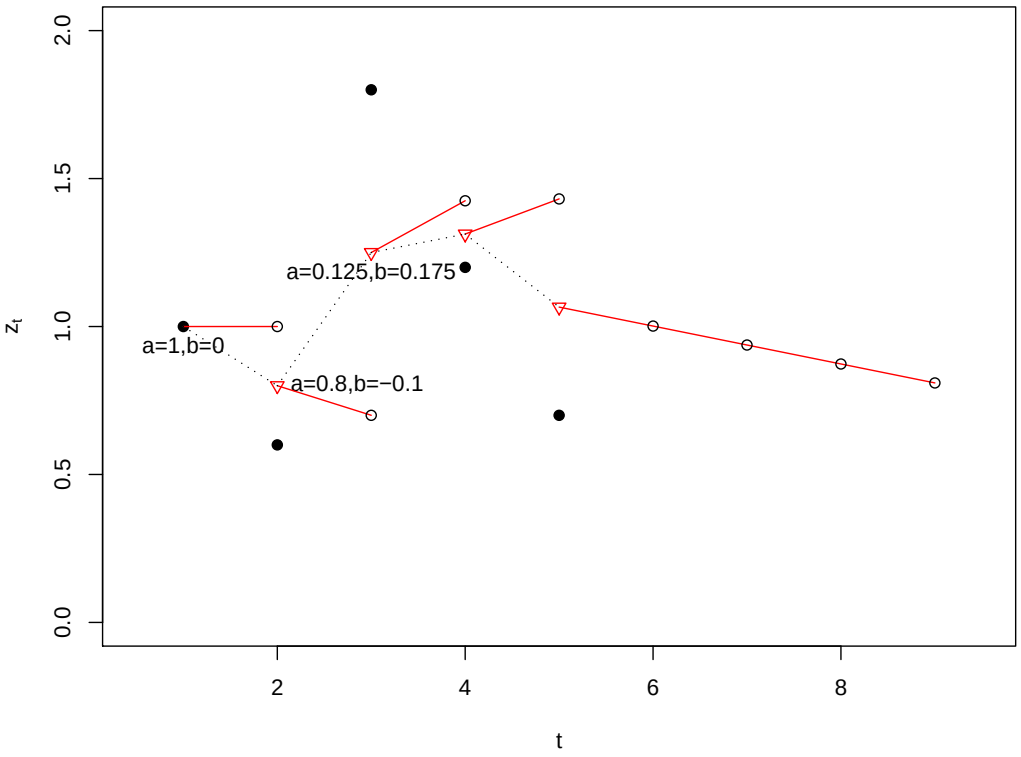
Meetodi töötamisega tutvumiseks vaatleme varasemast tuttavat näidet, kus aegrea väärtusteks on

$$z = (1, 0.6, 1.8, 1.2, 0.7).$$

Vaatleme prognoosimise protsessi juhul $\alpha = \beta = 0.5$ ning valime algväärtusteks $a_1 = 1, b_1 = 0$. Olgu meie eesmärgiks prognoosida Holti meetodiga järgmised 4 väärtust z_6, z_7, z_8, z_9. Selleks peame leidma a_5 ja b_5, milleks tuleb alustada valitud a_1 ja b_1 väärtustest ning liikuda mööda aegrida kuni ajani $t = 5$, arvutades iga ajamomendi jaoks vastavad taseme a ja tõusu b väärtused. Protsessi on graafiliselt kujutatud joonisel 2.1. Kõigepealt lähtume ajale $t = 1$ vastavast trendijoonest (sirge punktist $(1, a)$ tõusuga b), millele vastavaks prognoosiks $\hat{z}_2$ on 1 (joonisel kujutatud ringikesena). Ajale $t = 2$ vastava taseme väärtuse leiame nüüd prognoosi ja tegeliku väärtuse keskmisena (kuna $\alpha = 0.5$, siis leiame aritmeetilise keskmise), mis on joonisel kujutatud kolmnurgaga. Uue trendi leidmiseks võtame keskmise kaaluga β prognoosi leidmisel kasutatud tõusust (praegu 0) ja taseme a väärtuste muudust; leitud uus tõusukordaja b_2 vastab sirgele, mis läheb eelmise trendijoone ja punktiirina kujutatud a väärtusi ühendava sirge vahelt. Pärast a_2 ja b_2 arvutamist kordame varasemat protseduuri: leiame prognoosi $\hat{z}_3$ vastavalt trendijoonele, seejärel a_3 väärtuse prognoosi ja tegeliku väärtuse keskmisena ning lõpuks uue tõusukordaja b_3 jne. Ajamomendiks $t = 5$ oleme leidnud

$$a_5 = 1.065625,\ b_5 = -0.0640625$$

ning seega huvipakkuvad prognoosid leiame vastavalt sirgele, mis läbib punkti $(5, a_5)$ tõusuga b_5.



Joonis 2.1: Lokaalse regressiooni abil aegrea silumine

2.2 Holt-Wintersi meetod sesoonse aegrea prognoosimiseks

Järgnevalt eeldame, et aegreal on perioodiline komponent perioodiga s. Holt-Wintersi meetodil on kaks versiooni sõltuvalt sellest, kas me eeldame, et perioodiline komponent on korrutatud trendiga (multiplikatiivne mudel) või liidetud trendile (aditiivne mudel). Mõlemal juhul hinnatakse ennustamiseks jooksvat taset a, trendikõvera tõusu b ning sesoonset (perioodilist) komponenti S ning lisaks algväärtustele tuleb valida kolm silumistegurit α, β ja γ.

2.2.1 Multiplikatiivne Holt-Wintersi meetod

Ennustusvalemiks on sel juhul

$$\hat{z}_{t+p|t} = (a_t + p\, b_t) S_{t+p-s},\ \ p = 1, \ldots, s,$$

kus

$$a_t = \alpha \frac{z_t}{S_{t-s}} + (1-\alpha)(a_{t-1} + b_{t-1}),$$
$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1},$$
$$S_t = \gamma \frac{z_t}{a_t} + (1-\gamma)S_{t-s}.$$

Meetodi kasutamiseks alates ajamomendist $t = s+1$ tuleb ette anda $a_s, b_s, S_1, \ldots, S_s$ ning määrata sobivad kordajad α, β, γ. Kui kasutada erinevates tarkvarapakettides realiseeritud Holt-Wintersi meetodit samade andmete ja automaatse parameetrivaliku korral, siis võivad prognoosivead vähemalt alguses olla küllaltki erinevad, kuna etteantavate parameetrite automaatne valik on realiseeritud neis erinevalt.

2.2.2 Aditiivne Holt-Wintersi meetod

Ennustusvalemiks on sel juhul

$$\hat{z}_{t+p|t} = (a_t + p\, b_t) + S_{t+p-s},\ p = 1, \ldots, s,$$

kus

$$a_t = \alpha(z_t - S_{t-s}) + (1-\alpha)(a_{t-1} + b_{t-1}),$$
$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1},$$
$$S_t = \gamma(z_t - a_t) + (1-\gamma)S_{t-s}.$$

2.3 Prognoosimudeli headuse mõõdikud

Selleks, et võrrelda konkreetse andmestiku korral omavahel erinevaid prognoosimeetodeid, arvutatakse sageli mingit tüüpi keskmist prognoosiviga ühesammuliste prognooside jaoks. Mõned tuntumatest on

- Keskmine absoluutne viga (*mean absolute deviation*):

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |z_i - \hat{z}_i|$$

- Keskmine ruut viga (*mean square error*):

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (z_i - \hat{z}_i)^2$$

- Ruutkeskmine viga:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (z_i - \hat{z}_i)^2} = \sqrt{MSE}$$

- Keskmine suhteline viga (*mean absolute presentage error*):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|z_i - \hat{z}_i|}{z_i}$$

Mõningaid täiendavaid headuse mõõdikuid vaatleme hiljem. Vaadeldud näitajad annavad aimu sellest, kui hästi vaadeldav aegrida mingi meetodi korral prognoositav võib olla. Samas tuleb aga suhtuda arvutatud näidikutesse ettevaatusega, seda eriti juhul, kui meetodis kasutatavad parameetrid on leitud sama andmestiku põhjal, mille korral näidikuid arvutatakse (parameetrite valikuga võib saavutada hea kooskõla selleks kasutatava andmestikuga, kuid see kooskõla ei pruugi edasi kanduda uute andmete peale, nn. ülesobitamise efekt). Samuti tuleb uurida ennustusvigade juhuslikkust, sest kui ennustusvead ei ole omavahel sõltumatud, siis ei ole mingit garantiid, et mineviku põhjal arvutatud näitajad tuleviku kohta midagi ütlevad ning kindlasti on sel juhul võimalik prognoose avastatud sõltuvust kasutades parandada. Teisalt, kui prognoosivead on sõltumatud nullkeskmisega juhuslikud suurused, siis mudel on vähemalt ruutkeskmise vea suhtes parim võimalik. Kahjuks ei ole sõltumatust lõpliku aegrea baasil võimalik kindlaks teha, kuid see-eest on mitmeid teste, mis võimaldavad sõltuvust kindlaks teha. Nii et praktikas ennustusmeetodi valikul tuleb alati kontrollida, kas prognoosijäägid on sõltumatud. Kui tuleb välja, et ei ole, siis on (vähemalt teoreetiliselt) kindlasti võimalik leida parem prognoosimeetod. Aegridade analüüsimisel on üheks tähtsamaks sõltuvuse tüübiks ajaline sõltuvus, mille kindlakstegemisest tuleb juttu järgnevas peatükis.

2.4 Aegridade mudelid

Selleks, et oleks võimalik leida veapiire arvutatud prognoosidele veapiire ning tuletada prognoosimeetodite sobivuse kindlakstegemiseks ja omavaheliseks võrdlemiseks matemaatiliselt põhjendatud protseduure, tuleb teha eeldused selle kohta, kuidas juhuslikkus mõjutab vaadeldava aegrea väärtuseid, st tuleb kirjeldada aegrea mudel. Selleks, et minevikuväärtuste põhjal oleks võimalik midagi öelda ka tulevikus tekkivate juhuslike häirituste kohta, peab neil häiritustel olema korduv iseloom (ideaalis on nad sõltumatud ja sama jaotusega juhuslikud suurused). Järgnevas eeldame, et juhuslikkus aegrea väärtuste $z_1, \ldots, z_n$ tekkimisel on tulenenud (enamasti sõltumatute ja sama jaotusega) juhuslike suuruste $A_1, \ldots, A_n$ väärtustest $a_1, \ldots, a_n$. Seega me eeldame, et andmete tekkimise taga on juhuslik protsess (Z_t), kus Z_t võib sõltuda varasematest $Z_i,\ i < t$ väärtustest ning kuni ajani t tekkivatest juhuslikest suurustest $A_i,\ i \leq t$; meie näeme andmetena selle protsessi ühte võimalikku käitumist, mis vastab juhuslike suuruste A_i väärtustele a_i.

Definitsioon 2 *Aegrea mudelit nimetatakse olekuruumi mudeliks, kui protsess Z_t esitub kujul*

$$Z_t = w(X_{t-1}) + r(X_{t-1})A_t,$$
$$X_t = f(X_{t-1}) + g(X_{t-1})A_t,$$

kus (A_t) *in sõltumatud, sama jaotusega ja tsentreeritud juhuslikud suurused ning* $X_t = (X_{1,t}, \ldots, X_{m,t})$ *on olekuvektor. Olekuruumi mudelit nimetatakse lineaarseks, kui funktsioonid* $w()$ *ja* $f()$ *on lineaarsed funktsioonid,* $g()$ *on konstantne vektor ja* $r(X_{t-1}) = 1$*.*

Märkus 3 *Eelnevalt defineeritud olekuruumimudel on tuntud kui ühe veaallikaga mudel (*Single Source of Error model*). Laialdaselt on kasutatav ka mitme veaallikaga mudel (*Multiple Source of Error model*), mille korral igal ajasammul mõjutab nii olekuruumi kui ka järgmist vaatlust mitmemõõtmelise juhusliku vektori* A *väärtus,* $g()$ *ja* $r()$ *on sel juhul sobivate mõõtmetega maatriksid.*

Harjutus 7 *Leida Holti meetodi esitus olekuruumi mudelina eeldusel, et ühesammulise prognoosi viga on* A_t*:*

$$Z_t = a_{t-1} + b_{t-1} + A_t.$$

Definitsioon 4 *Aegrea mudelit nimetatakse ARIMA tüüpi mudeliks, kui* Z_t *avaldub lõpliku arvu varasemate* Z_i*,* $i < t$ *ja lõpliku arvu sõltumatute sama jaotusega häirituste* A_i*,* $i \leq t$ *lineaarkombinatsioonina kujul*

$$Z_t = \phi_0 + \sum_{i=1}^{p} \phi_i Z_{t-i} + A_t + \sum_{i=1}^{q} \psi_i A_{t-i}.$$

Harjutus 8 *Näidata, et ARIMA tüüpi mudel esitub lineaarse olekuruumi mudelina.*

Osutub aga, et lineaarne olekuruumi mudel esitub samuti ARIMA tüüpi mudeli kujul. Käesolevas kursuses keskendume põhiliselt lineaarsetele mudelitele ning lähtume peamiselt ARIMA tüüpi esitlusest.

Peatükk 3

Statsionaarsed aegread

3.1 Statsionaarsuse mõiste. Autokorrelatsioonifunktsioon

Selleks, et teadaolevate aegrea väärtuste põhjal saaks tulevikku ennustada ja ka prognoosivigu arvutada, peab tegema mingid eeldused, mis garanteerivad tulevikukäitumise ja tulevikus tekkivate juhuslikke häiritusi iseloomustava info sisalduvuse mineviku andmetes. Üheks aegridade teoorias sageli kasutatavaks eelduseks on nn statsionaarsuse nõue. Definitsiooni sõnastamiseks on aga vaja teada järgnevat mõistet.

Definitsioon 5 *Juhusliku vektori* $(X_1, \ldots, X_m)$ *järguga* k *momentideks nimetatakse suuruseid*

$$\mathbf{m}_{\alpha_1,\alpha_2,\ldots,\alpha_m} = E(X_1^{\alpha_1} X_2^{\alpha_2} \cdots X_m^{\alpha_m}),$$

kus $\alpha_i,\ i = 1, \ldots, m$ *on mittenegatiivsed täisarvud ning* $\sum_{i=1}^{m} \alpha_i = k$.

Näiteks vektoril (X_1, X_2, X_3) on kuus erinevat teist järku momenti, milleks on

$$\mathbf{m}_{2,0,0} = E(X_1^2),\ \mathbf{m}_{0,2,0} = E(X_2^2),\ \mathbf{m}_{0,0,2} = E(X_3^2),$$
$$\mathbf{m}_{1,1,0} = E(X_1X_2),\ \mathbf{m}_{1,0,1} = E(X_1X_3),\ \mathbf{m}_{0,1,1} = E(X_2X_3).$$

Definitsioon 6 *Juhuslikku protsessi* $(Z_t)_{t\in\mathbb{Z}}$ *nimetatakse (tugevalt) statsionaarseks, kui iga täisarvude komplekti* $t_1, \ldots, t_m$ *ja iga täisarvu* q *korral on juhuslikud vektorid* $(Z_{t_1}, \ldots, Z_{t_m})$ *ning* $(Z_{t_1+q}, \ldots, Z_{t_m+q})$ *sama jaotusega. Kui iga täisarvude komplekti* $t_1, \ldots, t_m$ *ja iga täisarvu* q *korral on juhuslike vektorite* $(Z_{t_1}, \ldots, Z_{t_m})$ *ning* $(Z_{t_1+q}, \ldots, Z_{t_m+q})$ *kõik kuni* k *järku momendid võrdsed, siis nimetatakse protsessi* Z_t k*-järku nõrgalt statsionaarseks.*

Statsionaarse protsessi näiteks on protsess, mis koosneb sõltumatutest sama jaotusega juhuslikest suurustest. Olgu meil tegemist teist järku nõrgalt statsionaarse protsessiga, siis juhul $m = 1$ järeldub definitsioonist, et

$$E(Z_t) = \mu,\ DZ_t = \sigma^2\ \forall t$$

mingite konstantide μ ja σ korral. Samuti järeldub, et suuruste Z_t ja Z_{t+p} kovariatsioon ja korrelatsioon (täpsemalt autokorrelatsioon ja autokorrelatsioon, kuna tegemist on sama protsessi eri ajamomentidele vastavate juhuslike suuruste kovariatsiooni ja korrelatsiooniga) sõltub ainult ajamomentide vahest p, väärtuste hulkasid

$$\gamma_p = \mathrm{cov}(Z_t, Z_{t+p}),\ \rho_p = \mathrm{cor}(Z_t, Z_{t+p}) = \frac{\gamma_p}{\sigma^2},\ p \in \mathbb{Z}$$

nimetatakse vastavalt protsessiZ_t autokovariatsioonifunktsiooniks ja autokorrelatsioonifunktsiooniks.

Harjutus 9 *Näidata, et sõltumatute sama jaotusega juhuslike suuruste ε_t abil defineeritud protsess*

$$Z_1 = \varepsilon_1;\ Z_t = Z_{t-1} + \varepsilon_t,\ t > 1$$

ei ole statsionaarne.

Harjutus 10 *(*, lisapunktide saamiseks esitada 18.10.2012) Näidata, et sõltumatute standardse normaaljaotusega juhuslike suuruste ε_t abil defineeritud protsess*

$$Z_t = \varepsilon_t + \frac{1}{2}\varepsilon_{t-2}$$

on statsionaarne.

Kui meil on teada statsionaarsele protsessile vastavad aegrea väärtused ajamomentidel $t = 1, 2, \ldots, N$, siis arvutatakse empiiriliste autokovariatsiooni ja autokorrelatsioonifunktsiooni väärtused tavaliselt valemitega

$$\begin{aligned}
\bar{\mu} &= \frac{1}{N}\sum_{t=1}^{N} z_t, \\
c_p &= \frac{1}{N}\sum_{t=1}^{N-p}(z_t - \bar{\mu})(z_{t+p} - \bar{\mu}),\ p = 0, 1, \ldots, N-1, \\
r_p &= \frac{c_p}{c_0}.
\end{aligned}$$

Harjutus 11 *(*, lisapunktide saamiseks esitada 18.10.2012) Näidata, et kui statsionaarse protsessi korral kehtib $\lim_{p\to\infty} |\gamma_p| = 0$, siis*

$$P(|\bar{\mu} - \mu| \geq \varepsilon) \underset{N\to\infty}{\to} 0$$

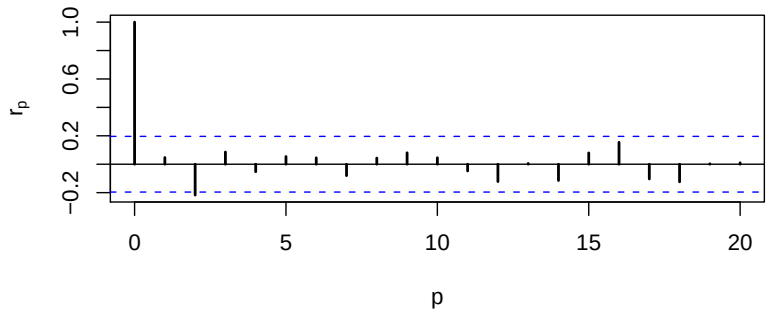
iga $\varepsilon > 0$ korral.

Tähtis on aru saada, et empiirilisi autokorrelatsioone ja autokovariatsioone saame me arvutada suvaliste andmeridade põhjal, kuid mittestatsionaarse rea korral ei iseloomusta saadavad numbrid mingite konkreetsete juhslike suuruste korrelatsioone ja kovariatsioone, seega nende mõistlik tõlgendamine on väga raske.

Kui meil on tegemist sõltumatutest sama jaotusega juhuslikest suurustest koosneva protsessiga, siis kõik teoreetilised korrelatsioonid $\rho_p,\ p > 0$ on võrdsed nulliga, kuid

protsessile vastava lõpliku pikkusega aegrea põhjal arvutatud hinnangud $r_p,\ p > 1$ on üldiselt nullist erinevad. Seetõttu on väga oluline teada mingeid kriteeriume, mille põhjal otsustada konkreetse aegrea empiirilise autokorrelatsioonifunktsiooni abil, kas aegrida võib vastata täiesti juhuslikule protsessile. Selleks tutvume kahe tulemusega.

Esiteks, on teada (vt [6]), et sõltumatutele sama jaotusega juhuslikele suurustele vastava aegrea korral on suurused $r_p,\ p > 1$ asümptootiliselt (vaatluste arvu N kasvades) normaaljaotusega, kusjuures keskväärtus on suurte N väärtuste korral ligikaudu $-\frac{1}{N}$ ja standardhälve $\frac{1}{\sqrt{N}}$. Seega piisavalt suure N korral peaks iga konkreetse $p > 0$ korral jääma tõenäosusega 0.95 vahemikku $[-\frac{1}{N} - \frac{2}{\sqrt{N}}, -\frac{1}{N} + \frac{2}{\sqrt{N}}]$. Tavaliselt statistikatarkvaras on empiirilise autokorrelatsioonifunktsiooni graafilisel esitamisel vastavad piirid ka joonisel välja toodud.



Joonis 3.1: Saja sõltumatu normaaljaotusega juhusliku suuruse väärtuste põhjal arvutatud autokorrelatsioonid

Eelnev tulemus kehtib iga üksiku autokorrelatsioonikordaja suhtes. Samas arvutatakse neid kordajaid tavaliselt mitu ning näiteks 20 kordaja arvutamisel on loomulik, et keskmiselt ühe kordaja väärtus satub väljapoole 95% veapiire. Näiteks joonisel 3.1 on kujutatud sõltumatute juhuslike suuruste väärtuste rea põhjal arvutatud autokorrelatsioonikordajad ning nendest r_2 väärtus on väljaspool 95% veapiire. Seetõttu oleks hea teada, kas terve väljaarvutatud autokorrelatsioonikordajate komplekt võib vastata sõltumatutele juhuslikele suurustele. Selleks sobib näiteks Ljung-Box test, mis põhineb suurusel

$$Q = N(N+2)\sum_{p=1}^{m}\frac{r_p^2}{N-p},$$

kus m näitab, kui mitmest esinevast autokorrelatsioonist koosnevat rühma testitakse. On teada, et see statistik on asümptootiliselt m vabadusastmega χ^2 jaotusega ning selle põhjal saab hinnata tõenäosust, et vaadeldavad kordajad vastavad sõltumatutele juhuslikele suurustele.

3.2 Periodogramm ja spekter

Kui me vaatleme aegrida väärtustega $z_1, \ldots, z_N$, siis see kujutab endast vektorit N-mõõtmelises ruumis, tähistame seda vektorit kujul $\mathbf{z}$. Lineaaralgebrast on teada, et iga selline vektor on esitatav üheselt N lineaarselt sõltumatu vektori $\mathbf{v}_1, \ldots, \mathbf{v}_n$ lineaarkombinatsioonina kujul

$$\mathbf{z} = \sum_{i=1}^{N} c_i \mathbf{v}_i,$$

kus $c_i,\ i = 1, \ldots, N$ on reaalarvulised kordajad. Kui vektorid $\mathbf{v}_i,\ i = 1, \ldots, N$ on hästi valitud, siis võivad leitud kordajad anda meile olulist informatsiooni vaadeldava rea omaduste kohta.

Signaaliteoorias on tavaks esitada signaale erinevate sagedustega siinuste ja koosinuste (nn harmoonikute) summana. Eeldame järgnevalt lihtsuse mõttes, et vaatluste arv N on paaris, siis me saame aegrea väärtused z_t esitada kujul

$$z_t = a_0 + \sum_{i=1}^{\frac{N}{2}} \left(a_i \cos(\frac{2\pi i t}{N}) + b_i \sin(\frac{2\pi i t}{N}) \right),$$

kus kordajad on arvutatud järgmiselt:

$$\begin{aligned}
a_0 &= \frac{1}{N} \sum_{t=1}^{N} z_t, \\
a_i &= \frac{2}{N} \sum_{t=1}^{N} z_t \cos(\frac{2\pi i t}{N}),\ b_i = \frac{2}{N} \sum_{t=1}^{N} z_t \sin(\frac{2\pi i t}{N}),\ i = 1, \ldots, \frac{N}{2} - 1, \\
a_{N/2} &= \frac{1}{N} \sum (-1)^t z_t,\ b_{N/2} = 0.
\end{aligned}$$

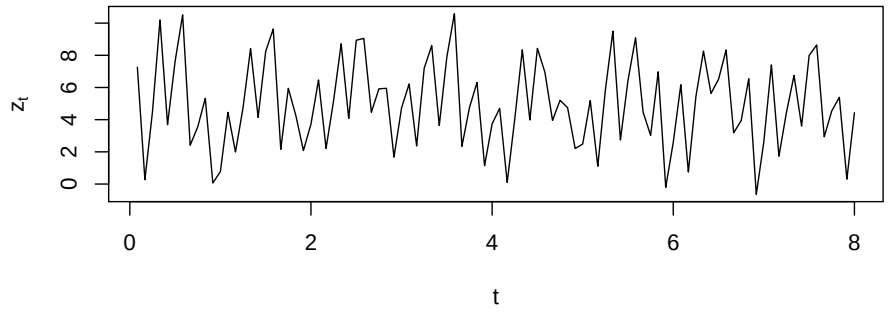
Periodogrammi väärtusteks on sel juhul

$$I(i/N) = \frac{N}{2}(a_i^2 + b_i^2),\ i = 1, \ldots, \frac{N}{2} - 1,\ I(\frac{1}{2}) = N a_{N/2}.$$

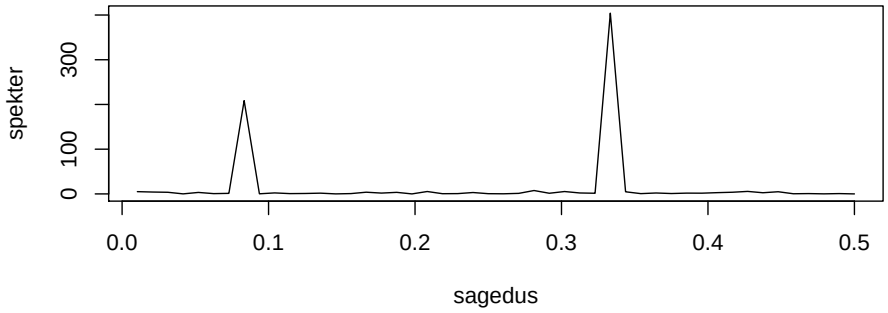
Kui a_i ja b_i definitsioonis asendada suurus $\frac{i}{N}$ arvuga $f,\ 0 < f \leq \frac{1}{2}$, siis aegrea **spektriks** nimetatakse suurust

$$I(f) = \frac{N}{2}(a_f^2 + b_f^2).$$

Spektri suur väärtus mingi f korral näitab, et andmestikus on oluliselt esindatud perioodiline komponent parioodiga $\frac{1}{f}$. Samas, kui andmed vastavad sõltumatutele juhuslikele suurustele, siis on tegemist nn valge müraga ning spektris peaks kõik sagedused olema üsna võrdselt esitatud. Nende omaduste baasil on loodud mitmeid teste perioodilise komponendi olemasolu kindlakstegemiseks ning samuti juhuslikkuse kindlakstegemiseks. Vaatleme näitena joonisel 3.2 toodud aegrida. Visuaalselt on selle käitumise kohta raske midagi öelda. Leides aga periodogrammi väärtused



Joonis 3.2: Aegrida spektri kasutamise näite jaoks



Joonis 3.3: Näiteaegrea spekter

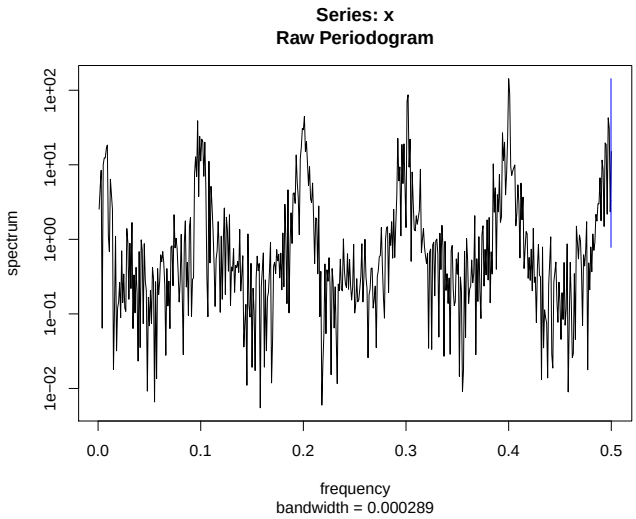
ning kujutades neid graafiliselt, saame joonisel 3.3 kujutatud graafiku. Siit paistab, et andmetes on olulisel määral esindatud kaks sagedust (üks alla 0.1 ja teine umbes 0.33). Kui uurida andmeid täpsemalt, siis vastavad sagedused on $\frac{1}{12}$ ja $\frac{1}{3}$. Kui tegemist on näiteks igakuiste andmetega, siis see vastab aastasele perioodile ja kvartaalsele (kolmekuulisele) perioodile. Tegelikkuses oli aga vastav aegrida genereeritud kujul

$$z_t = 5 + 3\sin(\frac{2\pi t}{3}) - 2\cos(\frac{2\pi t}{12}) + a_t,$$

kus a_t vastasid standardse normaaljaotusega juhuslike suuruste väärtustele. Seega võib öelda, et spektri uurimisega oli võimalik tuvastada andmetes peituva signaali omadusi.

Samuti võib saada kasulikku infot ka juhul, kui deterministlikku signaali reas ei esine, kuid on olemas tugev sõltuvus mingi ajalise nihke tagustest minevikuväärtustest. Vaatleme näiteks järgnevat R tarkvara abil arvutatud periodogrammi.

Siit on näha, et vastavas reas on kõige tugevamalt esitatud perioodiga 10 (sage-



dus 0.1), perioodiga 5 ja veel kahe väiksema perioodiga võnkumised. Osutub, et mingi nihkega minevikuandmetest sõltumine tekitab sageli ka täiendavaid perioodilisi võnkeid, mille perioodiks on nihkele vastava perioodi jagatis mingi täisarvuga. Vaadeldud periodogramm vastas tegelikult mudeli

$$Z_t = 0.8Z_{t-1} + A_t$$

abil genereeritud andmetele, nii et sisuline sõltuvus on perioodiga 10 ja väiksema perioodiga võnkumised on juhuslikult tekkinud.

Kasulikke rakendusi on spektraalanalüüsil palju ja tegelikult on kogu info, mis on aegrea autokorrelatsioonides, olemas ka tema spektris. Käesolevas kursuses me aga spektri ja periodogrammi kasutamist aegrea mudelite sobivuse kindlakstegemiseks ei kasuta.

Peatükk 4

Lineaarsed mudelid ühemõõtmelise aegrea jaoks

Käesolevas peatükis vaatleme selliseid aegrea mudeleid, kus aegrea hetkeväärtus avaldub lineaarse kombinatsioonina selle minevikuväärtustest ning juhusliku häirituse hetkeväärtusest ja minevikuväärtustest. Selliseid mudeleid nimetatakse lineaarseteks mudeliteks. Kuna lõpliku hulga andmete põhjal on võimalik leida ainult lõplik arv mudeli parameetreid, siis pakuvad erilist huvi sellised protsessid, mis on kirjeldatavad lõpliku arvu parameetrite abil.

4.1 Üldine lineaarne protsess, selle esitused, statsionaarsus ja pööratavus

On küllalt loomulik, et enamiku huvipakkuvate juhuslike protsesside korral hetkeväärtus sõltuv väga vähe selle protsessi kauge mineviku väärtustest ning seega võib öelda, et hetkeväärtus on sisuliselt määratud ainult juhuslikest häiritustest, mis minevikus on toimunud. Matemaatiliselt on kõige lihtsam uurida selliseid protsesse, kus sõltuvus häiritustest on lineaarne. Anname sellele kirjeldusele matemaatiliselt korrektse definitsiooni.

Definitsioon 7 *Üldiseks lineaarseks protsessiks nimetatakse protsesse, mis on esitatavad kujul*

$$\tilde{Z}_t = Z_t - \mu = A_t + \sum_{i=1}^{\infty} \psi_i A_{t-i}, \tag{4.1}$$

kus ψ_i on mingid reaalarvud, μ on suuruste Z_t keskväärtus ning $(A_t)_{t\in\mathbb{Z}}$ on vähemalt teist järku statsionaarne tsentreeritud ning mittekorreleeritud väärtustega protsess.

Eelnevat definitsiooni motiveerib järgmine tulemus, mida nimetatakse Woldi lahutuseks.

Teoreem 8 *(Woldi lahutus, vt. [7]) Iga statsionaarne protsess Z_t on esitatav kujul*

$$Z_t = \sum_{i=0}^{\infty} a_i \xi_{t-i} + \eta_t,$$

kus ξ_t on mittekorreleeritud protsess ning η_t on deterministlik protsess.

Nii et üldised lineaarsed protsessid on sellised statsionaarsed protsessid, mille Woldi lahtutuses on suurused ξ_t sõltumatud (või vähemalt teist järku statsionaarsed) ja mille deterministlik osa on konstantne.

Käesolevas peatükis eeldame, et A_t on sõltumatud ja sama jaotusega juhuslikud suurused keskväärtusega 0 ja standardhälbega σ_A. Selleks, et toodud lõpmatu summa defineeriks korrektselt juhusliku suuruse, peavad kordajad ψ_i rahuldama mingeid tingimusi. Täpsemalt, selleks piisab (vt [7], Teoreem 7.6.1), kui kehtib

$$\sum_{i=1}^{\infty} \phi_i^2 < \infty.$$

Harjutus 12 *(*, lisapunktide saamiseks esitada kuni 25.10.2012) Näidata, et kui protsess A_t on nullkeskmisega (tsentreeritud), konstantse dispersiooniga ning mittekorreleeritud (st $E(A_t A_{t+p}) = 0 \ \forall p \neq 0$), siis eelneva tingimuse täidetuse korral on protsess Z_t nõrgalt teist järku statsionaarne protsess.*

Aegridade mudelite esitamisel ja uurimisel on kasulik tuua sisse mõned tähistused. Esiteks, defineerime tagasi- ja edasinihke operaatorid

$$BZ_t = Z_{t-1}, \ FZ_t = Z_{t+1} \ \forall t.$$

Tehniline märkus. *Matemaatiliselt korrektne protseduur on järgmine: vaatleme protsesside ruumi $(Z_t)_{t\in\mathbb{Z}}$ normiga $\|(Z_t)_{t\in\mathbb{Z}}\| = \sup_{t\in\mathbb{Z}} EZ_t^2$. Operaator B teisendab siis ühe selle ruumi elemendi (protsessi) teiseks (kusjuures tegelikult oleks õige kirjutada $(BZ)_t = Z_{t-1}$, st B teisendab protsessi Z uueks protsessiks, mille t-s element on esialgse protsessi väärtus kohal $t-1$) ning $\|B\| = 1$.*

Paneme tähele, et $F = B^{-1}$ ning $B^j Z_t = Z_{t-j}$. Seega üldine lineaarne protsess on esitatav kujul

$$\tilde{Z}_t = (1 + \sum_{i=1}^{\infty} \psi_i B^i) A_t.$$

Edaspidises on meil kasulik defineerida funktsioonid operaatorist B.

Definitsioon 9 *Olgu f mingi reaalarvuliste väärtustega reaalmuutuja funktsioon, mis on esitatav punkti 0 ümbruses koonduva astmereana, st*

$$f(x) = \sum_{i=0}^{\infty} c_i x^i, \ |x| \leq \delta$$

mingi $\delta > 0$. Olgu M mingi pidev lineaarne operaator mingil Banachi ruumil Y. Siis $f(M)$ tähistab (formaalselt) operaatorit

$$f(M) = \sum_{i=0}^{\infty} c_i M^i.$$

Lihtne on näidata, et kui $\|M\| \leq \delta$, siis eelnevalt toodud formaalne definitsioon omab mõtet, st see summa koondub mingiks ruumil Y tegutsevaks pidevaks lineaarseks operaatoriks.

Defineerime funktsiooni

$$\psi(x) = 1 + \sum_{i=1}^{\infty} \psi_i x^i,$$

siis eelneva definitsiooni kohaselt võime üldise lineaarse protsessi kirjutada kujul

$$\tilde{Z}_t = \psi(B) A_t.$$

Siin tekib huvitav küsimus: millal me saame protsessi Z väärtusi teades teha kindlaks, millised häiritused A süsteemi on saabunud. Täpsemalt, defineerime mainitud omaduse matemaatiliselt.

Definitsioon 10 *Üldist lineaarset protsessi* (4.1) *nimetatakse* pööratavaks*, kui selle protsessi saab esitada autoregressiivsel kujul*

$$\tilde{Z}_t = \sum_{i=1}^{\infty} \pi_i \tilde{Z}_{t-i} + A_t.$$

Osutub, et pööratavuse jaoks saab anda üsna lihtsa piisava tingimuse.

