

Greške u proceni aproksimativne entropije

Dragana Bajić

Departman za energetiku, elektroniku i telekomunikacije
DEET
Fakultet tehničkih nauka
Novi Sad, Srbija
dragana.bajic@gmail.com

Nina Japundžić-Žigon

Katedra za farmakologiju, kliničku farmakologiju i
toksikologiju
Medicinski fakultet
Beograd, Srbija
nina.japundzic@gmail.com

Sažetak—Rad se bavi analizom grešaka pri proceni aproksimativne i uzoračke entropije. Pokazuje da osnovni problem nekonzistentnih procena predstavlja grubo kvantovanje verovatnoća sličnosti koje predstavljaju osnove za procenu entropije. Analiza je urađena na nizu nezavisnih odmeraka sa uniformnom raspodelom za koje je moguće analitički odrediti referentne vrednosti, i proverena na kardiovaskularnim signalima snimljenim na laboratorijskim pacovima.

Cljučne riječi—aproximativna entropija, entropija uzorka, verovatnoća sličnosti, greške procene;

I. UVOD

Entropija kao mera neuređenosti makrosistema predstavlja jednu od najzamerljivijih funkcija termodinamike i statističke mehanike [1]. Vežu između klasičnog koncepta i mere neuređenosti pri prenosu signala kroz komunikacioni kanal sa šumom iz Šenonovog fundamentalnog rada [2] uspostavio je fon Nojman „Nazovi je entropija. Kao prvo, tvoja funkcija neodređenosti se pod tim imenom koristi u statističkoj mehanici, tako da već ima ime. Kao drugo i mnogo važnije, niko zapravo ne zna šta je entropija, pa ćeš u diskusiji uvek imati prednost” [2].

Kolmogorov je prilagodio Šenonovu statičku entropiju dinamičkim sistemima, uz razradu koju je uradio Sinai [4]. Kolmogorov-Sinai (KS) entropija je teorijski veoma zahtevna što otežava primenu u realnim sistemima. Jednostavnu implementaciju je omogućio Pinkus aproksimativnom entropijom ($ApEn$) koja je zaobišla stroge matematičke zahteve, zato je i nazvana ‘aproximativna’ [5]. Entropija uzorka ($SampEn$) uvedena je kao robusna modifikacija $ApEn$ [6]. I $ApEn$ i $SampEn$ prepoznate su kao važni istraživački alati u gotovo svim naučnim disciplinama, a naročito biomedicinskim za koje su inicijalno i namenjene. Njihova citiranost se meri u hiljadama, što je postalo svojevrsan fenomen [7].

Poboljšanja originalnih koncepata $ApEn$ i $SampEn$ su brojna, na primer [8], [9], [10], [11], a postoje i brojni alternativni načini za aproksimaciju entropije [7], [12], [13], [14], [15]. Uprkos tome, $ApEn$ je ostala na vrhu kojem se približava i $SampEn$, a njihove dobre osobine su sistematizovane u preglednom radu [16].

U osnovi obe entropije je poređenje vrlo kratkih segmenata vremenskog niza po principu ‘svaki sa svakim’, pri čemu je $ApEn$ manje robusna od $SampEn$. Smatra se da je tome uzrok

takozvana ‘samosličnost’ tj. poređenje segmenta sa samim sobom [6]. Samosličnost izaziva pristrasnost, veliku zavisnost od dužine posmatranog vremenskog niza i nekonzistentnost, to jest za različite skupove parametara daje nesaglasne rezultate, što je pokazano i u [17], [18]. Ovi zaključci o samosličnosti i $ApEn$ se ponavljaju u svim radovima koji se bave metodologijom procene entropije zasnovane na Pinkusovom radu [5].

U ovom radu analiziramo razloge uticaja samosličnost na smanjene robusnosti $ApEn$ u odnosu na $SampEn$. Pokazujemo da su osnovni uzrok greške u proceni verovatnoća koje su, za kratke vremenske nizove, veoma velike. Dokazali smo, takođe, da greške ne utiču podjednako na $ApEn$ i $SampEn$ i to zbog razlike u njihovim definicijama: u $ApEn$ se usrednjavaju logaritmi verovatnoća, što, prema Šenonu, odgovara usrednjavanju količinâ informacijâ koje nose partikularni ishodi događaja, a u $SampEn$ se usrednjavaju verovatnoće, pa tek nakon toga logaritmuje, što odgovara količini informacije koju nosi usrednjavanje ishoda događaja.

