Identifiointiprosessi

Alustavia kokeita

- Koesuunnittelu, identifiointikoe
- Mittaustulosten / datan esikäsittely
- Ei-parametriset menetelmät:

  \- Transientti-, korrelaatio-, taajuus-, Fourier- ja spektraalianalyysi
  => askel-, impulssi- ja taajuusvaste, ajatuksia esim. kertaluvuista

- Parametrien estimointi – **mallirakenteen (rakenneparametrit) valinta**
- Mallin arviointi:

  \- Simulointi, navat, erilaiset vasteet, jne...

- Mallin validointi:

  \- Parametrien merkitsevyys, residuaalianalyysi, jne...

Mallirakenteen valinta

- Rakenteellinen – black box?
  - Rakenteellinen vaikea tuottaa, mutta usein vähemmän estimoitavaa => ”tarkempia” tuloksia
- Black-box: mallirakenne & kertaluvut?
  - Estimointimenetelmä (ARX –PNS, muut vaativat iteratiivisia)
  - Kertaluvut?
    - ARMA-, ARIMA-mallit oma tarinansa, ks. Mat-2.3128
  - Niukkuusperiaate (parsimony principle): valitse kahdesta tasaväkisestä mallista se, jossa on vähemmän parametreja
  - Miksi? Voidaan osoittaa, että ylimääräiset parametrit lisäävät ennustevirheen varianssia
- Käyttötarkoitus; mallinnusosaaminen; algoritminen kompleksisuus;...

Mallirakenteiden vertaaminen

- Helppoa! verrataan vaan jäännösneliösummia!
- Vai hetkinen... kun lisään tarpeeksi monta parametria, siitä tulee nolla!??
  - Tietyn rajan jälkeen malli alkaa sovittautua datan kohinaan
- Kupletin juoni ei olekaan estimointidataan sovittaminen vaan ilmiön kuvaaminen
- Ristiinvalidointi:
  - Mallin suorituskyvyn (esim. ennustuskyvyn) arviointi riippumattomalla datalla
- Kvantitatiivinen analyysi (vrt. lineaariset regressiomallit, ks. esim. Mat-2.2104):
  - Tilastollinen analyysi: pieneneekö hyvyysfunktio merkitsevästi kun siihen lisätään parametri?
  - Informaatiokriteerit: sakotetaan parametrien lisäämisestä
  - ”Selittäjien” kovarianssin singulaarisuus

F-testi

- Olkoot M_1 ja M_2 kaksi hierarkkista mallirakennetta (ts. M_1 saadaan M_2:sta poistamalla parametri/parametreja)
- Ol. että kumpikin mallirakenne kuvaa systeemin niin, että ennustevirheet ovat valkoista kohinaa
- Tällöin testisuure $\chi = N(V_N^1 - V_N^2)/V_N^2$ on asymptoottisesti χ^2- jakautunut vapausteilla p_2-p_1
  - p_i= parametrien lkm mallissa i
  - V_N^i hyvyyskriteerin arvo mallilla i
  - N havaintojen lukumäärä
- Idea: poikkeaako testisuure nollasta tilastollisesti merkitsevästi kun siirrytään mallista 1 malliin 2?

Informaatiokriteerien idis

- Hyvyyskriteerin optimiarvo vs. parametrien lukumäärä d
- Tyypillinen funktionaalinen muoto: $W_N = \beta(d,N)V_N$
  - V_N = parametreistä (d kpl) riippuva ennustevirheiden neliösumma
  - β kasvaa kun d kasvaa ja vähenee kun N kasvaa
- Minimoi $W_n$in d:n ja parametrien suhteen

  =>

  Valitaan malli, jolle W_N on pienin

Informaatiokriteereitä

- Akaiken informaatiokriteeri (1972):
  $AIC=(1+2d/N)V_N$
  – Kullback-Leibler –etäisyysdellä (informaatioetäisyys) voidaan mitata kahden tn-jakauman etäisyyttä
  – AIC:n minimoiva malli minimoi mallin ja systeemin informaatioetäisyyden
- Final Prediction Error (Akaike, 1969):
  $FPE=(1+d/N)/(1-d/N)V_N/N$
  – Keskimääräinen estimaatti ennustevirheen varianssille, kun estimaatti lasketaan muulla kuin mallin identifointiin käytetyllä datalla
- Rissasen minimipituuskriteeri (1978):
  $MDL=(1+2d\ \log N/N)V_N$
  – Kuvaa tilaa, joka tarvitaan mallin sisältämän informaation (mallin parametrit + ennustevirheet) tallentamiseksi pienimpään mahdolliseen tilaan

Hankel-matriisin testaus 1/2

- N:nnen kertaluvun järjestelmän impulssivaste h(t) riippuu lineaarisesti n:stä edellisestä arvosta h(t-1),...,h(t-n)
- Esim. $y(t)=-a_1y(t-1)-a_2y(t-2)+b_1u(t-1)+b_2u(t-2)$
- u(0)=1,u(i)=0 kaikilla i >0; y(0)=y(-1)=y(-2)=0

$\Rightarrow$ Vaste $h(1)=b1$; $h(2)=-a_1h(1)+b2$; $h(3)=-a_1h(2)-a_2h(1)$;...;$h(t)=-a_1h(t-1)-a_2h(t-2)$ kaikilla t>2

Määritellään Hankel-matriisi:

$$H(l,k)=\begin{bmatrix} h(k) & h(k+1) & \dots & h(k+l-1) \\ h(k+1) & h(k+2) & \dots & h(k+l-2) \\ \vdots & \vdots & \ddots & \vdots \\ h(k+l-1) & h(k+l-2) & \dots & h(k+2l-2) \end{bmatrix}$$

Hankel-matriisin testaus 2/2

- Järjestelmän kertaluku n => H(l,k):n rangi = n kaikilla $l \geq n$
- Testaus: det (H(l,k))=0 kun l>n
- Kohinaa => testataan perättäisten determinanttien suhteita: $D_l = |\det(H(l,k))| / |\det(H,l+1,k))|$
- Determinantit voidaan keskiarvottaa yli k:n
- l=n maksimoi D_l:n
- Käytännössä:

  Transsienttianalyysi => impulssi/askelvaste => Hankel-matriisi => D_l => n

Tulomomenttimatriisin testaus

- Olkoot järjestelmän kertaluku n ja u pe vähintään n+1
- Määr. U(l):

$$U(l)=\begin{bmatrix} y(l) & \cdots & y(1) & | & u(l) & \cdots & u(1) \\ y(l+1) & \cdots & y(2) & | & u(l+1) & \cdots & u(2) \\ \vdots & \ddots & \vdots & | & \vdots & \ddots & \vdots \\ y(N) & \cdots & y(N-l+1) & | & u(N) & \cdots & u(N-l+1) \end{bmatrix}$$

- Hae suurin l = l* jolla U(l) vielä täyttä (rivi)rangia => n=l*
  - hae suurin l=l* jolla $U(l)^T U(l)$ (tulomomenttimatriisi) täyttä rangia => n=l*
- Käytännössä:
  - 1) Tee koe sopivalla herätteellä
  - 2) Etsi l* eli n