Ali je signal stacionaren?

Uvod

Na obravnavi konkretnega primera biološkega signala si bomo ogledali dva pristopa k ovrednotenju stacionarnosti naključnega procesa, ki temeljita na analizi ene same vzorčne funkcije tega procesa. Teoretično ozadje statističnih metod, ki jih pri tem uporabimo, je preobsežno za obravnavo na tem mestu. Podrobnejšo razlago lahko najdete v učbenikih in priročnikih, ki obravnavajo verjetnost in statistiko ali zahtevnejšo obdelavo naključnih signalov.

Teoretično ozadje

Stacionarnost signalov

Ko se lotimo obdelave neznanega signala, ki ni deterministične narave, se moramo na začetku vedno vprašati, ali se statistične lastnosti signala s časom spreminjajo ali ne. Od odgovora na to vprašanje je odvisna izbira orodij za obdelavo signala. Če namreč nestacionaren signal obravnavamo kot stacionarnega, so rezultati analize lahko nepravilni ali vsaj zavajajoči. Na nek način moramo torej oceniti stacionarnost signala. To je še posebej pomembno, če želimo iz obdelave časovnega signala (vzorčne funkcije) dobiti podatke o lastnostih mehanizma (naključnega procesa), ki je ta signal generiral. V splošnem je nemogoče preveriti, ali je signal stacionaren v strogem smislu, lahko pa preverjamo stacionarnost v širšem smislu. Obstajata dva splošna pristopa k ugotavljanju stacionarnosti v širšem smislu. S prvim testom, ki ga tu obravnavamo, imenujmo ga test I, opravimo neposredno statistično analizo podatkov. Pri drugem testu (test II) pa dejansko ugotavljamo ponovljivost posameznih statističnih veličin signala.

Test I: ocenjevanje stacionarnosti v širšem smislu

V angleški literaturi ta test imenujejo *the runs test* ali *the run test* [1,2]. Glede na pomen tega imena bi poslovenjeni naziv lahko bil *test potekov* ali *test tendenc* ali kaj podobnega. S tem testom ocenjujemo časovno nespremenljivost statističnih veličin (momentov) signala. Postopek lahko izvedemo za poljuben moment, v primeru preverjanja stacionarnosti signala v širšem smislu na primer preverjamo časovno nespremenljivost srednje vrednosti in variance. Ne glede na moment, ki ga obravnavamo, je postopek enak in je opisan v naslednjih korakih.

1. Signal dolžine N vzorcev razdelimo v urejeno zaporedje k neprekrivajočih se enako dolgih segmentov dolžine M. Če N ni celoštevilsko deljiv s k, pač izpustimo tistih nekaj vzorcev signala iz nadaljne obravnave. Vrednost k mora biti dovolj velika, da je statistični test smiselen (na primer $k \geq 10$) in hkrati dovolj majhna, da je znotraj vsakega segmenta dovolj točk za smiselen izračun ocen momentov.

2. Za vsakega od k segmentov izračunamo iz pripadajočih M vzorcev oceno momenta, ki nas zanima, v našem primeru za srednjo vrednost $\hat{\mu}_i$in varianco $\hat{\sigma}_i^2$.

$$\begin{aligned} & & \hat{\mu}_i &= \frac{1}{M}\sum_j^M x_j & & \\ i\text{-ti segment} \rightarrow & & & & ; 1 \leq i \leq k & \quad (1) \\ & & \hat{\sigma}_i^2 &= \frac{1}{M-1}\sum_j^M (x_j - \hat{\mu}_i)^2 & & \end{aligned}$$