Rad je organizovan na sledeći način: sledeće poglavlje ukratko definišu $ApEn$ i $SampEn$ i obavezno je u svim radovima koji se bave ovom problematikom. U poglavlju se, dodatno, uvode greške procene. Motivacija za ovu analizu data je u poglavlju III, gde se analizira uticaj grešaka procene verovatnoća potrebnih za procenu entropije. U poglavlju IV prikazali smo rezultate na realnim vremenskim nizovima – kardiovaskularnim signalima zdravih laboratorijskih Wistar pacova u normalnim, tj. kontrolnim, uslovima. Na rezultate se, u istom poglavlju, nadovezuju zaključna razmatranja.

II. ENTROPIJE I GREŠKE PROCENE

Posmatra se vremenski niz $x_i \in \mathbf{X}$, $i = 1, \dots, N$. Niz se razlaže na $N - m + 1$ vektora (segmenata) dužine m :

$$\mathbf{X}_i^{(m)} = [x_i, x_{i+1}, \dots, x_{i+m-1}], \quad i = 1, \dots, N - m + 1. \quad (1)$$

Vektori $\mathbf{X}_i^{(m)}$ and $\mathbf{X}_j^{(m)}$ su slični ako je maksimalno apsolutno rastojanje između njihovih skalarnih komponenta manje od praga r :

$$d(\mathbf{X}_i^{(m)}, \mathbf{X}_j^{(m)}) = \max_{k=0, \dots, m-1} |x_{i+k} - x_{j+k}| \leq r, \quad i, j = 1, \dots, N - m + 1, \quad (2)$$

gde d označava rastojanje. Vektor $\mathbf{X}_i^{(m)}$ je “obrazac”, a vektori $\mathbf{X}_j^{(m)}, j = 1, \dots, N - m + 1$ koji ispunjavaju uslov (2) su “slični” obrascu. Obrazac upoređen sam sa sobom je “samosličan”.

Sušтина $ApEn$ i $SampEn$ je u prebrojavanju sličnih vektora. Za i -ti obrazac definiše se broj sličnih vektora u vremenskom nizu:

$$C_i^{(m)}(r) = \sum_{j=1}^{N-m+1} I\{d(\mathbf{X}_i^{(m)}, \mathbf{X}_j^{(m)}) \leq r\},$$

$$i = 1, \dots, N - m + 1, \text{ za } ApEn; \quad (3)$$

$$B_i^{(m)}(r) = \sum_{j=1, j \neq i}^{N-m} I\{d(\mathbf{X}_i^{(m)}, \mathbf{X}_j^{(m)}) \leq r\},$$

$$i = 1, \dots, N - m, \text{ za } SampEn. \quad (4)$$

U jednačinama (3) i (4) izraz $I\{\text{uslov}\}$ označava indikatorsku funkciju koja je jednaka jedinici ako je uslov ispunjen, inače je jednaka nuli. U našem slučaju, sumiranje indikatorskih funkcija predstavlja prebrojavanje vektora koji su slični obrascu. Taj broj se razlikuje za $ApEn$ i $SampEn$, jer $SampEn$ isključuje samosličnost ($i = j$) i ne uzima u obzir poslednji vektor br. $N-m+1$ ni kao obrazac, ni kao sličan vektor.

Broj sličnih vektora podeljen ukupnim brojem testiranih vektora odgovara klasičnoj definiciji verovatnoće kao relativne učestanosti. Verovatnoća sličnosti se definiše kao verovatnoća da obrazac nađe sličan vektor:

$$\hat{p}_{Ai}^{(m)}(r) = \frac{C_i^{(m)}(r)}{N-m+1}, \quad i = 1, \dots, N - m + 1, \text{ za } ApEn; \quad (5)$$

$$\hat{p}_{Si}^{(m)}(r) = \frac{B_i^{(m)}(r)}{N-m}, \quad i = 1, \dots, N - m, \text{ za } SampEn, \quad (6)$$

gde “ $\hat{}$ ” označava procenu.