Lemma 11 *Kui $\psi(x)$ astmerida koondub $|x| \leq 1$ korral ning funktsioon $\pi(x) = \frac{1}{\psi(x)}$ on esitatav astmereana, mis samuti koondub $|x| \leq 1$ korral, siis on üldine lineaarne protsess* (4.1) *pööratav ning kehtib võrdus*

$$\pi(B)\tilde{Z}_t = A_t,$$

kus $\tilde{Z}_t$ on protsess (4.1) *.*

Olgu funktsiooni $\pi(x)$ astmereaks

$$\pi(x) = 1 - \sum_{i=1}^{\infty} \pi_i x^i,$$

siis pööratav üldine lineaarne protsess on esitatav ka kujul

$$\tilde{Z}_t = \sum_{i=1}^{\infty} \pi_i \tilde{Z}_{t-i} + A_t.$$

Näide 12 *Vaatleme protsessi*

$$Z_t = A_t - \theta_1 A_{t-1},$$

kus $|\theta_1| < 1$. Sel juhul geomeetrilise rea summa valemi kohaselt

$$\pi(x) = \frac{1}{1 - \theta_1 x} = \sum_{i=0}^{\infty} \theta_1^i x^i, \ |x| < \frac{1}{|\theta_1|}.$$

Seega kehtib võrdus

$$Z_t = -\sum_{i=1}^{\infty} \theta_1^i Z_{t-i} + A_t.$$

Kui näiteks $\theta_1 = -0.2$, siis kahanevad suure nihkega Z väärtuste kordajad väga kiiresti nulli ning seetõttu on lõplike andmemahtude juures praktika seisukohalt peaaegu võimatu teha kindlaks, kas andmed vastavad mudelile

$$Z_t = A_t + 0.2A_{t-1}$$

või hoopis mudelile

$$Z_t = 0.2Z_{t-1} - 0.04Z_{t-2} + A_t.$$

Sellises situatsioonis eelistame tulevikus kindlasti esimest mudelit, sest selle sobitamisel andmetega tuleb leida ainult üks tundmatu parameeter (θ_1) teise mudeli kahe parameetri asemel.

Harjutus 13 *Näidata, et üldise lineaarse protsessi autokovariatsioonid on antud valemiga*

$$\gamma_k = \sigma_A^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}.$$

4.1.1 Osaautokorrelatsioonid

Lisaks autokorrelatsioonidele on etteantud aegreale sobiva mudeli valikul suureks abiks osaautokorrelatsioonid. Defineerime selle mõiste. Selleks aga on eelnevalt vaja veel ühte mõistet.

Definitsioon 13 *Olgu X ning $Y_1, \ldots, Y_k$ mingid lõpliku dispersiooniga juhuslikud suurused. Suuruse X projektsiooniks suurustega $Y_1, \ldots, Y_k$ määratud alamruumile nimetatakse suurust kujul*

$$\bar{X} = \sum_{i=1}^{k} c_i Y_i,$$

mille korral $E((X - \bar{X})^2)$ on minimaalne. Operaatorit $P : X \to \bar{X}$ nimetatakse vähimruutude projektoriks suurustega $Y_1, \ldots, Y_k$ määratud alamruumile.

Eelnev definitsioon tugineb teadmisele, et selline projektsioon on ühene. Neile, kes on tuttavad funktsionaalanalüüsi põhitulemustega, võib meelde tuletada, et Hilberti ruumides (milleks on ka lõpliku dispersiooniga juhuslike suuruste ruum, kus normiks on juhusliku suuruse ruudu keskväärtus) kehtib üldisem tulemus, et projektsioon igale kinnisele kumerale hulgale on üheselt määratud. Samas ei ole selles ka kuigi raske ise veenduda.

Harjutus 14 *Näidata, et kui X, W_1 ja W_2 on lõpliku dispersiooniga juhuslikud suurused, mille korral kehtivad tingimused*

$$E[(X - W_1)^2] = E[(X - W_2)^2],\ E[(W_1 - W_2)^2] > 0,$$

siis

$$E[(X - \frac{W_1 + W_2}{2})^2] < E[(X - W_1)^2].$$

Tõestada selle omaduse abil, et eelnevad definitsioonis kirjeldatud projektsioon on ühene (sest tingimusest $E[(W_1 - W_2)^2] = 0$ *järeldub, et* $W_1 = W_k$ *peaaegu kindlasti, st tõenäosusega 1).*

Juhuslike suuruste kontekstis vastab eelnevalt defineeritud projektor juhusliku suuruse X parimale suuruste $Y_1, \ldots, Y_k$ lineaarkombinatsiooni kujul avalduvale prognoosile.

Harjutus 15 *Näidata, et kui* $X, Y_1, \ldots, Y_k$ *on keskväärtusega 0 ja* $\bar{X}$ *on suuruse* X *vähimruutude projektsioon suurustega* $Y_1, \ldots, Y_k$ *määratud ruumile, siis* $E(\bar{X}) = 0$ *ning* $\mathrm{cov}(X - \bar{X}, Y_i) = 0 \ \forall i$.

Paneme tähele, et eelneva harjutuse põhjal saab suuruse X projektsiooni leida kordajate $c_i,\ i = 1, \ldots, k$ leidmise teel võrrandisüsteemist

$$\sum_{j=1}^{k} c_j \mathrm{cov}(Y_i, Y_j) = \mathrm{cov}(Y_i, X),\ i = 1, \ldots, k.$$

Juhul, kui sellel süsteemil on mitu lahendit, siis võib võtta suvalise nendest, kuna saab näidata, et $\bar{X} = \sum_{i=1}^{k} c_i Y_i$ on sel juhul kõikide lahendite korral sama.

Definitsioon 14 *Juhuslike suuruste* X_1 *ja* X_2 *osakorrelatsiooniks pärast suuruste* $Y_1, \ldots, Y_k$ *mõju eemaldamist nimetatakse suuruste* $X_1 - PX_1$ *ja* $X_2 - PX_2$ *vahelist korrelatsiooni, kus* P *on vähimruutude projektor suurustega* $Y_1, \ldots, Y_k$ *määratud alamruumile.*

Tuletame meelde, et statsionaarse protsessi Z korral tähistab $\tilde{Z}$ vastavat tsentreeritud protsessi, st $\tilde{Z}_t = Z_t - EZ_t = Z_t - \mu$.

Definitsioon 15 *Statsionaarse protsessi* Z k*-ndat järku osaautokorrelatsioonikordajaks nimetatakse suuruste* $\tilde{Z}_t$ *ja* $\tilde{Z}_{t-k}$ *osakorrelatsiooni pärast suuruste* $\tilde{Z}_{t-1}, \ldots, \tilde{Z}_{t-(k-1)}$ *mõju eemaldamist.*

Definitsioonis järeldub, et protsessi Z esimest järku osaautokorrelatsioon on võrdne suurusega ρ_1.

Harjutus 16 *Leida otse definitsioonist lahtudes statsionaarse protsessi teist järku osakorrelatsiooni avaldis autokorrelatsioonide kaudu.*

Osaautokorrelatsioonikordaja definitsioonist lähtuvalt on võimalik näidata, et k-ndat järku osaautokorrelatsioonikordaja ϕ_{kk} on teadaolevate autokorrelatsioonide

põhjal leitav Yule-Walkeri võrrandisüsteemi

$$\begin{aligned}
\phi_{k1} + \rho_1\phi_{k2} + \ldots + \rho_{k-1}\phi_{kk} &= \rho_1,\\
\rho_1\phi_{k1} + \phi_{k2} + \ldots + \rho_{k-2}\phi_{kk} &= \rho_2,\\
&\ldots\\
\rho_{k-1}\phi_{k1} + \rho_{k-2}\phi_{k2} + \ldots + \phi_{kk} &= \rho_k,
\end{aligned}$$

lahendamise teel.

Harjutus 17 *(*, lisapunktide saamiseks esitada 01.11.2012) Näidata osaautokorrelatsiooni definitsioonist lähtudes, et k-ndat järku osaautokorrelatsioon on avaldub suurusena ϕ_{kk} Yule-Walker võrrandites.*

4.1.2 Lõpliku arvu parameetritega määratud lineaarsete protsesside klassid

Praktiliseks kasutamiseks on väga oluline, et aegrea mudelis oleks lõplik (ja võimalikult väike) arv parameetreid, mida andmete põhjal on vaja hinnata. Seetõttu pakuvad erilist huvi järgmised protsesside klassid.

Definitsioon 16 *Eeldame, et juhuslikud suurused A_t, $t \in \mathbb{Z}$ on tsentreeritud sõltumatud sama jaotusega juhuslikud suurused. Olgu μ mingi reaalarv. Kasutame tähistust $\tilde{Z}_t = Z_t - \mu$. Lõpliku arvu kordajatega lineaarsete protsesside klassid on järgmised:*

- *Järguga p autoregressiivseteks protsessideks ehk AR(p) protsessideks nimetatakse protsesse kujul*

  $$\tilde{Z}_t = \sum_{i=1}^{p} \phi_i \tilde{Z}_{t-i} + A_t.$$

  Kui defineerida funktsioon

  $$\phi(x) = 1 - \sum_{i=1}^{p} \phi_i x^i,$$

  siis see protsess on esitatav kujul

  $$\phi(B)\tilde{Z}_t = A_t.$$

- *Järku q liikuva keskmisega protsessideks ehk MA(q) protsessideks nimetatakse protsesse kujul*

  $$\tilde{Z}_t = A_t - \sum_{i=1}^{q} \theta_i A_{t-i}.$$

Kui defineerida funktsioon

$$\theta(x) = 1 - \sum_{i=1}^{q} \theta_i x^i,$$

siis see protsess on esitatav kujul

$$\tilde{Z}_t = \theta(B)A_t.$$

- *ARMA(p,q) protsessideks nimetatakse protsesse kujul*

$$\tilde{Z}_t = \sum_{i=1}^{p} \phi_i \tilde{Z}_{t-i} + A_t - \sum_{i=1}^{q} \theta_i A_{t-i}.$$

Eelnevaid tähiseid kasutades võib sellise protsessi kirja panna kujul

$$\phi(B)\tilde{Z}_t = \theta(B)A_t.$$

Saab näidata, et kui eelnevalt defineeritud protsessid on statsionaarsed, siis definitsioonis kasutatud parameeter μ on juhuslike suuruste Z_t keskväärtuseks.
Edaspidi uurime nende protsesside omadusi lähemalt.

4.2 Autoregressiivsed protsessid

Uurime lähemalt protsessi

$$\tilde{Z}_t = \sum_{i=1}^{p} \phi_i \tilde{Z}_{t-i} + A_t, \ t \in \mathbb{Z} \tag{4.2}$$

omadusi. Nagu näeme, on selle protsessi käitumine seotud polünoomi

$$\phi(x) = 1 - \sum_{i=1}^{p} \phi_i x^i$$

nullkohtadega.

4.2.1 Autokorrelatsioonifunktsioon ja statsionaarsus

Oletame kõigepealt, et protsess $\tilde{Z}$ on statsionaarne. Korrutades võrrandi (4.2) mõlemaid pooli suurusega $\tilde{Z}_{t-k}$ ning võttes keskväärtuse, saame

$$\gamma_k = \sum_{i=1}^{p} \phi_i \gamma_{k-i}, \ k > 0$$

kust pärast suurusega γ_0 läbijagamist saame võrduse

$$\rho_k = \sum_{i=1}^{p} \phi_i \rho_{k-i}, \ k > 0.$$

See tähendab, et autokorrelatsioonikordajad rahuldavad p-ndat järku lineaarset rekurrentset võrrandit. Selliste võrrandite kohta on teada, et mingite kordajate $c_i, i = 1, \ldots, p$ kehtib

$$\rho_k = \sum_{i=1}^{p} c_i d_{ik},$$

kus jadad $d_{ik},\ k = 1, 2, \ldots$ on defineeritud funktsiooni $\phi(x)$ juurte abil järgmiselt: kui x_j on funktsiooni ϕ m-kordne nullkoht (komplekstasandil), siis m jadadest c_{ik} on kujul $k^\ell x_j^{-k},\ 0 \le \ell \le m-1$. Sellest esitusest järeldub, et AR(p) protsesside korral on lõpmatult paljud suurustest ρ_k nullist erinevad. Lisaks sellele, autokorrelatsioonikordajad $\rho_1, \ldots, \rho_p$ ei ole suvalised, vaid rahuldavad Yule-Walker võrrandeid

$$\rho_1 = \phi_1 + \phi_2\rho_1 + \ldots + \phi_p\rho_{p-1}, \tag{4.3}$$

$$\rho_2 = \phi_1\rho_1 + \phi_2 + \ldots + \phi_p\rho_{p-2}, \tag{4.4}$$

$$\ldots \tag{4.5}$$

$$\rho_p = \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \ldots + \phi_p \tag{4.6}$$

mistõttu on kordajad $c_1, \ldots, c_p$ üheselt määratud. Samas autokorrelatsioonikordajad peavad definitsiooni kohaselt olema vahemikus $[-1, 1]$, mistõttu statsionaarsuse eeldus ei saa olla täidetud, kui mõni polünoomi ϕ nullkohtadest on mooduli poolest ühest väiksem ja vastav kordaja autokorrelatsioonide esituses nullkohtade kaudu on nullist erinev. Kui aga polünoomi ϕ kõik nullkohad komplekstasandil on mooduli poolest suuremad kui 1, siis on $\psi(x) = \frac{1}{\phi(x)}$ esitatav ühikringis koonduva astmereana ning seega on ka operaator $\phi(B)$ pööratav ja vastav protsess statsionaarne. Selle väite kehtivuses võib veenduda mitmel moel. Üheks võimaluseks on kasutada kompleksmuutuja funktsioonide teooriat: Cauchy valemi kohaselt kehtib

$$\psi(z) = \frac{1}{2\pi} \oint_{|\xi|=r} \frac{\psi(\xi)}{\xi - z}\, d\xi,\ |z| < r,$$

kus r on ühest suurem arv, mille korral kõik ϕ nullkohad jäävad väljapoole komplekstasandi ringi raadiusega r ning integreerimine toimub üle komplestasandi ühikringi. Siit valemist järeldub (kuidas ?), et $|\psi^k(0)| \le \text{const.}\, k! r^{-k}$, mistõttu vastav Taylori rida koondub ringis raadiusega r, seega ka ühikringis. Teiseks võimaluseks on esitada ψ nullkohtade abil määratud osamurdude summana:

$$\psi(x) = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{c_{ij}}{(x_i - x)^j},$$

kus $x_1, \ldots, x_k$ on polünoomi ϕ nullkohad (komplekstasandil) ning $m_1, \ldots, m_k$ on nende nullkohtade kordsused. Kuna

$$\begin{aligned}
\frac{1}{(x_i - x)^j} &= \frac{1}{(j-1)!} \frac{d}{dx^{j-1}} \left(\frac{1}{x_i - x} \right) \\
&= \frac{1}{x_i\,(j-1)!} \frac{d}{dx^{j-1}} \left(\frac{1}{1 - x_i^{-1}x} \right) = \frac{1}{x_i\,(j-1)!} \frac{d}{dx^{j-1}} \left(\sum_{\ell=0}^{\infty} x_i^{-\ell} x^\ell \right) \\
&= \sum_{\ell=0}^{\infty} \frac{\ell!}{(j-1)!(\ell - j + 1)!} x_i^{-\ell-j} x^\ell,
\end{aligned}$$

siis kõikidele osamurdudele vastavad astmeread koonduvad ühikringis ning seega ka $\psi(x)$ astmerida (kui koonduvate astmeridade summa) koondub ühikringis.

Viimast lähenemist saab kasutada ka autoregressiivsel kujul oleva protsessi esitamisel üldise lineaarse protsessina. Selleks

1. Leiame funktsiooni $\psi(x) = \frac{1}{\phi(x)}$ esituse osamurdudena (st leiame vastavad kordajad c_{ij}.
2. Esitame iga liidetava astmerea kujul.
3. Leiame astmeridade summa. Selle summa x^ℓ kordaja on ψ_ℓ vaadeldava rea esituses üldise lineaarse protsessina.

Samas, kui meil ei ole vaja leida kordajate ψ_ℓ üldkuju, vaid ainult fikseeritud arvu esimeste kordajate väärtusi, siis võib leida need ka seosest

$$\phi(x)(1 + \sum_{i=1}^{\infty} \psi_i x^i) = 1,$$

kirjutades välja vasaku poole erinevate x astmete kordajad ning võrdsustades parema pool (st praegusel juhul konstantse polünoomi 1) kordajatega. Näiteks x kordajast saame

$$-\phi_1 + \psi_1 = 0,$$

x^2 kordajast saame

$$-\phi_2 - \phi_1\psi_1 + \psi_2 = 0$$

jne. Neid seoseid rakendades saab lihtsalt leida suvalise lõpliku arvu kordajaid.

Harjutus 18 *Leida protsessi*

$$\tilde{Z}_t = 1.2\tilde{Z}_{t-1} - 0.35\tilde{Z}_{t-2} + A_t$$

esitus üldise lineaarse protsessi kujul (st, leida suurused ψ_i, $i = 1, 2, \ldots$). Kontrollimiseks leida esimesed neli kordajat ka alternatiivsel teel (korrutise $\phi(x)\psi(x)$ muutuja x astmete abil saadud seoseid kasutades).

Juht, kus mõni funktsiooni ϕ nullkohtadest on mooduli poolest võrdne ühega, vajab eraldi uurimist, kuid selle käsitlemine on käesoleva kursuse mahtu arvestades ebaotstarbekas. Käesolevas kursuses kasutame teadmist, et statsionaarsuse jaoks on tarvilik ja piisav, et funktsiooni ϕ nullkohad on kõik mooduli poolest ühest suuremad.

4.2.2 Osaautokorrelatsioonid

AR(p) protsesside korral on sobiva mudeli kindlakstegemisel suur kasu järgmisest tulemusest.

Lemma 17 *Olgu Z statsionaarne AR(p) protsess. Siis tema osaautokorrelatsioonikordajad on võrdsed nulliga alates järgust $p+1$.*

Tõestus. Olgu $k > p$. Olgu P vähimruutude projektor suurustega $\tilde{Z}_{t-1}, \ldots, \tilde{Z}_{t-k+1}$ määratud alamruumile. Kuna A_t on sõltumatu suurustest $\tilde{Z}_{t-1}, \ldots, \tilde{Z}_{t-k+1}$, siis on lihtne veenduda, et $A_t = \tilde{Z}_t - P\tilde{Z}_t$. Selleks näitame, et

$$P\tilde{Z}_t = \sum_{i=1}^{p} \phi_i \tilde{Z}_{t-i}.$$

P definitsiooni kohaselt peame me selleks näitama, et

$$E[(\tilde{Z}_t - \sum_{i=1}^{k-1} c_i \tilde{Z}_{t-i})^2] \geq E[(\tilde{Z}_t - \sum_{i=1}^{p} \phi_i \tilde{Z}_{t-i})^2]$$

kõikide kordajate $c_1, c_2, \ldots, c_{k-1}$ korral.
Tähistame kirjapaneku lihtsustamise huvides

$$X = \sum_{i=1}^{k-1} c_i \tilde{Z}_{t-i}, \; Y = \sum_{i=1}^{p} \phi_i \tilde{Z}_{t-i},$$

siis kasutades suuruse A_t tsentreeritust ning sõltumatust suurustest X ja Y saame

$$E[(\tilde{Z}_t - X)^2] = E[(A_t - Y - X)^2] = E[A_t^2] - 2E[A_t(X+Y)] + E[(X+Y)^2] = E[A_t^2] + E[(X+Y)^2] \geq E[A_t^2]$$

Samas

$$E[(\tilde{Z}_t - Y)^2] = E[A_t^2],$$

seetõttu on suurus Y tõepoolest võrdne suuruse $\tilde{Z}_t$ projektsiooniks vähimruutude mõttes.
Kasutades jällegi suuruse A_t sõltumatust varasematest protsessi Z väärtustest saame nüüd

$$\operatorname{cov}(\tilde{Z}_t - P\tilde{Z}_t, \tilde{Z}_{t-k} - P\tilde{Z}_{t-k}) = \operatorname{cov}(A_t, \tilde{Z}_{t-k} - P\tilde{Z}_{t-k}) = 0.$$

Osaautokorrelatsioonikordaja definitsiooni kohaselt on seega protsessi $\tilde{Z}$ k-ndat järku osaautokorrelatsioonikordaja võrdne nulliga.□ Yule-Walkeri võrrandeid võib kasutada osaautokorrelatsioonide hindamiseks, asendades võrrandites teoreetilised autokorrelatsioonid nende hinnangutega. Praktikas on samuti kasulik teadmine (vt [8], valem 3.2.35), et AR(p) protsessi korral on osaautokorrelatsioonikordajate hinnangud alates järgust $k = p+1$ ligikaudu sõltumatud, keskväärtusega 0 ning standardhälbega $\frac{1}{\sqrt{n}}$.

4.2.3 AR(1) tüüpi mudelid

Vaatleme mudeleid kujul

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + A_t.$$

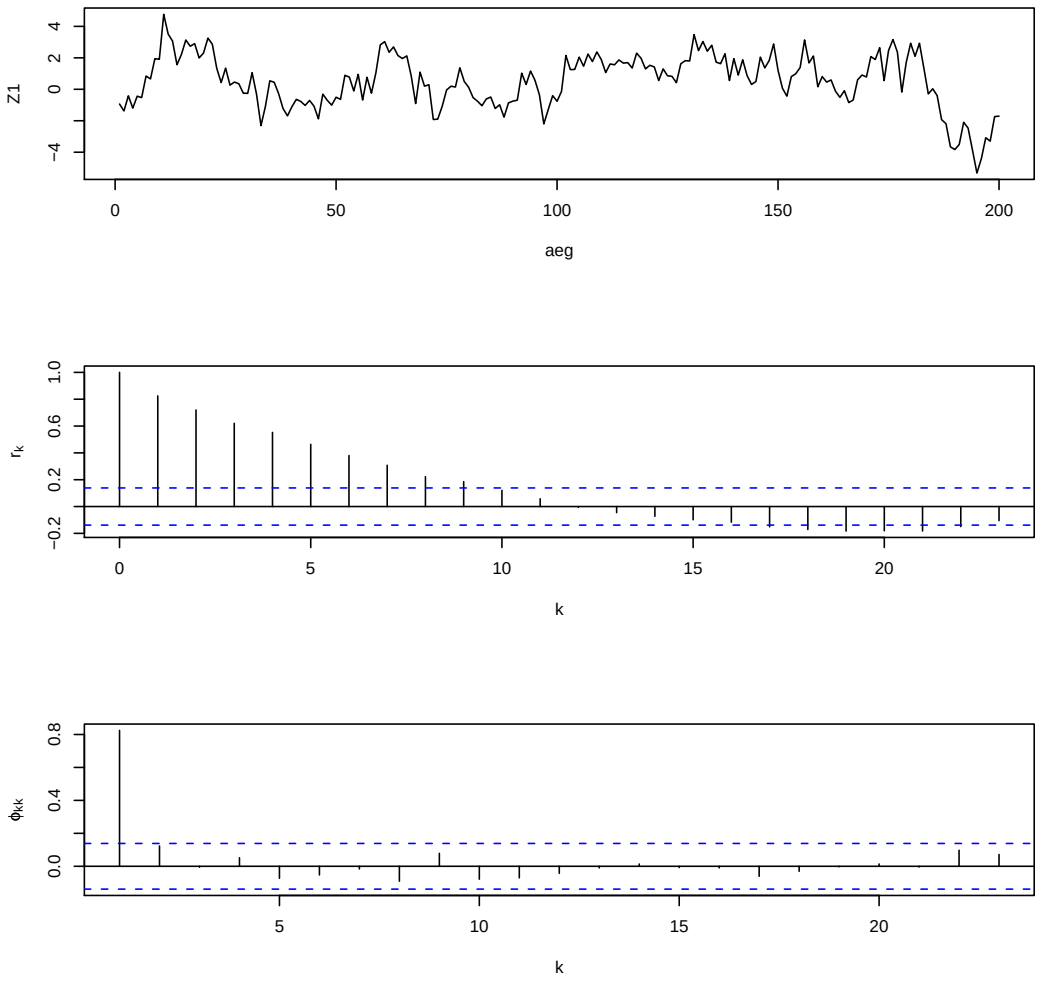
Kuna selle mudeli korral $\phi(x) = 1 - \phi_1 x$, mille ainsaks nullkohaks on $x_1 = \frac{1}{\phi_1}$, siis statsionaarsuse jaoks on vajalik tingimuse $|\phi_1| < 1$ täidetus. Kuna autokorrelatsioonid rahuldavad eelneva põhjal seost

$$\rho_k = \phi_1 \rho_{k-1}, \ k > 0,$$

siis

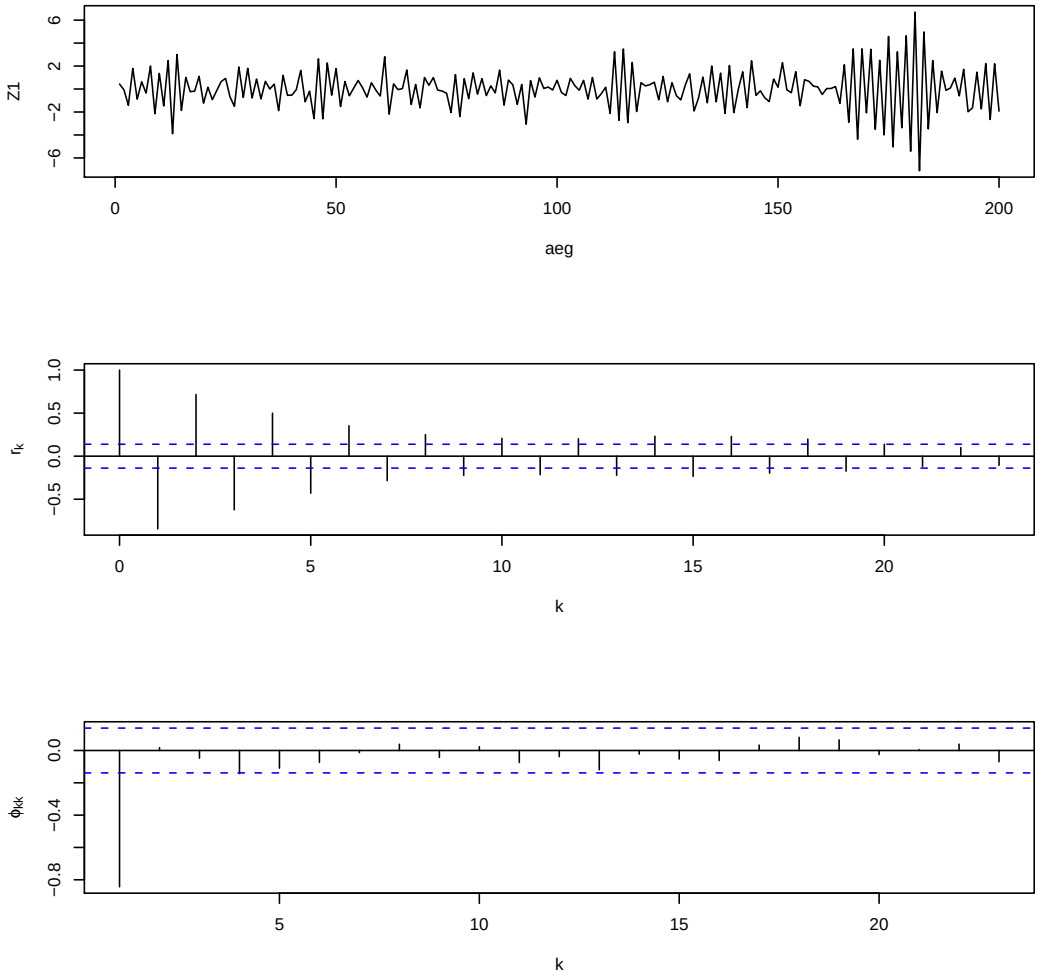
$$\rho_k = \phi_1^k, \ k = 1, 2, \ldots.$$

Seega kahanevad autokorrelatsoonide absoluutväärtused eksponentsiaalselt, kusjuures juhul $\phi_1 > 0$ on nad sama märgiga ning juhul $\phi_1 < 0$ vahelduvate märkidega. Osaautokorrelatsioonid on alates järgust 2 võrdsed nulliga ning esimest järku osaautokorrelatsioon on (nagu alati) võrdne ρ_1-ga. Näited vastavate aegridade käitumi-



Joonis 4.1: AR(1) tüüpi aegrea 200 väärtust juhul $\phi_1 = 0.8$ ning autokorrelatsioonide ja osaautokorrelatsioonide hinnangud

sest koos empiiriliste autokorrelatsioonide ja empiiriliste osaautokorrelatsioonidega juhul $\phi_1 = 0.8$ ja $\phi_1 = -0.8$ on toodud vastavalt joonistel 4.1 ja 4.2.



Joonis 4.2: AR(1) tüüpi aegrea 200 väärtust juhul $\phi_1 = -0.8$ ning autokorrelatsioonide ja osaautokorrelatsioonide hinnangud

4.2.4 AR(2) tüüpi mudelid

Vaatleme mudeleid kujul

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + A_t.$$

Selleks, et funktsiooni $\phi(x) = 1 - \phi_1 x - \phi_2 x^2$ nullkohad oleks väljaspool komplekstasandi ühikringi, peavad kordajad ϕ_1 ja ϕ_2 rahuldama geomeetriliselt küllaltki lihtsalt kirjeldatavaid tingimusi.

Harjutus 19 *(*, lisapunktide saamiseks esitada 13.11.2012) Näidata, et statsio-*

naarsuse tingimus on samaväärne võrratustega

$$\begin{aligned}\phi_1 + \phi_2 &< 1,\\ \phi_2 - \phi_1 &< 1,\\ \phi_2 &> -1,\end{aligned}$$

st punkti (ϕ_1, ϕ_2) *peab paiknema tippudega (0,1), (-2,-1) ja (2,-1) määratud kolmnurgas.*

Autokorrelatsioonikordajad saab arvutada vastavalt rekurrentsele seosele

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2},\ k > 1,$$

lähtudes väärtustest $\rho_0 = 1$ ja $\rho_1 = \frac{\phi_1}{1-\phi_2}$. Viimane tuleneb Yule-Walkeri esimesest võrrandist:

$$\rho_1 = \phi_1 + \phi_2\rho_1.$$

Lahendades Yule-Walkeri võrrandid ϕ_1 ja ϕ_2 suhtes, saame

$$\phi_1 = \frac{\rho_1(1-\rho_2)}{1-\rho_1^2},\ \phi_2 = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2},$$

mille abil on võimalik empiirilistest autokorrelatsioonidest r_1 ja r_2 arvutada kordajate ϕ_1 ja ϕ_2 hinnangud.

Teoreetilised osaautokorrelatsioonid on nullid alates järgust 3, esimest järku osaautokorrelatsion ϕ_{11} on võrdne ρ_1-ga ning teist järku osaautokorrelatsioon on võrdne kordajaga ϕ_2 (miks?), seega on võimalik kordajaid hinnata ka osaautokorrelatsioonide hinnangute abil.

4.3 Liikuva keskmise protsessid

Järgnevas vaatleme MA(q) protsesside

$$\tilde{Z}_t = A_t - \sum_{i=1}^{q} \theta_i A_{t-i}$$

omadusi. Varasemast teame, et sellised protsessid (sõltumatute ja sama jaotusega juhuslike suuruste A_t korral) on alati statsionaarsed.

4.3.1 Autokorrelatsioonid ja pööratavuse tingimused

Arvestades, et üldise lineaarse protsessi korral

$$\gamma_k = \sigma_A^2 \sum_{j=0}^{\infty} \psi_j\psi_{j+k}$$

ning et MA(q) protsessi korral

$$\psi_0 = 1;\ \psi_i = -\theta_i,\ i = 1, \ldots, q; \psi_k = 0,\ k > q$$

saame, et MA(q) protsessi autokorrelatsioonikordajad avalduvad kujul

$$\rho_k = \frac{-\theta_k + \sum_{i=1}^{p-k} \theta_i \theta_{i+k}}{1 + \sum_{i=1}^{q} \theta_i^2},\ k = 1, \ldots, q$$

ning $\rho_k = 0,\ k > q$. Nagu me ka hiljem näeme, on $MA(q)$ protsesside korral võimalik, et täpselt samad autokorrelatsioonid (ja seega ka osaautokorrelatsioonid) vastavad erinevatele parameetritele $\theta_1, \ldots, \theta_q$. Osutub aga, et ainult üks parameetrite θ valik vastab pööratavale protsessile ning aegrea vaatluste põhjal saame parima tuleviku prognoosi, kasutades pööratavat mudelit.

Varasema põhjal teame, et pööratavuseks on vajalik, et polünoomi $\theta(x) = 1 - \sum_{i=1}^{q} \theta_i x^i$ nullkohad oleks kõik mooduli poolest ühest suuremad, kuna sel juhul on funktsioon $\pi(x) = \frac{1}{\theta(x)}$ avaldatav astmereana, mis koondub $|x| \leq 1$ korral.

4.3.2 MA(1) protsessi omadused.

Vaatleme protsessi kujul

$$\tilde{Z}_t = A_t - \theta_1 A_{t-1}.$$

Selle protsessi korral

$$\rho_1 = -\frac{\theta_1}{1 + \theta_1^2}$$

ning $\rho_k = 0,\ k > 1$. Kuna me saame ρ_1 avaldise kirjutada (juhul $\rho_1 \neq 0$) ka kujul

$$\rho_1 = -\frac{1}{\frac{1}{\theta_1} + \theta_1},$$

siis on selge, et täpselt samasugused autokorrelatsioonid on ka protsessil

$$\tilde{Z}_t = A_t - \frac{1}{\theta_1} A_{t-1}.$$

Samas, kui $|\theta_1| > 1$, siis võime defineerida suurused $\bar{A}_t$ kujul

$$\bar{A}_t = \sum_{i=0}^{\infty} \theta_1^{-i} \tilde{Z}_{t-i},$$

mille korral kehtib

$$\tilde{Z}_t = \bar{A}_t - \frac{1}{\theta_1} \bar{A}_{t-1}.$$

Kuna suurused $\bar{A}_t$ on tsentreeritud, mittekorreleeritud ning konstantse dispersiooniga (vt järgnev harjutus), siis võime ilma üldsust kitsendamata eeldada, et vaadeldava protsessi korral on kordaja θ_1 absoluutväärtuselt ühest väiksem ning et tegemist on pööratava protsessiga.

Harjutus 20 *Veenduda, et juhul $|\theta_1| > 1$ defineeritud juhuslikud suurused $\bar{A}_t$ on mittekorreleeritud ning konstantse dispersiooniga (seega tegemist on vähemalt teist järku nõrgalt statsionaarse protsessiga).*

Harjutus 21 *Näidata, et juhul $|\theta_1| \neq 1$ kehtib võrratus $|\rho_1| < \frac{1}{2}$.*

Osaautokorrelatsioonide leidmiseks paneme tähele, et vaadeldaval juhul tuleb Yule-Walker võrrandisüsteemi kohaselt leida kolmediagonaalse võrrandisüsteemi

$$\begin{aligned}
\phi_{k1} + \rho_1\phi_{k2} &= \rho_1,\\
\rho_1\phi_{k1} + \phi_{k2} + \rho_1\phi_{k3} &= 0,\\
&\dots\\
\rho_1\phi_{k,k-2} + \phi_{k,k-1} + \rho_1\phi_{kk} &= 0,\\
\rho_1\phi_{k,k-1} + \phi_{kk} &= 0
\end{aligned}$$

puhul suuruse ϕ_{kk} väärtus. Kui siit elimineerida teisest võrrandist esimese võrrandi abil tundmatu ϕ_{k1}, kolmandast võrrandist saadud teise võrrandi abil ϕ_{k2} jne, siis jõuame kahediagonaalse süsteemini

$$\begin{aligned}
b_1\phi_{k1} + \rho_1\phi_{k2} &= f_1,\\
b_2\phi_{k2} + \rho_1\phi_{k3} &= f_2,\\
&\dots\\
b_{k-1}\phi_{k,k-1} + \rho_1\phi_{kk} &= f_{k-1},\\
b_k\phi_{kk} &= f_k,
\end{aligned}$$

kus $b_1 = 1,\ f_1 = \rho_1$ ning

$$b_{i+1} = 1 - \frac{\rho_1^2}{b_i},\ f_{i+1} = -\frac{\rho_1}{b_i}f_i,\ i = 2, \dots, k.$$

Arvestades, et $|\rho_1| < \frac{1}{2}$ saame $b_i > \frac{1}{2}\ \forall i$ ning seega

$$|\phi_{kk}| = \frac{|f_k|}{b_k} \leq (2\rho_1)^k,$$

seega osaautokorrelatsioonikordajad on küll kõik nullist erinevad, kuid kahanevad eksponentsiaalselt.