3. Dobili smo dva *urejena* niza, ki vsebujeta vsak po k zaporednih ocen srednje vrednosti in variance. Za vsakega od tako dobljenih *urejenih* nizov ocen momentov izračunamo njegovo mediano in nato tvorimo dva nova *urejena* niza, ki predstavljata odstopanje posamezne ocene momenta od mediane vseh ocen:

$$\begin{aligned} \{\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_k\} \Rightarrow \hat{\mu}_{med} \Rightarrow \{d_{\mu 1}, d_{\mu 2}, \cdots, d_{\mu k}\};\ d_{\mu i} = \hat{\mu}_i - \hat{\mu}_{med} \\ \{\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_k^2\} \Rightarrow \hat{\sigma}_{med}^2 \Rightarrow \{d_{\sigma 1}, d_{\sigma 2}, \cdots, d_{\sigma k}\};\ d_{\sigma i} = \hat{\sigma}_i^2 - \hat{\sigma}_{med}^2 \end{aligned} \quad (2)$$

4. Za vsako od urejenih sekvenc $\{d_{\mu i}\}$ in $\{d_{\sigma i}\}$ preštejemo, kolikokrat se spremeni predznak pripadajočih elementov. Tako dobljeno število povečano za 1 se v angleščini imenuje *the number of runs* (poslovenjeno morda število potekov ali število tendenc) in je ključno za ugotavljanje časovne nespremenljivosti momenta. Vsak obravnavani moment ima svoje število tendenc, ki je vedno med 1 in k. V našem primeru naj bosta števili tendenc $n_{T\mu}$ in $n_{T\sigma}$:

$$\begin{aligned} n_{T\mu} &= 1 + \text{število sprememb predznaka v } \{d_{\mu 1}, d_{\mu 2}, \cdots, d_{\mu k}\} \\ n_{T\sigma} &= 1 + \text{število sprememb predznaka v } \{d_{\sigma 1}, d_{\sigma 2}, \cdots, d_{\sigma k}\} \end{aligned} \quad (3)$$

   Obstajajo tabele, ki za različne vrednosti k povejo, znotraj katerih meja mora ležati *število tendenc* za izbrani moment, da lahko z določeno mero zaupanja trdimo, da je izbrani moment *časovno nespremenljiv*. Uporaba tabele 1 je sledeča: če *število tendenc* za izbrani moment ne leži med spodnjo in zgornjo mejo, potem zaključimo, da ta moment NI stacionaren oziroma NI časovno nespremenljiv. Za stacionarnost signala v širšem smislu morata tako $n_{T\mu}$ kot $n_{T\sigma}$. biti znotraj meja. Zavedati pa se moramo, da tudi če obe *števili tendenc* ležita znotraj meja zaupanja, s tem nismo dokazali stacionarnosti v širšem smislu, ampak smo le pokazali, da signal verjetno je stacionaren (zanesljivost ocene nikoli ne more biti 100%).

5. Zaključek: naš signal je (verjetno) stacionaren v širšem smislu, če kažeta na stacionarnost tako srednja vrednost kot varianca, če torej oba $n_{T\mu}$ in $n_{T\sigma}$ ležita med pripadajočima mejama iz tabele 1.

Število ocen momenta (k)	Spodnja meja za n_T	Zgornja meja za n_T
10	3	8
12	3	10
14	4	11
20	6	15
50	19	32
100	42	59

Tabela 1: meje zaupanja za test tendenc pri nivoju statistične značilnosti $\alpha = 0,05$.

Test II: ocenjevanje stacionarnosti v širšem smislu

Pri drugem načinu testiranja stacionarnosti razdelimo signal na dve polovici in izračunamo različne statistične veličine za vsako polovico posebej. Nato preverimo enakost teh veličin za obe polovici signala. Če med statističnimi veličinami obeh polovic ni pomembnih razlik, lahko z določenim zaupanjem zaključimo, da je signal na opazovanem območju stacionaren. Na ta način lahko testiramo na primer enakost srednje vrednosti, variance, avtokorelacijske funkcije in tudi na primer spektra močnostne gostote. Vsaka od teh veličin zahteva svoj postopek. Ker nas zanima stacionarnost v širšem smislu, si bomo ogledali postopka za testiranje enakosti za srednjo vrednost ter varianco. Ta dva klasična postopka sta podrobneje opisana v večini učbenikov za statistiko.