Za $ApEn$ i $SampEn$ su potrebne i procene verovatnoće sličnosti za obrasce dužake $m+1$:

$$\hat{p}_{Ai}^{(m+1)}(r) = \sum_{j=1}^{N-m} I\{d(\mathbf{X}_i^{(m+1)}, \mathbf{X}_j^{(m+1)}) \leq r\} =$$

$$\frac{C_i^{(m+1)}(r)}{N-m}, \quad i = 1, \dots, N - m, \text{ za } ApEn; \quad (7)$$

$$\hat{p}_{Si}^{(m+1)}(r) = \sum_{j=1, j \neq i}^{N-m} I\{d(\mathbf{X}_i^{(m+1)}, \mathbf{X}_j^{(m+1)}) \leq r\} =$$

$$\frac{C_i^{(m+1)}(r)-1}{N-m} = \frac{A_i^{(m)}(r)}{N-m}, \quad i = 1, \dots, N - m, \text{ za } SampEn; \quad (8)$$

gde je $A_i^{(m)}(r)$ broj sličnih vektora dužine $m+1$. Oznake za broj događaja $A_i^{(m)}(r), B_i^{(m)}(r),$ i $C_i^{(m)}(r)$ su iz izvornih radova [5] i [6]. U ovom radu njih zamenjujemo verovatnoćama koje su matematički prihvatljivije. Takođe, sada je jasno zašto je isključen poslednji vektor u jednačini (4): time se izjednačava broj procenjenih verovatnoća za m i za $m+1$.

$ApEn$ i $SampEn$ se procenjuju prema sledećim formulama:

$$\hat{ApEn}(m, r, N) = \hat{\Phi}^{(m)}(r, N) - \hat{\Phi}^{(m+1)}(r, N), \quad (9)$$

$$\hat{SampEn}(m, r, N) = \hat{\Psi}^{(m)}(r, N) - \hat{\Psi}^{(m+1)}(r, N), \quad (10)$$

gde su

$$\hat{\Phi}^{(m)}(r, N) = \frac{1}{N-m+1} \cdot \sum_{i=1, \hat{p}_{Si}^{(m)}(r) \neq 0}^{N-m+1} \ln(\hat{p}_{Ai}^{(m)}(r)), \quad (11)$$

$$\hat{\Psi}^{(m)}(r, N) = \ln\left(\frac{1}{N-m} \cdot \sum_{j=1, j \neq i}^{N-m} \hat{p}_{Si}^{(m)}(r)\right), \quad (12)$$

$$\hat{\Phi}^{(m+1)}(r, N) = \ln\left(\frac{1}{N-m} \cdot \sum_{j=1, j \neq i}^{N-m} \hat{p}_{Si}^{(m+1)}(r)\right). \quad (13)$$

Greške u proceni verovatnoća su:

$$\varepsilon_{Ai}^{(m)}(r) = p_i^{(m)}(r) - \hat{p}_{Ai}^{(m)}(r), \quad (14)$$

$$\varepsilon_{Si}^{(m)}(r) = p_i^{(m)}(r) - \hat{p}_{Si}^{(m)}(r). \quad (15)$$

Lako se pokazuje da se za sume Φ i Ψ dobija:

$$\hat{\Phi}^{(m)}(r, N) = \frac{1}{N-m+1} \cdot \sum_{i=1}^{N-m+1} \ln(p_i^{(m)}(r) - \varepsilon_{Ai}^{(m)}(r)) =$$

$$= \Phi^{(m)}(r, N) + \frac{1}{N-m+1} \cdot \sum_{i=1}^{N-m+1} \ln\left(1 - \frac{\varepsilon_{Ai}^{(m)}(r)}{p_i^{(m)}(r)}\right) =$$

$$= \Phi^{(m)}(r, N) + \varepsilon_{\Phi}^{(m)}(r, N), \quad (16)$$

$$\hat{\Psi}^{(m)}(r, N) = \ln\left(\sum_{i=1}^{N-m} (p_i^{(m)}(r) - \varepsilon_{Si}^{(m)}(r))\right) =$$

$$= \Psi^{(m)}(r, N) + \ln\left(1 - \frac{\sum_{i=1}^{N-m} \varepsilon_{Si}^{(m)}(r)}{\sum_{i=1}^{N-m} p_i^{(m)}(r)}\right) =$$

$$= \Psi^{(m)}(r, N) + \varepsilon_{\Psi}^{(m)}(r, N). \quad (17)$$

Greške u proceni entropije su:

$$\varepsilon_{ApEn}^{(m)}(r, N) = \varepsilon_{\Phi}^{(m)}(r, N) - \varepsilon_{\Phi}^{(m+1)}(r, N), \quad (18)$$

$$\varepsilon_{SampEn}^{(m)}(r, N) = \varepsilon_{\Psi}^{(m)}(r, N) - \varepsilon_{\Psi}^{(m+1)}(r, N). \quad (19)$$