4.3.3 MA(2) protsessi omadused.

Vaatleme protsessi kujul

$$\tilde{Z}_t = A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2}.$$

Selle protsessi korral

$$\begin{aligned}
\rho_1 &= \frac{-\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2},\\
\rho_2 &= \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2}
\end{aligned}$$

ning $\rho_k = 0,\ k > 2$. Küllalt lihtne on veenduda, et kui x_1, x_2 on polünoomi

$$\theta(x) = 1 - \theta_1 x - \theta_2 x^2$$

nullkohad, siis täpselt samad autokorrelatsioonikordajad on kõikidel MA(2) protsessidel, millele vastavate polünoomide nullkohad on kujul x_1^i, x_2^j, kus $i, j \in \{-1, 1\}$. Samas ainult üks nendest protsessidest rahuldab pööratavuse tingimust ning jällegi võime üldsust kitsendamata eeldada, et meid huvitav protsess rahuldab pööratavuse tingimust. Analoogiliselt AR(2) protsesside statsionaarsuse tingimustega saame nüüd, et pööratava MA(2) protsessi kordajad θ_1 ja θ_2 peavad rahuldama tingimusi

$$\theta_2 + \theta_1 < 1,\ \theta_2 - \theta_1 < 1,\ \theta_2 > -1.$$

4.4 ARMA(p,q) protsessid

Vaatleme protsesse kujul

$$\tilde{Z}_t = \sum_{i=1}^{p} \phi_i \tilde{Z}_{t-i} + A_t - \sum_{i=1}^{q} \theta_i A_{t-i}.$$

See protsess on statsionaarne, kui polünoomi $\phi(x) = 1 - \sum_{i=1}^{p} \phi_i x^i$ nullkohad on väljaspool komplekstasandi ühikringi ning pööratavuse tingimus on täidetud, kui polünoomi $\theta(x) = 1 - \sum_{i=1}^{q} \theta_i x^i$ nullkohad on väljaspool komplekstasandi ühikringi. Arvutades kovariatsiooni $\tilde{Z}_t$ ja $\tilde{Z}_{t-k}$ vahel juhul $k > q$, saame

$$\gamma_k = \sum_{i=1}^{p} \phi_i \gamma_{k-i},$$

mistõttu autokorrelatsioonikordajad ρ_k rahuldavad rekurrentset võrrandit

$$\rho_k = \sum_{i=1}^{p} \phi_i \rho_{k-i}$$

alates $k = q + 1$.

4.5 ARMA(1,1) protsessid

Vaatleme protsesse kujul

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + A_t - \theta_1 A_{t-1}.$$

Leiame avaldised selle protsessi autokovariatsioonidele ja autokorrelatsioonidele. Kõigepealt paneme tähele, et

$$\operatorname{cov}(\tilde{Z}_t, A_t) = \sigma_A^2.$$

Seejärel saame leida

$$\mathrm{cov}(\tilde{Z}_t, A_{t-1}) = (\phi_1 - \theta_1)\sigma_A^2.$$

Leides nüüd kovariatsiooni $\tilde{Z}_t$ avaldise parema poole ja $\tilde{Z}_t$ vahel, saame

$$\gamma_0 = \phi_1\gamma_1 + (1 - \theta_1(\phi_1 - \theta_1))\sigma_A^2$$

ning kovariatsioon $\tilde{Z}_t$ avaldise parema poole ja $\tilde{Z}_{t-1}$ vahel annab

$$\gamma_1 = \phi_1\gamma_0 - \theta_1\sigma_A^2.$$

Siit saame

$$\gamma_0 = \frac{(1 - 2\theta_1\phi_1 + \theta_1^2)\sigma_A^2}{1 - \phi_1^2},$$
$$\gamma_1 = \frac{(\phi_1 - \theta_1)(1 - \theta_1\phi_1)\sigma_A^2}{1 - \phi_1^2}$$

ja seega

$$\rho_1 = \frac{(\phi_1 - \theta_1)(1 - \theta_1\phi_1)}{1 - 2\theta_1\phi_1 + \theta_1^2}.$$

Kuna alates $k = 2$ kehtib

$$\rho_k = \phi_1\rho_{k-1},$$

siis autokorrelatsioonid kahanevad eksponentsiaalselt alates järgust 2.

4.6 Lineaarsed mudelid mittestatsionaarsete aegridade jaoks. Prognoosimine ja parameetrite hindamine

4.6.1 ARIMA mudelid

Sageli ei vasta aegrida statsionaarsuse nõuetele, kuna keskmine on ajas muutuv, kuid selle rea muudud või muutude muudud käituvad kooskõlas statsionaarsuse eeldustega. Järgnevas vaatlemegi selliste protsesside mudeleid.

Kui $Z_t,\ t \in \mathbf{R}$ on mingi protsess, siis muutude protsessi $Z_t - Z_{t-1}$ võime kirjutada kujul $(1 - B)Z_t$ ning muutude muutude protsessi kujul $(1 - B)^2 Z_t$.

Definitsioon 18 *ARIMA(p,d,q) protsessiks nimetatakse juhuslikku protsessi Z_t, mille d-ndat järku muudud (ehk diferentsid) $W_t = (1 - B)^d Z_t$ esituvad kujul*

$$\tilde{W}_t = \sum_{i=1}^{p} \phi_i \tilde{W}_{t-i} + A_t - \sum_{i=1}^{q} \theta_i A_{t-i},$$

kus $\tilde{W}_t = W_t - E(W_t)$, juhuslikud suurused A_t on sõltumatud, sama jaotusega, tsentreeritud ning on sõltumatud ka suurustest $\tilde{W}_{t-i}, i = 1, 2, \ldots$.

Vaatleme juhtu $W_t = (1-B)Z_t$, siis

$$Z_t = Z_{t-1} + W_t = Z_{t-2} + W_{t-1} + W_t = Z_{t-k} + \sum_{j=k+1}^{t} W_t.$$

Summa on aga vaadeldav tükiti konstantse funktsiooni integraalina. Üldisemalt, kui $W_t = (1-B)^d Z_t$, siis tuleb Z leidmiseks protsessist W rakendada summeerimist d korda, mis on samastatav d-kordse integraali leidmisega. Seetõttu nimetataksegi ARIMA(p,d,q) protsesse integreeritud ARMA protsessideks.

Harjutus 22 *Olgu antud suurused* $c_i = (1-B)^i Z_1,\ i = 0, \ldots, d-1$ *ning protsessi* $W_t = (1-B)^d Z_t$ *väärtused* $w_2, \ldots, w_n$. *Avaldada* Z_n *väärtus* z_n $c_0, \ldots, c_{d-1}, w_2, \ldots, w_n$ *kaudu.*

4.6.2 Aegridade prognoosimine ARIMA mudelite korral

Olgu meil antud aegrida $z_1, z_2, \ldots, z_n$ ning oletame, et me teame, et see vastab ARIMA(p,d,q) tüüpi protsessile teadaolevate parameetritega μ (suuruse $(1-B)^d Z_t$ keskväärtus),$\phi_1, \ldots, \phi_p$ ja $\theta_1, \ldots, \theta_q$, kusjuures eeldame, et protsess on pööratav ning et kaalud $\phi_i,\ i = 1, \ldots, p$ rahuldavad statsionaarsuse tingimust. Esituse lihtsuse mõttes eeldame samuti, et $(1-B)^d Z_t$ keskväärtus on null. Järgnevas uurime, kuidas sel juhul leida minimaalse ruutkeskmise veaga prognoose suurusele $Z_{n+p},\ p \geq 1$ ning prognoosivigade standardhälbeid.

Edasises kasutame oluliselt aegrea erinevaid esitusi. Pööratavusest järeldub, et me saame vaadeldava ARIMA(p,d,q) protsessi esitada autoregressiivsel kujul

$$Z_t = \sum_{i=1}^{\infty} \pi_i Z_{t-1} + A_t,$$

kus kaalud π_i on on määratavad võrdusest

$$1 - \sum_{i=1}^{\infty} \pi_i x^i = \frac{\phi(x)(1-x)^d}{\theta(x)},$$

kus

$$\phi(x) = 1 - \sum_{i=1}^{p} \phi_i x^i,\ \theta(x) = 1 - \sum_{i=1}^{q} \theta_i x^i.$$

Samuti teame, et statsionaarse protsessi saab esitada üldise lineaarse protsessi kujul, seega leiduvad kordajad ψ_i, mille korral

$$(1-B)^d Z_t = \left(\sum_{i=0}^{\infty} \psi_i B^i \right) A_t.$$

Kolmandaks esituseks on protsessi definitsioonis olev kuju, mis sisaldab lõpliku arvu eelnevaid Z ja A väärtuseid.

Harjutus 23 *Esitada ARIMA(0,1,1) protsess*

$$Z_t = Z_{t-1} + A_t - \frac{1}{2}A_{t-1}$$

autoregressiivsel kuju.

Autoregressiivse kuju kordajaid saab leida ka samm-sammult, võrdsustades seose

$$\pi(x)\theta(x) = \phi(x)$$

paremal ja vasakul pool x astmete ees olevaid kordajaid.

Harjutus 24 *Leida protsessi*

$$Z_t = A_t - 0.5A_{t-1} + 0.3A_{t-2}$$

autoregressiivse esituse kordaja π_4.

Ruutkeskmise vea mõttes parima prognoosi leidmine

Kui juhusliku suuruse kohta ei ole mingit lisainformatsiooni, siis tema parimaks prognoosiks on keskväärtus.

Harjutus 25 *Olgu X lõplikku dispersiooni omav keskväärtusega μ juhuslik suurus ning olgu a prognoos selle juhusliku suuruse väärtuse jaoks. Näidata, et ennustusviga $E((X-a)^2)$ on minimaalne juhul, kui $a = \mu$.*

Kui juhusliku suuruse ennustamisel on kasutada mingit lisainformatsiooni, siis saab näidata, et parimaks ennustuseks selle informatsiooni põhjal on tinglik keskväärtus. Käesolevas kursuses me tingliku keskväärtuse üldist definitsiooni sisse ei too, küll aga kasutame teadaolevaid tulemusi selle omaduste kohta.

Lemma 19 *Olgu I mingi lõplik või loenduv indeksite hulk ning olgu Z ja $Y_i,\ i \in I$ juhuslikud suurused. Siis suuruse Z tinglik keskväärtus tingimusel, et $Y_i,\ i \in I$ on teada on juhuslik suurus $E(Z \mid Y_i,\ i \in I)$, mis rahuldab järgmisi omadusi:*

1. *Kui Z on sõltumatu juhuslikest suurustest $Y_i,\ i \in I$, siis*

$$E(Z \mid Y_i,\ i \in I) = EZ.$$

2. *Kui $Z = \alpha Z_1 + \beta Z_2$, siis*

$$E(Z \mid Y_i,\ i \in I) = \alpha E(Z_1 \mid Y_i,\ i \in I) + \beta E(Z_2 \mid Y_i,\ i \in I).$$

3. *Kui $Z = Y_{i_0}$ mingi $i_0 \in I$ korral, siis*

$$E(Z \mid Y_i,\ i \in I) = Z.$$

Üldisemalt, kui $Z = f(Y)$, kus $Y = (Y_i)_{i \in I}$, siis

$$E(Z \mid Y_i\ i \in I) = Z.$$

Tingliku keskväärtuse korrektse definitsiooni ja omaduste tõestused võib leida näiteks raamatust [9]. Vastaku protsess Z ARIMA(p,d,q) mudelile, kusjuures järgnevas eeldame, et $E((1-B)^d Z_t) = 0$ ning samuti eeldame mudeli protsessi $W_t = (1-B)^d Z_t$ statsionaarust ja pööratavust. Tähistame

$$\hat{Z}_{\ell|k} = E(Z_\ell \mid Z_{k-i},\ i = 0, 1, 2, \ldots),$$

siis tingliku keskväärtuse omadustest järeldub

$$\hat{Z}_{\ell|k} = Z_\ell \text{ kui } \ell \leq k.$$

Kasutades protsessi esitust autoregressiivsel kujul ning teadmist, et $E(A_{k+j} \mid Z_{k-i},\ i \geq 0) = 0$, saame juhul $\ell \geq 1$

$$\begin{aligned}\hat{Z}_{k+\ell|k} &= \sum_{i=1}^{\infty} \pi_i \hat{Z}_{k+l-i|k} \\ &= \sum_{i=1}^{\ell-1} \pi_i \hat{Z}_{k+l-i|k} + \sum_{i=\ell}^{\infty} \pi_i Z_{k+\ell-i}.\end{aligned}$$

Pikemate prognooside arvutamisel saab seda valemit kasutada samm-sammult: kõigepealt arvutame $\hat{Z}_{k+1|k}$ ajaks k teadaolevate Z väärtuste abil, seejärel kasutame saadud tulemust $\hat{Z}_{k+2|k}$ arvutamiseks jne. Saadud tulemuse rakendamisel on aga kaks probleemi. Esiteks, tegemist on lõpmatu summaga, mille täpne arvutamine on põhimõtteliselt raskendatud kui mitte lausa võimatu. Teiseks, praktikas on alati teada ainult lõplik arv Z minevikuväärtusi, nii et lõpmatu summa tuleb igal juhul asendada lõpliku summaga ja see toob kaasa mõningase prognoosivea. Samas aga pööratava mudeli korral kahanevad kordajad π_i eksponentsiaalselt ning seda kiiremini, mida suurem on mooduli poolest vähim polünoomi $\theta(x)$ nullkoht. Seega enamikel juhtudel lähenevad kordajad π_i kiiresti nullile ning lõpmatu summa on väga hästi lähendatav küllalt väikese arvu liidetavatega lõpliku summaga.

Alternatiivne moodus tulevikuväärtuste ennustamiseks põhineb otseselt ARIMA(p,d,q) mudeli kujul, mis sisaldab lõpliku arvu liidetavaid. Paneme tähele, et me võime selle mudeli esitada kujul

$$\tilde{\phi}(B) Z_t = \theta(B) A_t,$$

kus

$$\tilde{\phi}(x) = \phi(x)(1-x)^d,\ \phi(x) = 1 - \sum_{i=1}^{p} \phi_i x^i,\ \theta(x) = 1 - \sum_{i=1}^{q} \theta_i x^i.$$

Olgu polünoomi $\tilde{\phi}$ esituseks

$$\tilde{\phi}(x) = 1 - \sum_{i=1}^{p+d} \tilde{\phi}_i x_i,$$

siis võib vaadeldava mudeli kirjutada ka kujul

$$Z_t = \sum_{i=1}^{p+d} \tilde{\phi}_i Z_{t-i} + A_t - \sum_{i=1}^{q} \theta_i A_{t-i}.$$

Seega juhul $\ell \le q$ korral kehtib

$$\hat{Z}_{k+\ell|k} = \sum_{i=1}^{p+d} \tilde{\phi}_i \hat{Z}_{k+\ell-i|k} - \sum_{i=\ell}^{q} \theta_i A_{k+\ell-i}$$

ning juhul $\ell > q$ on prognoosid arvutatavad valemist

$$\hat{Z}_{k+\ell|k} = \sum_{i=1}^{p+d} \tilde{\phi}_i \hat{Z}_{k+\ell-i|k}.$$

Nagu näha, määrab polünoomiga $\tilde{\phi}$ määratud rekurrentne võrrand prognooside käitumise alates $\ell > q$.

Eelneva prognoosivalemi kasutamine nõuab A_t väärtuste teadmist $k - q \le t \le k$ korral. Teoreetiliselt ei valmista see probleeme, kuna pööratava mudeli korral on A-d Z-de kaudu leitavad, kuid praktiliselt on probleemiks see, et meil on teada ainult lõplik arv Z_t-de väärtuseid ning isegi kui oleks teada kogu minevik, oleks lõpmatute summade leidmine tülikas. Hädast päästab aga meid järgnevas harjutuses toodud tulemus, mille kohaselt võime piisavalt pika aegrea korral leida A_t realiseerinud väärtused vastavalt võrrandile

$$\tilde{a}_t = z_t - \sum_{i=1}^{p+d} \tilde{\phi}_i z_{t-i} + \sum_{i=1}^{q} \theta_i \tilde{a}_{t-i},\ p+d < t \le k, \tag{4.7}$$

kus

$$\tilde{a}_t = 0,\ p+d-q+1 \le t \le p+d. \tag{4.8}$$

Harjutus 26 *Olgu $z_t,\ 1 \le t$ mingi ARIMA(p,d,q) tüüpi protsessi väärtused ning olgu $a_t,\ t \ge 1$ nendele väärtustele vastavad protsessi A_t väärtused. Olgu $\tilde{a}_t$ vastavalt võrrandile (4.7) ja algväärtustele (4.8) arvutatud suurused. Näidata, et*

$$\lim_{n\to\infty} |a_n - \tilde{a}_n| = 0.$$

Prognoosivea standardhälbe leidmine

Defineerime $\theta_i = 0,\ i > q$, siis võime parima prognoosi kirjutada kujul

$$\hat{Z}_{k+\ell|k} = \sum_{i=1}^{p+d} \tilde{\phi}_i \hat{Z}_{k+\ell-i} - \sum_{i=\ell}^{\infty} \theta_i A_{k+\ell-i}.$$

Kasutades seda valemit ning ARIMA mudeli kuju (kus samuti käsitluse lihtsuse mõttes summeerime θ-dega liikmeid kuni lõpmatuseni), saab matemaatilise induktsiooni abil küllaltki lihtsalt näidata, et kehtib võrdus

$$Z_{k+\ell} - \hat{Z}_{k+\ell|k} = \sum_{i=0}^{\ell-1} \psi_i A_{k+\ell-i},$$

kus

$$\psi_0 = 1,\ \psi_k = \sum_{i=1}^{\min(k,p+d)} \tilde{\phi}_i \psi_{k-i} - \theta_k,\ k \ge 1.$$

Harjutus 27 *(*) Tõestada, et prognoosiveas esinevad kaalud ψ_k avalduvad eelpool toodud kujul.*

Seega on prognoosivea kaalud lihtsalt arvutatavad ning nende abil avaldub ℓ-sammulise prognoosi vea dispersioon kujul

$$D(Z_{k+\ell} - \hat{Z}_{k+\ell|k}) = \sigma_A^2 \sum_{i=0}^{\ell-1} \psi_i^2.$$

Harjutus 28 *Olgu Z protsess, mis vastab ARIMA mudelile*

$$(1 + 0.1\,B + 0.4\,B^2)(1 - B)Z_t = (1 + 0.2B)A_t,$$

kus B on tagasinihke operaator ($BZ_t = Z_{t-1}$). Leidke sellel mudelil põhinevate parimate 1,2,3,4-sammuliste prognooside vigade standardhälbed eeldusel, et $D(A_t) = 1$.

Eelnevas vigade standardhälbe arvutuses läks tegelikult vaja ainult, et suurused A_t on mittekorreleeritud ja konstantse dispersiooniga. Tavaliselt väljastavad programmid ka usaldusintervalle, mis kehtivad juhul, kui suurused A_t on normaaljaotusega ja sõltumatud (sest siis on ka prognoosivead normaaljaotusega).

4.6.3 ARIMA mudeli parameetrite hindamine

Parameetrite hindamisel on mitmeid võimalikke lähenemisi. Kuna tarkvara pakub sageli võimalust nende vahel valida, siis oleks hea teada, mille poolest nad erinevad. Parameetreid hindame protsessile $W_t = (1 - B)^d Z_t$ vastava aegrea andmete $w_1, w_2, \ldots, w_n$ põhjal. Lihtsuse mõttes eeldame ka, et $E(W_t) = 0$.

Tingimusliku ruutude summa minimiseerimine

Oletame, et meil on teada $z_0, z_{-1}, \ldots, z_{1-p-d}$ ning $a_0, a_{-1}, \ldots, a_{-q}$ tegelikud väärtused; siis saame fikseeritud parameetrite θ ja ϕ korral arvutada $a_1 = z_1 - \hat{z}_{1|0}$, $a_2 = z_2 - \hat{z}_{2|1}$, $\ldots, a_n = z_n - \hat{z}_{n|n-1}$. Seega võime parameetreid valida näiteks nii, et minimiseerime prognoosivigade ruutude summat

$$\sum_{i=1}^{n} a_i^2.$$

Sama valikukriteeriumini jõuame ka siis, kui eeldame, et suurused A_t on sõltumatud ja normaaljaotusega ning leiame parameetrid θ ja ϕ nii, et maksimiseerime aegrea $z_1, \ldots, z_n$ tõepära (ehk vektori $(a_1, \ldots, a_n)$ tõepära). Siin tuleb aga aru saada, et tegemist on tingliku tõepäraga; tingimuseks on see, et meil on teada lõigu alguses toodud z ja a eelnevad väärtused. Seetõttu nimetatakse seda parameetrite valiku reeglit tingliku ruutude summa minimiseerimiseks ehk tinglikuks suurima tõepära meetodiks.

Praktikas kasutatakse mitmeid erinevaid varasemate väärtuste fikseerimise mooduseid, millest lihtsaim vastab kõikide eelnevate z ja a väärtuste võrdsustamisele nulliga. Kui vaatlusi on väga palju, siis selline lähenemine annab praktiliselt sama tulemuse, kui suurima tõepära meetod; suhteliselt lühikeste aegridade korral võivad tulemused olla oluliselt erinevad. Käesolevas alapeatükis eeldame, et juhuslikud suurused A_t on sõltumatud ning normaaljaotusega.

Suurima tõepära meetod

Kuna protsessi W_t väärtused on esitatavad sõltumatute normaaljaotusega juhuslike suuruste A_t lineaarkombinatsioonidena, siis on vektor $W_1, \ldots, W_n$ mitmemõõtmelise normaaljaotusega. Mitmemõõtmelise tsentreeritud normaaljaotuse tihedusfunktsioon esitub kujul

$$f(\mathbf{x}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\frac{\mathbf{n}}{\mathbf{2}}}|\mathbf{\Sigma}|^{\frac{\mathbf{1}}{\mathbf{2}}}} \exp(-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{x}'\mathbf{\Sigma^{-1}}\mathbf{x}),$$

kus $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, Σ on vastava juhusliku vektori kovariatsioonimaatriks ning $|\Sigma|$ tähistab maatriksi determinanti. Kui $(W_1, \ldots, W_n)$ vastab ARMA(p,q) protsessile, siis saab kovariatsioonimaatriksi kirjutada kujul

$$\Sigma = \sigma_A^2 \Omega(\phi, \theta),$$

kus Ω ei sõltu juhuslike suuruste A_t dispersioonist. Suurima tõepära meetodi korral maksimiseeritakse tavaliselt tõepära logaritmi, mis avaldub kujul

$$l(\phi, \theta, \sigma_A) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma_A^2 - \ln|\Omega(\phi,\theta)| - \frac{1}{2\sigma_A^2}\mathbf{w}'\Omega(\phi,\theta)^{-1}\mathbf{w}.$$

Kuna σ_A järgi see funktsioon saavutab maksimumi kohal

$$\sigma_A^2 = \frac{1}{n}\mathbf{w}'\Omega(\phi,\theta)^{-1}\mathbf{w},$$

siis suurima tõepära hinnangud parameetritele $\phi = (\phi_1, \ldots, \phi_p)$ ja $\theta = (\theta_1, \ldots, \theta_q)$ leidmiseks tuleb maksimiseerida (pärast konstantse liidetava ärajätmist) avaldist

$$-\frac{n}{2}\ln(\mathbf{w}'\Omega(\phi,\theta)^{-1}\mathbf{w}) - \ln|\Omega(\phi,\theta)|.$$

Tingimusteta ruutude summa meetod

Inglise keeles *method of unconditional sum of squares*. Selle meetodi puhul jäetakse suurima tõepära avaldises vaatluse alt välja determinandiga liige ning minimiseeritakse avaldist

$$\mathbf{w}'\Omega(\phi,\theta)^{-1}\mathbf{w}.$$

Arvutuslikult on see veidi lihtsam, kuid ei oma märkimisväärseid eeliseid suurima tõepära meetodi eest.

4.7 ARIMA tüüpi mudelite valikust

Üldine lähtekoht on see, et mida vähem on mudelis parameetreid, seda parem (muidugi tingimusel, et mudel andmetega sobib). Mudeli sobivuse üle otsustatakse prognoosivigade sõltumatuse kontrolli põhjal; teoreetiliselt peaksid prognoosivead vastama sõltumatutele juhuslike suuruste väärtustele. Kuna täielikku sõltumatust on väga raske kindlaks teha, siis aegridade puhul keskendutakse autokorrelatsioonide uurimisele (mis peaks sõltumatute vigade puhul olema teoreetiliselt nullid).

Omaette küsimus on see, kuidas teha valikut erinevate mudelite vahel, mis kõik rahuldavad sobivuse kriteeriume, seda eriti juhul, kui parameetrite arvud on erinevad (või mudelid kuuluvad erinevatesse klassidesse). Naiivseks lähenemiseks on see, et kui me sobitame mudeleid tõepära maksimiseerides, siis sobivaima mudeli korral peaks vaadeldava aegrea tekkimise tõepära olema suurim. Selle lähenemise puuduseks on see, et lõpliku arvu andmete olemasolul saame me suurema parameetrite arvuga mudelit olemasolevate andmetega alati paremini sobitada isegi juhul, kui tegelikkuses selline mudel sobiv ei ole (nn ülesobitamine). Seetõttu tuleb sobivuse võrdlemisel kindlasti arvestada ka parameetrite arvu. Üheks selliseks sobivuse mõõdikuks, mis arvestab nii parameetrite arvu kui ka tõepära, on nn Akaike informatsioonikriteerium, mis avaldub kujul

$$AIC = 2k - 2\ln L,$$

kus k on mudeli parameetrite arv ja L on aegrea tõepära sobitatud mudeli korral.

4.8 Sesoonsed ARIMA mudelid

Sageli on andmetes mingi loomulik periood (näiteks aasta), mille korral aegrea järgnevat väärtust mõjutavad lisaks hiljutistele väärtustele ka perioodi või isegi mitme perioodi võrra minevikus olevad väärtused. Üheks võimaluseks sellise efekti modelleerimiseks on lihtsalt lisada perioodile vastavate nihetega autoregressiivseid ja/või liikuva keskmise liikmeid üldisesse mudelisse, kuid sageli on tulemusi lihtsam interpreteerida, kui perioodilist käitumist kirjeldav mudel esitada sesoonsete ja mittesesoonsete tegurite korrutise teel.

Perioodiga s multiplikatiivseteks ARIMA(p,d,q)x(P,D,Q)$_s$ tüüpi mudeliteks nimetatakse mudeleid kujul

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Z_t = \theta(B)\Theta(B^s)A_t,$$

kus

$$\phi(x) = 1 - \sum_{i=1}^{p} \phi_i x^i,\ \Phi(x) = 1 - \sum_{i=1}^{P} \Phi_i x^i,$$
$$\theta(x) = 1 - \sum_{i=1}^{q} \theta_i x^i,\ \Theta(x) = 1 - \sum_{i=1}^{Q} \Theta_i x^i.$$

Siin tuleb tähele panna, et tegemist on ARIMA tüüpi mudelite alamklassiga. Saadava mudeli puhul eeldatakse, et $\phi(x)$ ja $\Phi(x)$ rahuldavad statsionaarsuse tingimusi ning et $\theta(x)$ ja $\Theta(x)$ rahuldavad pööratavuse tingimusi; sel juhul on vastavad tingimused täidetud ka vaadeldaval ARIMA tüüpi mudelil.

Lihtsalt mõistetavaks multiplikatiivseks sesoonseks ARIMA tüüpi mudelite erijuhuks on ARIMA(p,d,q)x(0,1,0)$_s$ tüüpi mudelid, kus s tähistab vaatluste arvu perioodis (tavaliselt aasta), kuna sel juhul vastavad aastased muudud ARIMA tüüpi mudelile.

Mudelite identifitseerimiseks on kasulik teada mõningatel lihtsamatel juhtudel autokorrelatsioonikordajate teoreetilist käitumist. Vaatleme näitena ARIMA(0,0,1)x(0,0,1)$_s$ autokorrelatsioonikordajate leidmist juhul $s \geq 3$. Olgu

$$Z_t = (1 - \theta_1 B)(1 - \Theta_1 B^s) A_t$$

ehk

$$Z_t = A_t - \theta_1 A_{t-1} - \Theta_1 A_{t-s} + \theta_1 \Theta_1 A_{t-s-1}.$$

Siit $A_k,\ k \in \mathbb{Z}$ sõltumatuse tõttu saame

$$\gamma_0 = (1 + \theta_1^2 + \Theta_1^2 + \theta_1^1 \Theta_1^2)\sigma_A^2.$$

Kuna Z_{t-1} avaldises

$$Z_{t-1} = A_{t-1} - \theta_1 A_{t-2} - \Theta_1 A_{t-s-1} + \theta_1 \Theta_1 A_{t-s-2}$$

on Z_t avaldisega võrreldes samade indeksitega A_{t-1} ja A_{t-s-1}, siis

$$\gamma_1 = \operatorname{cov}(Z_t, Z_{t-1}) = (-\theta_1 - \theta_1 \Theta_1^2)\sigma_A^2,$$

seega

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{-\theta_1(1 + \Theta_1^2)}{1 + \theta_1^2 + \Theta_1^2 + \theta_1^1 \Theta_1^2}.$$

Seejärel on $k \geq 2$ korral Z_{t-k} ja Z_t avaldises samu liikmeid ainult siis, kui $k = s - 1, s, s + 1$; muudel juhtudel on kõik liikmed erinevad ning vastavad autokovariatsioonid ja autokorrelatsioonid võrdsed nulliga. Juhul $k = s - 1$ avaldub Z_{t-k} kujul

$$Z_{t-s+1} = A_{t-s+1} - \theta_1 A_{t-s} - \Theta_1 A_{t-2s+1} + \theta_1 \Theta_1 A_{t-2s},$$

seega

$$\gamma_{s-1} = \theta_1 \Theta_1 \sigma_A^2,\ \rho_{s-1} = \frac{\theta_1 \Theta_1}{1 + \theta_1^2 + \Theta_1^2 + \theta_1^1 \Theta_1^2}.$$

Analoogiliselt leiame

$$\gamma_s = (-\Theta_1 - \theta_1^2 \Theta_1)\sigma_A^2,\ \rho_s = -\frac{\Theta_1(1 + \theta_1^2)}{1 + \theta_1^2 + \Theta_1^2 + \theta_1^1 \Theta_1^2},$$

$$\gamma_{s+1} = \theta_1 \Theta_1 \sigma_A^2,\ \rho_{s+1} = \frac{\theta_1 \Theta_1}{1 + \theta_1^2 + \Theta_1^2 + \theta_1^1 \Theta_1^2}$$

Kõik ülejäänud autokorrelatsioonid on nullid. Seega vaadeldava mudeli tunnuseks on üks madalat järku nullist erinev autokorrelatsioon ning kolm nullist erinevat autokorrelatsiooni nihke s ümbruses, kusjuures perioodile s eelnev ja järgnev autokorrelatsioon on teoreetiliselt võrdsed.

Harjutus 29 *Leida ARIMA(0,0,2)x(0,0,1)$_s$ mudeli autokorrelatsioonikordajad juhul* $s \geq 5$

Harjutus 30 *Leida ARIMA(0,0,1)x(1,0,0)$_s$ mudeli autokorrelatsioonikordajad juhul* $s = 6$.

Peatükk 5

Mitmemõõtmelised ja mittelineaarsed mudelid

Siiani vaatlesime aegrea tulevikuväärtuste prognoosimist juhul, kus kasutada oli ainult vaadeldava aegrea minevikuväärtused. Sageli aga on võimalik prognoose tunduvalt täpsustada, kui kasutada lisaks vaadeldava aegrea minevikuväärtustele veel teisi andmeid, näiteks teiste juhuslike suuruste minevikuväärtuseid. Näiteks on loomulik arvata, et majanduse üldseisundi näitajate minevikuväärtused mõjutavad oluliselt siseturismiga seotud suuruseid. Käesolevas peatükis vaatleme mõningaid mooduseid, kuidas selliseid sõltuvusi matemaatiliselt modelleerida ning kuidas vastavaid mudeleid praktikas kasutada.

5.1 Mitmene lineaarne regressioon ARIMA tüüpi vigadega

Olgu Z_t meid huvitava tunnuse väärtus ajal t ning $(X_1(t), \ldots, X_m(t))$ argumenttunnuste vektor, mida saab kasutada suuruse Z_t prognoosimiseks. Mitmese lineaarse regressiooni mudeliks on mudel kujul

$$Z_t = \beta_0 + \sum_{i=1}^{m} \beta_i X_i(t) + \varepsilon_t,$$

kus vead ε_t on sama jaotusega, sõltumatud ja tsentreeritud, kordajate hinnanguteta vigade tuletamisel eeldatakse ka vigade normaaljaotusele vastavust. Aegridade puhul enamasti selline mudel (eriti vigade sõltumatuse eeldus) ei kehti, mistõttu standardsete lineaarse regressiooni vahendite kasutamine ning saadud mudeli põhjal prognoosimine võib viia vägagi valedele tulemustele. Sageli aga sobivad aegridade puhul andmetega mudelid, kus vead ε_t vastavad mingile ARIMA tüüpi protsessile, sel juhul räägitakse lineaarsest regressioonist ARIMA tüüpi vigadega ehk ARIMAX mudelist. Mudeli sobitamise protseduur on kaheetapiline: kõigepealt sobitame tavalise regressiooni abil andmetele mitmese regressioonimudeli ja analüüsime vigade käitumist. Vigade käitumise põhjal valime ARIMA mudeli kuju suuruste ε_t jaoks ning seejärel leiame valitud ARIMAX tüüpi mudeli parameetrid (nii β-d kui ARIMA kordajad) näiteks suurima tõepära meetodil või siis tinglike prognoosivigade

ruutude summa minimiseerimise teel. Leitud mudeli headuse kriteeriumiks on prognoosivigade sõltumatus, mida testitakse autokorrelatsioonide sõltumatute juhuslike suuruste väärtustele vastavuse testimise abil.

5.2 Ülekandefunktsiooni mudelid

Inglise keeles *transfer function models*. Vaatleme juhtu, kus meil on kaks protsessi Z_t ja X_t, millele vastavate aegridade väärtused on meil olemas. Lihtsuse mõttes eeldame, et mõlemad protsessid on statsionaarsed ning tsentreeritud (vastasel juhul võib proovida leida mõlemast sobivat järku diferentsid ja maha lahutada nende keskmised, et saada soovitud omadustega ridu). Meie eesmärgiks on kindlaks teha, milliseid X minevikuväärtuseid (ja võib-olla ka Z minevikuväärtuseid) regressoritena kasutada nii, et saada võimalikult häid ennustusi protsessi Z jaoks. Täpsemalt, vaatleme ARIMAX mudelit kujul

$$Z_t = \beta_0 + \sum_{i=b}^{\infty} \beta_i X_{t-i} + \varepsilon_t, \tag{5.1}$$

kus $b \geq 1$ ja suurused ε_t vastavad mingile $ARMA$ protsessile

$$\phi(B)\varepsilon_t = \theta(B)A_t,$$

kusjuures eeldame, et suurused A_t on sõltumatud ka protsessi X_t väärtustest. Sel juhul on ka suurused ε_t sõltumatud suurustest X_t. Selleks, et parameetreid oleks lõplik arv ning et ennustamiseks kasutataks ainult X_t minevikuväärtuseid, otsime sobivat mudelit selliste hulgast, kus funktsioon

$$\beta(x) = \sum_{i=b}^{\infty} \beta_i x^i$$

on esitatav lõpliku arvu parameetrite abil kujul

$$\beta(x) = x^b \frac{v(x)}{\delta(x)} = x^b \frac{\sum_{i=0}^{s} v_i x^i}{1 - \sum_{i=1}^{r} \delta_i x^i}.$$

Sellise mudeli võib kirjutada ka kujul

$$Z_t = \sum_{i=1}^{r} \delta_i Z_{t-i} + \sum_{i=0}^{s} v_i X_{t-b-i} + \eta_t,$$

kus suurused η_t vastavad ARMA protsessile kujul

$$\phi(B)\eta_t = \delta(B)\theta(B)A_t.$$

Funktsiooni $\beta(x)$ nimetatakse ülekandefunktsiooniks, kuna ta kirjeldab, kuidas X omandatud väärtused mõjuvad ehk kanduvad üle suuruste Z väärtustele. Mudeli sobitamise etapid on järgmised:

1. Leiame hinnangud suurustele β_i
2. Suuruste β_i hinnangute põhjal määrame kindlaks sobiva nihke b ning kasutades teadmist sellest, kuidas erinevate r ja s väärtuste korral peaks suurused β teoreetiliselt käituma, leiame hinnangud ka parameetritele r ja s
3. Leiame sobiva mudeli vigade ε_t jaoks
4. Hindame mudeli parameetreid suurima tõepära meetodil
5. Kontrollime jääkvigade sõltumatust
6. arvutame prognoosid (kuni b ajaperioodi ette).

Vaatleme lähemalt mõningaid nendest etappidest

5.2.1 Suuruste β_i hindamine

Tähistame kujul $\gamma_{xx}(k)$ ja $\gamma_{zz}(k)$ protsesside X ja Z k-ndat järku autokovariatsioone ning defineerime ristkovariatsioonid kujul

$$\gamma_{xz}(k) = \operatorname{cov}(X_t, Z_{t+k});$$

vastavad autokorrelatsioonid olgu $\rho_{xx}(k)$, $\rho_{yy}(k)$ ning $\rho_{xz}(k)$. Korrutades võrrandit (5.1) suurusega X_{t-k} ning leides keskväärtuse (ehk arvutades võrrandi parema ja vasaku poole kovariatsiooni suurusega X_{t-k}) saame

$$\gamma_{xz}(k) = \sum_{i=0}^{\infty} \beta_i \gamma_{xx}(i-k).$$

Kui nüüd eeldada, kordajad β_i on praktiliselt võrdsed nulliga alates mingist järgust K, siis saame võrrandisüsteemi suuruste $\gamma_{xz}(k),\ k = 0, \ldots, K-1$ määramiseks. Samas on nende võrrandite lahendamisel (asendades teoreetilised auto- ja ristkovariatsioonid empiirilistega) saadud hinnangud küllalti ebatäpsed, seetõttu on võimaluse korral parem kasutada nn eelvalgendamise (inglise keeles *prewhitening*) tehnikat.