1. Signal dolžine N vzorcev razdelimo na dva enako dolga segmenta dolžin N_1 in N_2. Če je N liho število vsebuje eden od segmentov en vzorec več od drugega.

2. Izračunamo srednji vrednosti in varianci za oba segmenta:

$$\begin{aligned} \hat{\mu}_1 &= \frac{1}{N_1}\sum_{i=1}^{N_1} x_i; \quad \hat{\sigma}_1^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1}(x_i - \hat{\mu}_1^2)^2 \\ \hat{\mu}_2 &= \frac{1}{N_2}\sum_{i=1}^{N_2} x_i; \quad \hat{\sigma}_2^2 = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2}(x_i - \hat{\mu}_2^2)^2 \end{aligned} \tag{4}$$

3. Statistična veličina, s katero testiramo enakost dveh srednjih vrednosti je tako imenovana statistika T, ki je definirana kot:

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s_{12}^2}}, \tag{5}$$

pri čemer je s_{12}^2 ocena skupne (združene) variance ocen za srednjo vrednost, ki jo izračunamo kot:

$$s_{12}^2 = (\frac{1}{N_1} + \frac{1}{N_2})(\frac{N_1\hat{\sigma}_1^2 + N_2\hat{\sigma}_2^2}{N_1 + N_2 - 2}) \tag{6}$$

4. Statistika T ima tako imenovano Studentovo t porazdelitev s številom prostostnih stopenj $\nu = N_1 + N_2 - 2$ (angl. *degrees of freedom*). Pri izbranem nivoju zaupanja α (običajno 0,05) lahko zaključimo, da se srednji vrednosti $\hat{\mu}_1$ in $\hat{\mu}_2$ *ne razlikujeta statistično značilno*, če je vrednost T med določeno spodnjo in zgornjo mejo:

$$\text{za } \nu > 200 \text{ in } \alpha = 0,05 : \begin{matrix} t_{\nu,1-\alpha/2} = -1,96 \\ t_{\nu,\alpha/2} = 1,96 \end{matrix} \tag{7}$$

   Povedati moramo, da je zgornji test enakosti srednjih vredosti veljaven le, če se varianci obeh segmentov signala ne razlikujeta statistično pomembno. Torej moramo enakost varianc preveriti iz dveh razlogov: da preverimo stacionarnost variance same in da upravičimo veljavnost testa enakosti srednjih vrednosti.

5. Statistična veličina, s katero testiramo enakost dveh varianc, je tako imenovana Fisherjeva statistika ali statistika F, ki je za prostostni stopnji $\nu_1 = N_1 - 1$ in $\nu_2 = N_2 - 1$ definirana kot:

$$F = \frac{\hat{\sigma}_1^2(\frac{N_1}{\nu_1})}{\hat{\sigma}_2^2(\frac{N_2}{\nu_2})} \tag{8}$$

6. Podobno kot prej tudi tu pri izbranem nivoju zaupanja α (običajno 0,05) zaključimo, da se varianci $\hat{\sigma}_1^2$ in $\hat{\sigma}_2^2$ *ne razlikujeta statistično* značilno, če je vrednost F med mejama:

$$F_{\nu_1,\nu_2;1-\alpha/2} \leq F \leq F_{\nu_1,\nu_2;\alpha/2} \tag{9}$$

   Meji $F_{\nu_1,\nu_2;1-\alpha/2}$ in $F_{\nu_1,\nu_2;\alpha/2}$ odčitamo iz tabel, kjer so meje podane za različne kombinacije prostostnih stopenj ν_1 in ν_2 in za različne nivoje zaupanja α. V našem primeru bomo uporabili meji:

$$\text{za } \nu_1 = \nu_2 = 5000 \text{ in } \alpha = 0,05 : \begin{matrix} F_{\nu,1-\alpha/2} = 0,9461 \\ F_{\nu,\alpha/2} = 1,0570 \end{matrix} \tag{10}$$

Zaključek: naš signal je (verjetno) stacionaren v širšem smislu, če kažeta na stacionarnost tako srednja vrednost, kot varianca. Torej ne sme biti statistično značilne razlike ne med $\hat{\sigma}_1^2$ in $\hat{\sigma}_2^2$ (to je hkrati tudi pogoj za veljavnost primerjave srednjih vrednosti), ne med $\hat{\mu}_1$ in $\hat{\mu}_2$.