III. VEROVATNOĆE SLIČNOSTI

Iz jednačina (1) i (2) je očigledno da su $ApEn$ i $SampEn$ parametarske metode – zavise od dužine vektora m , praga r i dužine vremenskog niza N . Prag se smatra najkritičnijim i detaljno je razmatran u radovima [19], [20], [21], uključujući i naše [17], [18]. Dužina vremenskog niza N je takođe kritična, što se naglašava u brojnim radovima još od inicijalnog rada o $SampEn$ [6]. Dužina vektora m se po pravilu postavlja na vrednost $m = 2$ na osnovu analiza iz [22], a i zbog porasta potrebnog procesorskog vremena za veće vrednosti.

Radovi u kojima se metodološki analiziraju procene entropije navode samosličnost kao uzrok pristrasnosti i nekonzistentnosti $ApEn$, ali bez suštinske analize kako do problema zapravo dolazi. U radu [23] je, međutim, naznačen jedan mogući uzrok. Motivaciju za dalju razradu ideje iz [23] predstavlja eksperiment čiji su rezultati prikazani na Slici 1. Naime, da je samosličnost doista jedini razlog za nekonzistenciju $ApEn$, njeno uklanjanje bi rešilo problem. Stoga

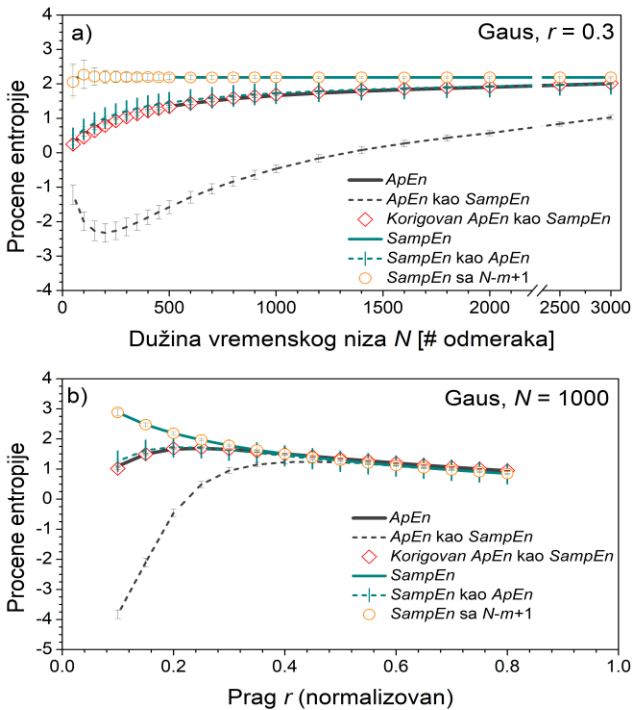
smo za svaki skup parametara (N, r) generisali po 1000 vremenskih nizova sačinjenih od nezavisnih i identično raspodeljenih odmeraka. Iz njih smo procenili $ApEn$, $SampEn$ i njihove modifikacije na kojima smo ispitali uticaj samosličnosti. tako što smo verovatnoće (5) i (7) koje važe za $ApEn$ ubacili u jednačinu (10) koja važi za $SampEn$, a verovatnoće (6) i (8) koje važe za $SampEn$ ubacili u jednačinu (9) koja važi za $ApEn$ i iz sume (11) izbacili sve događaje koji dovode do logaritma od nule ($\ln(0)$). U trećoj modifikaciji smo korigovali sumu (11) tako da imenilac odgovara broju sumiranih događaja nakon izbacivanja $\ln(0)$, što predstavlja korekciju koja je uvedena za unakrsnu entropiju u radu [18]. Korigovana suma (11) je:

$$\hat{\Phi}_{KORIG}^{(m)}(r, N) = \frac{1}{N - m - N_0} \cdot \sum_{i=1, \hat{p}_{Si}^{(m)}(r) \neq 0}^{N-m} \ln(\hat{p}_{Si}^{(m)}(r)),$$

$$N_0 = \sum_{i=1}^{N-m} I\{\hat{p}_{Si}^{(m)}(r) = 0\} \quad (20)$$

Konačno, mada ne preterano važno, proverili smo i da li izostavljanje poslednjeg vektora br. $N-m+1$ iz jednačine (4) ima ikakvog uticaja na $SampEn$.