Eelvalgendamise tehnika on rakendatav, kui protsess X vastab mingile pööratavale ARMA tüüpi mudelile. Oletame, et X vastab mudelile

$$\phi_x(B)X_t = \theta_x(B)\alpha_t,$$

kus suurused α_t on sõltumatud, sama jaotusega (ning sõltumatud suurustest $X_{t-1}, X_{t-2}, \ldots$. Eelnevate eelduste põhjal on nad ka sõltumatud suurustest η_t. Pööratavuse tõttu saame

$$\alpha_t = \theta_x(B)^{-1}\phi_x(B)X_t.$$

Rakendades nüüd operaatorit $\theta_x(B)^{-1}\phi_x(B)$ võrduse (5.1) mõlemale poole, saame

$$W_t = \sum_{i=b}^{\infty} \beta_i \alpha_{t-i} + \xi_t,$$

kus

$$W_t = \theta_x(B)^{-1}\phi_x(B)Z_t,\ \xi_t = \theta_x(B)^{-1}\phi_x(B)\varepsilon_t.$$

Leides nüüd eelneva võrduse mõlema poole kovariatsiooni suurusega α_{t-k} saame

$$\gamma_{\alpha w}(k) = \sigma_\alpha^2 \beta_k,$$

kust saame kordaja β_k avaldada. Praktilise arvutuse seisukohalt on suurused α_k leitavad suuruste ühesammuliste prognooside vigadena leitud mudeli abil suuruste X_t prognoosimisel ning suurused W_t vastavad täpsel sama mudeli kasutamisel suuruste Y_t ennustamisel tekkivatele ühesammulistele prognoosivigadele.

5.2.2 Mudeli kuju parameetrite b,r ja s valik

Kui kordajad β_k on hinnatud, siis b valiku kriteeriumiks on tingimus $\beta_i \approx 0,\ i = 0, 1, \ldots, b-1$. Kui ainult (küllalt väike) lõplik arv kordajatest β_k on nullist erinevad, siis võime võtta $r = 0$ ja $s = k_0 - b$, kus k_0 vastab viimase nullist erineva β_k indeksile. Muudel juhtudel aga saab analoogiliselt ARIMA tüüpi mudelite analüüsiga näidata, et alates järgust $k = b + s + 1$ rahuldavad kordajad β_k rekurrentset võrrandit

$$\beta_k = \sum_{i=1}^{r} \delta_i \beta_{k-i}.$$

Teades selliste rekurrentsete võrrandite lahendite käitumist on võimalik püstitada hüpoteese sobiva r (ja ka s) väärtuse kohta.

Harjutus 31 *Näidata, et juhul* $r = 1$ *kehtib*

$$\beta_k = \delta_1^{k-b-s}\beta_{b+s}\ \forall k > b + s.$$

15-da loengu lõpp

5.3 Mitmemõõtmeline ARMA mudel

Vaatleme juhtu, kus meil on komplekt k erinevast aegreast, millele vastav k-mõõtmeline protsess olgu $\mathbf{Z}_t$ (siin $\mathbf{Z}_t$ on k-mõõtmeline vektor). Eeldame, et see protsess on statsionaarne ning (lihtsuse mõttes) tsentreeritud. Kui protsess rahuldab võrrandit

$$\mathbf{Z}_t = \sum_{i=1}^{p} \Phi_i \mathbf{Z}_t + \mathbf{A}_t - \sum_{i=1}^{q} \Theta_i \mathbf{A}_{t-i},$$

kus Φ_i ja Θ_i on $k \times k$ maatriksid ning A_t on sõltumatud k-mõõtmelised juhuslikud suurused, mis on sõltumatud ka varasematest $\mathbf{Z}$ väärtustest, siis öeldakse, et $\mathbf{Z}$ on k-mõõtmeline ARMA(p,q) protsess.

5.4 Garch mudelid

Sageli on majanduslike aegridade puhul võimalik täheldada seda, et suuremate võnkumistega rahutumad perioodid vahelduvad suhteliselt stabiilsete perioodidega ning sageli jääb see efekt alles ka prognoosivigade puhul pärast parima ARIMA tüüpi mudeli sobitamist. See aga tähendab, et vähemalt prognoosivigade arvutamise tulemused ei ole usaldusväärsed, kuna seal lähtutakse üleüldisest keskmisest vigade standardhälbest, mis huvipakkuva hetke jaoks võib olla liiga suur (kui parajasti on tegemist rahulikuma perioodiga) või liiga väike (kui on tegemist rahutuma perioodiga).

Lihtsalt vaatluse abil ei ole alati lihtne eristada juhuslikult tekkivaid sõltumatute juhuslike suuruste keskmiselt suuremate ja keskmiselt väiksemate rühmade teket sellisest, kus sellised rühmad on seotud muutuva varieeruvusega. Samas ARIMA mudelite sobitamisel võime teha lihtsa testi: vaatleme prognoosivigade ruutude autokorrelatsioone ja osaautokorrelatsioone. Kui on tegemist mudeliga, kus häiritused A_t on sõltumatud, on ka prognoosivigade ruutude teoreetilised autokorrelatsioonid ja osaautokorrelatsioonid nullid ning seega peaks empiirilised autokorrelatsioonid ja osaautokorrelatsioonid jääma vastavatesse veapiiridesse. Kui see nii aga ei ole, siis ei pruugi leitud mudel olla sugugi parim ning kindlasti tuleb veapiiridesse suhtuda suure ettevaatusega.

Üheks mudelite klassiks, mille korral varieeruvus muutub ajast sõltuvalt sissetulnud häiritustest, on ARIMA mudelid GARCH (*Generalised Autoregressive Conditional Heteroskedasticity*) häiritusega kujul

$$\begin{aligned} Z_t &= \sum_{i=1}^{p} \phi_i Z_{t-1} + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \\ \varepsilon_t &= \sqrt{\sigma_t} A_t, \\ \sigma_t^2 &= \omega + \sum_{i=1}^{q_1} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p_1} \beta_i \sigma_{t-i}^2, \end{aligned}$$

kus suurused A_t on sõltumatud, sama jaotusega ning tsentreeritud. Kui parameeter σ_t sõltub ainult eelnevatest häiritustest (st kui β-dega liikmeid pole), siis nimetatakse seda mudelit ARCH mudeliks.

Mudeli sobitamise protseduur on järgmine:

1. Leiame andmestikule parima ARIMA tüüpi mudeli

2. Vaatleme prognoosivigade ruutude autokorrelatsioone ja osaautokorrelatsioone. Kui mõned autokorrelatsioonid on veapiiridest selgelt väljas, siis on mõistlik katsetada GARCH tüüpi mudelitega. Esialgseks hüpoteesiks parameetri q_1 osas võib võtta nullist erinevate osaautokorrelatsioonide arvu; samas tasub vaadelda ka madalamat järku GARCH mudeleid.

3. Mudeli loeme sobivaks siis, kui nn. normaliseeritud prognoosivead (st vead, mis on jagatud hetkele vastava σ_t väärtusega) ja nende ruudud ei ole oluli-

selt korreleeritud (st Ljung-Box testi p-väärtused on nii vigade kui ka vigade ruutude korral piisavalt suured).

Kirjandus

[1] Statistikaamet, tarbijahinna indeks,
http://pub.stat.ee/px-web.2001/Database/Majandus/04HINNAD/04HINNAD.asp

[2] Statistikaamet, majutatud turistide arv,
http://pub.stat.ee/px-web.2001/Database/Majandus/23Turism_ja_majutus/02Majutus/02Majut

[3] http://en.wikipedia.org/wiki/Local_regression

[4] R. B. Cleveland, W. S. Cleveland, J.E. McRae, and I. Terpenning (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. Journal of Official Statistics, 6, 3–73.

[5] http://www.census.gov/srd/www/x12a/

[6] M. G. Kendall, A. Stuart, *The Advanced Theory of Statistics*, Vol. 3, 1966, London, Griffin

[7] T. W. Andersson, *The statistical analysis of time series* J. Wiley & Sons, 1971

[8] G. E. P. Box, G. M. Jenkins, *Time series analysis: forecasting and control*, Holden-Day 1969.

[9] D. Williams, Probability with Martingales, Cambridge University Press 1991.

Aegridade analüüs
Praktikum nr. 1, 2012

Käsitletavad teemad:

- Aegrea andmete hankimine ning mitmesugustes formaatides andmete analüüsikeskkondadesse sisselugemine
- Aegrea graafiline kujutamine

Aegrida on teatud sagedusega (nt kord minutis, kord päevas, kord kuus, kord kvartalis, kord aastas) mõõdetud andmete kogum. Aegrea andmestikus peaks olema nii mõõdetud andmed kui info selle kohta, kuidas igale mõõtmisele vastavat ajamomenti kindlaks teha. Osades andmestikes on kirjas kommentaar esimese vaatluse aja ja sageduse kohta ning ülejäänud andmestik koosneb ainult mõõtmistulemustest. Osades andmestikes on ajainfo ühes tulbas (nt kujul 2010-09-01) ning vastav vaatlustulemus teises tulbas. Sageli on aga ajainfo mitme välja vahel laiali ning mõnikord esitatakse andmestik ka tabeli kujul (kus read vastavad näiteks aastatele ning veerud kuudele). Erinevused võivad olla veel kasutatavates väljaeraldajates (tühik, tabulaator, koma, semikoolon jms). Seega on tähtis andmestiku sisuga kõigepealt visuaalselt tutvuda ning vaatlusest saadud info põhjal osata andmeid õigel kujul analüüsikeskkonda sisse lugeda.

Käesolevas praktikumis vaatleme andmete sisselugemist tekstifailidest, kuna muul kujul olevaid andmeid on enamasti teksti kujul (nt csv failina) küllalt lihtne salvestada.

Analüüsikeskkondadena kasutame SAS tarkvara ning R tarkvara.

Ül. 1 Andmestike hankimine. Internetis on saadaval suurel hulgal aegridade andmestikke. Hangime ühe andmestiku: Kinnisvara ostu/müügitehingute summad Eestis (alustades lehelt `http://www.stat.ee/`, pärast mitmeid valikuid andmestik kv015, formaadiks valida tabelieraldusega pealkirjata tekst). Aja kokkuhoiu mõttes on ülejäänud andmestikud (USA töötuse määr lehe `http://www.bls.gov/` vahendusel, Cisco aktsia ajaloolised sulgemishinnad aadressilt `http://www.google.com/finance` ja tarbijahinnaindeks statistikaameti lehelt) saadavad kursuse veebilehelt
`http://moodle.ut.ee/course/view.php?id=1742`.
Kopeerige andmed oma kodukataloogi (või selle mingisse alamkataloogi) ning tutvuge (näiteks notepad vahendusel) nende failide sisuga.

Ül. 2 Andmete sisselugemine SAS keskkonda. Protseduur on järgmine:

1. Anname andmestikule nime ja alustame sisselugemise protseduuri käsuga `data nimi;`
2. Kirjeldame sisendfaili käsuga `infile failinimi suvandid;`, kus tähtsamateks suvanditeks on:
   `firstobs=n` - andmed hakkavad failis alates n-dast reast
   `delimiter="s"` - vaikimisi on eraldajateks tühikud. Kui s koosneb mitmest sümbolist, siis igaüks neist loetakse eraldajaks. Kui eraldajaks on tabulaator, siis selle saab kirjeldada kujus `delimiter='09'x`
   `dsd` - vajalik, kui kaks järjestikust eraldajat (näiteks koma) tähendavad, et mingi muutuja väärtus on puudu.
3. andmestikku kirjutatavate muutujate määramine: käsuga `keep nimekiri;`, kus nimekiri on tühikutega eraldatud muutujate nimed

4. Ajainfo jaoks kasutatava abimuutuja tekitamine. Kuupäevadega tegelemiseks on SAS-is palju võimalusi. Aegridadega tegelemisel on käesoleva kursuse jaoks kõige lihtsam teha andmete vaatlusel kirja vaatluste alguskuupäev ja intervall ning tekitada kuupäevainfo neid kasutades, ignoreerides muud failis olevat kuupäevainfot. Selleks aga tekitame abimuutuja, mis loendab vaatlusi käsuga `retain abi 0;`. Selle muutuja abil saame arvutada jooksva vaatluse kuupäeva käsuga `date=intnx(interval,'alguskuup'd,abi)`, kus intervall on tavaliselt 'day','month','qtr' või 'year' ning `alguskuup` on alguskuupäev kujul 02feb07 (st päeva number, inglisekeelse kuunimetuse kolm tähte ning aastanumbri kaks viimast kohta. Siinjuures tuleb muutuja abi väärtust suurendada pärast igat `output;` käsku ning ajas tagurpidi oleva andmestiku korral tuleb kasutada viimast kuupäeva ning muutuja abi väärtusi kahandada.

5. Määrame andmete vaatamiseks aja näitamise formaadi käsuga `format date;`. Kvartaalsete andmete sobivaks formaadiks on `yyqc4.`, igakuiste andmete korral `mmyyc5.`, igapäevaste korral aga `date7.`. Kasutada on võimalik palju muid formaate, mille kohta leiab infot SAS spikrit kasutades.

6. Kasutame andmete sisselugemiseks `input` käsku. Kui igas faili reas on ainult üks aegrea väärtus, siis `input` käsu taha kirjutame loetelu väljanimedest, mis on tühikutega eraldatud, kusjuures iga tekstilise välja nime taga tuleb pärast tühikut kirjutada $. Seejärel arvutame kuupäeva, kasutame `output` käsku andmete andmestikku kirjutamiseks ning järgmisel real muudame abimuutuja väärtust. Kui aga igas reas on mitu ühe aegrea väärtust, siis on üks võimalik tegevuseeskiri järgmine: 1) loeme sisse eelnevad muutujad kujul `input nimekiri @;`, kus `@` tähendab, et järgnevad `input` käsud jätkavad samast faili reast lugemist. 2) Kirjutame tsükli, mis loeb sisse aegrea väärtused:

```
do i=1 to n;
   input nimi @;
   date=intnx(interval,'alguskuup'd,abi);
   output;
   abi=abi+1;
end;
```

   Siin n on vaatluste arv reas.

7. Kirjutame lõppu `run;` ning jooksutame eelnevat koodi. Andmestik peaks seejärel olema valmis.

8. Puuduvate andmetega kirjete eemaldamine andmestiku lõpust. Selleks, et olevasoleva andmestiku põhjal uut teha, võib kasutada andmesammu (*data step*) kujul

   `data` andmestiku2_nimi;
   `set` andmestiku1_nimi;

   sel juhul tehakse järgnevad käsud andmestiku nr 1 iga rea korral läbi ja tulemused kirjutatakse andmestikku nr 2. Kui kasutada mõlema jaoks sama nime, siis kirjutatakse vanad andmed üle (või tekitatakse sama nimega tabel, kus on lisaks vanadele ka uued muutujad). Selleks, et eemaldada kõik kirjed, kus mingi numbrilise muutuja väärtus puudub, sobib käsk

   `if muutuja^=.;`

   eelneva käsu tulemusena kirjutatakse väljundandmesse ainult kirjed, kus vastaval muutujal on väärtus olemas. Andmestiku lõpust saab kuupäeva järgi kirjeid eemaldada kujul `if date<'kuupäev'd;`

kus kuupäev peaks olema hilisem kui viimane vaatlus ja varasem rea lõpus olevate puuduvate vaatluste kuupäevadest. Siin on tähtis teada, et näiteks kvartaalsete andmete korral loetakse vaikimisi kuupäevaks kvartali esimese kuu esimest kuupäeva.

9. Selleks, et andmete analüüs toimuks õigesti, on vaja andmestik viia ajas kavava järjestuse kujule. Kui vaatlustel on juures õiges järjestuses kuupäevad, siis sobib selleks protseduur `sort`, mille lihtsaim kuju on järgmine:
   ```
   proc sort data=andmestiku_nimi;
   by date;
   run;
   ```
   Siin on eeldatud, et andmestikus on väli nimega `date`, kus on kuupäevainfo. See tegevus on näiteks vajalik Cisco andmete korral.
10. Graafiku tekitamiseks on `gplot` protseduur, mille süntaks lihtsaimal kujul on järgmine:
    ```
    proc gplot data=andmestiku_nimi;
    symbol i=spline v=dot h=0.3;
    plot y_muutuja*x_muutuja;
    run;
    ```

Ül. 3 Aegrea sisselugemine R keskkonda. Lugege R keskkonda sisse ja kujutage graafiliselt neli esimeses ülesandes hangitud aegrida. Protseduur on järgmine:

1. Andmete sisselugemine. Selleks sobiv käsk on `andmed=read.table(failinimi,...)`, kus kolme punkti asemel võib anda loetelu mitmetest suvanditest. Mõned olulisemad:
   `skip=x` - ignoreerida x rida faili algusest. Vaikimisi 0
   `sep="c"` - andmeväljade eraldajaks on sümbol c. Vaikimisi tühikud või tabulaatorid
   `header=T` - esimene sisseloetav rida sisaldab veergude nimesid
   `dec=","` - kümnenderaldajaks on koma. Vaikimisi on punkt
   `na.strings="s"` - millega on puuduvad väärtused tähistatud. Vaikimisi on NA.
2. Aegrea väärtuste väljaeraldamine. Kui aegrida paikneb ühes veerus, siis sobivad käsud kujul `rida=andmed$nimi` (kus nimi on vastava veeru nimi) või `rida=andmed[,nr]`, kus nr on veeru järjekorranumber. Kui vaatlused on esitatud tabeli kujul nii, et iga rida vastab näiteks aastale ja veerud kuudele või kvartalitele, siis saab õigel kujul olevad arvud kätte käsuga `rida=c(t(andmed[,a:b]))`. Siin a on veeru number, kus on esimene vaatlus ning b on veeru number, kus on viimane vaatlus selles reas. Seletuseks: `andmed[,a:b]` eraldab andmestikust näidatud veergudega määratud osa, `t()` transponeerib selle ning `c()` väljastab tulemuse ühe vektorina, kuhu argumendiks oleva maatriksi väärtused lähevad veergude kaupa järjest.
3. Aegrea lõpus olevate puuduvate väärtuste eemaldamine. Sageli on andmestiku lõpus mõned puuduvad väärtused lihtsalt selle tõttu, et need vastavad tulevikukuupäevadele ja seega ei ole tegemist andmetega, mis sisuliselt puudu on (nagu näiteks mingi vahelejäänud mõõtmistulemus minevikus). Need tuleks ennem aegrea uurimist kõrvaldada. Üks võimalus on kasutada käsku kujul `rida=rida[1:(length(rida)-x)]`, kus x on rea lõpust eemaldatavate väärtuste arv. Kui puuduvad väärtused on ainult real lõpus, siis võib eemalda need ka käsuga `rida=rida[!is.na(rida)]`, mis eemaldab kõik puuduvad väärtused.

4. Andmete paigutamine ajas kasvavalt. Osades andmestikes on kõige uuemad andmed eespool; aegrea jaoks oleks vaja, et mida edasi, seda uuemad andmed. Kui on vaja järjekorda muuta, siis seda saab teha käsuga `rida=rida[length(rida):1]`
5. Mõõtmistulemustest aegrea tegemine. Selleks, et oleks väärtuste rida aegreana käsitleda, tuleb lisada ka ajainformatsioon selle kohta, kui sageli tulemusi mõõdeti. Aegrea tegemine käib kujul `aegrida=ts(rida,...)`, kus kolme punkti asemel võib olla loetelu suvanditest. Mõned tähtsamad:
   `frequency=n` - mitu korda ajaühikus mõõdetakse. Vaikimisi 1, kvartaalsete andmete puhul 4, igakuiste andmete puhul 12.
   `start=x` - esimese mõõtmise aeg. Kui mõõtmisi on üks kord ajaühikus või kui esimene väärtus on esimese ajaühiku esimene mõõtmine, siis x peaks olema esimese ajaühiku (nt aasta) number; Kui andmestik ei alga perioodi algusest (nt esimene tulemus on 2000 aasta mais), siis tuleb x asemel kirjutada paar ajaühikust ja mõõtmise numbrist kujul `c(aasta,nr)`.
6. Aegrea graafiku tekitamine käib käsuga `plot(aegrida)`. Kui on soovi kuvada ainult osa aegreast, siis seda saab teha käsku `window()` kasutades kujul `plot(window(aegrida,start=algus,end=lopp)`, kus algus ja lopp võivad olla lihtsalt perioodi (tavaliselt aasta) numbrid või paarid perioodi numbrist ja mõõtmise järjekorranumbrist.

```
#skripti puhul on eeldatud, et andmefailid on h kettal kataloogis aegread_praktikum
###########kinnisvara ost-müük
#1)andmestiku sisselugemine. Eraldajaks on tabulaator, mis on kuulub vaikimisi määratud
eraldajate hulka
#seega eraldajat ei pea ütlema, kuid võib kirjutada sep="\t" (\t tähistab R-is tabulaatorit)
andmed1=read.table("h:/aegread_praktikum/kv015m.csv",na.strings="..")
#2)arvulised tulemused on kolmandas tulbas, mille nimi on V3
rida1=andmed1$V3
#3)puuduvate vaatluste eemaldamine lõpust
rida1=rida1[1:(length(rida1)-2)] #või rida1=rida1[!is.na(rida1)]
#4)ajas ümberpööramine- ei ole vaja
#5)aegrea tegemine. Tegemist on kvartaalsete andmetega
kv=ts(rida1,start=1997,frequency=4)
#6)graafiku joonistamine
plot(kv)
#saab joonistada ka ainult osa reast:
plot(window(kv,start=c(2000,3)))

###########USA töötuse määr
#1)andmestiku sisselugemine: andmed hakkavad seitsmeteistkümnendast reast. Eraldajaks on koma
andmed2=read.table("h:/aegread_praktikum/unemployment2012.txt",na.strings=" ",sep=',',skip=16)
#2)arvulised tulemused on tulpades 2-13 (esimeses on aasta, viimases aga aasta keskmine)
rida2=c(t(andmed2[,2:13]))
#3)puuduvate vaatluste eemaldamine lõpust
rida2=rida2[!is.na(rida2)]
#4)andmete järjekord on õige
#5)aegrea tegemine: tegemist on igakuiste andmetega, st 12 tulemust aasta kohta
rate=ts(rida2,start=c(1948,1),frequency=12)
#6)graafiku joonistamine
plot(rate)
plot(window(rate,start=c(2000,3)))

###########Cisco andmed.
#1) sisselugemine. Kasutame esimeses reas olevaid veerunimesid. Eraldajaks on koma
andmed3=read.table("h:/aegread_praktikum/cisco2012.csv",sep=',',header=T)
#2) Vaatleme sulgemishindasid, st tulpa nimega Close
rida3=andmed3$Close
#3) puuduvaid andmeid ei ole
#4) andmed on tagurpidi ajalises järjestuses, keerame ümber
rida3=rida3[length(rida3):1]
#5) aegrea tegemine. Andmed on igapäevased. Finantsmatemaatikas on tavaks arvestada
# ainult tööpäevi, vaadeldavas aastas on 251 tööpäeva (andmestik on aasta kohta ning seal on
251 rida)
# 2012 aastasse jääb 173 vaatlust; eeldame, et sama palju oli 2011 tööpäevi ennem esimest
kirjet
cisco=ts(rida3,start=c(2011,173),frequency=251)
#graafiku joonistamine
plot(cisco)

##########tarbijahinnaindeks
#1)andmete sisselugemine. Andmed hakkavad neljandast reast
andmed4=read.table("h:/aegread_praktikum/IA02s.csv",na.strings="..",skip=4,sep=";")
#2) andmed on 12 kaupa ridades, esimeses veerus lihtsalt tühi string ning teises on
aastanumber. Paneme nad õigesti ritta
rida4=c(t(andmed4[3:14]))
#3) eemaldame puuduvad andmed. Nii saab, kuna ainsad puuduvad andmed on lõpus
rida4=rida4[!is.na(rida4)]
```

```
#4) ümber pöörata ei ole vaja
#5) Teeme aegrea
indeks=ts(rida4,start=1998,frequency=12)
#joonistame graafiku
plot(indeks)

#väga lihtne on vajadusel aegread uuesti tekitada, jooksutades sisselugemise skripte. Soovi
korral võib neid aga ka salvestada
#ning ennem kasutamist sisse lugeda. Selleks on save() ja load() käsud. Kasutamise näide
#save(cisco,kv,indeks,rate,file="h:/aegread_praktikum/prax1andmed.RData")
#siis saab neid aegridu hiljem sisse lugeda käsuga
load("h:/aegread_praktikum/prax1andmed.RData")
#pärast load käsku on aegread cisco,kv,indeks ja rate defineeritud ning neid saab edasiseks
analüüsiks kasutada
```

```
*praktikumi nr. 1 skript SAS-is;
libname aegread "h:/aegread_praktikum";*määrab nime andmekaustale
ning selle kausta tegeliku asukoha;
data aegread.kv;
*andmete tekitamine ja teisendamine käib nn "data step" abil.
esimesel real antakse uue andmestiku nimi kujul
kausta_nimi.andmestiku_nimi. Kui kasuta nime ei anta, siis läheb
andmestik ajutisse kasuta nimega work;
infile "h:/aegread_praktikum/kv015m.csv" firstobs=1 delimiter="09"x
dsd;
*sisendfail, mitmendas reas andmed ning mis on andmeväljade
eraldajaks. Tabulaatori puhul andtakse klahvi kood 16-süsteemis,
seda tähistab see
x pärast koodi "09". tavalise sümboli puhul tuleb lihtsalt sümbol
kirjutada jutumärkide või ülakomade vahele.;
keep date tehingud;
*me võime defineerida mitmeid muutujaid ning andmevälju, kuid
väljundisse ei ole kõiki vaja. Käsu keep taga näitame, mida
tulemusandmestikku kirjutada;
retain abi 0; *abimuutuja kuupäeva tekitamiseks;
date=intnx("qtr","01jan97"d,abi); *iga rea sisselugemisel tekitame
uue kuupäeva, mis on antud kuupäevast suurusega abi määratud
kvartalite võrra kaugemal;
input tmp1 $ tmp2 $ tehingud; *näitame, mitu ja mist tüüpi välja ($
nime järel tähistab tekstilist välja) tuleb igast failireast sisse
lugeda;
output; *kirjutame hetkeseisu väärtused andmestikku;
format date yyqc4.; *see on mugavuse jaoks. Määrab, mis kujul
kuupäevi meile näidatakse, kui me andmeid vaatame;
abi=abi+1; *suurendama muutuja abi väärtust ennem järgmise kirje
lugemist;
run;
*eemaldame kõik puuduvad vaatlused;
data aegread.kv;
set aegread.kv;
if tehingud^=.; *sas tekitab väljundit ainult siis, kui hetkel
vaadeldava rea puhul on tingimus täidetud;
*seda kuju ei ole hea kasutada siis, kui andmestiku keskelt on
midagi puudu. Siis peaks kuupäeva järgi algusest ja lõpust
eemaldama,
kuid vahepealsed puuduvad väärtused alles jätma.;
run;
*kuna andmed on kuupäeva järgi kasvavas järjestuses, siis ei ole
vaja neid ümber järjestada;
*viimane samm on visuaalne vaatlus graafiku tekitamise teel. Kui
graafik on mõistlik, on lootus, et andmete sisselugemine õnnestus.;
proc gplot data=aegread.kv;*data taga näidatakse kasutatav
andmestik;
symbol i=spline v=dot h=0.3;*i=spline käsib joonistada pideva joone,
st andmepunktid ühendatakse graafikul joonega.
v=dot ütleb, et andmestikus olevatele väärtustele vastavad punktid
tuuakse joonisel välja suurema pallikesena.
h=0.3 abil saab näidata pallikese suurust;
plot tehingud*date;*y-koordinaat ennem ja x-koordinaat pärast;
run;
*loeme sisse sulgemishindade andmed failist cisco2012.csv;
data aegread.cisco;
```

```
infile "h:/aegread_praktikum/cisco2012.csv" firstobs=2
delimiter="," dsd;
*andmed hakkavad teisest reast (esimeses on väljade nimed),
eraldatud on komaga;
keep date Close;*meid huvitab praegu ainult sulgemishindade aegrida,
ülejäänu on ebaoluline;
retain abi 0;
date=intnx("weekday","7sep12"d,abi);*kauplemine toimub tööpäevadel;
input tmp1 $ tmp2 tmp3 tmp4 Close; *igast reast peame lugema välju
kuni huvipakkuva tulbani (mis on viies);
output;
format date date7.;
abi=abi-1; *aeg on tagurpidi;
run;
*puuduvaid kirjeid ei ole;
*andmed on ajas tagurpidi, nii et pöörame ümber;
proc sort data=aegread.cisco;
by date;
run;
*vaatame tulemust jooniselt;
proc gplot data=aegread.cisco;
symbol i=spline v=dot h=0.3;
plot Close*date;
run;
*tabeli kujul oleva aegrea sisselugemine;
data aegread.indeks;
infile "h:/aegread_praktikum/IA02s.csv" firstobs=5 delimiter=";"
dsd;
*aegrea andmed hakkavad viiendast reast, eraldatud semikooloniga;
keep date indeks;
retain abi 0;
input tmp1 $ tmp2 $ @;
*igas reas on kõigepealt kaks aegritta mittekuuluvat teksti kujul
olevat kirjet.  @ tähendab, et pärast nende sisselugemist töötame
edasi
sama andmereaga;
do i=1 to 12; * edasi tulevad tarbijahinna indeksi erinevatele
kuudele vastavad väärtused, mida on 12 tükki;
      input indeks@;
      date=intnx("month","1jan98"d,abi);
      output;
      abi=abi+1;
end;
format date mmyyc5.;
run;
*eemaldame puuduvad kirjed;
data aegread.indeks;
set aegread.indeks;
if indeks^=.;
run;
*kontrollime visuaalselt;
proc gplot data=aegread.indeks;
symbol i=spline v=dot h=0.3;
plot indeks*date;
run;
*USA töötuse määra andmestik;
data aegread.unemployment;
```

```
infile "h:/aegread_praktikum/unemployment2012.txt" firstobs=17
delimiter="," dsd;
*aegrea andmed hakkavad viiendast reast, eraldatud komaga;
keep date rate;
retain abi 0;
input tmp1 @;
*igas reas on kõigepealt aastanumber, mis ei kuulu aegritta.;
do i=1 to 12; * edasi tulevad töötuse määra erinevatele kuudele
vastavad väärtused, mida on 12 tükki;
      input rate @;
      date=intnx("month","1jan48"d,abi);*kuude kohta algava andmed
48-st aastast;
      output;
      abi=abi+1;
end;
format date mmyyc5.;
run;
```

Aegridade analüüs
Praktikum nr. 2, 2012

Käsitletavad teemad: Andmete töötlemine, graafiline esitamine, lihtsamad silumisvõtted.

Sageli on andmestikuga vaja ennem põhjalikumat analüüsi teostada mitmesuguseid tegevusi, näiteks rakendada mingeid teisendusi, arvutada uusi väärtusi, siluda jms. Kuigi seda on sageli võimalik teha mingite keerukamate andmetöötlusfunktsioonide suvandeid kasutades, ei pruugi kõik vajaminevad tegevused nii teostatavad olla. Seetõttu vaatleme mõningaid võimalusi, kuidas lihtsamaid teisendusi ise rakendada.

Praktikumi tegevused SAS-is (eeldusel, et eelmise praktikumi andmestikud on SAS-i sisse loetud):

1. Lugege sisse turistide andmestik (saadaval aine kodulehelt Moodle's)

2. Tehke graafik majutatud turistide andmetest. Graafikult on näha, et kuudevaheline varieeruvus kasvab ühes majutatute arvu kasvuga. Selline käitumine on vastuolus mitmete aegrea mudelitega, mistõttu võib olla vajalik andmeid ennem mudelite rakendamist teisendada. Üheks populaarseks teisenduseks on logaritmimine, kuigi kasutusel on ka mitmeid teisi teisendusfunktsioone. Lisage andmestikku logaritmitud turistide arvu tulp. Selleks tuleb jooksutada andmesammu (`data step`) nii, et sisendandmeteks on juba sisseloetud andmed (näidatud käsureaga `set andmestiku_nimi;`) ning omistada uuele veerunimele logaritm (funktsioon `log`)turistide arvust. Kas logaritmimine muutis varieeruvuse amplituudi ajas püsivamaks?

3. Lihtsa libiseva keskmise arvutamine.

   - Vaatleme näitena kolmekuuse lihtsa sümmeetrilise libiseva keskmise lisamist töötuse määra andmetele. Üheks võimaluseks on jaotada tegevus mitmeks etapiks:
   - Tekitame ajutise andmestiku (st anname talle ilma punktita nime), kus on kuupäev (NB! kasutame sama nime, mis lähteandmestikus!) ja kolme vaatluse sümmeetriline lihtne keskmine. Siin tuleb arvestada, et SAS-i andmesamm töötab andmetega ridade kaupa ning eelmiste ridade andmete kättesaamiseks peab kasutama erikäske. Tuleviku väärtusi ei olegi võimalik kätte saada, kuid see ei ole probleem, lihtsalt peame arvestama, et kui arvutame keskmise praegusest, eelmisest ja üleeelmisest väärtusest, siis sümmeetrilise keskmise korral tuleb vastav kuupäev võtta ka eelmiselt realt. Varasemate ridade andmete kättesaamiseks on funktsioonid kujul `LAGx()`, kus `x` näitab, kui mitu sammu varasemat väärtust saada tahetakse Seega kolmepäevase keskmise arvutamiseks saame kasutada rida kujul
     ```
     keskmine=(rate+lag1(rate)+lag2(rate))/3;
     ```
     Õige kuupäeva kirjutamiseks tuleb muutuja `date` ümber defineerida käsuga `date=lag1(date);`
     Selleks, et tulemusandmestikus ei oleks üleliigseid andmeid, võib loetleda väljundisse kirjutatavad muutujad kujul `KEEP nimi1 nimi2 jne;` või siis loetleda sinna mitteminevad muutujad käsuga `DROP` (kasutada ainult üht nendest!).
   - Paneme tekitatud keskmised väärtused esialgsele andmestikule juurde. Selleks saab andmesammu sees kasutada SAS käske kujul
     ```
     merge andemstik1 andmestik2;
     by date;
     ```
   - Joonistame ühel graafikul originaalsed andmed ja keskmistatud andmed. Selleks saab kasutada samuti gplot protseduuri, defineerides seal symbol1 ja symbol2 nii, et neis oleks erinev color=värv osa ning plot käsu taga tuleb siis anda kaks muutujatepaari ning nende taga `/ overlay;`.

4. Lisage 9 kvartali minevikuandmeid (vaadeldavale hetkele vastavat ning kaheksat minevikust) kasutav libisev keskmine kinnisvara tehingute andmetele.

5. Leiame eksponentsiaalselt silutud aegrea aktsiahindade korral, kasutades silumisparameetrit $\alpha = 0.3$. Selleks läheb vaja teada, et aegrea $(z_t)_{t=1,\dots,n}$ eksponentsiaalselt silutud versiooni y_t leidmisel saab kasutada valemit kujul
$$y_t = \alpha z_t + (1-\alpha) y_{t-1}$$
sobivalt defineeritud y_0 korral. Tavaliselt võetakse $y_0 = z_1$ (mis tagab ka $y_1 = z_1$; seetõttu sageli y_0 andmestikku ei lisata). SAS-is realiseerimiselt läheb vaja juba eelmises praktikumis saadud teadmist, et

juba varem arvutatud (või defineeritud) muutuja väärtust saab meeles pidada käsuga

```
retain muutujanimi algväärtus;
```

ning y algväärtuse käsitsi sissekirjutamise asemel võime käsitleda erinevalt esimest sisseloetavat rida ning hilisemaid ridu konstruktsiooniga

```
if _n_=1 then do;
    esimese rea tegevused;
end;
else do;
    teiste ridadega tehtavad käsud;
end;
```

Hilisemates praktikumides vaatleme eksponentsiaalse silumise kasutusvõimalusi teiste SAS protseduuride vahendusel.

6. Eelpool vaadeldud teisendusi (ja palju muud) saab teostada ka SAS/ETS protseduuri EXPAND vahendusel. Keskmiste arvutamiseks kasutatav kuju on järgmine:

```
proc expand data=sisendandmestik out=väljundandmestik;
convert sisendmuutuja=väljundmuutuja/method=none transform=(nimi params);
id date;
run;
```

Suvandi `transform` taha sobivateks teisenduste nimedeks on näiteks `movave` (libisev keskmine minevikuandmetest), `cmovave` (tsentreeritud libisev keskmine, st võimalusel võrdne arv vaatlusi tulevikust ja minevikust) ja `ewma` (eksponentsiaalne keskmine). Esimese kahe järele võib kirjutada kasutatavate vaatluste arvu (sel juhul arvutatakse lihtne libisev keskmine) või siis kaalud kujul `(w0 w1 ... wn)`, kus kaalud tuleb anda ajas tagurpidi järjekorras (st kõige uuema vaatluse kaal kõige ennem). Hea on teada, et kaalude kirjapanekul võib nende normeerimise (st garanteerimise, et summa oleks 1) jätta SAS hooleks, kuna kaalud jagatakse vajadusel läbi nende summaga. Eksponentsiaalse keskmistamise korral tuleb parameetriks kirjutada kaalu α väärtus. Lisage USA töötuse määra andmestikule eksponentsiaalselt keskmistatud rida kaaluga $\alpha = 0.4$ ning perioodiga 12 sobiv modifitseeritud lihtne sümmeetriline keskmine (esimene ja viimane kaal poole väiksemad ülejäänutest). Kujutage esialgne rida koos tekitatud keskmistatud ridadega ühel graafikul.