Naloge

1. Naloga 1 - Ocenjevanje stacionarnosti po I. postopku S spletne učilnice si prenesite datoteko s podatki za tretjo vajo. V Matlab naložite datoteko s svojim imenom (`v5_xxxxxxxx.mat`). V datoteki so tri posnetki utrujanja mišice z EMG. Signali izvirajo iz različnih mišic treniranih plavalcev. Dolžina signalov ja 10000 vzorcev, vzorčna frekvenca pa je bila 2 kHz, kar ustreza 5 s meritve.

- Izrišite signale in *na oko* ocenite, ali so stacionarni za srednjo vrednost in varianco.
- Napišite kodo, ki bo preverila stacionarnost signalov po I. postopku. Rezultat naj bo število tendenc (*number of runs*) za srednjo vrednost in varianco. Razdelite originalne signale na $k = 20$ neprekrivajočih se enako dolgih segmentov in uporabite postopek, ki je opisan višje.
- Preverite stacionarnost signalov s prejšnjo metodo. Ali se ocena stacionarnosti po I. postopku ujema z vašo oceno *na oko*?

2. Naloga 2 - Ocenjevanje stacionarnosti po II. postopku

   - Napišite kodo, ki bo preverila stacionarnost signalov po II. postopku. Rezultat naj bo vrednost T in F statistike ter sklep ali je signal stacionaren za srednjo vrednost in za varianco.

3. Naloga 3 - Izbor dela signala, ki je stacionaren

   - Če kateri izmed signalov ni stacionaren v širšem smislu, preverite, ali lahko znotraj signala najdete krajši odsek, ki je stacionaren, pri čemer smiselno izberite časovno okno, tako da bo dobljeni signal čim daljši.
   - Pri krajšanju signala je treba prilagoditi meje za statistiko F, za kar uporabite matlabovo funkcijo `finv(P,V1,V2)`, ki izračuna inverz porazdelitve verjetnosti za F porazdelitev s številom prostostnih stopenj `V1` in `V2`. Kot vrednost `P` uporabite $1 - \alpha/2$ za spodnjo in $\alpha/2$ za zgornjo mejo.

Priročni MATLAB ukazi

Spodaj imate navedene MATLAB ukaze, ki vam lahko zelo olajšajo delo pri vaji. Ne pozabite na pomoč (tipka F1)!

```
x_med = median(x)
```

Mediana za vrednosti, ki so zbrane v vektorju `x`. Mediana predstavlja sredino populacije. Polovica elementov x je večjih (ali enakih) in polovica manjših (ali enakih) od mediane.

```
mu_x = mean(x)
```

Srednja povprečna vrednost elementov iz `x`. Funkcija na matrikah deluje tako da izračuna srednjo povprečno vrednost vsakega stolpca.

```
s_x = std(x)
```

Standardni odklon elementov iz `x`.

```
var_x = var(x)
```

Varianca (kvadrat standardne deviacije) elementov iz `x`. Tako kot funkciji `mean` in `median` pri matrikah izračuna vrednosti vsakega stolpca.

`predznaki = sign(x)`
Za vsak poztivni element v vektorju `x` vsabuje število 1. Za negativne predznake vsebuje število -1. Uporabno za avtomatsko štetje sprememb predznaka pri postopku testa I.

`X=finv(P,V1,V2)`
Funkcija izračuna inverz porazdelitve verjetnosti F porazdelitve s številom prostostnih stopenj `V1` in `V2`. Kot vrednost `P` uporabite $1 - \alpha/2$ za spodnjo in $\alpha/2$ za zgornjo mejo.