Slike 1a i 1b prikazuju procene entropija uz navedene eksperimentalne modifikacije, i to slika 1a u funkciji dužine vremenskog niza N , a 1b u funkciji normalizovanog praga r .



Slika 1. Procene $ApEn$, $SampEn$ i modifikacije koje ispituju samosličnost u funkciji dužine vremenskog niza N (a) i normalizovanog praga r (b). Testirani vremenski nizovi se sastoje od nezavisnih i identično raspodeljenih odmeraka sa Gausovom raspodelom

Izbacivanje poslednjeg vektora ($N-m+1$) iz jednačine (4) ne utiče na procenu $SampEn$. Nadalje, ubacivanje samosličnosti u jednačine za $SampEn$ doista svodi njegove vrednosti na vrednosti $ApEn$. Međutim, obrnuto ne važi: korigovane vrednosti za $ApEn$ sa ubačenim verovatnoćama iz $SampEn$ i dalje su iste kao originalne vrednosti za $ApEn$. Nekorigovane

$ApEn$ vrednosti pokazuju značajna odstupanja, usled neadekvatnog usrednjavanja, što je prema [18] i očekivano. Iz toga može da se zaključi da samosličnost nije jedini faktor koji utiče na odstupanja između $ApEn$ i $SampEn$.

Slika 1 pokazuje i odavno poznat rezultat da se nekonzistentnost smanjuje sa porastom N i r . Pored toga, iako je srednja vrednost 1000 ponovljenih procena $SampEn$ stabilna za sve dužine vremenskog niza N , standardna devijacija je značajno veća nego za $ApEn$. Slika 2 pokazuje da $SampEn$ može da ostupa i do 40% u odnosu na srednju vrednost, dok su, na istim uzorcima, odstupanja $ApEn$ do 12%.

Za dalju analizu je najpogodnije koristiti vremenske nizove sa nezavisnim i identično raspodeljenim odmercima poznate raspodele $f(x)$ jer omogućavaju tačno izračunavanje verovatnoće sličnosti $p_i^{(m)}(r)$ kao što je ilustrovano na slici 3: prvo se izračuna verovatnoća da pojedinačni odmerak $x_i \in \mathbf{X}$ nađe sebi sličan odmerak:

$$p_i(r) = \int_{x_i-r}^{x_i+r} f(x) \cdot dx, \quad (21)$$

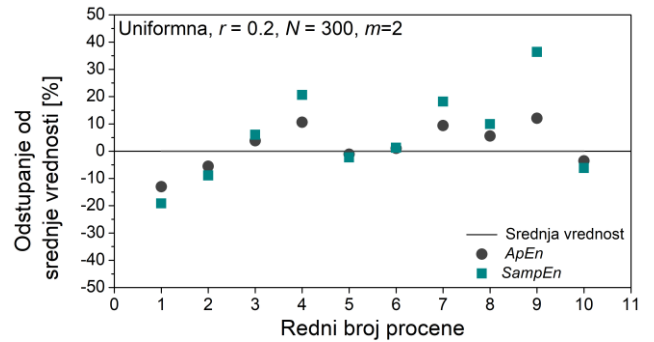
a zatim i verovatnoća sličnosti za obrazac $[x_i x_{i+1} \dots x_{i+m-1}]$:

$$p_i^{(m)}(r) = \prod_{k=0}^{m-1} p_{i+k}(r), \quad (22)$$

što omogućava njeno poređenje sa verovatnoćom sličnosti procenjenom iz istog vremenskog niza.