Tegevused R-is:

1. loeme R keskkonda turistide andmestiku.

2. Libisevat keskmist saab R keskkonnas arvutada `filter` käsuga. Kui on tegemist lõpliku arvu nullist erinevate kordajatega (nt lihtsa libiseva keskmisega), siis tuleb kaalud anda ajas tagurpidi järjekorras (st esimesena kõige hilisema vaatluse kaal, siis eelmise jne) suvandiga `filter=kaalude_vektor`. Üheks võimaluseks kaalude vektor tekitada on kaalud kirja panna funktsiooni `c()` argumentidena, näiteks `filter=c(1/3,1/3,1/3)`. Ainult minevikuväärtusi kasutava keskmistamise korral tuleb lisaks kasutada veel suvandit `sides=1`. Leidke 11-päevane lihtne sümmeetriline libisev keskmine cisco aktsia sulgemishindade aegreast.

3. Mitme aegrea kujutamine ühel joonisel: kõigepealt joonistada ühe aegrea graafik plot käsuga, kasutades lisaks suvandit `col="värvi_nimi"` ning seejärel lisada teiste aegridade graafikud käsuga `lines(rida,col="värv")`. Kujutada ühel graafikul keskmistatud aktsiahinnad ja esialgsed aktsiahinnad.

4. Arvutame eksponentsiaalselt silutud kinnisvara müügiandmed juhtudel $\alpha = 0.1, 0.3, 0.6, 0.9$ ja kujutame silutud aegridu ühel joonisel. Eksponentsiaalseks keskmistamiseks on kasulik käsk
   `y=filter(x,filter=w,method="recursive",init=algväärtused)`.
   Selle käsuga arvutatakse tulemusrea y väärtused valemiga

$$y_i = x_i + w_1 y_{i-1} + \ldots + w_p y_{i-p},$$

kus p on kaalude vektori komponentide arv ja algväärtused annavad vektori y esimesed p väärtust. Eksponentsiaalse keskmistamise saamiseks peab olema ainult üks kaal $w = (1-\alpha)$, algväärtuseks võib võtta keskmistatava rea esimese väärtuse ning `x` asemel eelpooltoodud käsus `filter` tuleb anda $\alpha \cdot esialgne\_rida$.

```
####harjutus 1
#loeme andmed sisse
#1)andmete sisselugemine. Andmed hakkavad esimesest reast, puuduvad andmed tähistatud kahe
punktiga
andmed=read.table("h:/aegread_praktikum/turistid.csv",na.strings="..",skip=0)
#2) andmed on 12 kaupa ridades, esimeses kahes veerus on ebaoluline info. Paneme nad õigesti
ritta
rida=c(t(andmed[3:14]))
#3) eemaldame puuduvad andmed. Nii saab, kuna ainsad puuduvad andmed on lõpus
rida=rida[!is.na(rida)]
#4) ümber pöörata ei ole vaja
#5) Teeme aegrea
turistid=ts(rida,start=1996,frequency=12)
#vaatame graafikut
plot(turistid)

###harjutus 2
#kõigepealt tuleb sisse lugeda cisco andmed
#Võib jooksutada esimese praktikumi vastavaid käske
#kui vastavad read said salvestatud mingisse Rdata faili, siis võib need lihtsalt
#sisse lugeda
load("h:/aegread_praktikum/prax1andmed.rdata")
#vaikimisi kasutab filter käsk sümmeetrilist keskmist
lslk11=filter(cisco,filter=c(1,1,1,1,1,1,1,1,1,1,1)/11)

#######harjutus 3

#joonistame graafikud
plot(cisco)
lines(lslk11,col="red")

###harjutus 4
alfa=0.1
ewma0p1=filter(alfa*kv,filter=1-alfa,method="recursive",init=kv[1])
alfa=0.3
ewma0p3=filter(alfa*kv,filter=1-alfa,method="recursive",init=kv[1])
alfa=0.6
ewma0p6=filter(alfa*kv,filter=1-alfa,method="recursive",init=kv[1])
alfa=0.9
ewma0p9=filter(alfa*kv,filter=1-alfa,method="recursive",init=kv[1])
plot(kv,col="black")
lines(ewma0p1,col="red")
lines(ewma0p3,col="blue")
lines(ewma0p6,col="green")
lines(ewma0p9,col="brown")
#mida väiksem on alfa, seda "siledam" on tulemus, kuid seda kauem on muudatustele
reageerimise aeg
```

```
*praktikumi nr. 2 skript SAS-is;
libname aegread "h:/aegread_praktikum";*määrab nime andmekaustale
ning selle kausta tegeliku asukoha;
*******harjutus 1*****************;
data aegread.turistid;
infile "h:/aegread_praktikum/turistid.csv" firstobs=1
delimiter="09"x dsd;
keep date turistid;
retain abi 0;
input tmp1 $ tmp2 $ @;
do i=1 to 12;
      input turistid @;
      date=intnx("month","1jan96"d,abi);
      output;
      abi=abi+1;
end;
format date mmyyc5.;
run;
*eemaldame puuduvad kirjed;
data aegread.turistid;
set aegread.turistid;
if turistid^=.;
run;
**************harjutus 2********************;
*kontrollime visuaalselt;
proc gplot data=aegread.turistid;
symbol i=spline v=dot h=0.3;
plot turistid*date;
run;
*lisame logaritmitud andmetulba;
data aegread.turistid;
set aegread.turistid;
logturiste=log(turistid);
run;
*vaatleme tulemust;
proc gplot data=aegread.turistid;
symbol i=spline v=dot h=0.3;
plot logturiste*date;
run;
*pärast logaritmimist on varieeruvus ühtlasem, nii et tasub kaaluda
selle teisenduse kasutamist;

************************harjutus 3******************;
data ajut;
set aegread.unemployment;
drop rate;
keskmine=(rate+lag1(rate)+lag2(rate))/3;
date=lag1(date);
if keskmine^=.;*eemaldame puuduvate väärtustega kirjed;
output;
run;
*ühendame keskmistega ajutise faili ja esialgse faili kuupäevade
abil;
data aegread.unemployment;
merge ajut aegread.unemployment;
by date;
run;
```

```
*tulemuste joonistamine;
proc gplot data=aegread.unemployment;
symbol1 i=spline v=dot h=0.1 color="blue";
symbol2 i=spline color="green";
plot rate*date keskmine*date /overlay;
where date>"01jan2001"d;*kogu graafik liiga tihe, vaatame osa;
run;

**********harjutus 4**********;
*kõigepealt ajutine andmestik;
data ajut;
set aegread.kv;
drop tehingud;
keskmine=(tehingud+lag1(tehingud)+lag2(tehingud)+lag3(tehingud)+lag4
(tehingud)+lag5(tehingud)+lag6(tehingud)+lag7(tehingud)+lag8
(tehingud))/9;
*date=lag4(date); *kui oleks sümmeetriline keskmine;
if keskmine^=.;*eemaldame puuduvate väärtustega kirjed;
output;
run;
*lisame tulemuse kinnisvara andmetele;
data aegread.kv;
merge ajut aegread.kv;
by date;
run;
*vaatleme huvi pärast tulemust;
proc gplot data=aegread.kv;
symbol1 i=spline v=dot h=0.1 color="blue";
symbol2 i=spline color="green";
plot tehingud*date keskmine*date /overlay;
run;

**********harjutus 5 **********;
*kuna ei ole vaja kuupäevi muuta, ei pea kasutama ajutist
andmestikku, kuigi soovitatav oleks seda ikka teha;
data aegread.cisco;
set aegread.cisco;
retain keskmine alfa 0.3; *alfa väärtus on kogu aeg 0.3, keskmist
tuleb meelde jätta järgmise rea arvutamiseks;
drop alfa;
if _n_=1 then do;*esimese rea jaoks teistsugune valem;
    keskmine=close;
    end;
else do;
    keskmine=alfa*close+(1-alfa)*keskmine;
    end;
output;
run;
*vaatame tulemust;
proc gplot data=aegread.cisco;
symbol1 i=spline v=dot h=0.1 color="blue";
symbol2 i=spline color="green";
plot close*date keskmine*date /overlay;
run;

**********harjutus 6**********;
*mitu convert rida võib olla järjest;
```

```
*perioodile 12 vastab modifitseeritud lihtne sümmeetriline keskmine,
kus kasutatakse 13 vaatlust, esimene ja viimane poole väiksema
kaaluga;
proc expand data=aegread.unemployment out=aegread.unemployment;
convert rate=ewma0p4/method=none transform=(ewma 0.4);
convert rate=modlslk/method=none transform=(cmovave (1 2 2 2 2 2 2 2
2 2 2 2 1));
id date;
run;
* vaatame tulemusi;
proc gplot data=aegread.unemployment;
symbol1 i=spline v=dot h=0.1 color="blue";
symbol2 i=spline color="green";
symbol3 i=spline color="brown";
plot rate*date modlslk*date ewma0p4*date /overlay;
where date>"01jan2001"d;*kogu graafik liiga tihe, vaatame osa;
run;
```

Aegridade analüüs
Praktikum nr. 3, 2012

Käsitletavad teemad: Globaalsete trendide leidmine ja eemaldamine. Aegridade osadeks lahutamine.

Tutvume lihtsamate globaalse trendi leidmise võtetega ning aegrea osadeks lahutamise meetoditega.

Olgu meil antud vaatlused $z_1, z_2, \ldots, z_n$. Edasises vaatleme nii SAS kui R tarkvara abil parameetrite suhtes lineaarse trendi leidmist kujul

$$\text{trend}(t) = c_0 + c_1 \cdot f_1(t) + c_2 \cdot f_2(t) + \ldots + c_m \cdot f_m(t),$$

kus $f_i,\ i = 1, \ldots, m$ on mingid fikseeritud funktsioonid või teiste aegridade väärtused. Samuti vaatleme mõningaid parameetrite suhtes mittelineaarselt esituvate trendikõverate leidmise võimalusi.

Tegevused R tarkvara abil. Lineaarse trendi parameetrite leidmiseks on R keskkonnas mitmeid võimalusi. Kui meie eesmärgiks on ainult vastava trendikõvera sobitamine, siis on kõige õigemaks käsuks `lm()` (*linear models*). Samas on aegridade analüüsimisel vastava trendi leidmine ainult üks paljudest tegevustest, mida on aktsepteeritava mudeli saamiseks vaja teha ning seetõttu on mõistlikum ka vaadeldaval kujul trendi leida selliste vahendite abil, mis lubavad lihtsalt mudeli hilisemat täiendamist. Seetõttu vaatleme parameetrite suhtes lineaarse trendi leidmist käsuga `arima()`, mis on ka täiendavate aegridade mudelite sobitamisel kasutusel. Vaadeldaval kujul oleva trendi leidmiseks tuleb tekitada maatriks, mille veergudes on funktsioonide $f_i, i = 1, \ldots, m$ väärtused ajamomentidel $t = 1, 2, \ldots, n$ näiteks käsuga `regressorid=cbind(F1,...,Fm)`, kus vektorid `Fi` sisaldavad funktsioonid f_i väärtusi ning kasutada seejärel käsku kujul

```
glob_trend=arima(z,xreg=regressorid)
```

Siin kolm punkti tähistavad kõigi liikmete väljakirjutamist ning nimetuse `glob_trend` asemel võib kasutada suvalist enda antud nime. Aegreast trendi lahutamisel saadava rea saab kätte kujul `residuals(glob_trend)` (või kujul `glob_trend$residuals`), olemasolevatele väärtustele vastavaid trendi väärtuseid saab arvutada käsuga `z-residuals(glob_trend)`, tulevikuväärtuste arvutamisel on käsu kujuks

```
predict(glob_trend,n.ahead=mitu,newxreg=cbind(uued_F1_vaartused,...,uued_Fm_vaartused))
```

Otsitavate parameetrite suhtes mittelineaarse trendi sobitamisel saab aga kasutada käsku `nls()`. Käsu kasutamiseks lugege vastavat R abifaili (trükkides käsureale `?nls`).

1. Sobitage lineaarne ja ruuttrend tarbijahinna indeksi andmetele. Selleks leidke selle rea vaatluste arv `n` käsuga `length()`, defineerige vektorid `F1=1:n` ja `F2=F1*F1` ning seejärel leidke trendile vastavad mudelid.

2. Tehke lineaarse trendi ennustustele vastav aegrida ning kujutage see koos tarbijahinna indeksiga ühel joonisel.

3. Sobitage tarbijahinna indeksile mittelineaarne trend kujul $c_1 \cdot e^{c_2 t}$. Lisage ka see eelnevale graafikule.

4. Vaatleme ka aegrea osadeks lahutamist. Klassikaline meetod on R tarkvaras realiseeritud käsuga `\decompose()`. Leidke kinnisvara andmestiku nii multiplikatiivne kui ka aditiivne osadeks lahutus.

5. Katsetage osadeks lahutamisel ka käsu `stl()` võimalusi. Selleks lugege õpetust ning leidke nii täielikult perioodilise sesoonse osaga lahutus kui ka muutuva sesoonse osaga lahtus kinnisvara andmete jaoks.

Praktikumi tegevused SAS-is. Nagu ka R tarkvaras, on SAS puhul võimalik kasutada mitmeid protseduure lineaarsete ja mittelineaarsete (leitavate parameetrite suhtes) trendikõverate leidmiseks. Vaatleme siiski ainult lineaarse trendi leidmist protseduuriga arima, mida me kasutame ka keerulisemate mudelite sobitamiseks. Kui otsitavas kujus olevad funktsioonid f_i on mingid deterministlikud funktsioonid ja mitte mingi juhusliku protsessi väärtused, siis on võimalik tegevuste järjekord selline:

- Tekitame ajutise andmestiku, kus on lisaks meid huvitavale aegreale trendifunktsioonide väärtused.
- Kasutame protseduuri arima järgmiselt:

```
proc arima data=sisendandmed;
identify var=aegrida crosscorr=(F1 F2 ... Fm);
estimate input=(F1 F2 ... Fm);
forecast id=date lead=0 out=tulemustabel;
run;
quit;
```

  Siin peab var= taha kirjutama tulba nime, kus on analüüsitav aegrida ning F1,F2 jne asemel kirjutama tulpade nimed, kus on funktsioonide $f_i,\ i = 1, \ldots, m$ väärtused. Pärast sellise käsu jooksutamist on tabelis nimega tulemustabel esialgse rea väärtused, kuupäevad (eeldus on, et nimega date on sisendandmetes olemas), tulp nimega forecast leitud trendifunktsiooni väärtustega, tulp nimega residual, kus on trendi mahalahutamisel jääv osa ja veel mõned tulbad. Seega on lihtne teha huvipakkuvaid jooniseid. Hinnatud parameetrite väärtuseid saab vaadata tulemuste (Results) aknas

Harjutusülesanded:

1. Leidke lineaarne trend tarbijahinna indeksi andmetele ning tehke joonis, kus on trendi ja indeksi väärtused.
2. Korrake sama ruuttrendiga. Võrrelge leitud kordajaid R poolt leitutega. Kas on erinevusi?
3. Aegrea osadeks lahutamiseks vaatleme SAS-is ainult klassikalist dekompositsiooni. Seda on lihtne teha sarnaselt keskmiste leidmisele protseduur expand käsuga, kasutades seal `convert` käskudes suvandi `transform=` taga paare kujul `(osa periood)`, kus võimalikeks aegrea osadeks on `cd_tc` (trendi osa), `cd_s` või `cda_s` (sesoonne osa vastavalt multiplikatiivses ja aditiivses esituses, `cd_i` või `cda_i` (ebaregulaarne osa multiplikatiivses ja aditiivses esituses). Tekitage andmestik, kus oleks kinnisvara tehingu andmed koos selle osadega multiplikatiivse ja aditiivse lahutuse korral.

```
setwd("h:/aegread_praktikum/")
#loen sisse varem tekitatud andmed
load("h:/aegread_praktikum/prax1andmed.rdata")
#tarbijahinna indeks on salvestatud aegreana nimega indeks
#harjutus 1
n=length(indeks)
lin=1:n
lin_trend=arima(indeks,xreg=lin)
#vaatame leitud kordajaid
lin_trend
#trendi väärtused
trend1=indeks-lin_trend$residuals
ruut=lin*lin
regressorid=cbind(lin,ruut)
ruut_trend=arima(indeks,xreg=regressorid)
trend2=indeks-ruut_trend$residuals
#harjutus2
plot(indeks)
lines(trend1,col="red")
lines(trend2,col="blue")
#harjutus 3
t=lin
mittelin_trend=nls(indeks~c1*exp(c2*t),start=list(c1=100,c2=0.01))
trend3=indeks-residuals(mittelin_trend)
lines(trend3,col="green")
#harjutus 4
lahutus1=decompose(kv)
plot(lahutus1)
lahutus2=decompose(kv,type="multiplicative")
plot(lahutus2)
#harjutus 5
#perioodilise sesoonse osaga, muud valikud vaikeväärtused
stl_lahutus1=stl(kv,s.window="periodic")
plot(stl_lahutus1)
#muutuva perioodilise osaga. Suurem s.window väärtus tähendab aeglasemat muutumist
stl_lahutus2=stl(kv,s.window=9,s.degree=1)
plot(stl_lahutus2)
```

```
*praktikumi nr. 3 skript SAS-is;
libname aegread "h:/aegread_praktikum";*määrab nime andmekaustale
ning selle kausta tegeliku asukoha;
*******harjutus 1*****************;
*tekitame andmestiku ajut, kus on kuupäev, tarbijahinna indeks ning
lineaarse ja ruutfunktsiooni väärtused;
data ajut;
set aegread.indeks;
lin=_n_;
ruut=_n_*_n_;
output;
run;
proc arima data=ajut;
identify var=indeks crosscorr=lin;
estimate input=lin;
forecast lead=0 id=date out=tulemus1;
quit;
proc gplot data=tulemus1;
symbol1 i=spline color="blue";
symbol2 i=spline color="green";
plot indeks*date forecast*date/overlay;
run;
proc gplot data=tulemus1;
symbol1 i=spline color="blue";
plot residual*date;
run;
proc arima data=ajut;
identify var=indeks crosscorr=(lin ruut);
estimate input=(lin ruut);
forecast lead=0 id=date out=tulemus2;
quit;
proc gplot data=tulemus1;
symbol1 i=spline color="blue";
symbol2 i=spline color="green";
plot indeks*date forecast*date/overlay;
run;
proc expand data=aegread.kv out=lahutus;
convert tehingud=sm/method=none transform=(cd_s 4);*sesoonne
komponent, multiplikatiivne lahutus;
convert tehingud=trend/method=none transform=(cd_tc 4);*trend;
convert tehingud=irregular/method=none transform=(cd_i 4);
*ebaregulaarne osa, multiplikatiivne lahutus;
convert tehingud=sa/method=none transform=(cda_s 4);*sesoonne
komponent, aditiivne lahutus;
convert tehingud=irregulara/method=none transform=(cda_i 4);
*ebaregulaarne osa, multiplikatiivne lahutus;
id date;
run;
```

Aegridade analüüs

Praktikum nr. 4, 2012

Käsitletavad teemad: Eksponentsiaalse silumise, Holti meetodi ja Holt-Wintersi meetodi rakendamine aegrea prognoosimiseks; prognoosivigade mõõdikute arvutamine.

Silumisvõtteid kasutatakse sageli ka prognoosimisel. Tähistame kujul

$$\hat{z}_{t+p|t}$$

aegrea väärtuse prognoosi ajamomendi $t+p$ jaoks, mis on arvutatud ainult ajamomendil t teadaolevate andmete põhjal. Kui ennustame aegrea järgmist väärtust (st $p = 1$), siis kasutame ka lihtustatud tähist $\hat{z}_{t+1}$.

Kui eeldame, et aegreal ei ole trendi ja perioodilist komponenti, siis on üheks loomulikuks mooduseks järgmise ajahetke väärtuse arvutamisel olemasolevate andmete eksponentsiaalne keskmine. Arvutusi on sel juhul hea organiseerida järgmiselt: defineerime $\hat{z}_1 = z_1$ ning arvutame seejärel

$$\hat{z}_{t+1} = \alpha z_t + (1-\alpha)\hat{z}_t,\ t = 1, \ldots, n.$$

Parameetri α võib valida ise või määrata sobivaim väärtus olemasolevate andmete põhjal nii, et mineviku jaoks leitud ühesammulised ennustused oleks võimalikult lähedased olemasolevate andmetega.

Ajas muutuva trendiga, kuid ilma perioodilise komponendita aegrea väärtuste prognoosimiseks vaatleme Holti meetodit, mille korral nii trendikõvera hetkeväärtus kui ka tõus leitakse andmetest eksponentsiaalse keskmistamise teel. Valemid on järgmised:

$$\hat{z}_{t+p|t} = a_t + b_t\, p,$$

kus trendikõvera hetkeväärtuse a_t arvustamisel kasutatakse z_t väärtust ning eelnevate andmete põhjal tehtud prognoosi:

$$a_t = \alpha\, z_t + (1-\alpha)\hat{z}_t = \alpha\, z_t + (1-\alpha)(a_{t-1} + b_{t-1})$$

ning trendikõvera tõusu b_t arvutamisel kasutatakse eelmise väärtuse ning a muutuse keskmist:

$$b_t = \beta\,(a_t - a_{t-1}) + (1-\beta)b_{t-1}.$$

Meetodi kasutamiseks tuleb valida või andmete põhjal hinnata väärtused a_1, b_1, α, β. Sageli valitakse $a_1 = z_1,\ b_1 = 0$ või siis kasutada nendeks teatud arvu esimeste z väärtuste jaoks leitud lineaarse regressioonikõvera vastavaid väärtusi.

Perioodilise komponendi olemasolul tuleb kõigepealt otsustada, kas trend ja perioodiline komponent on omavahel korrutatud või liidetud. Korrutamise eeldus sobib positiivsete väärtustega ridadele juhul, kui perioodiliste võngete ulatus kasvab koos väärtuste kasvuga. Käesolevas praktikumis vaatleme sesoonsete ridade prognoosimist eksponentsiaalsel keskmistamisel põhineva Holt-Wintersi aditiivse ja multiplikatiivse meetodiga. Vastavad valemid on toodud loengukonspektis, kuid intuitiivselt on tähtis aru saada, et leitud parameetritest α kirjeldab seda, kui tugevalt me arvestame viimast väärtust jooksvale ajahetkele vastava trendikõvera väärtuse leidmisel ($\alpha = 1$ puhul loeme selleks viimast vaatlust, $\alpha = 0$ korral on aga trendikõver ajas muutumatu sirge), β kirjeldab lineaarse trendi tõusu muutumise kiirust juhul, kui leitud trendikõvera muutus üle viimase perioodi erineb prognoositud muutuses (ehk eelmisel ajamomendil prognoositud trendikõvera tõusust) ning γ kirjeldab seda, kui kiiresti muutub sesoonne komponent ($\gamma = 1$ puhul arvestame järgmise perioodi vaatluse sesoonse osa puhul ainult käesolevat aastat, $\gamma = 0$ puhul aga sesoonne komponent on ajas muutumatu).

Praktikumi tegevused SAS-is (eeldusel, majutatud turistide andmestik on SAS-i sisse loetud):

1. Kõigepealt tutvume aegridade ennustamise graafilise kasutajaliidesega. Selleks:
   (a) valige menüüst solutions->analysis->time series forecasting system
   (b) Seejärel avanevas aknas valige andmestikuks turistide andmestik ning klikake nupule „Develop Models“ ning aegrea muutujaks võtke majutatud turistide arvu sisaldav tunnus.
   (c) Klikake aknasse, kus esialgu on „no models“ ning valige sealt „fit smoothing model“
   (d) Looge eksponentsiaalse silumise, Holti, Holt-Wintersi aditiivsele ning multiplikatiivsele meetodile vastavad mudelid, võrrelge saadud parameetreid ning headusemõõdikuid R keskkonnas saadud tulemustega.
2. Graafiline keskkond sobib juhul, kui on harva vaja mõnda üksikut aegrida prognoosida ja selles keskkonnas realiseeritud valikud võimaldavad soovitud tegevusi sooritada. SAS-i täielikku funktsionaalsust on võimalik aga kasutada ainult skriptide abil. Holt-Wintersi meetodi abil aegrea prognoosimist on võimalik teha protseduuri `forecast` vahendusel (oleks võimalik ka protseduuri `hpf` vahendusel, mis sisaldab ka kordajate optimaalse valiku võimalusi, kuid vastavat SAS moodulit ei ole meil võimalik kasutada). Näide kasutamisest:

```
proc forecast data=aegread.turistid method=winters
seasons=month interval=month weight=(0.24 0.01 0.99)
out=tulem outactual outresid outest=statistikud;
var turiste;
id date;
run;
```

Mõned seletused: meetodiks võib olla expo (eksponentsiaalne silumine), winters või addwinters; viimased kaks on vastavalt multiplikatiivne ja aditiivne Holt-Wintersi meetod. Kui jätta parameeter `seasons=` andmata, on tulemuseks Holti meetod. Selleks, et tulemusi oleks võimalik uurida, on mõistlik anda väljundandmestike nimed käskudega `out=` ja `outest=`, viimases on kõikvõimalike statistikute väärtused. Parameeter `var` määrab, millise sisendandmestiku muutuja väärtusi ennustatakse. Täpsemat abi tuleb lugeda SAS manuaalist.
Ülesanne: kasutage protseduuri forecast selleks, et ennustada turistide andmed järgnevaks aastaks eksponentsiaalse keskmistamise, Holti meetodi ja Holt-Wintersi meetodi abil.

3. (*) Kuigi optimaalsete parameetrite leidmist ei ole meetodis forecast realiseeritud, ei ole seda kuigi raske ise protseduuri `nlin` abiga teha. Lugege protseduuri `nlin` õpetust ning katsuge selle abil leida SAS-is optimaalsed parameetrid Holti meetodi jaoks.

Tegevused R-is:

1. Käesoleva praktikumi ennustusmeetodite rakendamiseks on R-is käsk `HoltWinters`. Vaikimisi kasutab R aditiivset meetodit ja leiab parameetritest α, β ja γ need, mis ei ole käsus otseselt ette antud, ennustusvigade ruutude summat minimiseerides. Näiteid kasutamisvõimalustest olemasoleva aegrea `x` korral:
   `m1=HoltWinters(x)`-sobitab aditiivse meetodi reale `x`,
   `m1=HoltWinters(x,seasonal="multiplicative")`-sobitab multiplikatiivse meetodi reale `x`,
   `m1=HoltWinters(x,gamma=FALSE)`-sobitab Holti meetodi reale `x`,
   `m1=HoltWinters(x,beta=FALSE,gamma=FALSE)`-eksponentsiaalse keskmistamise meetod.
   Soovi korral võib optimiseerimise alustamiseks ise sobivad parameetrite väärtused ette anda. Selleks tuleb käsu `HoltWinters` argumendina edastada
   `optim.start=c(alpha=a0,beta=b0,gamma=g0)`,
   kus $a0, b0$ ja $g0$ on mingid konkreetsed numbrid nulli ja ühe vahel. Detailsemat infot käsu võimaluste kohta saab `help` käsuga. Käsu tulemuse põhjal saab tekitada graafiku `plot` käsuga, leida ennustusvigu, ennustada tulevikuväärtusi `predict` käsuta jpm. Tekitage joonised, kus on majutatud turistide andmestiku esialgse aegrea ning eksponentsiaalse silumise, Holti meetodi ja Holt-Wintersi aditiivse ja multiplikatiivse meetodi poolt leitud ühesammuliste ennustuste graafikud. Siin tuleb kasuks teadmine, et käsu `HoltWinters` tulemuseks on andmestruktuur (objekt), mille komponendis nimega x on esialgne aegrida ning komponendis nimega `fitted` on maatriks, mille veerus nimega xhat on ennustused. Seega ühesammulised ennustused on kättesaadavad kujul `m1$fitted[,"xhat"]`, kus `m1` on käsu `HoltWinters` tulemus.

2. Meetodi headusest aimusaamiseks ja erinevate meetodite omavaheliseks võrdlemiseks arvutatakse mitmesuguseid ennustusvigade mõõdikuid. Arvutage eelmises punktis mainitud meetodite korral keskmine absoluutne viga MAD, ruutjuur vigade ruutude keskmisest RMSE ja keskmine absoluutne suhteline viga MAPE (vt. loengumaterjale). Selleks on hea teada, et pärast mudeli sobitamist kujul `mudel=HoltWinters(...)` saab esialgse rea väärtusi kätte kujul `mudel$x` ning ühesammuliste prognooside vigasid kujul `residuals(mudel)`. Samuti saab keskmiste arvutamisel kasutada R käsku `mean`, mis leiab etteantud vektori aritmeetilise keskmise.

3. Leidke parima RMSE-ga meetodi korral ennustus järgnevaks neljaks kuuks ning ka ennustusvigade hinnangud. Vajadusel uurige selleks käsu `pridict.HoltWinters` abimaterjale.

```
####harjutus 1
#loeme andmed sisse
#1)andmete sisselugemine. Andmed hakkavad esimesest reast
andmed=read.table("h:/aegread_praktikum/turistid.csv",na.strings="..",skip=0)
#2) andmed on 12 kaupa ridades, esimeses kahes veerus on ebaoluline info. Paneme nad õigesti
ritta
rida=c(t(andmed[3:14]))
#3) eemaldame puuduvad andmed. Nii saab, kuna ainsad puuduvad andmed on lõpus
rida=rida[!is.na(rida)]
#4) ümber pöörata ei ole vaja
#5) Teeme aegrea
turistid=ts(rida,start=1996,frequency=12)
#joonistame graafiku
plot(turistid)

#aditiivne meetod
m1=HoltWinters(turistid)
plot(m1) #kuidas järgmise väärtuse prognoosid on kooskõlas tegelike väärtustega
predict(m1,4) #ennustame 4 järgmist vaatlust
m1 #info leitud parameetrite kohta
#multiplikatiivne meetod
m2=HoltWinters(turistid, seasonal="multiplicative")
plot(m2)
m2
#Holti meetod (selle rea puhul tegelikult ei ole mõistlik kasutada)
m3=HoltWinters(turistid,gamma=F)
plot(m3)
#eksponentsiaalne keskmine
m4=HoltWinters(turistid,gamma=F,beta=F)
plot(m4)
#sobitamise käigus leitakse tegelikult suur hulk andmeid, mis lisatakse objektile m1
summary(m1) #info m1 koostisosadest, mida saab kätte kujul m1$nimi
m1$SSE #ühesammuliste prognoosivigade ruutude summa

#mõõdikute arvutamiseks tuleb leida prognoosivead ja seejärel teha vastavaid arvutusi
#edaspidise lihtsustamiseks loome funktsiooni
mõõdikud=function(mudel){ #mudel peab olema loodud käsuga HoltWinters
    vead=residuals(mudel)  #mudel$x on esialgse aegrea andmed
    MAD=mean(abs(vead)) #keskmine absoluutviga, SASis MAE
    MSE=mean(vead**2) #keskmine ruutviga
    RMSE=sqrt(MSE)
    MAPE=mean(abs(vead/mudel$x)) #keskmine suhteline viga, sageli korrutatakse 100ga, et
    saada protsentides
    return(list(MAD=MAD,MSE=MSE,RMSE=RMSE,MAPE=MAPE))
}
mõõdikud(m1)
mõõdikud(m2)
mõõdikud(m3)
mõõdikud(m4)
#parim on m2 ehk multiplikatiivne HoltWinters
predict(m2,4) #prognoosid nelja järgmise kuu jaoks
```

```
*praktikumi nr. 4 skript SAS-is;
libname aegread "h:/aegread_praktikum/";
/*ennustame turistide aegrea väärtusi HoltWintersi meetodi abil.
proc forecast ei oska kaalusid sobitada (kui ette ei anna, siis
võetakse
mingid fikseeritud arvud, andmete põhjal optimiseerimist ei toimu*/
*multiplikatiivne Holt-Winters;
proc forecast data=aegread.turistid method=winters
seasons=month interval=month weight=(0.26 0.001 0.96)
out=tulem outactual outresid outest=statistikud; /*kirjutama
ajutisse andmestikku nimega tulem lisaks prognoosidele ka esialgse
aegrea väärtused ning prognoosivead*/
var turistid;
id date;
run;
*ka tekitatud väljundit saab mõistlikult joonistada järgneva
konstruktsiooni abil;
proc gplot data=tulem;
symbol1 i=spline color="red";
symbol2 i=spline color="blue";
plot turistid*date=_type_; *tulbas nimega turistid on mitu rida, mis
on eristatavad tulbas nimega _type_ oleva info baasil;
where _type_ ^= "RESIDUAL"; *prognoosivigu ei ole mõtet samal
graafikul joonistada;
run;
*aditiivne Holt-Winters;
proc forecast data=aegread.turistid method=addwinters
seasons=month interval=month weight=(0.26 0.001 0.96)
out=tulem outactual outresid outest=statistikud; /*kirjutama
ajutisse andmestikku nimega tulem lisaks prognoosidele ka esialgse
aegrea väärtused ning prognoosivead*/
var turistid;
id date;
run;
*Holti meetod, seasons= suvand tuleb maha võtta;
proc forecast data=aegread.turistid method=winters
interval=month weight=(0.9 0.7)
out=tulem outactual outresid outest=statistikud; /*kirjutama
ajutisse andmestikku nimega tulem lisaks prognoosidele ka esialgse
aegrea väärtused ning prognoosivead*/
var turistid;
id date;
run;
*eksponentsiaalne keskmistamine;
proc forecast data=aegread.turistid method=expo
interval=month weight=(0.9)
out=tulem outactual outresid outest=statistikud; /*kirjutama
ajutisse andmestikku nimega tulem lisaks prognoosidele ka esialgse
aegrea väärtused ning prognoosivead*/
var turistid;
id date;
run;
```

Aegridade analüüs, praktikum nr. 5, 2012

Käsitletavad teemad: autokorrelatsioonifunktsioon, prognoosimudeli sobivuse uurimine. AR ja MA tüüpi aegridade mudelitele vastavate protsesside simuleerimine.

Aegridade ennustamismeetodit võib lugeda sobivaks, kui tekkivad ennustusvead ei ole süstemaatilised, st ei sisalda enam mingit informatsiooni, mida saaks (lihtsalt) kasutada tulevikuprognooside täpsustamiseks. Näiteks kui enamasti järgnevad positiivsetele vigadele positiivsed vead ja negatiivsetele vigadele negatiivsed vead, siis saaks viimase ennustuse vea abil järgmise perioodi prognoosi parandada ning seetõttu ei ole kasutatav prognoosimeetod kindlasti parim. Sellist vigade vahelist sõltuvust saab uurida nn. autokorrelatsioonifunktsiooniga, mis väljendab hetkevea ja k ajasammu minevikus olnud prognoosivea sõltuvust. Mida suurem on selle kordaja absoluutväärtus, seda täpsemalt saab k sammu tagasi arvutatud vea abil hetkeviga ennustada.

Arvestada tuleb aga seda, et isegi täiesti juhuslike ennustusvigade puhul võivad andmete põhjal arvutatud autokorrelatsioonid olla nullist erinevad, seetõttu võetakse sobivuse üle otsustamisel arvesse ka seda, kui suured võivad vastavad autokorrelatsioonid juhuslike ennustusvigade puhul olla. Üheks sobivuse kriteeriumiks on praktikas näiteks autokorrelatsioonide jäämine piiridesse, kuhu juhuslike vigade puhul arvutatud autokorrelatsioonid jääksid 95% juhtudest. Samas arvutatakse korraga paljudele erinevatele nihetele vastavaid autokorrelatsioone ning mida suurem on vaadeldavate autokorrelatsioonide hulk, seda suurem on võimalus, et mõni neist juhuslikult veapiiridest välja jääb. Seetõttu on kasulik kontrollida tekkinud autokorrelatsioonide rühma kui terviku vastavust juhuslike vigade eeldusega. Üheks sobivaks vahendiks on Ljung-Box test, mille puhul tuleb ette anda korraga vaadeldavate autokorrelatsioonide arv, mille põhjal leitakse sobivalt kaalutud autokorrelatsioonide summa ning leitakse tõenäosus, et vähemalt nii suur väärtus saadakse täiesti juhuslike andmete korral.

Praktikumis vaatleme prognoosivigade sõltumatuse uurimist Holt-Wintersi meetodi baasil arvutatud ennustuste korral.

Lisaks vaatleme mudelile vastavate aegridade väärtuste tekitamist. Simuleeritud aegridade uurimine võimaldab aru saada, kuidas mingile mudelile vastavad read peaks käituma ning samuti on võimalik uurida, kui korrektselt õnnestub aegrea kordajaid leida juhul, kui rida vastab täpselt sobitatavale meetodile. Simuleeritud ridade analüüsist saadavad kogemused võimaldavad paremini käsitleda reaalsetele andmetele vastavaid aegridu.

Käesolevas praktikumis võtame vaatluse alla ARIMA mudelite erijuhud: autoregressiivsed ehk AR tüüpi mudelid kujul

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + A_t$$

ning liikuva keskmisega ehk MA (*moving average*) tüüpi mudelid kujul

$$Z_t = A_t - \sum_{i=1}^{q} \theta_i A_{t-i}.$$

Praktikumi tegevused SAS-is

1. Aegridu on küllaltki lihtne ise SAS skripti abil tekitada ning sel juhul on võimalik kontrollida, mida täpsemalt tehakse. Järgnev on näide AR(1) mudelile $Z_t = 0.8 * Z_{t-1} + A_t$ vastava aegrea tekitamisest:

```
data sim1; *ar(1) mudel;
n=100;*mitu väärtust genereerida;
phi1=0.8; Zt=0.5;
sigma=0.2; *häirituste standardhälve;
seed=10; *juhuslike suuruste genereerimise algväärtus, kui negat., siis kasutatakse kellaaega;
keep Zt date;
date=intnx('month','01jan1970'd,0);
output; *väljastame esimese rea algväärtusega;
do i=1 to n;*tekitame n simuleeritud väärtust;
Zt_1=Zt; *viimati arvutatud väärtus muutub eelmise ajamomendi omaks;
At=sigma*rannor(seed); *tekitame uue häirituse;
Zt=phi1*Zt_1+At;
date=intnx('month','01jan1970'd,i);
output;
end;
format date mmddyy8.;
run;
```

   Modifitseerige seda nii, et saaks genereerida AR2 mudelile $Z_t = 0.6Z_{t-1} - 0.2Z_{t-2} + A_t$ vastavaid väärtuseid.

2. Liikuva keskmise mudeleid on kõige lihtsam tekitada nii, et kõigepealt kirjutada ajutisesse andmestikku juhuslikud arvud A_t. Seejärel on võimalik selle andmestiku põhjal tekitada rida Z, kasutades LAGx funktsioone. Tekitage andmestik, kus on MA(2) rida juhul $\theta_1 = 0.5,\ \theta_2 = -1$.

3. Tekitatud ridade autokorrelatsioone on võimalik vaadelda käsuga `arima` kujul

```
proc arima data=sisendandmestik;
identify var=tulba_nimi;
run;
```

Vaadelge tekitatud ridade autokorrelatsioone. Kui palju nullist oluliselt erinevaid autokorrelatsioone on vaadeldavatel ridadel?

Tegevused R-is:

1. Tekitage käsuga `rnorm` 100 sõltumatut standardse normaaljaotusega juhusliku suuruse väärtust ning leidke nendest moodustatud aegrea autokorrelatsioonifunktsiooni väärtused (käsuga `acf`). Testige selle aegrea esimese kümne korrelatsioonikordaja vastavust sõltumatute juhuslike suuruste reale Ljung-Box testiga (käsk `Box.test`, kasutades suvandit `type="Ljung-Box"`).

2. Kontrollige eelmises praktikumis sobitatud eksponentsiaalse keskmistamise, Holti meetodi ja Holt-Wintersi meetodite sobivust prognoosimiseks ennustusvigade autokorrelatsioonide uurimise teel. Ljung-Box testi korral soovitatakse aegrea jääkide uurimise korral arvestada ka sobitamisel kasutatud parameetrite arvu, mida saab käsule `Box.test` edastada parameetriga `fitdf=param_arv`.

3. Kontrollige aditiivse ja multiplikatiivse Holt-Wintersi meetodi sobivust kinnisvara tehingumahtude prognoosimiseks.

4. Vaatleme kõigepealt AR mudelitele vastavate aegridade genereerimist ja nende ridade omadusi. Genereerimiseks kasutame kahte moodust: käsku `filter` ja käsku `arima.sim`. Näidetes eeldame, et suurused A_t on normaaljaotusega ja keskväärtusega 0.

   (a) Käsu `filter` põhiline kuju on järgmine:

   `filter(A,filter=c(phi_1,phi_2,...,phi_p),method="recursive", init=c(z1,...,zp))`

   kus A sisaldab suuruste A_t väärtusi ning $z1, \ldots, zp$ sisaldavad genereeritava rea esimesed p väärtust. Standardhälbega σ normaaljaotusega juhuslikke suurusi saab tekitada näiteks käsuga `sigma*rnorm(n)`, kus n on genereeritavate juhuslike suuruste arv.

   (b) Käsu `arima.sim` kasutamise põhikuju on järgmine:
   `arima.sim(n=100,model=list(ar=c(phi_1,...,phi_p),sd=sigma)`.
   Sellise kuju puhul kasutatakse normaaljaotusega juhuslikke suurusi, mille standardhälve on antud parameetriga `sd`. Samuti on võimalik ette anda juhuslike suuruste A_t väärtused kujul `innov=A`, kus A peab sisaldama n väärtust.

   Ülesanded:

   - Vaatleme juhtu $p = 1$ ning ϕ_1 väärtuseid $-1.5, -1, -0.8, 0, 0.8, 1, 1.5$. Eeldame, et häiritused A_t on normaaljaotusega ning standardhälbega $\sigma = 0.2$. Genereerige 100 häirituste rea A väärtust ning kasutage neid väärtuseid ning erinevaid ϕ_1 väärtuseid selleks, et tekitada vastavad AR tüüpi read. Iga rea puhul vaadelge graafikut ning püüdke otsustada, kas vastav rida võib vastata fikseeritud keskmise ümber muutumisele nii, et ka varieeruvus muutub mingi fikseeritud keskmise ümber, samuti vaadelge autokorrelatsioonide käitumist ning püüdke märgata iseloomulikke omadusi.
   - Vaatleme juhtu $\phi_1 = 0.7,\ \phi_2 = 0.1$ ning olgu juhuslike suuruste A_t standardhälve $\sigma = 0.2$. Genereerige 200 vastavat A väärtust ning defineerige seda kasutades käsuga `filter` aegrida Z. Vaadelge selle aegrea graafikut (kas vastab teie ettekujutusele statsionaarsest aegreast?), leidke autokorrelatsioonid ning uurige tekitatud aegrea viimase väärtuse sõltuvust etteantud algväärtustest. Katsetage ka käsku `arima.sim`.
   - Vaatleme juhtu $\phi_1 = -0.5,\ \phi_2 = 0.66$. Korrake eelmise ülesande tegevusi filter käsku kasutades. Kas tegemist võib olla statsionaarse reaga?

5. MA(q) mudelitele $Z_t = A_t - \sum_{i=1}^{q} \theta_i A_{t-i}$ vastavate aegridade genereerimine:

   (a) Käsu `filter` kuju on järgmine:
   `filter(A,filter=c(1,-theta_1,...,-theta_q),sides=1)`.

   (b) Käsu `arima.sim` kasutamise põhikuju on järgmine:
   `arima.sim(n=100,model=list(ma=c(-theta_1,...,-theta_p),sd=sigma)`.

   Ülesanded:

   - Simuleerige MA(1) rea käitumist juhul $\theta_1 = -1.25, -1, -0.8, 0.8, 1, 1.25$ ning uurige autokorrelatsioonikordajae käitumist samu A väärtusi kasutades. Kui palju erinevad juhtudele $\theta_1 = -1.25$ ning $\theta_1 = -0.8$ vastavad autokorrelatsioonikordajad?
   - Simuleerige MA(2) rea käitumist juhul $\theta_1 = 0.5,\ \theta_2 = -1$. Uurige autokorrelatsioonikordajate käitumist.

```
#ülesanne 1
A=rnorm(100)
plot(A)
acf(A)
Box.test(A,10,type="Ljung-Box")
#harjutus 2
#loeme andmed sisse
andmed=read.table("c:/portable/math/aegread2012/turistid.csv",na.strings="..",skip=0)
rida=c(t(andmed[3:14]))
rida=rida[!is.na(rida)]
turistid=ts(rida,start=1996,frequency=12)
#aditiivne meetod
m1=HoltWinters(turistid)
#multiplikatiivne meetod
m2=HoltWinters(turistid, seasonal="multiplicative")
#Holti meetod (selle rea puhul tegelikut ei ole mõistlik kasutada)
m3=HoltWinters(turistid,gamma=F)
#eksponentsiaalne keskmine
m4=HoltWinters(turistid,gamma=F,beta=F)
#analüüsime sobivust
#leiame prognoosivead
vead1=residuals(m1)
plot(vead1)
acf(vead1,40)#sesoonse rea puhul võiks olla näha vähemalt 3 perioodi
#korrelatsioonid ei jää lubatud piiridesse
#ei saa lugeda sõltumatuteks
#näite mõttes kasutame ka Ljung-Box testi
#sobitatud parameetrite arv on 3
Box.test(vead1,10,type="Ljung-Box",fitdf=3)
#sesoonsete ridadega soovitatakse testida vähemalt kahekordsele perioodile vastaavt rühma
Box.test(vead1,24,type="Ljung-Box",fitdf=3)
vead2=residuals(m2)
plot(vead2)
acf(vead2)
#Ikkagi autokorrelatsioonid selgelt piiridest väljas
#multiplikatiivne mudel ka ei sobi
Box.test(vead2,24,type="Ljung-Box",fitdf=3)
vead3=residuals(m3)
plot(vead3)
#graafikult on näha, et Holti meetodi vead ei ole kindlasti juhuslikud
acf(vead3)
#ka autokorrelatsioonid kinnitavad seda. Eriti suur on autokorrelatsioon aasta taguste
vigadega, st vigadel on tugev perioodiline komponent
Box.test(vead3,24,type="Ljung-Box",fitdf=2)
vead4=residuals(m4)
plot(vead4)
#graafikult on näha, et eksponentsiaalse keskmistamise meetodi vead ei ole kindlasti
juhuslikud
acf(vead4)
#ka autokorrelatsioonid kinnitavad seda
Box.test(vead4,24,type="Ljung-Box",fitdf=2)

#harjutus 3
###########kinnisvara ost-müük
#1)andmestiku sisselugemine. Eraldajaks on tabulaator, mis on kuulub vaikimisi määratud
eraldajate hulka
andmed1=read.table("c:/portable/math/aegread2012/kv015m.csv",na.strings="..")
```

```
#2)arvulised tulemused on kolmandas tulbas, mille nimi on V3
rida1=andmed1$V3
#3)puuduvate vaatluste eemaldamine lõpust
rida1=rida1[1:(length(rida1)-2)]
#4)ajas ümberpööramine- ei ole vaja
#5)aegrea tegemine. Tegemist on kvartaalsete andmetega
kv=ts(rida1,start=1997,frequency=4)
#aditiivne HW
m1=HoltWinters(kv)
vead1=residuals(m1)
plot(vead1)
acf(vead1)
#vead on ajas sõltuvad
Box.test(vead1,24,type="Ljung-Box",fitdf=3)
#ka Ljung-Box kinnitab sõltuvust
m1=HoltWinters(kv,seasonal="multiplicative")
vead1=m1$x-m1$fitted[,"xhat"]
plot(vead1)
acf(vead1)
#ei sobi
Box.test(vead1,20,type="Ljung-Box",fitdf=3)
#harjutus 4
#tekitame juhuslikud häiritused A
n=100
sigma=0.2
A=sigma*rnorm(n)
phi1=-1.5
rida1=filter(A,filter=phi1,method="recursive",init=0)
plot(rida1)
#varieeruvus kasvab ajas
acf(rida1)
#autokorrelatsioonid on vahelduvate märkidega
phi2=-1
rida2=filter(A,filter=phi2,method="recursive",init=0)
plot(rida2)
#Tegemist on piiripealse juhuga, nii et lühikese rea (100 vaatlust)
#korral ei pruugi täiesti selge olla, mis toimub
#Samas peaks olema näha, varieeruvus (st kaugus lähimineviku keskmisest) ei kõigu mingi
mõistliku
#keskmise varieeruvuse ümber ja ei pruugi olla tõkestatud
acf(rida2)#nulli kahanemine väga aeglane
phi3=-0.8
rida3=filter(A,filter=phi3,method="recursive",init=0)
plot(rida3)#Varieeruvus paistab juhuslikult aeg-ajalt suuremaks minevat, kuid
#pärast seda kahaneb jälle kiiresti. Kõigub nn keskmise varieeruvuse ümber
acf(rida3)
phi4=0
rida4=filter(A,filter=phi4,method="recursive",init=0)
plot(rida4)
acf(rida4)
phi5=0.8
rida5=filter(A,filter=phi5,method="recursive",init=0)
plot(rida5)
#keskmise juurde tagasijõudmise aeg võib olla suhteliselt suur, kuid
#siiski toimub muutumine selgelt ümber fikseeritud keskmise
acf(rida5)
phi6=1
```

```
rida6=filter(A,filter=phi6,method="recursive",init=0)
plot(rida6)
#fikseeritud keskmist ei ole
acf(rida6)
#autokorrelatsioonid kahanevad väga aeglaselt
n=1000
sigma=0.2
A=sigma*rnorm(n)
rida6=filter(A,filter=phi6,method="recursive",init=0)
plot(rida6)
acf(rida6)
#AR(2) mudelid
n=200
A=sigma*rnorm(n)
rida7.1=filter(A,filter=c(0.7,0.1),method="recursive",init=c(0,0))
plot(rida7.1)
#võib küll olla statsionaarne
acf(rida7.1,40)
#autokorrelatsioonid kahanevad nulli küllalt aeglaselt, kuid siiski küllalt kiiresti
#jõuavad veapiiride lähedale
#alternatiivne tekitamine, uued juhuslikud arvud
rida7.2=arima.sim(n,model=list(ar=c(0.7,0.1)),sd=0.2)
plot(rida7.2)
acf(rida7.2,40)
#alternatiivne, samad juhuslikud arvud kui rida7.1 korral
rida7.3=arima.sim(n,model=list(ar=c(0.7,0.1)),innov=A)
plot(rida7.3) #käitumine ei ole täpselt sama, kuna algväärtused on teised
rida8=filter(A,filter=c(-0.5,0.66),method="recursive",init=c(0,0))
plot(rida8)
#kindlasti ei ole statsionaarne, varieeruvus kasvab ajas
################################33
#harjutus 5
#MA tüüpi ridade simuleerimine
#kõik read on statsionaarsed, st fikseeritud keskmisega ning tõkestatud varieeruvusega
sigma=0.2
n=500
A=sigma*rnorm(n)
theta=-1.25
rida1=filter(A,filter=c(1,-theta),sides=1,method="convolution")
plot(rida1)
#filter käsuga tekib praegu esimeseks vaatluseks NA, kuna A[0] väärtust pole
#acf käsule tuleb öelda, et puuduvaid vaatluseid tuleb ignoreerida
acf(rida1,na.action=na.pass)
#esimene kordaja selgelt piiridest väljas, ülejäänud piirides või väga lähedal
abi1=acf(rida1,na.action=na.pass) #salvestan väärtused edasiseks võrdluseks
theta=-0.8
rida2=filter(A,filter=c(1,-theta),sides=1,method="convolution")
plot(rida2)
acf(rida2,na.action=na.pass)
abi2=acf(rida2,na.action=na.pass)
summary(abi2) #autokorrelatsioonid on komponendis acf
abi1$acf-abi2$acf #erinevused väga väikesed (teoreetiliselt nullid)
#järeldus: ainult autokorrelatsioonide põhjal ei saa MA tüüpi rea kordajaid kindlaks teha ...
theta=0.8
rida3=arima.sim(n,model=list(ma=-theta,sd=sigma))
plot(rida3)
acf(rida3)
```

```
#jällegi ainult esimene atukorrelatsioon selgelt veapiiridest väljas
theta=1.25
rida4=arima.sim(n,model=list(ma=-theta,sd=sigma))
plot(rida4)
acf(rida4)
#jällegi ainult esimene kordaja oluliselt nullist erinev.
#Kõigil MA(1) mudelitel on ainult esimene kordaja nullist erinev
#MA(2) mudeli simuleerimine
rida5=arima.sim(n,model=list(ma=c(-0.5,1),sd=sigma))
plot(rida5)
acf(rida5)
#kaks esimest kordajat selgelt nullist erinevad
```

```
*praktikumi nr. 5 skript SAS-is;
libname aegread "h:/aegread_praktikum/";
data sim1; *ar(1) mudel;
phi1=0.8; Zt=0.5;
sigma=0.2;
seed=10; *juhuslike suuruste genereerimise algväärtus;
mitu=100; *mitu väärtust tekitada;
keep Zt date;
date=intnx('month','01jan1970'd,0);
output; *väljastame esimese rea algväärtusega;
do i=1 to mitu;*tekitame 100 simuleeritud väärtust;
Zt_1=Zt;*viimane väärtus muutub esimeseks;
At=sigma**rannor(seed);*arvutame uue häirituse;
Zt=phi1*Zt_1+At;
date=intnx('month','01jan1970'd,i);
output;
end;
format date mmddyy8.;
run;
data sim2; *ar(2) mudel;
phi1=0.3;phi2=-0.2;Zt=0; *Zt on siin esimese vaatluse väärtus;
sigma=0.2;
seed=10; *juhuslike suuruste genereerimise algväärtus;
mitu=100;
keep Zt date;
date=intnx('month','01jan1970'd,0);
output; *väljastame esimese rea algväärtusega;
Zt_1=Zt;Zt=0; *Zt on nüüd teine vaatlus;
date=intnx('month','01jan1970'd,1);
output; *väljastame teise rea algväärtusega;
do i=2 to mitu+1;*tekitame n=mitu simuleeritud väärtust;
Zt_2=Zt_1;Zt_1=Zt;*uus ajamoment, olemasolevad väärtused muutuvad
vanemateks;
At=sigma*rannor(seed);
Zt=phi1*Zt_1+phi2*Zt_2+At;
date=intnx('month','01jan1970'd,i);
output;
end;
format date mmddyy8.;
run;
* MA tüüpi mudeli väärtuste genereerimine;
* teeme kõigepealt juhuslikud arvud;
data ajut;
keep A;
sigma=0.5;
seed=151;
do i=1 to 100;
A=sigma*rannor(seed);
output;
end;
run;
data sim3; *MA(2) mudel;
set ajut;
retain i 0 theta1 0.5 theta2 0.8;*kui toimub andmete sisselugemine,
siis tuleb kasutada retain väärtuste meelespidamiseks;
keep date Zt;
date=intnx('month','01jan1970'd,i);
```

```
Zt=A-theta1*lag1(A)-theta2*lag2(A);
i=i+1;
if Zt^=.;
output;
format date mmddyy8.;
run;
*alternatiiv, kui teha samuti nagu AR mudelitega
data sim2; *ar(2) mudel;
data sim4;
theta1=0.5;theta2=0.8;
sigma=0.5;
seed=151; *juhuslike suuruste genereerimise algväärtus;
*kaks A-d on vaja ennem tekitada, kui saab Z arvutama hakata;
At_1=sigma*rannor(seed); *A_1 väärtus;
At=sigma*rannor(seed); *A_2 väärtus;
mitu=98;
keep Zt date;
do i=2 to mitu+1;*tekitame n=mitu simuleeritud väärtust;
At_2=At_1;At_1=At;*uus ajamoment, olemasolevad väärtused muutuvad
vanemateks;
At=sigma*rannor(seed);
Zt=At-theta1*At_1-theta2*At_2;
date=intnx('month','01jan1970'd,i);
output;
end;
format date mmddyy8.;
run;

ods graphics off;
proc arima data=sim1;
identify var=Zt;
run;
proc arima data=sim2;
identify var=Zt;
run;
proc arima data=sim3;
identify var=Zt;
run;
proc arima data=sim4;
identify var=Zt;
run;
```

Aegridade analüüs
Praktikum nr. 6, 2012

Käsitletavad teemad: aegridade sisselugemine failidest, näide spektraalanalüüsi kasutamisest, sobiva mudelite tüübi esialgne hindamine autokorrelatsioonide ja osaautokorrelatsioonide põhjal.

Spektraalanalüüs tugineb etteantud rea esitamisel erineva perioodiga siinuste ja koosinuste summana. Kui reas on lisaks mürale peidus mingi kindla sagedusega signaal, siis vastava sagedusega siinuste ja/või koosinuste kordajad peaksid olema oluliselt suuremad, kui muude sageduste kordajad (kuna muud sagedused tekivad juhuslikult müra toimel ja müra ei tohiks ühtegi sagedust spetsiaalselt eristada). Kui rea jooksvad väärtused sõltuvad oluliselt eelmise perioodi vastavatest väärtustest, siis peaks ka sellele perioodile vastava sageduse kordajad müra poolt põhjustatud kordajatest suuremad olemaa. Kui rida vastab ainult mürale, siis peaksid kõik sagedused enam-vähem võrdväärselt esindatud olema ning sellele faktile tuginedes on loodud ka sõltumatute häirituste (nn valge müra) kindlakstegemise teste.

Käesolevas praktikumis (ja kogu kursuses) tutvume ainult väga põgusalt spektraalanalüüsi võimalustega.

1. Lugege SAS keskkonda andmestikust `praktikum6.txt` kolme aegrea andmed.
2. Üks neist aegridadest vastab nn valgele mürale (sõltumatud sama jaotusega juhuslikud suurused), üks sisaldab siinuste ja/või koosinuste kujul olevat signaali lisaks mürale ning ühel on küllalt tugev sõltuvus teatud nihkega minevikuväärtustega. Proovige SAS protseduuri `spectra` kasutades teha iga rea puhul kindlaks, milline neist kolmest see on ning samuti määrake signaali sisaldava rea puhul signaalis esinevate komponentide perioodid ja ajalise nihkega minevikuväärtustest sõltuva rea puhul selle nihke suurus. Perioodi arvutamisel on kasulik teadmine, et periood=1/sagedus.
3. Korrake analüüsi R keskkonnas. Vastavaks funktsiooniks on `spectrum`

Tänases praktikumis vaatleme lihtsamat tüüpi aegridade identifitseerimist. Kõigepealt, liikuva keskmisega mudelid MA(q) on kujul

$$Z_t - \mu = A_t - \sum_{i=1}^{q} \theta_i A_{t-1},$$

kus A_t on vähemalt teist järku statsionaarsed, tsentreeritud ja mittekorreleeritud. Küllalt lihtne on veenduda, et sellise rea teoreetilised autokorrelatsioonid võrduvad nulliga alates järgust $q+1$ (sest Z_{t-k} ja Z_t avaldises on $k \geq q+1$ korral erinevate indeksitega A-d ning need on mittekorreleeritud).

Autoregressiivsed mudelid $AR(p)$ esituvad kujul tähistades $\bar{Z}_t = Z_t - EZ_t$)

$$\bar{Z}_t = \sum_{i=1}^{p} \phi_i \bar{Z}_{t-i} + A_t,$$

kus A_t rahuldavad samu nõudeid, mis MA(q) mudeli korral ning on mittekorreleeritud ka kõikide varasemate Z_t väärtustega. Selliste ridade puhul annavad rohkem infot osaautokorrelatsioonid, mille leidmisel eemaldatakse Z_t-st ja Z_{t-k}-st vahepealsete Z väärtuste

mõju ning leitakse korrelatsioon nii saadud suuruste vahel. Kui eemaldamisel kasutatakse vähemalt p vahepealset väärtust, siis nende mõju Z_t-le (ehk parim ennustus Z_t väärtusele vahepealseid Z väärtuseid kasutades) on $\sum_{i=1}^{p} \phi_i \bar{Z}_{t-i}$, mille eemaldamisel jääb järele A_t. Et aga A_t on mittekorreleeritud kõikide varasemate Z väärtustega, siis teoreetiline osaautokorrelatsioon on alates järgust $p+1$ võrdne nulliga.

Mudeli valikul veendutakse kõigepealt, et tegemist võib olla statsionaarse aegreaga ning siis valitakse alustuseks sellist tüüpi mudel, millel on kõige vähem parameetreid. Käesolevas praktikumis tutvumegi esialgse mudeli valiku protsessiga. Failis praktikum6.csv on 5 erinevat aegrida. Nendest aegridadest osad on mittestatsionaarsed, osad on genereeritud mitmesuguste AR tüüpi mudelite põhjal või MA tüüpi mudelite põhjal ning osad on genereeritud ARMA mudeli (mudel, kus on nii eelmised Z väärtused kui eelmised A väärtused kasutusel) põhjal. Ülesandeks on need andmed nii R-i kui SAS-i sisse lugeda ning graafiku, autokorrelatsioonide ja osaautokorrelatsioonide põhjal püstitada hüpotees iga aegrea korral, mis tüüpi aegreaga võib olla tegemist. R-is saab osaautokorrelatsioone arvutada funktsiooniga `pacf()`, SAS-is kasutame protseduuri `arima` kujul

```
proc arima data=andmetabel;
identify var=veeru_nimi nlags=10;
run;
```

muutuja nlags ütleb, mitu autokorralatsioonikordajat ja osaautokorrelatsioonikordajat arvutada.

```
#periodogrammi ehk spektri kasutamise näited
andmed=read.table("h:/aegread2012/praktikum6.txt",header=T,sep=" ")
z1=ts(andmed$z1)
plot(z1)
#visuaalselt tundub mingi võnkumine sees olevat, seega arvatavasti ei ole päris juhuslik
#statsionaarsusga ei tundu vastuolu olevat
spectrum(z1)
#maksimum sagedusel 0.2, vastav periood on 5
z2=ts(andmed$z2)
plot(z2)
spectrum(z2)
#pilt on iseloomulik valgele mürale
#ei ole eelistatud sagedusi
z3=ts(andmed$z3)
plot(z3)
spectrum(z3)
#kaks sagedust oluliselt suuremad, kui teised
abi=spectrum(z3)
summary(abi) #info muutujas abi sisalduvate komponentide kohta
max(abi$spec) #maksimaalne väärtus, umbes 31
abi$freq[abi$spec>28]
#kahele suurimale väärtusele vastavad sagedused
1/abi$freq[abi$spec>28]
#z3 sisaldab nn signaali (kahe sagedusega), perioodidega 4 ja 9
#z2 on valge müra
#z1 sisaldab sõltuvust 5 perioodi tagustest andmetest

andmed=read.table("h:/aegread2012/praktikum6.csv",header=T,sep=",")
z6=ts(andmed$Z6,start=1870,frequency=4)
plot(z6)
#ei ole statsionaarne
acf(z6) #autokorrelatsioonid samuti ei kahane nullini
#praegu edasi ei analüüsi
z7=ts(andmed$Z7,start=1870,frequency=4)
plot(z7) #paistab olevat statsionaarne
acf(z7)
#esimene autokorrelatsioon selgelt veapiiridest väljas, teised enam-vähem piirides
#võib sobida MA(1) mudel
pacf(z7)
#4 veapiiridest väljas
#võib sobida ka AR(4)
#hüpotees:MA(1)
#kuna enamasti on vaja kõiki kolme graafikut koos vaadata, siis teeme funktsiooni,
#mis neid kõiki koos näitab
graafikud=function(z,mitu=30){
    layout(1:3) #jaotab graafikaekraani kolmeks võrdseks osaks
    plot(z) #esimesesse ossa aegrea graafik
    acf(z,mitu) #teiseks autokorrelatsioonid kuni etteantud nihete arvuni
    pacf(z,mitu) #osaautokorrelatsioonid
    layout(1) #graafikaaken edasisteks käskudeks jälle ühes osas
}
graafikud(z7,mitu=40)
z8=ts(andmed$Z8,start=1870,frequency=4)
graafikud(z8)
#hüpotees:MA(2)
z9=ts(andmed$Z9,start=1870,frequency=4)
graafikud(z9)
```

```
#hüpotees:AR(2)
z10=ts(andmed$Z10,start=1870,frequency=4)
graafikud(z10)
#kui valida AR ja MA mudelite vahel, siis AR(5) või AR(7)
#enamasti sellisel juhul saab leida lihtsama segamudeli (ARMA)
```

```
*praktikumi nr. 6 skript SAS-is;
data praktikum6_1; *loeme andmed ajutisse andmestikku;
infile "h:/aegread2012/praktikum6.txt" firstobs=2 delimiter=" " dsd;
keep date z1-z3; *nii tähistatakse muutujaid z1 z2 z3;
retain abi 0;
format date yyqc6.;
input z1-z3;
date=intnx("year","01jan00"d,abi);
output;
abi=abi+1;
run;
*eelnevalt võiks ridu ka visuaalselt vaadelda;
proc gplot data=praktikum6_1;
symbol i=spline;
plot z1*date;
where date>"01jan2450"d; *vaatame lõpuosa, et käitumist selgemalt
näha;
run; *võib sisaldada mingit perioodilist osa, miinimumid paistavad
korduvat küllalt regulaarselt 5 vaatluse tagant;
proc gplot data=praktikum6_1;
symbol i=spline;
plot z2*date;
where date>"01jan2420"d;
run; *visuaalselt mingit seaduspära ei erista;
proc gplot data=praktikum6_1;
symbol i=spline;
plot z3*date;
where date>"01jan2420"d;
run;*siin on samuti raske midagi seaduspära eristada;
proc spectra data=praktikum6_1 p out=spektrid; *p tähendab
periodogrammi väljastamist;
var z1-z3;
run;
*mugavam on vaadelda andmeid järjestatuna perioodi pikkuse järgi;
proc sort data=spektrid ;
by period;
run;
*vaatleme spektreid;
proc gplot data=spektrid;
symbol i=spline;
plot P_01*period;
where period<50;
run;
*paistab olema kaks sagedust, üks perioodiga umbes 5 ja teine 2.5.
Samas mingi nihkega minevikuväärtusest sõltumine võib tekitada
andmetesse ka 2*, 4* jne  väiksema perioodiga võnkumisi;
*igal juhul ei ole tegemist juhusliku reaga;
*vaatleme ka numbriliselt suurimatele väärtustele vastavaid
perioode;
proc print data=spektrid;
where P_01>50;
run; *tulemus tekib output aknasse;
proc gplot data=spektrid;
symbol i=spline;
plot P_02*period;
where period<50;
run;
```

```
*palju perioode esindatud, kuid kõik küllalt sarnaselt (erinevused
ei ole kümnete ja sadade kordsed);
*see on tüüpiline käitumine puhta müra korral;
proc gplot data=spektrid;
symbol i=spline;
plot P_03*period;
where period<50;
run;
*siin jällegi kaks perioodi olulisemalt esindatud, 5 ja 9;
*erinevused nende vahel väiksemad, seega teadaoleva info põhjal
peaks ütlema, et see rida sisaldab kahe erineva perioodiga;
*trigonomeetrilisi funktsioone, eelmine on nn valge müra ja esimene
kirjeldab rida, kus jooksev väärtus sõltub 5 sammu tagasi;
*olnud väärtusest;
*võib vaadelda ka numbriliselt suurematele väärtustele vastavaid
kirjeid;
proc print data=spektrid;
where P_03>30;
run;
data praktikum6_2; *loeme andmed ajutisse andmestikku;
infile "h:/aegread2012/praktikum6.csv" firstobs=2 delimiter="," dsd;
keep date z6-z10;
retain abi 0;
format date yyqc6.;
input tmp1  tmp $ z6-z10; *alguses on aasta ja kvartal, seejärel
tulevad aegridade andmed;
date=intnx("qtr","01jan1870"d,abi);
output;
abi=abi+1;
run;
*visuaalne vaatlemine;
proc gplot data=praktikum6_2;
symbol i=spline;
plot z6*date;
run; *varieeruvus paistab selgelt kasvavat, mittestatsionaarne;
proc gplot data=praktikum6_2;
symbol i=spline;
plot z7*date;
run;
*see võib küll olla statsionaarne;
proc gplot data=praktikum6_2;
symbol i=spline;
plot z8*date;
where date>"01jan1940"d; *igaks juhuks täpsemalt mingi osa reast;
run;
*võib olla statsionaarne;
proc gplot data=praktikum6_2;
symbol i=spline;
plot z9*date;
where date>"01jan1940"d; *igaks juhuks täpsemalt mingi osa reast;
run;
*võib olla statsionaarne;
proc gplot data=praktikum6_2;
symbol i=spline;
plot z10*date;
run;
*ei oska kindlalt öelda, keskmine võib olla kahanev, kuid võib ka
```

```
mitte olla;
proc arima data=praktikum6_2;
identify var=z6 nlags=40;
run; *autokorrelatsioonid ei kahane nulli ka nihkega 40, loeme
mittestatsionaarseks;
identify var=z7 nlags=20;
run; *vaadata tuleb autokorrelatsioone ja osaautokorrelatsioone,
hüpotees MA(1);
identify var=z8 nlags=20;
run; *hüpotees MA(2);
identify var=z9 nlags=20;
run;*hüpotees AR(2);
identify var=z10 nlags=20;
run;*ei paista head AR või MA mudelit, katsetada võib AR(5)-st
alates (äkki hilisemad on juhuslikud;
*tegelikult sellisel juhul on sageli võimalik leida oluliselt
väiksema parameetrite arvuga segamudel,
kus on nii AR kui MA liikmed;
quit;
```

Aegridade analüüs
Praktikum nr. 7, 2012

Käsitletavad teemad: AR, MA ja ARMA tüüpi mudelite valimine aegrea kirjeldamiseks, sobivuse diagnostika

Sobiva mudeli valimise protseduur on järgmine:

1. Vaatleme aegrida, autokorrelatsioonifunktsiooni ja osaautokorrelatsioonifunktsiooni, et teha kindlaks statsionaarsuse eelduse täidetus ja sobiva mudeli esialgne kuju. Mittestatsionaarse aegrea korral tuleb kindlaks teha sobiva teisenduse iseloom; seda vaatleme järgnevas praktikumis. Sobiva ARMA(p,q) tüüpi mudeli valikut abistavad faktid on järgmised:
   - AR(p) mudeli (ehk ARMA(p,0) mudeli) tunnuseks on see, et osaautokorrelatsioonid võrduvad nulliga alates järgust $p + 1$. AR(1) mudeli puhul on lisakontrolliks teadmine, et sel juhul autokorrelatsioonid peavad kahanema nagu absoluutväärtuselt ühest väiksema arvu astmed.
   - MA(q) mudeli tunnuseks on autokorrelatsioonide võrdumine nulliga alates järgust $q + 1$. MA(1) mudeli puhul on lisakontrolliks osaautokorrelatsioonide kahanemine nagu mingi absoluutväärtuselt ühest väiksema arvu astmed.
   - ARMA(1,q) mudeli tunnuseks on autokorrelatsioonide kahanemine vastavalt ühest väiksema arvuga korrutamisele alates järgust q+1 (st q+1 järku autokorrelatsioon on nagu mingi arv korda q-s autokorrelatsioon jne). ARMA(1,1) mudeli puhul peaks nii autokorrelatsioonid kui osaautokorrelatsioonid kahanema alates järgust 2 vastavalt mingile absoluutväärtuselt ühest väiksema arvuga korrutamisele.
2. Sobitame tarkvara abil valitud tüüpi aegrea (st laseme hinnata vajalikud parameetrid).
3. Uurime leitud mudeli sobivust. Põhiliseks sobivuse kriteeriumiks on ühesammuliste prognoosivigade juhuslikkus; tüüpiliseks testiks on prognoosivigade autokorrelatsioonide nulliga võrdumise hüpoteesi kontroll. Selleks analüüsime prognoosivigade autokorrelatsioonikordajaid (kas jäävad etteantud veapiiridesse) ning leiame Ljung-Box statistiku põhjal tõenäosuse, et juhuslike prognoosivigade korral tekivad sellised autokorrelatsioonid, nagu andmete põhjal näha. Kui Ljung-Box testi põhjal saadud p-väärtused on alates mingist rühma suurusest kõik piisavalt suured (näiteks suuremad kui 0.05), siis loeme mudeli adekvaatseks; vastasel korral sobitame uue mudeli ja kordame diagnostikat.

Tegevused SAS tarkvara kasutades

1. Kasutame SAS protseduuri `arima` selleks, et, leida sobivad mudelid eelmise praktikumi andmestikus `praktikum6.csv` olevatele statsionaarsetele aegridadele. Selleks tuleb anda kõigepealt käsk `proc arima data=andmestiku_nimi;` ning seejärel läbida iga aegrea korral identifitseerimise ja parameetrite hindamise etapid. Identifitseerimine toimub käskudega
   ```
   identify var=nimi; run;
   ```
   ja parameetrite hindamine käskudega

```
estimate p=number1 q=number2; run;
```
Viimasel juhul p tähistab autoregressiivse osa järku ja q tähistab liikuva keskmisega osa järku ning ette anda tuleb ainult nullist erinevad väärtused.

2. Pärast sobitamist tuleb kontrollida prognoosivigade sõltumatust. SAS-is saab Ljung-Box statistiku väärtuseid näha tabelis nimega "autocorrelation check of residuals" tulbas nimega "Pr>ChiSq".

3. Sobiva mudeli puhul kirjutada välja selle kuju.

Praktikumi tegevused R-is:

1. Leiame R vahenditega sobivad mudelid andmestikus `praktikum7.dat` olevatele andmestikele. Selleks lugege kõigepealt andmed sisse ja moodustage vastavad aegread.