Za uniformnu raspodelu sa nultim očekivanjem i jediničnom varijansom, jednačina (21) je vrlo jednostavna:

$$p_i(r) = \frac{r + \min(r, \sqrt{3} - |x_i|)}{(2 \cdot \sqrt{3})}. \quad (23)$$



Slika 2. Relativno odstupanje pojedinačnih vrednosti procene entropije od vrednosti usrednjene na 1000 uniformnih vremenskih nizova za parametre koji su preporučeni u inicijalnim radovima [5] i [6]

Uniformna raspodela je pogodna i zato što se za nju lako analitički procenjuje prosečna verovatnoća sličnosti za jedan obrazac. Ako je uniformna raspodela definisana kao:

$$f(x) = \begin{cases} \frac{1}{2 \cdot a}, & -a \leq x \leq a, \quad a = \sqrt{3} \\ 0, & \text{drugde} \end{cases}, \quad (24)$$

tada je prosečna verovatnoća da je apsolutna razlika dva uniformno raspodeljena odmerka x i y manja od praga $r > 0$ ravna:

$$p = Pr\{|x - y| \leq r\} = \frac{4 \cdot a \cdot r - r^2}{4 \cdot a^2} = \frac{4 \cdot \sqrt{3} \cdot r - r^2}{12}, \quad (25)$$

$$\begin{aligned}
\text{jer je: } p &= \int_{-a}^{-a+r} f(x) \cdot \left(\int_{-a}^{x+r} f(y) \cdot dy \right) \cdot dx + \\
&+ \int_{-a+r}^{a-r} f(x) \cdot \left(\int_{x-r}^{x+r} f(y) \cdot dy \right) \cdot dx + \\
&+ \int_{a-r}^a f(x) \cdot \left(\int_{x-r}^a f(y) \cdot dy \right) \cdot dx = \\
&= \frac{1}{4 \cdot a^2} \cdot \int_{-a}^{-a+r} (x+r+a) \cdot dx + \int_{-a+r}^{a-r} 2 \cdot r \cdot dx + \\
&+ \int_{a-r}^a (a+r-x) \cdot dx = \frac{4 \cdot \sqrt{3} \cdot r - r^2}{12} \quad (26)
\end{aligned}$$

Verovatnoća da je apsolutna razlika svih m parova odmeraka za jedan obrazac manja od praga r ravna je:

$$p_i^{(m)}(r) = p^m = \left(\frac{4 \cdot \sqrt{3} \cdot r - r^2}{12} \right)^m. \quad (27)$$

Lako se pokazuje da uvrštavanje prosečne verovatnoće sličnosti u jednačine (11), (12) i (13) dovodi do iste vrednosti i za $ApEn$, i za $SampEn$:

$$ApEn = SampEn = -\ln \left(\frac{4 \cdot \sqrt{3} \cdot r - r^2}{12} \right). \quad (28)$$

Ova vrednost može da se koristi kao tačna vrednost ("ground truth") za procene entropije.

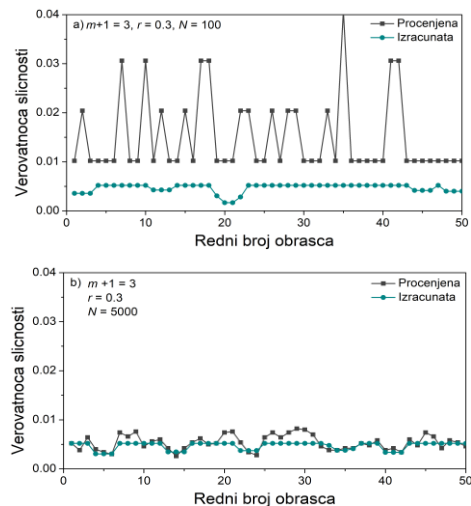
Ilustracije radi, slika 3 prikazuje procenjene (5) i izračunate (22) verovatnoće sličnosti za nekoliko uzastopnih obrazaca, i za dve dužine vremenskih nizova, $N = 100$ i $N = 5000$. U drugom slučaju je slaganje daleko veće.