2. Mudeli sobitamiseks on käsk `arima` ning diagnostikaks käsk `tsdiag`. Siin on kasulik teada, et käsus `tsdiag` ei arvestata vähemalt praegustes R versioonis Ljung-Box statistiku väärtuste arvutamisel sobitatud mudeli parameetrite arvu, mistõttu leitud p-väärtused on suuremad kui korrektse lähenemise korral. Samas suuremate autokorrelatsioonirühmade korral see erinevus kahaneb ning seetõttu on käesoleva kursuse raames lubatud otsuste tegemisel kasutada selle käsu tulemusi. Leidke neid käske kasutades vähimat arvu parameetreid sisaldavad sobivad mudelid kõikidele statsionaarsetele aegridadele. Käsule `arima` antakse mudeli kuju ette kolmest numbrist koosneva vektoriga `order`, millest esimene on autoregressiivse osa järk p, teine on AR, MA ja ARMA mudelite korral 0 ja kolmas on MA osa järk q. Parima mudeli kuju pange paberile kirja.

```
graafikud=function(z,mitu=30){
    layout(1:3) #jaotab graafikaekraani kolmeks võrdseks osaks
    plot(z) #esimesesse ossa aegrea graafik
    acf(z,mitu) #teiseks autokorrelatsioonid kuni etteantud nihete arvuni
    pacf(z,mitu) #osaautokorrelatsioonid
    layout(1) #graafikaaken edasisteks käskudeks jälle ühes osas
}
andmed=read.table("c:/portable/math/aegread2012/praktikum7.dat",header=T,sep=";")
Z=ts(andmed$Z11,start=1600)
graafikud(Z) #hüpotees:AR(2)
mudel=arima(Z,order=c(2,0,0))#esimene on p, teine d(=0 statsionaarse rea korral), kolmas q
tsdiag(mudel,20) #sobib
mudel
#Kui jääme selle juurde, siis on kujuks
# (Z_t-0.0260)=0.2502*(Z_{t-1}-0.0260)-0.0202*(Z_{t-2}-0.0260)+A_t
#ehk Z_t=0.2002+0.2502*Z_{t-1}-0.0202*Z_{t-2}+A_t
#ilma vabaliikmeta
mudel=arima(Z,order=c(2,0,0),include.mean=FALSE)
mudel#mudel on kujul Z_t=0.2507*Z_{t-1}-0.0199*Z_{t-2}+A_t
###############
Z=ts(andmed$Z12,start=1600)
graafikud(Z) #hüpotees:MA(2)
mudel=arima(Z,order=c(0,0,2))#esimene on p, teine d(=0 statsionaarse rea korral), kolmas q
tsdiag(mudel,20) #sobib
mudel #Kui jätta vabaliige,siis on mudeliks Z_t-0.0067=A_t+0.5063*A_{t-1}+0.3390*A_{t-2}
#ehk Z_t=0.0067+A_t+0.5063*A_{t-1}+0.3390*A_{t-2}
#R-is on kordajad toodud selle märgiga, kuidas nad võrduse paremal pool avalduvad
#ilma vabaliikmeta, kuna vabaliige ebaoluline
mudel=arima(Z,order=c(0,0,2),include.mean=FALSE)
mudel #loeme selle sobivaimaks.

######################
Z=ts(andmed$Z13,start=1600)
graafikud(Z) #hüpotees:MA(3), võib-olla ka MA(2), kuna kolmas on vähe väljas
mudel=arima(Z,order=c(0,0,3))#esimene on p, teine d(=0 statsionaarse rea korral), kolmas q
tsdiag(mudel,20) #sobib
mudel
#proovin ka MA(2)
proov=arima(Z,order=c(0,0,2))
tsdiag(proov,20) #ei sobi, jään MA(3) juurde
mudel
#ilma vabaliikmeta, kuna vabaliige ebaoluline
mudel=arima(Z,order=c(0,0,3),include.mean=FALSE)
mudel #loeme selle sobivaimaks, kuna väiksem aic
##############################
Z=ts(andmed$Z14,start=1600)
graafikud(Z) #ei ole statsionaarne (keskmine puudub, ka autokorrelatsioonid ei kahane nullini)
#ei analüüsi praegu edasi
######################
Z=ts(andmed$Z15,start=1600)
graafikud(Z) #hüpotees:MA(1)või AR(1)
#AR(1)
mudel=arima(Z,order=c(1,0,0))#esimene on p, teine d(=0 statsionaarse rea korral), kolmas q
tsdiag(mudel,20) #sobib
mudel
mudel1=arima(Z,order=c(1,0,0),include.mean=F)
mudel1
```

```
#proovin ka MA(1)
mudel2=arima(Z,order=c(0,0,1))
tsdiag(mudel2,20) #samuti sobib
mudel2
#ilma vabaliikmeta, kuna vabaliige ebaoluline
mudel2=arima(Z,order=c(0,0,1),include.mean=FALSE)
mudel2 #loeme sobivaimaks AR(1) mudeli, kuna väiksem aic
```

```
*praktikumi nr. 7 skript SAS-is;
data praktikum6; *loeme andmed ajutisse andmestikku;
infile "h:/aegread2012/praktikum6.csv" firstobs=2 delimiter="," dsd;
keep date z6-z10;
retain abi 0;
format date yyqc6.;
input tmp1 $ tmp2 $ z6-z10;
date=intnx("qtr","01jan1870"d,abi);
output;
abi=abi+1;
run;
*praktikumis nr. 6 leiti, et rida z6 ei ole statsionaarne, mudelit
ei sobita.;
proc gplot data=praktikum6; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z7*date;
run;
proc arima data=praktikum6;
identify var=z7 nlags=40;
run; *MA(1) mudel tundub sobivat;
estimate q=1; run;
*sobitame ka ilma keskmiseta;
estimate q=1 noint; run; *tulemus eelmisega väga sarnane, sobib;
*üheks võimalikuks otsuse tegemise aluseks sel juhul AIC (mida
väiksem, seda parem);
* sobitatud mudel Z_t= (1 + 0.67623 B**(1))A_t=A_t+0.67623*A_{t-1};
quit;
proc gplot data=praktikum6; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z8*date;
run;
proc arima data=praktikum6;
identify var=z8 nlags=40;
run; *MA(2) mudel tundub sobivat;
estimate q=2; run;
*ei sobi, see võib olla tingitud ka arvutusmeetodi
mittekoondumisest;
*proovime leida paremat mudelit;
*AR tüüpi mudelid nõuavad oluliselt rohkem parameetreid
*kuna jääkidel on kolm esimest autokorrelatsiooni oluliselt nullist
suuremad, siis
on mõistlik proovida MA(3) mudelit;
estimate q=3; run;*sobib
*proovime segamudeleid;
estimate q=2 p=1; run; *võib samuti sobida, kuid AIC oluliselt
halvem kui eelmisel;
estimate q=1 p=2; run; *ei sobi;
estimate q=3 noint; run;*ilma keskmiseta, sobib ja parima AICga;
quit;
proc gplot data=praktikum6; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z9*date;
run;
proc arima data=praktikum6;
identify var=z9 nlags=40;
run; *AR(2) mudel tundub sobivat;
```

```
estimate p=2; run;*sobib hästi;
*keskmine ebaoluline, proovime ära jätta;
estimate p=2 noint; run;*parem AIC, selle võtame sobivaimaks
mudeliks;
* mudeli kuju:  (1 - 0.05175 B**(1) - 0.41875 B**(2))Z_t=A_t ehk;
* Z_t=0.05175*Z_{t-1}+0.41875*Z_{t-2}+A_t;
quit;
proc gplot data=praktikum6; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z10*date;
run;
proc arima data=praktikum6;
identify var=z10 nlags=40;
run; *AR ja MA tüüpi mudelid nõuavad palju parameetreid;
*katsetama ARMA(1,1);
estimate p=1 q=1; run;*sobib hästi;
*keskmine ebaoluline, proovime ära jätta;
*näitena mudelist: kui keskmist ära ei jäta, siis mudeliks on;
*(1 - 0.83276 B**(1))*(Z_t-0.827111)=(1 + 0.76193 B**(1))*A_t;
*ehk Z_t=0.138326+0.83276*Z_{t-1}+A_t+0.76193*A_{t-1};
estimate p=2 noint; run;*parem AIC, selle võtame sobivaimaks
mudeliks;
quit;
```

Aegridade analüüs
Praktikum nr. 8, 2012

Käsitletavad teemad: ARIMA mudelite sobitamine mittestatsionaarsetele ja perioodilist (sesoonset) komponenti mitteomavatele aegridadele, tulevikuväärtuste prognoosimine ARIMA mudelite korral

Sobiva mudeli valimise protseduur on järgmine:

1. Vaatleme aegrida ja autokorrelatsioonifunktsiooni. Mittestatsionaarsuse tunnusteks on keskmise ja/või varieeruvuse selge muutumine ajas (ka sarnane käitumine erinevatel tasemetel) ning autokorrelatsioonide väga aeglane kahanemine järgu kasvades ning väga suur (praktiliselt ühega võrdne) esimene autokorrelatsioon.

2. Mittestatsionaarsuse olemasolul tuleks teha otsus selle kohta, et kas on mõistlik eeldada globaalse (lineaarse, ruutfunktsioonina või mingi teise funktsiooni abil kirjeldatava) trendi olemasolu. Kui on, siis eemaldame selle ennem sobiva ARIMA mudeli kuju kindlakstegemist. Eemaldamiseks võib sobitada andmetele vastava trendikõvera vähimruutude meetodil.

3. Kui globaalset trendi ei ole mõistlik eeldada või pärast selle eemaldamist on ikkagi tegemist mittestatsionaarse aegreaga, siis järgmisena uuritakse aegrea muutusid ehk diferentsitud aegrida. Diferentsimise vajadusele on rea ebaselge käitumise korral võimalik saada kinnitust ka mitmete nn ühikjuurte testidega; meie vaatleme Phillips-Perroni testi. Viimase nullhüpoteesiks on see, et esialgne rida ei ole statsionaarne ka pärast võimaliku lineaarse trendi eemaldamist, kuid diferentsitud rida on statsionaarne. Kui ka diferentsitud rida ei vasta statsionaarsuse tingimustele, siis vaadeldakse kaks korda diferentsitud aegrida jne. NB! Lihtsalt igaks juhuks diferentside leidmine on vale lähenemine, see vähendab olemasolevat info hulga ja muudab prognoose ebatäpsemaks.

4. Kui mingi arvu d korda diferentsitud aegrea korral paistavad statsionaarsuse tingimused olevat täidetud, siis leiame selle jaoks sobiva ARMA(p,q) mudeli; kokkuvõttes tähendab see esialgse rea jaoks ARIMA(p,d,q) tüüpi mudeli sobitamist. Kui globaalne trend on eemaldatud või kui selle olemasolu esialgse rea jaoks ei ole mõistlik eeldada, siis diferentsitud reale mudeli sobitamisel on õige eeldada nullkeskmist. Kui d korda diferentsitud rea sobitamisel lubada nullist erinevat keskmist, siis see tähendab tegelikult d järku polünoomiga kirjeldatava globaalse trendi olemasolu lubamist.

Tegevused SAS tarkvara kasutades

1. Kasutame SAS protseduuri `arima` selleks, et, leida sobivad mudelid failis `praktikum8.txt` olevatele aegridadele. Selleks, et uurida diferentsitud rida, tuleb identify käsus muutuja taga näidata, mis nihkega ja mitu korda diferentsida, näiteks
   ```
   identify var=nimi(1); run;
   ```
   esimest järku diferentsi korral või
   ```
   identify var=nimi(1 1); run;
   ```
   teist järku diferentsi kasutamise jaoks. Statsionaarsuse testi teostamiseks tuleb `identify` käsus lisada `stationarity=(pp)`. Parameetrite hindamisel tuleks juhul, kui globaalse trendi olemasolu eeldus ei ole õigustatud, nõuda, et keskmine oleks null, lisades sõna noint estimate käsule, nt
   ```
   estimate p=n1 q=n2 noint; run;
   ```
   Samas, kui d järku diferentsi korral on mõistlik eeldada d järku trendi olemasolu, siis võib selle sobitamist lubada nii, et noint jätta lisamata.

   Sobitatud mudeli abil aegrea tulevikuväärtuste ennustamine käib `forecast` käsuga. Selle käsu parameetriga `lead` saab ette anda soovitud tulevikuperioodide arvu.

Praktikumi tegevused R-is:

1. Tutvume kõigepealt mittestatsionaarsete ARIMA tüüpi aegridade ning nende diferentsitud versioonide võimaliku käitumisega. Diferentsitud aegrida saab tekitada käsuga `diff`. Phillips-Perroni testi saab teostada käsuga `PP.test`.

   - Genereerige käsu `arima.sim(n=500,model=list(order=(0,1,1),ma=0))` abil aegrida (see on lihtne juhuslik ekslemine). Kas on selge, millega põhjendada globaalse trendi puudumist vaadeldud rea korral? Kas Phillips-Perroni test aitab jõuda õigele otsusele diferentsi leidmise vajalikkuse osas? Korrake tegevust mitu korda, varieerides veidi ka `ma` parameetri väärtust või lisades mingi väikese `ar` parameetri.
   - Korrake eelnevaid tegevusi ARIMA(p,2,q) tüüpi mudelite korral (sel juhul `order` parameetri teine komponent on 2).

2. Uurime globaalset trendi sisaldavate ridade omadusi ning trendi eemaldamist. Trendi eemaldamine R-is on oluline, kuna ARIMA(p,d,q) mudeli sobitamisel $d > 0$ korral eeldatakse, et diferentsitud rea keskmine on null.

   - Moodustage lineaarsele trendile vastav aegrida käsuga `linear=ts(1:500)` ning tekitage trendiga aegrida kujul
     `z=arima.sim(n=500,list(ar=0.6))+0.05*linear`.
     Vaadelge rea ning selle diferentsitud versioonide graafikuid, autokorrelatsioone ja osaautokorrelatsioone. Kas Phillips-Perroni test aitab selle rea puhul teha õige otsuse?
   - Trendi eemaldamiseks sobitame andmetele trendikõvera käsuga `arima` kujul
     `trend=arima(z,xreg=trendi_rida)`
     ning lahutame reast trendi:
     `trendita_z=residuals(trend)`.
     Pärast trendita rea uurimist on kõige otstarbekam sobitada mudel esialgsele reale nii, et `arima` käsk leiaks koos nii trendi kui ka mudeli parameetrid. Selleks tuleb trendile vastav rida anda ette parameetriga `xreg=trendi_rida`.
   - Korrake eelnevat juhul, kui trend on liidetud Teie poolt valitud ARIMA(p,1,q) tüüpi mudelile.

3. Rakendage eelnevaid teadmisi andmestikus `praktikum8.txt` olevate aegridadele sobiva mudeli leidmiseks ja nelja järgneva väärtuse ennustamiseks. Kui arima mudeli sobitamisel on trend eemaldatud suvandit `xreg=trendi_rida` kasutades, siis prognoosimiseks tuleb predict käsuga anda ette trendi uued väärtused suvandiga `newxreg=uued_väärtused`.

```
graafikud=function(z,mitu=30){
    layout(1:3) #jaotab graafikaekraani kolmeks võrdseks osaks
    plot(z) #esimesesse ossa aegrea graafik
    acf(z,mitu) #teiseks autokorrelatsioonid kuni etteantud nihete arvuni
    pacf(z,mitu) #osaautokorrelatsioonid
    layout(1) #graafikaaken edasisteks käskudeks jälle ühes osas
}
#tekitame juhuslikule ekslemisele vastava rea
z=arima.sim(n=500,model=list(order=c(0,1,1),ma=0))
#uurime sellise mittestatsionaarse rea ja selle diferentsitud versioonide
#kuju, autokorrelatsioone ja osaautokorrelatsioone.
#siin tasub silmas pidada, et see on ilma trendita aegrida
#ning õige oleks sobitada mudeleid üks kord diferentsitud reale
graafikud(z)
PP.test(z) #PP testi kohaselt võib diferentsi leida
graafikud(diff(z))#statsionaarne
PP.test(diff(z)) #ka PP.test annab info, et edasi diferentsitud rida ei ole praegusest
rohkem statsionaarne,
#st edasi diferentsimine võib vastata ülediferentsimisele
#kordame eelnevat teise mudeli korral
z=arima.sim(n=500,model=list(order=c(0,1,1),ma=0.5))
graafikud(z)
PP.test(z) #PP testi kohaselt peaks diferentsi leidma
graafikud(diff(z))#statsionaarne
PP.test(diff(z)) #ka PP.test annab info, et edasi diferentsitud rida ei ole praegusest
rohkem statsionaarne
#vaatleme ka teist järku mittestatsionaarset aegrida
z=arima.sim(n=500,model=list(order=c(0,2,1),ma=0.5))
graafikud(z)
#mittestatsionaarsus on ilmne, seega pole mõistlik PP testi kasutada
graafikud(diff(z))#ei tundu olevat statsionaarne
#aga väga kindel ei ole
PP.test(diff(z)) #PP testi kohaselt võib veel kord diferentsitud rea lugeda statsionaarseks,
seega diferentsime veel
graafikud(diff(diff(z)))#visuaalselt vastab statsionaarsusele, sellisel juhul pole PP testi
vaja teha. Proovime siiski:
PP.test(diff(diff(z))) #PP testi kohaselt on praegune 2 korda diferentsitud rida
"statsionaarsem" kui veel kord diferentsitud rida, seega
#ka test annab kinnitust, et edasi pole vaja diferentse leida.
#trendiga read
linear=ts(1:500)
z=0.05*linear+arima.sim(n=500,model=list(ar=0.6)) #sel real on globaalne trend, kuid
diferentsid ei tule kasuks(sest pärast trendi eemaldamist
#jääb järgi tavaline AR(1) rida, mida ei tohiks diferentsida
graafikud(z) #rida on mittestatsionaarne ning lineaarse trendi olemasolu on ka ainult
graafikut vaadates usutav
PP.test(z) #PP testi kohaselt diferentsimine ei tule kasuks
graafikud(diff(z)) #ülediferentsimine toob enamasti kaasa suure arvu osaautokorrelatsioonide
veapiiridest välja jäämise
trend=arima(z,xreg=linear)
trendita_z=residuals(trend)
graafikud(trendita_z)# AR(1) mudel
#sobitamine
mudel=arima(z,order=c(1,0,0),xreg=linear)
tsdiag(mudel)
mudel
#ennustamine
```

```
predict(mudel,5,newxreg=(length(z)+1):(length(z)+5))

#######
andmed=read.table("h:/aegread2012/praktikum8.txt",header=T,sep=" ")
y=ts(andmed$Y1,start=1900,frequency=12)
graafikud(y) #arvatavasti mittestatsionaarne
PP.test(y) #ka testi tulemuse kohaselt võiks diferentsi leida
graafikud(diff(y))#statsionaarne, hüpotees arima(2,1,0)
#autokorrelatsioonid ei ole suured positiivsed, seega edasi diferentsida ei ole vaja
PP.test(diff(y)) #kuigi testi pole sel juhul vaja teha, veendume, et
#testi tulemus on selline, nagu peaks olema
mudel=arima(y,order=c(2,1,0))
tsdiag(mudel,20)
mudel
y=ts(andmed$Y2,start=1900,frequency=12)
graafikud(y) #selgelt mittestatsionaarne, ilma globaalse trendita
#sellise rea puhul ei ole mõtet PP testi otsuse tegemiseks kasutada
#ning test võib anda meile näiliselt vale info
PP.test(y) #R-is olev testi versioon ütleb, et diferentsimine ei anna meile
#paremini statsionaarsusele vastavat rida kui esialgne. See võib aga tähendada, et
#on vaja diferentse leida rohkem kui üks kord
graafikud(diff(y))#ei tundu olevat statsionaarne, kuid täiesti võimatu ei ole
#siiski sellise autokorrelatsioonide käitumise korral pigem PP testi ei kasutata, vaid
leitakse
#järgmine diferents
PP.test(diff(y)) #ka PP testi kohaselt võib diferentsi leida
graafikud(diff(diff(y))) #statsionaarne, sobib ehk MA(2)
mudel=arima(y,order=c(0,2,2)) #sobib, kuid mitte ideaalselt (üsna 0.05 piiri lähedal)
mudel
#alternatiiv: AR(3)
mudel=arima(y,order=c(3,2,0))
tsdiag(mudel,20)
mudel #eelmine parem, võtame selle aluseks

y=ts(andmed$Y3,start=1900,frequency=12)
graafikud(y) #mittestatsionaarne, globaalne lin. trend võimalik
#R-is tuleb trend analüüsimisel eelnevalt eemaldada
lineaarne=1:length(y)
trend=arima(y,xreg=lineaarne)
trendita=residuals(trend)
graafikud(trendita)# võib olla mittestatsionaarne, autokorrelatsioonid püsivad kaua suured
PP.test(trendita) #PP testi kohaselt võiks ikka diferentsi leida
graafikud(diff(trendita))#paistab olevat statsionaarne hüpotees: arima(1,1,1)
#kuid autokorrelatsioonid on alguses suhteliselt suured
#võime igaks juhuks kontrollida täiendava diferentsimise mõistlikkust
PP.test(diff(trendita)) #ka test ütleb, et ei ole mõistlik rohkem diferentse leida
mudel=arima(y,order=c(1,1,1),xreg=lineaarne) #leiame mudeli, lastes eelnevalt trendi eemaldada
tsdiag(mudel,20)# sobib väga hästi

y=ts(andmed$Y4,start=1900,frequency=12)
graafikud(y) #mittestatsionaarne, ilma globaalse trendita
graafikud(diff(y))#ei ole statsionaarne, vaja diferents leida
graafikud(diff(diff(y))) #ikka ei ole statsionaarne
graafikud(diff(y,1,3)) #arvatavasti statsionaarne. Igaks juhuks PP test
PP.test(diff(y,1,3)) # ei tule diferentsi enam leida
#3 korda diferentsitud rea mudeliks võiks sobida AR(4) või mõni segamudel (näiteks ARMA(1,1))
mudel=arima(y,order=c(4,3,0))
```

```
tsdiag(mudel,20) #sobib, kuid vaatleme ka väiksema parameetrite arvuga alternatiive
mudel
mudel=arima(y,order=c(1,3,1))
tsdiag(mudel,20) #ei sobi
mudel=arima(y,order=c(1,3,2))
tsdiag(mudel,20) #sobib, kuid p-väärtused suhteliselt piiri lähedal
#proovime korrektseid Ljung-Box väärtuseid
Box.test(residuals(mudel),10,type="L", fitdf=3)
Box.test(residuals(mudel),15,type="L", fitdf=3)
#ka korrektsed väärtused on ülalpoor kriitilist piiri;
#seega loeme korrektseks ja seni parimaks
#ja vaatleme kolme parameetriga alternatiive
mudel
mudel=arima(y,order=c(2,3,1))
tsdiag(mudel,20) #ei sobi
mudel
```

```
*praktikumi nr. 8 skript SAS-is;
data praktikum8; *loeme andmed ajutisse andmestikku;
infile "h:/aegread2012/praktikum8.txt" firstobs=2 delimiter=" " dsd;
keep date y1 y2 y3 y4;
retain abi 0;
format date date9.;
input tmp1 $ tmp2 $ y1-y4;
date=intnx("month","01jan1900"d,abi);
output;
abi=abi+1;
run;
proc gplot data=praktikum8; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot y1*date;
run;
proc arima data=praktikum8;
identify var=y1 nlags=40 stationarity=(pp=(5));
run;
*PP testi on mõistlik kasutada, kui ei ole selge, kas esialgne rida
on ise statsionaarne või siis tuleb diferentsida;
*PP tulemuses tasub vaadata viimast rida (trend);
*kui p-väärtus piisavalt suur, siis diferentsitud rida võib olla
statsionaarne ja ülediferentsimist ei teki;
*sulgudes on nn truncation parameter, mis R valib vaikimisi 4*
(n/100)^0.25 täisosa, st praegu 5;
*autokorrelatsioonid ei kahane nulli ka nihkega 40, loeme
mittestatsionaarseks;
identify var=y1(1) nlags=40 stationarity=(pp=(5));run;
*hüpotees:AR(2) diferentsitud reale ehk ARIMA(2,1,0) mudel;
*PP test samuti ütleb, edasine diferentsimine ei ole mõistlik;
estimate p=2 noint;run;
*sobib;
forecast lead=4; run;
quit;
proc gplot data=praktikum8; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot y2*date;
run;
*jooniselt selge, et rida ei ole statsionaarne, tuleb diferentsida;
proc arima data=praktikum8;
identify var=y2 nlags=40;
run; *mittestatsionaarne, diferentsimine vajalik, pp testi ei ole
siin mõistlik kasutada;
identify var=y2(1) nlags=40 stationarity=(pp=(5));run;
*ei ole statsionaarne, diferentsimine vajalik;
identify var=y2(1 1) nlags=40 stationarity=(pp=(5));run;
*statsionaarne, hüpotees AR(3) (või MA(2)), kokku ARIMA(3,2,0) või
ARIMA(0,2,2);
*edasine diferentsimine selgelt ebavajalik;
estimate p=3 noint; run; *sobib, kuid mitte väga hästi;
estimate q=2 noint; run; *AIC põhjal veidi parem. võtame aluseks.;
quit;
proc gplot data=praktikum8; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot y3*date;
run;
```

```
*paistab mittestatsionaarne olevat;
*kuid võib olla statsionaarne võnkumine ümber sirge;
*siin kindlasti aitab pp test;
proc arima data=praktikum8;
identify var=y3 nlags=40 stationarity=(pp=(5));
run; *mittestatsionaarne, diferentsimine vajalik ka pärast lineaarse
trendi eemaldamist;
identify var=y3(1) nlags=40 stationarity=(pp=(5));
run; *sobib, hüpotees ARMA(1,1) (ehk ARIMA(1,1,1));
*kuna esialgsetes andmetes paistab olevat selge lineaarne trend,
siis lubame nullist erinevat keskmist;
estimate p=1 q=1; run;*sobib hästi;
quit;
proc gplot data=praktikum8; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot y4*date;
run;
*mittestatsionaarne, arvatavasti mitu korda vaja diferentse leida;
proc arima data=praktikum8;
identify var=y4 nlags=40;
run; *mittestatsionaarne, leiame diferentsi;
identify var=y4(1) nlags=40 stationarity=(pp=(5)); run;
*mittestatsionaarne, leiame teise diferentsi;
identify var=y4(1 1) nlags=40; run;
*ikka mittestatsionaarne, autokorrelatsioonid püsivad väga suured;
identify var=y4(1 1 1) nlags=40; run;
*tunduvalt parem kui eelmine, kuid autokorrelatsioonid ei kahane
veel väga kiiresti. Kontrollime PP testi abil, kas on;
*mõistlik veel diferentse leida;
identify var=y4(1 1 1) nlags=40 stationarity=(pp=(5)); run;
*rohkem diferentse pole vaja leida;
*AR(4) võib sobida. Võimalikud ka segamudelid. Proovin ARMA(1,1)
alustuseks;
estimate p=1 q=1 noint; run;*ei sobi;
estimate p=4 noint;run; *sobib, kuid mitte ideaalselt (alguses on p-
väärtused üsna kriitilise lähedal;
*segamudelite proovimise osas selgelt sobivad teadaolevad juhud
(peale arma(1,1) puuduvad;
*katsetan sama või vähema parameetrite arvuga segamudeleid;
estimate p=1 q=2 noint; run;*sobib paremini, kui eelmine;
estimate p=2 q=1 noint; run; *ei sobi;
estimate p=1 q=3 noint; run; *sobib;
estimate p=2 q=2 noint; run; *sobib, AIC parim, võtame aluseks;
quit;
```

Aegridade analüüs
Praktikum nr. 10, 2012

Käsitletavad teemad: osalised ning sesoonsed ARIMA tüüpi mudelid

Sageli ei sobi andmetele madalat järku ARIMA(p,d,q) tüüpi mudelid, kuna esinevad üksikud nullist erinevad küllaltki kõrget järku autokorrelatsioonid ja/või osaautokorrelatsioonid. Sageli esineavad need teatud kindla sammu tagant, mis on indikaator aegrea perioodilise (sesoonse) käitumise olemasolu kohta. Sellisel juhul tuleb kasutusele võtta osalised või multiplikatiivsed sesoonsed ARIMA mudelid.

Osalisteks ARIMA tüüpi mudeliteks nimetatakse selliseid ARIMA(p,d,q) tüüpi mudeleid, milles ainult osa kordajatest ϕ_1, ϕ_2, ϕ_p ja $\theta_1, \theta_2, \ldots, \theta_q$ on nullist erinevad (nende hulgas on ϕ_p ja θ_q).

Perioodiga s multiplikatiivseteks ARIMA(p,d,q)x(P,D,Q)$_s$ tüüpi mudeliteks nimetatakse mudeleid kujul

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Z_t = \theta(B)\Theta(B^s)A_t,$$

kus

$$\phi(x) = 1 - \sum_{i=1}^{p} \phi_i x^i, \;\; \Phi(x) = 1 - \sum_{i=1}^{P} \Phi_i x^i,$$
$$\theta(x) = 1 - \sum_{i=1}^{q} \theta_i x^i, \;\; \Theta(x) = 1 - \sum_{i=1}^{Q} \Theta_i x^i.$$

Lihtsalt mõistetavaks multiplikatiivseks sesoonseks ARIMA tüüpi mudelite erijuhuks on ARIMA(p,d,q)x(0,1,0)$_s$ tüüpi mudelid, kus s tähistab vaatluste arvu perioodis (tavaliselt aasta), kuna sel juhul vastavad aastased muudud ARIMA tüüpi mudelile. Kui p, q on oluliselt väiksemad suurusest s ning $D = 0$, siis on multiplikatiivne sesoonne mudel osalise ARIMA tüüpi mudeli erijuhuks, kuid multiplikatiivsel mudelil on vastava osalise ARIMA tüüpi mudeliga võrreldes vähem kordajaid.

Kõikide ARIMA tüüpi mudelite korral võib parameetrite hinnanguid leida mitmesuguste erinevate meetoditega. Levinuimad on tinglik ühesammuliste prognoosivigade ruutude summa (sõna tinglik tuleneb vajadusest liikuva keskmisega ridade puhul ette anda mõningad A_t väärtused, et prognoosimisega alustada ning edasised vead on leitud tingimusel, et etteantud väärtused on õiged) ning suurima tõepära meetod, mis baseerub häirituste normaaljaotusele vastamise eeldusel. Mõlemal juhul tuleb parameetrite leidmiseks minimiseerida mittelineaarset funktsiooni ning tulemus võib sõltuda algväärtuste valikust. Parameetrite leidmise meetoditest räägime lähemalt loengus.

R-is saab kasutatava meetodi ette anda arima käsus suvandiga `method`, võimalikud väärtused on `CSS` (tinglik ruutude summa), `ML` (suurima tõepära meetod) ning `CSS-ML` (esialgsed parameetrite hinnangud leitakse tinglike ruutude summa abil ning seejärel täpsustatakse suurima tõepära meetodil).

Praktikumi tegevused R-is:

1. Uurime katseliselt osaliste ARIMA tüüpi mudelite autokorrelatsioonide ja osaautokorrelatsioonide käitumist.

   - Genereerige `arima.sim` käsuga 1000 mudelile $Z_t = 0.8\, Z_{t-10} + A_t$ väärtust (kasutades vaikimisi määratud A_t standardhälvet). Uurige tekkinud rea autokorrelatsioonide ja osaautokorrelatsioonide graafikuid.

- Genereerige `arima.sim` käsuga 1000 mudelile $Z_t = A_t - 0.8A_{t-10}$ väärtust (kasutades vaikimisi määratud A_t standardhälvet). Uurige tekkinud rea autokorrelatsioonide ja osaautokorrelatsioonide graafikuid.
- Genereerige `arima.sim` käsuga 1000 mudelile $Z_t = 0.5\,Z_{t-1} + 0.3Z_{t-10} + A_t$ väärtust (kasutades vaikimisi määratud A_t standardhälvet). Uurige tekkinud rea autokorrelatsioonide ja osaautokorrelatsioonide graafikuid.
- Genereerige `arima.sim` käsuga 1000 mudelile $Z_t = A_t - 0.5A_{t-1} - 0.3A_{t-10}$ väärtust (kasutades vaikimisi määratud A_t standardhälvet). Uurige tekkinud rea autokorrelatsioonide ja osaautokorrelatsioonide graafikuid.

2. Leidke sobivad osalised või sesoonsed ARIMA tüüpi mudelid failis `praktikum10.txt` olevatele aegridadele. Osaliste ARIMA mudelite puhul tuleb R-is näidata parameetriga `fixed`, millised kordajad on nulliga võrdsed, kusjuures parameetri `fixed` väärtusteks peab olema vektor, mille pikkus on $p + q + 1$ (kui $d = 0$ ja parameeter `include.mean` väärtuseks on TRUE (mis on vaikeväärtuseks) ning $p + q$ vastasel korral. Selles vektori esimesed p komponenti vastavad AR kordajatele, järgmised q komponenti vastavad MA kordajatele ja seejärel tuleb vajadusel keskmisele μ vastav komponent. Need parameetrid, mille väärtusi otsitakse, tuleb tähistada NA-ga, ülejäänud peavad aga olema nullid. Näiteks mudeli

$$Z_t - \mu = \phi_4(Z_{t-4} - \mu) + A_t - \theta_1 A_{t-1} - \theta_3 A_{t-3}$$

sobitamiseks sobib käsk

`arima(Z,order=c(4,0,3),fixed=c(0,0,0,NA,NA,0,NA,NA))`.

Multiplikatiivse sesoonse mudeli sobitamiseks on aga lisaparameeter

`seasonal=list(order=(P,D,Q),period=s)`,

kus s on perioodi pikkus vaatluste arvuna.

```
graafikud=function(z,mitu=30){
    layout(1:3) #jaotab graafikaekraani kolmeks võrdseks osaks
    plot(z) #esimesesse ossa aegrea graafik
    acf(z,mitu) #teiseks autokorrelatsioonid kuni etteantud nihete arvuni
    pacf(z,mitu) #osaautokorrelatsioonid
    layout(1) #graafikaaken edasisteks käskudeks jälle ühes osas
}
#harjutus 1
n=1000
Z=arima.sim(model=list(ar=c(rep(0,9),0.8)),n)
graafikud(Z) #osaautokorrelatsioonidest üks (kümnes) oluliselt piiridest väljas
#autokorrelatsioonidest kümnes, kahekümnes jne piirdest väljas, kuid kahanevad nulli
Z=arima.sim(model=list(ma=c(rep(0,9),-0.8)),n)
graafikud(Z) #autokorrelatsioonidest üks (kümnes) oluliselt piiridest väljas
#osaautokorrelatsioonidest kümnes, kahekümnes jne piirdest väljas, kuid kahanevad nulli
Z=arima.sim(model=list(ar=c(0.5,rep(0,8),0.3)),n)
graafikud(Z) #Osaautokorrelatsioonidest ei ole nulli piiridest väljas sugugi ainult need
#millele vastavad liikmed on mudelis sees. Praegusel juhul on nulli piiridest väljas
esimene, üheksas
# ja kümnes. Seega osalise mudeli otsimisel ei ole alguses vaja lõpu poolt kõiki kordajaid
kohe sisse võtta
Z=arima.sim(model=list(ma=c(-0.5,rep(0,8),-0.4)),n) #sama jutt osalise MA mudeli kohta.
graafikud(Z)
#ülesanne 2
andmed=read.table("c:/portable/math/aegread2012/praktikum10.txt",header=T,sep=" ")
Z1=ts(andmed$Z1)
graafikud(Z1) #esimene ja üheksas autokorrelatsioon selgelt piiridest väljas
#proovime osalist MA(9) mudelit
m1=arima(Z1,order=c(0,0,9),fixed=c(NA,rep(0,7),NA,NA)) #viimane NA on keskväärtuse jaoks
#kesväärtus mitteoluline
m1=arima(Z1,order=c(0,0,9),fixed=c(NA,rep(0,7),NA,0))
tsdiag(m1,30) #sobib, kuid mitte väga kindlalt. Siiski võib võtta prognoosimisel aluseks (ja
edasi paremaid otsida ei ole vaja)
#proovime siiski lisada ka viienda A, st A[t-5]
m1.2=arima(Z1,order=c(0,0,9),fixed=c(NA,rep(0,3),NA,rep(0,3),NA,0))
m1.2 #AIC järgi parem
tsdiag(m1.2,30) #sobib hästi, võib võtta prognoosimisel aluseks
#proovime ka sesoonset mudelit, mis võib tekitada sarnase pildi
m1.3=arima(Z1,order=c(0,0,1),seasonal=list(order=c(0,0,1),period=8))
tsdiag(m1.3,30)#ei sobi
Z2=ts(andmed$Z2)
graafikud(Z2)
#kõigepealt kõigepealt sesoonset mudelit
m2.1=arima(Z2,order=c(1,0,0),seasonal=list(order=c(1,0,0),period=8))
tsdiag(m2.1,30)
m2.1 #vabaliige ei ole oluline
m2.1=arima(Z2,order=c(1,0,0),seasonal=list(order=c(1,0,0),period=8),include.mean=F)
m2.1
#sobib väga hästi, võib võtta prognoosimisel aluseks
#võrdluse mõttes proovime leida ka osalise arima mudeli
m2.2=arima(Z2,order=c(9,0,0),fixed=c(NA,rep(0,7),NA,NA))
tsdiag(m2.2) #ei sobi
#proovime lisada Z[t-8], sest kaheksa kordse nihkega autokorrelatsioonid nullist erinevad
m2.3=arima(Z2,order=c(9,0,0),fixed=c(NA,rep(0,6),NA,NA,NA))
tsdiag(m2.3,30)
m2.3 #keskmine ei ole oluline
m2.3=arima(Z2,order=c(9,0,0),fixed=c(NA,rep(0,6),NA,NA,0))
```

```
m2.3 #sobib, kuid AIC suurem, kui sesoonsel mudelil. Kui sesoonne sobib, siis
#ei ole tegelikult mõtet osalist mudelit otsida.
Z3=ts(andmed$Z3)
graafikud(Z3,40) #perioodile 12 vastavad autokorrelatsioonid ei lähe nulli
#andmetes ka perioodilisus selgelt näha.
# tuleb leida sesoonne diferents perioodiga 12
graafikud(diff(Z3,12),40)
#paistab statsionaarne. Võib sobida MA(4) või AR(5), kuid tegelikult vastab käitumine
#pigem segamudelile ARMA(1,1)
m3.1=arima(Z3,order=c(1,0,1),seasonal=list(order=c(0,1,0),period=12)) #üks sesoonne
diferents, muidu tavaline ARMA
tsdiag(m3.1,40) #mudel sobib
#viimane sobib paremini (AIC väiksem)
```

Aegridade analüüs
Praktikum nr. 11, 2012

Sesoonse ARIMA tüüpi mudelite sobitamine. Erinevate parameetrite leidmise meetodite katsetamine

Jätkame eelmise praktikumi tegevusi, seekord SAS tarkvara abil. Praktikumi ülesanneteks on SAS tarkvara abil leida sobivad mudelid viiele aegreale failis `prax11SAS.txt` ning uurida ka erinevate parameetri hindamise meetodite mõju saadavatele hinnangutele. SAS tarkvaras saab meetodit näidata arima protseduuri `estimate` käsu all, kirjutades sinna `method=ML`, `method=ULS` või `method=CLS`. Esimesel juhul hinnatakse parameetreid suurima tõepära meetodil eeldusel, et häiritused A_t on sõltumatud ja normaaljaotusega, viimasel juhul minimiseeritakse ühesammuliste prognoosivigade ruutude summat eeldusel, et olemasolevatele andmetele eelnevad prognoosimiseks vajalikud suurused on nullid ning ULS täpsema tähenduse kohta saab infot SAS abifailidest.

SAS tarkvara abil ARIMA(p,d,q)x(P,D,Q)$_s$ mudeli sobitamisel tuleb tuleb kõigepealt `identify` käsus ära näidata, milliseid diferentse tuleb leida; selleks kirjutatakse

```
identify var=z(1,...,1,s,...,s);
```

kus ühtesid on d tükki ja perioodi pikkuseid s on D korda; tavaliselt $d, D \leq 1$. Seejärel antakse `estimate` käsus p ja q abil ette otsitavad kordajad kujul

```
p=(1 2 ... p)(s 2s 3s ... Ps) q=(1 2 ... q)(s 2s ... Qs)
```

Näiteks kui andmed on igakuised ($s = 12$) ning me sobitame mudelit ARIMA(2,0,2)x(0,1,1)$_{12}$, siis on käskudeks

```
identify var=z(12);run;
estimate p=2 q=(1 2)(12); run;
```

Osalise Arima mudeli korral näidatakse p ja q taga sulgudes indeksid, millised kordajad võivad olla nullist erinevad.

```
*praktikumi nr. 11 skript SAS-is;
data praktikum11; *loeme andmed ajutisse andmestikku;
infile "h:/aegread2012/prax11SAS.txt" firstobs=2 delimiter="," dsd;
keep date z6-z10;
retain abi 0;
format date date9.;
input aasta kuu z6-z10;
date=intnx("month","01jan1900"d,abi);
output;
abi=abi+1;
run;
proc gplot data=praktikum11; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z6*date;
run; *pastab olevat statsionaarne, aastasele perioodile vastavaid
liikmeid ei tundu olevat, seega tavaline ARMA mudel.;
proc arima data=praktikum11;
identify var=z6 nlags=40; run; *madalat järku AR või MA ei sobi;
estimate p=6  noint;run;*sobib, kuid otsime väiksema kordajate
arvuga mudelit;
estimate p=1 q=1; run;*ei sobi;
estimate p=2 q=1; run; *sobib;
estimate p=1 q=2; run; *kindlalt kehvem kui eelmine;
estimate p=2 q=1 noint; run; *Parim kuju. Katsetame erinevaid
parameetrite hindamise meetodeid;
estimate p=2 q=1 noint method=ml; run; *suurima tõepära hinnangud.
AIC õigeks arvutamiseks oluline;
estimate p=2 q=1 noint method=cls; run; *see on vaikimisi meetod. ;
estimate p=2 q=1 noint method=uls; run; *vahepealne meetod, hindab
AIC väärtust täpsemalt kui cls;
quit;
proc gplot data=praktikum11; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z7*date;
where date>"01jan73"d;
run; *mittestatsionaarne, vaja leida diferents;
proc arima data=praktikum11;
identify var=z7 nlags=40; run; *ka siit paistab diferentsi leidmise
vajalikkus, autokorrelatsioonid ei kahane nulli;
identify var=z7(1) nlags=40; run; * väikeseid nihkeid vaadates
paistab AR(1), sammuga 12 autokorrelatsioone vaadates MA(1);
estimate p=1 q=(12) noint; run; *sobib hästi;
estimate p=1 q=(12) noint method=ml; run; *sobib hästi, kordajates
väiksed muutused. Kui vead võivad vastata normaaljaotusele, siis ml
hinnang parem;
proc gplot data=praktikum11; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z8*date;
where date>"01jan73"d; *perioodilisust näeb paremini, kui vaadelda
ainult osa tervest reast;
run; *tugev sesoonsus (aastane periood) tundub olevat olemas,
arvatavasti vaja leida sesoonne diferents;
proc arima data=praktikum11;
identify var=z8 nlags=50 ;run; *perioodile 12 vastavad
autokorrelatsioonid ei kahane nulli, seega vaja leida sesoonne
diferents;
```

```
identify var=z8(12) nlags=50 ;run; *madalat järku kordajate põhjal
MA(1), sesoonsete põhjal AR(1);
estimate p=(12) q=1; run; *sobib, kuid konstant on ebavajalik;
estimate p=(12) q=1 noint; run;

proc gplot data=praktikum11; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z9*date;
where date>"01jan73"d;
run; *sesoonsus (aastane periood) tundub olevat olemas, samuti
tundub keskmine ajas ujuvat;
proc arima data=praktikum11;
identify var=z9 nlags=50;run; *autokorrelatsioonid ei kahane nulli,
kõik suured. leiame diferentsi;
identify var=z9(1) nlags=50;run; *perioodile vastavad
autokorrelatsioonid ei kahane nulli, kõik suured. leiame sesoonse
diferentsi;
identify var=z9(1 12) nlags=50;run; *esimene osaautokorrelatsioon
selgelt piiridest väljas, mittesesoonne osa paistab AR(1)-le
vastavat;
estimate p=1; run; *ei sobi. Tundub, et on olemas ka perioodile
vastavad sõltuvused, võib sobida sesoonne MA(2);
estimate p=1 q=(12 24) method=ml;run; *sobib hästi, kuid keskmine ei
ole oluline. Eemaldame selle;
estimate p=1 q=(12 24) noint method=ml;run; *igaks juhuks proovime
ka sesoonset arma(1,1) ja ar(2);
estimate p=(1)(12) q=(12) noint; run; *AIC põhjal parem kui eelmine;
estimate p=(1)(12 24) noint; run;*samuti sobib, kuid AIC põhjal
eelmine parem.;
quit;

proc gplot data=praktikum11; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z10*date;
where date>"01jan73"d;
run;
*paistab olevat statsionaarne, perioodilist komponenti ja trendi ei
paista olevat;
proc arima data=praktikum11;
identify var=z10 nlags=50;run;*madalat järku MA(1), sesoonne osa MA
(2);
estimate q=(1)(12 24) method=ml;run;*sobib hästi, keskmine on
ebaoluline;
estimate q=(1)(12 24) noint method=ml;run;
quit;
```

Aegridade analüüs
Praktikum nr. 12, 2012

ARIMA tüüpi mudelite sobitamine praktikast pärinevatele aegridadele. Prognoosimine teisendatud reale sobitatud mudeli korral

Seni oleme enamasti sobitanud mudeleid andmetele, mis olid eelnevalt mingi ARIMA tüüpi mudeli abil genereeritud. Praktikast pärinevate ridade korral võib juhtuda, et tegelikult ükski mudel ei sobi päris hästi ning osa valikuid tuleb teha ka lähtudes konkreetse rea sisulisest tähendusest.

Lisaks vaatleme veel põgusalt prognoosimist juhul, kui me oleme sobitanud teisendatud rea jaoks. Olgu esialgne rida $y_t,\ t = 1, \ldots, n$ ning olgu $z_t = f(y_t)$, kus f on mingi pööratav funktsioon (näiteks logaritm). Olgu z_{t+1} prognoos ajal t kujul $\hat{z}_{t+1}$, siis

$$z_{t+1} = \hat{z}_{t+1} + A_{t+1},$$

kus A_{t+1} on tsentreeritud ning me teame ka tema standardhälvet. Sel juhul suuruse y_{t+1} nihketa hinnanguks ei ole enamasti $f^{-1}(\hat{z}_{t+1})$, vaid seda väärtust tuleb korrigeerida. Konkreetne korrektsioon sõltub funktsioonist f ja A_{t+1} jaotusest. Meie vaatleme ainult logaritmimise juhtu; siis on nihketa prognoosiks

$$\hat{y}_{t+1} = e^{\hat{z}_{t+1}} E(e^{A_{t+1}})$$

ning $E(e^{A_{t+1}})$ on paljude jaotuste puhul teada. Näiteks normaaljaotuse eeldusel saame nihketa prognoosiks

$$\hat{y}_{t+1} = e^{\hat{z}_{t+1} + \frac{\sigma_A^2}{2}}.$$

Praktikumi ülesandeks on leida parimad ARIMA tüüpi mudelid kõigi esimeses ja teises praktikumis vaadeldud praktikast pärinevate andmestike puhul ning kinnisvaratehingute andmete korral ka uurida, kas eelnev logaritmimine ja seejärel mudeli sobitamine annab parema tulemuse, kui otse esialgsete andmete modelleerimine.

```
graafikud=function(z,mitu=30){
    layout(1:3) #jaotab graafikaekraani kolmeks võrdseks osaks
    plot(z) #esimesesse ossa aegrea graafik
    acf(z,mitu) #teiseks autokorrelatsioonid kuni etteantud nihete arvuni
    pacf(z,mitu) #osaautokorrelatsioonid
    layout(1) #graafikaaken edasisteks käskudeks jälle ühes osas
}

#töökataloogi seadmine
setwd("c:/portable/math/aegread2012/")
###########USA töötuse määr
#1)andmestiku sisselugemine: andmed hakkavad kuueteistkümnendast reast. Eraldajaks on ,
andmed2=read.table("unemployment2012.txt",na.strings=" ",sep=',',skip=16)
#2)arvulised tulemused on tulpades 2-13 (esimeses on aasta, viimases aga aasta keskmine)
rida2=c(t(andmed2[,2:13]))
#3)puuduvate vaatluste eemaldamine lõpust
rida2=rida2[1:(length(rida2)-5)]
#4)aegrea tegemine: tegemist on igakuiste andmetega, st 12 tulemust aasta kohta
töötus=ts(rida2,start=c(1948,1),frequency=12)
graafikud(töötus)
#paistab mittestatsionaarne, leiame diferentsi
graafikud(diff(töötus,1),50) #aastasele perioodile vastavad autokorrelatsioonid ei lähe nulli,
#leiame ka aastase diferentsi
graafikud(diff(diff(töötus,1),12),50)
#sesoonset osa vaadates võib olla üks sesoonne MA liige
#madalat järku kordajaid vaadates kas AR(3) või ARMA(1,2)
m1=arima(töötus,order=c(3,1,0), seasonal=list(order=c(0,1,1),period=12))
tsdiag(m1,40) #mudel sobib
#võrdlema ARIMA(1,1,2)x(0,1,1)_12
m2=arima(töötus,order=c(1,1,2), seasonal=list(order=c(0,1,1),period=12))
tsdiag(m2,40) #samuti sobib
m1
m2 #see on AIC järgi parem
###############Kinnisvara
#1)andmestiku sisselugemine. Eraldajaks on tabulaator, mis on kuulub vaikimisi määratud
eraldajate hulka
#seega eraldajat ei pea ütlema, kuid võib kirjutada sep="\t" (\t tähistab R-is tabulaatorit)
andmed1=read.table("kv015m.csv",na.strings="..")
#2)arvulised tulemused on kolmandas tulbas, mille nimi on V3
rida1=andmed1$V3
#3)puuduvate vaatluste eemaldamine lõpust
rida1=rida1[1:(length(rida1)-2)] #või rida1=rida1[!is.na(rida1)]
#4)ajas ümberpööramine- ei ole vaja
#5)aegrea tegemine. Tegemist on kvartaalsete andmetega
kv=ts(rida1,start=1997,frequency=4)
graafikud(kv) #ei ole statsionaarne
graafikud(diff(kv)) #varieeruvus tekitab kahtlusi eelduste paikapidavuses
#osaautokorrelatsioone vaadates võib proovida sobitada AR(2) diferentsitud reale (ehk
ARIMA(2,1,0))
m2.1=arima(kv,order=c(2,1,0))
tsdiag(m2.1,30) #võib lugeda sobivaks
#proovime ka logaritmimist
logkv=log(kv)
graafikud(logkv)#mittestatsionaarne
graafikud(diff(logkv)) võib lugeda statsionaarseks
m2.2=arima(logkv,order=c(2,1,0))
tsdiag(m2.2,30)#enam-vähem sobib
```

```
m2.2
sum(residuals(m2.1)^2) #vaatleme prognoosivigade ruutude summat
ennustus2=exp(logkv-residuals(m2.2)+m2.2$sigma2/2)#teise meetodi korrigeeritud prognoosid
sum((kv-ennustus2)^2) #prognoosivigade ruutude summa teise meetodi puhul
#otse modelleerimine oli parem
```

Aegridade analüüs
Praktikum nr. 13, 2012

ARIMAX tüüpi mudelite sobitamine.

Sel korral kasutame praktikumis R tarkvara. Vajalikud andmed paiknevad failis `prax13.txt`.

Kõigepealt tuletame meelde, kuidas kasutada deterministlikku regressorit, st sobitame mudeleid kujul

$$Z_t = \beta_0 + \beta_1 X_t + \varepsilon_t,$$

kus $X_t = t$ ning ε_t on mingi ARIMA tüüpi protsess. Mudeli sobitamise protsess on järgmine:

i) Sobitame reale Z_t lineaarse regressioonimudeli regressoriga X_t ning leiame selle mudeli prognoosivead.

ii) Teeme kindlaks, mis tüüpi ARIMA mudel prognoosivigadele sobib

iii) Sobitame vastava ARIMAX mudeli andmetele ning testime saadud mudeli jääkide sõltumatust. Vajadusel muudame vigade jaoks valitud ARIMA mudeli kuju.

Kindlasti tasuks võrrelda leitud ARIMAX mudelit parima ARIMA mudeliga, mis regressoreid ei kasuta.

1) Sobitage lineaarse regressoriga ARIMAX mudel ridadele z1 ja z2. Tehke kindlaks, kas regressori lisamine on õigustatud. Meeldetuletus: lineaarse mudeli sobitamiseks on R-is käsk `arima(rida,xreg=regressor)`; regressioonimudelile vastavad väärtused saab kätte käsuga `residuals(mudel)`, kus mudel on eelneva käsu `lm` tulemus. Kui prognoosivigade analüüsimise tulemusel on leitud sobiv ARIMA mudel vigade jaoks, siis ARIMAX mudeli sobitamiseks saab kasutada `arima` käsku, kus lisaks tavalistele parameetritele on ette antud ka regressor parameetriga `xreg`, näiteks
   `arima(z,order=c(1,0,0),xreg=aeg)`,
   kus `aeg` sisaldab näiteks vaatluste järjekorranumbreid. Leidke samuti ennustused järgneva 5 perioodi jaoks; arimax mudelis tuleb selleks `predict` käsus anda regressorite vastavad väärtused parameetriga `newxreg`.

Järgnevalt vaatleme juhtu, kus regressoriteks on mingid täiendava aegrea andmed. Sobiva ARIMAX mudeli leidmise protseduur on sel juhul sama nagu varem.

2) Vaatleme juhtu, kus meil on teada, et rea $z3$ regressoriks peaks sobima rea $x3$ nihkega 6 väärtused. Seda, kuidas tegelikult sobivat nihet leida, vaatleme hiljem, kuid etteruttavalt võib öelda, et vähemalt statsionaarsete ridade korral annab sobiva nihke kohta informatsiooni ristkorrelatsioon (mida arvutatakse R-is käsuga `ccf`). Kuna esimese 6 z3 väärtuse jaoks meil regressoreid ei ole, jätame need vaatluse alt välja:
   `z31=window(Z3,start=7)`.
   Seejärel moodustame regressorite vektori $x3$ väärtustest (viimased 6 jäävad välja), leiame vastava lineaarse regressioonimudeli vigade põhjal sobiva ARIMA mudeli kuju ning seejärel sobitame ARIMAX mudeli valitud regressoritega. Pärast leitud mudeli sobivuse kindlakstegemist leidke ka prognoos järgmise kuue ajaperioodi jaoks.

3) Mitme regressori olemasolul tegevuste iseloom jääb samaks. Vaatleme juhtu, kus rea $z4$ jaoks kasutatakse regressoritena rea $x41$ nihkega 4 väärtusi ja rea $x42$ nihkega 2 väärtusi. Kõige suuremaks erinevuseks eelneva juhuga võrreldes on see, et `arima` käsus antakse regressorid ette kujul `xreg=data.frame(x1,x2)`, kus `x1,x2` on regressoritele vastavad aegread; samasugune süntaks on ka ennustamisel parameetri `newxreg` etteandmisel. Leidke sobiv mudel ja ennustage selle põhjal z3 järgnevad 2 väärtust.

Lõpuks vaatleme lühidalt küsimust, kuidas kindlaks teha, milliseid antud regressoraegrea X_t nihkeid huvipakkuva aegrea Z_t väärtuste ennustamisel kasutada. Lihtsuse mõttes eeldame, et mõlemad aegread on statsionaarsed ning nullkeskmisega, vastasel korral tuleb mõlemast leida diferentsid nii, et need eeldused oleks rahuldatud.

Eelnevalt nägime (nt ülesanne 2), et ristkorrelatsioonid võivad anda sobiva nihke kindlakstegemise jaoks olulist infot. Samas aga ei anna lihtsalt ridade Z_t ja X_t ristkorrelatsioonid tavaliselt kuigi selget infot selle kohta, mitu erinevat X_t nihet tuleks kasutada, et parimat ennustust leida. Oluliseks võtteks täpsema info saamiseks on nn eelvalgendamine (inglise keeles *prewhitening*). See seisneb selles, et kõigepealt leitakse sobiv ARIMA tüüpi mudel regressorrea X_t jaoks ning leitakse mudeli jäägid $A_{x,t}$. Need jäägid on mudeli kehtivuse korral sõltumatud ja sama jaotusega (ehk nn valge müra; sellest ka tehnika nimetus). Seejärel leitakse jäägid $A_{z,t}$, mis vastavad **sama** mudeli kasutamisele suuruste Z_t ennustamiseks ning arvutatakse leitud jääkide $A_{x,t}$ ja $A_{z,t}$ vahelised ristkorrelatsioonid. Osutub (täpsemalt räägitakse sellest loengus), et nende korrelatsioonide abil saab kindlaks teha, milliseid X nihkeid tuleks Z väärtuste ennustamisel kasutada. Erijuhul, kui nullist erinevaid ristkorrelatsioone on lõplik arv (ja need vastavad X minevikuväärtuste kasutamisele), siis tuleks kasutada täpselt selliste nihetega X_t väärtusi regressoritena.

- Kasutame eelpool kirjeldatud tehnikat, et leida parim prognoosimudel rea $z5$ tulevikuväärtuste ennustamiseks selle rea ja rea $x5$ teadaolevate väärtuste abil. Selleks leidke reale $x5$ sobiv ARIMA tüüpi mudel (olgu selle nimeks `m5x`). Leitud mudeli jäägid on kättesaadavad kujul `residuals(m5x)`. Sama mudeli kasutamisel rea $z5$ ennustamiseks tekkivate jääkide leidmiseks sobiama samade kordajatega mudeli sellele reale käsuga

  ```
  m5z=arima(z5,order=sama_mis_m5x,fixed=m5x$coef)
  ```

  Edasi leiame tekkinud jäägid ja nende ristkorrelatsiooni m5x jääkidega. Seejärel moodustame nullist erinevatele kordajatele vastavate regressoritega ARIMAX mudeli nii nagu eelnevalt, testime selle sobivust ja ennustame tulevikuväärtusi.

```
graafikud=function(z,mitu=30){
    layout(1:3) #jaotab graafikaekraani kolmeks võrdseks osaks
    plot(z) #esimesesse ossa aegrea graafik
    acf(z,mitu) #teiseks autokorrelatsioonid kuni etteantud nihete arvuni
    pacf(z,mitu) #osaautokorrelatsioonid
    layout(1) #graafikaaken edasisteks käskudeks jälle ühes osas
}
read=read.csv("h:/aegread2012/prax13.txt")
z1=ts(read$z1)
graafikud(z1)
n=length(z1)
aeg=1:n
trend=lm(z1~aeg)
trendita=z1-predict(trend)
graafikud(trendita)
graafikud(diff(trendita)) #kõik autokorrelatsioonid ja osaautokorrelatsioonid veapiirides
m1.1=arima(z1,c(0,1,0),xreg=aeg)
m1.1
tsdiag(m1.1,30)
#prognoosimine
predict(m1.1,5,newxreg=data.frame(aeg=(n+1):(n+5)))
#alternatiiv: ilma regressorita
graafikud(diff(z1))
m1.2=arima(z1,order=c(0,1,0)) #parem mudel AIC järgi
predict(m1.2,5)



z3=ts(read$z3)
x3=ts(read$x3)
ccf(z3,x3)
z=window(z3,start=7)#jätame ära esimesed kuus vaatlust#start peab andma seitsmenda vaatluse
aja
#nt start=c(1980,5)
x=window(x3,end=n-6)
trend=lm(z~x)
trendita=z-predict(trend)
graafikud(trendita)
m3.1=arima(z,order=c(0,0,2),xreg=x)
tsdiag(m3.1,30)
####
z4=ts(read$z4)
x41=ts(read$x41)
x42=ts(read$x42)
z=window(z4,start=5)
x1=window(x41,end=n-4)
x2=window(x42,start=3,end=n-2)
trend=lm(z~x1+x2)
trendita=z-predict(trend)
graafikud(trendita)
m4.1=arima(z,order=c(1,0,0),xreg=data.frame(x1=x1,x2=x2))
```

Aegridade analüüs
Praktikum nr. 14, 2012

ARIMAX tüüpi mudelite sobitamine. Ülekandefunktsiooni kuju määramine

Vaatleme juhtu, kus meil on kaks protsessi Z_t ja X_t, millele vastavate aegridade väärtused on meil olemas. Lihtsuse mõttes eeldame, et mõlemad protsessid on statsionaarsed ning tsentreeritud (vastasel juhul võib proovida leida mõlemast sobivat järku diferentsid ja maha lahutade nende keskmised, et saada soovitud omadustega ridu). Meie eesmärgiks on kindlaks teha, milliseid X minevikuväärtuseid (ja võib-olla ka Z minevikuväärtuseid) regressoritena kasutada nii, et saada võimalikult häid ennustusi protsessi Z jaoks. Täpsemalt, vaatleme ARIMAX mudelit kujul

$$Z_t = \beta_0 + \sum_{i=b}^{\infty} \beta_i X_{t-i} + \varepsilon_t, \tag{1}$$

kus $b \geq 1$ ja suurused ε_t vastavad mingile $ARMA$ protsessile

$$\phi(B)\varepsilon_t = \theta(B)A_t,$$

kusjuures eeldame, et suurused A_t on sõltumatud ka protsessi X_t väärtustest. Sel juhul on ka suurused ε_t sõltumatud suurustest X_t. Selleks, et parameetreid oleks lõplik arv ning et ennustamiseks kasutataks ainult X_t minevikuväärtuseid, otsime sobivat mudelit selliste hulgast, kus funktsioon

$$\beta(x) = \sum_{i=b}^{\infty} \beta_i x^i$$

on esitatav lõpliku arvu parameetrite abil kujul

$$\beta(x) = x^b \frac{v(x)}{\delta(x)} = x^b \frac{\sum_{i=0}^{s} v_i x^i}{1 - \sum_{i=1}^{r} \delta_i x^i}.$$

Sellise mudeli võib kirjutada ka kujul

$$Z_t = \sum_{i=1}^{r} \delta_i Z_{t-i} + \sum_{i=0}^{s} v_i X_{t-b-i} + \eta_t,$$

kus suurused η_t vastavad ARMA protsessile kujul

$$\phi(B)\eta_t = \delta(B)\theta(B)A_t.$$

Funktsiooni $\beta(x)$ nimetatakse ülekandefunktsiooniks, kuna ta kirjeldab, kuidas X omandatud väärtused mõjuvad ehk kanduvad üle suuruste Z väärtustele. Mudeli sobitamise etapid on järgmised:

1. Leiame hinnangud suurustele β_i. Selleks sobitame kõigepealt ARIMA mudeli reale X ning täpselt samade kordajatega mudeli reale Z. Seejärel leiame X jääkide ja Z jääkide vahelised ristkorrelatsioonid, mis ongi hinnangud β_i väärtustele.

2. Suuruste β_i hinnangute põhjal määrame kindlaks sobiva nihke b (mis vastab esimese nullist erineva β indeksile) ning kasutades teadmist sellest, kuidas erinevate r ja s väärtuste korral peaks suurused β teoreetiliselt käituma, leiame hinnangud ka parameetritele r ja s. Kui β_i hinnangutest on ainult mõni nullist erinev, siis s väärtuseks võtame nullist erinevate kordajate arv -1. Kui kordajatest on mitu nullist erinevat, kuid nad hakkavad kahanema vastavalt mingi arvuga korrutamisele, siis tasub proovida $r = 1$ ning s väärtus on määratud eelnevate nullist erinevate kordajate arvuga.

3. Leiame sobiva mudeli vigade ε_t jaoks
4. Hindame mudeli parameetreid suurima tõepära meetodil
5. Kontrollime jääkvigade sõltumatust
6. arvutame prognoosid (kuni b ajaperioodi ette).

Eesmärgiks on sobitada ülekandefunktsiooni mudel etteantud aegreale etteantud sisendaegrida kasutades nii R kui SAS tarkvara vahendusel. Andmed on kodulehel failides `prax15.txt`. Alustame SAS tarkvarast.

i) Hilisema võrdluse huvides leiame reale Z_t sobiva ARIMA tüüpi mudeli.

ii) Ülekandefunktsiooni kuju jaoks sobivate kordajate määramiseks asutame eelvalgendamise tehnikat. Selleks sobitame reale X kõigepealt ARIMA mudeli

iii) Leiame eelvalgendamisele vastavad ristkorrelatsioonid. SAS tarkvaras toimub see automaatselt, kui regressorile on eelnevalt ARIMA tüüpi mudel sobitatud. Vajalikuks käsuks on `identify var=Z crosscorr=X`. Ristkorrelatsioonide põhjal teeme eeldused b, s ja r kohta.

iv) Vigade jaoks sobiva mudeli kindlakstegemiseks eemaldame ülekandefunktsiooniga määratud sõltuvuse Z ja X vahel. Selleks sobiv käsu kuju on
`estimate input=(b$(1 ... s)/(1 ... r) X) plot;`,
kus kolm punkti tähistab lihtsalt kõikide vahepealsete arvude loetelu. Plot käsu tulemusena väljastatakse tekkinud jääkide autokorrelatsioonid ja osaautokorrelatsioonid, mille põhjal saab valida vigade jaoks sobiva ARIMA mudeli.

v) Sobitame lõpliku mudeli, täiendades eelnevat käsku sobivate p ja q väärtustega (ja jättes lõpust ära plot käsu). Kui saime sobiva mudeli (prognoosivigade põhjal), siis võib siin tegevuse lõpetada; vastasel korral aga tuleb eelnevas midagi täiendada (näiteks r ja s väärtusi katsuda muuta).

R tarkvaraga sobitamise protseduur on kirjeldatud praktikumis nr 13. Täienduseks niipalju, et vigade jaoks sobiva mudeli otsimisel ja lõpliku mudeli leidmisel tuleb arima käsus regressorite hulka lisaks X kasutatavate nihete lugeda ka $Z_{t-1}, \ldots, Z_{t-r}$.

```
*praktikumi nr. 14 skript SAS-is;
data prax14; *loeme andmed ajutisse andmestikku;
infile "h:/aegread2012/prax14.txt" firstobs=2 delimiter=" " dsd;
keep date x z;
retain abi 1;
input aasta x z;
date=abi;
output;
abi=abi+1;
run;
proc gplot data=prax14; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot x*aeg;
run;
proc gplot data=prax14; /*joonistame graafiku*/
symbol i=spline v=dot h=0.3;
plot z*date;
run;
proc arima data=prax14;
identify var=z;run;
estimate p=1 q=1; run;*ei sobi;
estimate q=3; run;*ei sobi;
estimate p=1 q=2;run;*paistab hästi sobivat, vea standardhälve 1.2;
identify var=x; run;
estimate q=1; run;
identify var=z crosscorr=x; run; *kas s=3,r=0 või s=1,r=1;
estimate input=(2$(1 2 3)X) plot;run; *esimene hüpotees AR(1);
estimate input=(2$(1 2 3)X) p=1; run; *ei sobi;
estimate input=(2$(1 2 3)X) p=1 q=1; run; *ikka ei sobi;
estimate input=(2$(1 2 3)X) p=4; run; *muidu sobib, kuid
ristkorrelatsioonid ei vasta juhuslikkusele;
*seega ei ole kogu X-s olev info kasutusel;
*teine versioon, s=1,r=1;
estimate input=(2$(1)/(1)X) plot;run;*esimene hüpotees AR(1);
estimate input=(2$(1)/(1)X) p=1;run;*ei sobi hästi;
estimate input=(2$(1)/(1)X) p=1 q=1;run;*ikka ei sobi;
estimate input=(2$(1)/(1)X) p=3;run;*paistab sobivat, ka
ristkorrelatsioonid on enam-vähem piirides;
*prognoosivea standardhälve 0.31, seega on tunduvalt paremaks
läinud;
```

Aegridade analüüs
Praktikum nr. 15, 2012

GARCH tüüpi mudelite sobitamine

Sageli on majanduslike aegridade puhul võimalik täheldada seda, et suuremate võnkumistega rahutumad perioodid vahelduvad suhteliselt stabiilsete perioodidega ning sageli jääb see efekt alles ka prognoosivigade puhul pärast parima ARIMA tüüpi mudeli sobitamist. See aga tähendab, et vähemalt prognoosivigade arvutamise tulemused ei ole usaldusväärsed, kuna seal lähtutakse üleüldisest keskmisest vigade standardhälbest, mis huvipakkuva hetke jaoks võib olla liiga suur (kui parajasti on tegemist rahulikuma perioodiga) või liiga väike (kui on tegemist rahutuma perioodiga).

Lihtsalt vaatluse abil ei ole alati lihtne eristada juhuslikult tekkivaid sõltumatute juhuslike suuruste keskmiselt suuremate ja keskmiselt väiksemate rühmade teket sellisest, kus sellised rühmad on seotud muutuva varieeruvusega. Samas ARIMA mudelite sobitamisel võime teha lihtsa testi: vaatleme prognoosivigade ruutude autokorrelatsioone ja osaautokorrelatsioone. Kui on tegemist mudeliga, kus häiritused A_t on sõltumatud, on ka ruutude teoreetilised autokorrelatsioonid ja osaautokorrelatsioonid nullid ning seega peaks empiirilised autokorrelatsioonid ja osaautokorrelatsioonid jääma vastavatesse veapiiridesse. Kui see nii aga ei ole, siis ei pruugi leitud mudel olla sugugi parim ning kindlasti tuleb veapiiridesse suhtuda suure ettevaatusega.

Üheks mudelite klassiks, mille korral varieeruvus muutub ajast sõltuvalt sissetulnud häiritustest, on ARIMA mudelid GARCH (*Generalised Autoregressive Conditional Heteroskedasticity*) häiritusega kujul

$$\begin{aligned} Z_t &= \sum_{i=1}^{p} \phi_i Z_{t-1} + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \\ \varepsilon_t &= \sqrt{\sigma_t} A_t, \\ \sigma_t^2 &= \omega + \sum_{i=1}^{q_1} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p_1} \beta_i \sigma_{t-i}^2. \end{aligned}$$

Kui parameeter σ_t sõltub ainult eelnevatest häiritustest (st kui β-dega liikmeid pole), siis nimetatakse seda mudelit ARCH mudeliks.

Mudeli sobitamise protseduur on järgmine:

1. Leiame andmestikule parima ARIMA tüüpi mudeli
2. Vaatleme prognoosivigade ruutude autokorrelatsioone ja osaautokorrelatsioone. Kui mõned autokorrelatsioonid on veapiiridest selgelt väljas, siis on mõistlik katsetada GARCH tüüpi mudelitega. Esialgseks hüpoteesiks parameetri q_1 osas võib võtta nullist erinevate osaautokorrelatsioonide arvu; samas tasub vaadelda ka madalamat järku GARCH mudeleid.
3. Mudeli loeme sobivaks siis, kui nn. normaliseeritud prognoosivead (st vead, mis on jagatud hetkele vastava σ_t väärtusega) ja nende ruudud ei ole oluliselt korreleeritud (st Ljung-Box testi p-väärtused on nii vigade kui ka vigade ruutude korral piisavalt suured).

Praktikumis kasutame R tarkvara, täpsemalt paketti `fGarch`. Mudeli sobitamine toimub käsuga `mudel=garchFit(rida~arma(p,q)+garch(q1,p1),data=rida)` ja olulisi sobitamise tulemusi, sealhulgas sobivustestide tulemusi näeb käsuga `summary(mudel)`. Ülesandeks on leida sobivaimad mudelid andmestikus `prax15.dat` olevatele aegridadele. Pärast nende ridade uurimist tuleks uurida GARCH mudelite kasutamise vajadust praktikumis 1 sisseloetud praktiliste aegridade korral.

```
graafikud=function(z,mitu=30){
    layout(1:3) #jaotab graafikaekraani kolmeks võrdseks osaks
    plot(z) #esimesesse ossa aegrea graafik
    acf(z,mitu) #teiseks autokorrelatsioonid kuni etteantud nihete arvuni
    pacf(z,mitu) #osaautokorrelatsioonid
    layout(1) #graafikaaken edasisteks käskudeks jälle ühes osas
}

#praktikum 15
#pakett fGarch peab olema paigaldatud
#kui ei ole admin õigusi, siis tuleks teha endale kataloog
#oma kettaruumi (näiteks kataloog h:/rlibs) ja siis anda allpool toodud käsud

#kui pakett on installeeritud ebastandardsesse kohta, siis peab ütlema R-le kus
#see on
#.libPaths("h:/rlibs")
#install.packages("fGarch",lib="h:/rlibs") #see käsk tuleb anda ainult üks kord
#laeme paketi fGarch
library("fGarch")

#töökataloogi seadistamine
setwd("c:/portable/math/aegread2012")
#ridade sisselugemine
andmed=read.table("prax15.dat", header=T,sep=";")
head(andmed)
z1=ts(andmed$z1) #jätan ajainfo panemata, see ei ole praegu oluline
graafikud(z1)#paistab statsionaarne, hüpotees AR(2)
m1.1=arima(z1,order=c(2,0,0))
tsdiag(m1.1,20)
#sobib, kuid varieeruvus paistab muutuvat "lainetena"
#uurime prognoosivigade ruutusid
vearuut=residuals(m1.1)^2
graafikud(vearuut)
#selge sõltuvus, paistab olevat AR(1) tüüpi sõltuvus
m1.2=garchFit(z1~arma(2,0)+garch(1,0),data=z1)
summary(m1.2)#sobib hästi, standardiseeritud vead ja vearuudud ei sisalda olulisi
#autokorrelatsioone,Ljung-Box testide p-väärtused on suured.
predict(m1.2,n.ahead=10,plot=T) #prognoosid
#####
#rida z2
z2=ts(andmed$z2)
graafikud(z2)#paistab olevat statsionaarne,MA(2) või ARMA(1,1)
m2.1=arima(z2,order=c(0,0,2))
tsdiag(m2.1,20)#sobib hästi, kuid jällegi on andmetes lained varieeruvuses
vearuut=residuals(m2.1)^2
graafikud(vearuut) #jällegi selge sõltuvus eelmisest
m2.2=garchFit(z2~arma(0,2)+garch(1,0),data=z2)
summary(m2.2)#ei tundu sobivat
#võib-olla siiski vaja teist liiget?
m2.2=garchFit(z2~arma(0,2)+garch(2,0),data=z2)
summary(m2.2)
#võime lugeda sobivaks
predict(m2.2,n.ahead=10,plot=T) #prognoosid
###########3
#rida z3
z3=ts(andmed$z3)
graafikud(z3)#paistab olevat statsionaarne,ARMA(1,1) või AR(2)
```

```
m3.1=arima(z2,order=c(2,0,0))
tsdiag(m3.1,30) #ei sobi
m3.1=arima(z2,order=c(1,0,1))
tsdiag(m3.1,30) #sobib hästi
vearuut=residuals(m3.1)^2
graafikud(vearuut) #jällegi selge sõltuvus eelmisest
m3.2=garchFit(z3~arma(1,1)+garch(1,0),data=z3)
summary(m3.2) #ei sobi,standardiseeritud prognoosivigade ruudud ei vasta mittekorreleerituse
nõuetele
m3.2=garchFit(z3~arma(1,1)+garch(1,1),data=z3)
summary(m3.2)#ikka ei sobi, kuid AIC parem
#otsin edasi arch tüüpi mudelit
m3.2=garchFit(z3~arma(1,1)+garch(2,0),data=z3)
summary(m3.2) #ikka ei sobi
m3.2=garchFit(z3~arma(1,1)+garch(3,0),data=z3)
summary(m3.2) #sobib väga hästi
predict(m3.2,n.ahead=10,plot=T)
```