Objašnjenje se nalazi na slici 4 koja prikazuje histograme verovatnoća sličnosti za vremenske nizove male dužine $N = 100$, ponovljene 100 puta. Radi lakšeg poređenja, histogrami su normalizovani tako da zbir njihovih amplituda odgovara jedinici, a za izračunate verovatnoće predstavljeni su linijom. Vidi se da na izračunatu verovatnoću sličnosti malo utiču rubni odmerci $|x_i| \geq \sqrt{3} - r$, tako da dominira vrednost $(r/\sqrt{3})^m$ (jednačine (22) i (23)). Međutim, procenjene verovatnoće sličnosti imaju vrednosti $k/(N-m)$, gde su vrednosti k za $ApEn$ od 1 do 9 za $m=2$, odnosno od 1 do 4 za $m=3$. Za $SampEn$, zbog isključivanja samosličnosti, vrednosti za k su od 0 do 8 i od 0 do 3.

Histogrami pokazuju i velika odstupanja između izračunate i procenjenih verovatnoća. Izračunata verovatnoća ima praktično samo jednu dominantnu vrednost, a procenjenih verovatnoća ima više, jer su neki obrasci našli 1, 2, 3, pa i više sličnih vektora u okviru posmatranog vremenskog niza.

Slike 4e i 4f predstavljaju histograme razlike procenjene i izračunate verovatnoće sličnosti – praktično, histogram greške procene verovatnoće (jednačine (4) i (5)). Iz ovih slika se može naslutiti razlog veće stabilnosti $SampEn$ u odnosu na $ApEn$.

Vertikalna linija deli slike 4e i 4f na segment sa pozitivnim, i segment sa negativnim greškama. Na slici 4f se naročito dobro vidi da su greške procene verovatnoća sličnosti za $ApEn$ isključivo negativne, a iste greške za $SampEn$ i pozitivne i



Slika 3. Procenjene (5) i izračunate (22) verovatnoće sličnosti za 50 uzastopnih obrazaca vremenskih nizova dužine $N=100$ (a) i $N = 5000$ (b). Na panelu a) jasno su izražene procenjene verovatnoće sličnosti $\frac{k}{N-m} \approx k \cdot 0.1$.

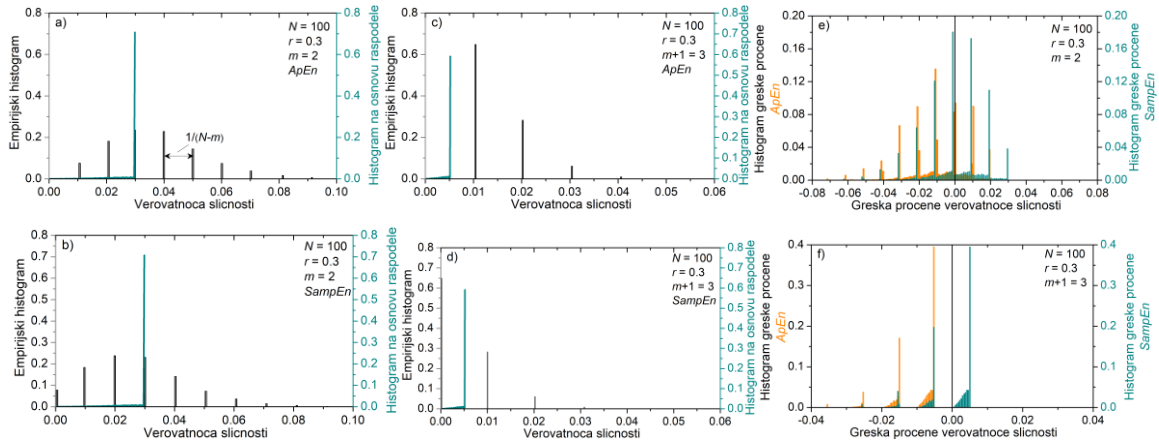
negativne. Na slici 4e se takođe uočava da su linije histograma za $SampEn$ ravnomerno raspoređene oko nulte ose, a da su za $ApEn$ uglavnom skoncentrirane u delu sa negativnim verovatnoćama. To je značajno, jer se na osnovu jednačine (17) greške procene verovatnoće za $SampEn$ izračunavaju kao:

$$\varepsilon_{\psi}^{(m)}(r, N) = \ln \left(1 - \frac{\sum_{i=1}^{N-m} \varepsilon_{Si}^{(m)}(r)}{\sum_{i=1}^{N-m} p_i^{(m)}(r)} \right), \quad (29)$$

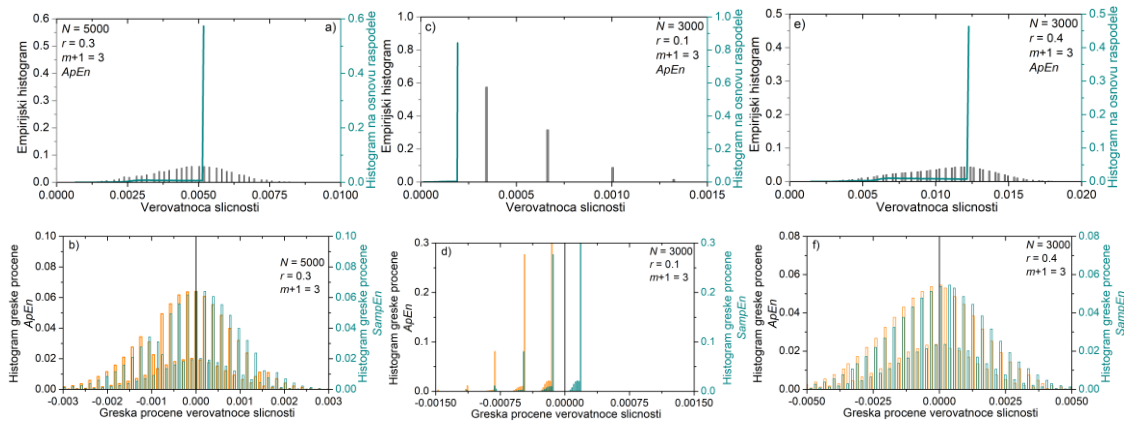
tako da se greške simetrično raspoređene oko nule sumiranjem $\sum_{i=1}^{N-m} \varepsilon_{Si}^{(m)}(r)$ praktično potiru.

Primeri koji pokazuju uticaj povećane dužine i uticaj praga na verovatnoće sličnosti i greške u njihovoj proceni prikazani su na slici 5. Paneli c) i d) pokazuju da je uticaj niskog praga r ekvivalentan uticaju male dužine vremenskog niza N . U oba slučaja je mogućnost da obrazac dužine m nađe sličnog parnjaka niska: kada je prag r mali, odmerci signala moraju da budu veoma sličnih vrednosti da bi se parnjaci našli, a kada je dužina N mala, nema mnogo mogućnosti da se takav događaj desi, zato je i prikazan slučaj $m+1=3$ kao gori od dva slučaja koja definišu $ApEn$ i $SampEn$ za $m=2$. U oba slučaja, broj sličnih vektora je mali pa je i procenjena verovatnoća sličnosti grubo kvantovana, što je prikazano histogramom na slikama 4a, 4b, 4c, 4d. Na slikama 5c i 5d, kvantovanje nije grubo, ali je broj događaja mali pa ne omogućava finu granulaciju verovatnoća sličnosti. Kao posledica, greška u proceni je velika. Definicija $SampEn$ je takva da se greške potiru, ali za $ApEn$ greške se akumuliraju i to dovodi do nekonzistentnih rezultata. Kada parametri N i r imaju velike vrednosti, procenjene verovatnoće sličnosti su fino kvantovane i vrednosti su bliske tačnim, tako da je procena entropije stabilna (slike 2b, 5b i 5f).

Da se ne bi sve svelo da vizualizaciju, izračunata je srednja relativna greška verovatnoće sličnosti, izražena u procentima, usrednjena po svih $(N-m) \cdot 100$ procenjenih verovatnoća. Relativna greška je prikazana na slici 6, za slučajeve sa slika 4 i 5, i to kao srednja vrednost \pm standardna greška (standard error



Slika 4. Histogrami procenjenih i izračunatih verovatnoća sličnosti, i histogrami greške u proceni verovatnoće sličnosti za 100 uzastopnih vremenskih nizova, $N = 100$ i $r = 0.3$. a) *ApEn*, $m=2$; b) *SampEn*, $m=2$; c) *ApEn*, $m+1=3$; d) *SampEn*, $m+1=3$; e) greška procene, $m=2$; f) greška procene, $m+1=2$; Na panelima a)-d) uočavaju se procenjene verovatnoće $\frac{k}{N-m} \approx k \cdot 0.1$, kao i izračunate verovatnoće među kojima dominira, zbog uniformnosti raspodele, samo jedna vrednost, ravna $\left(\frac{r}{\sqrt{3}}\right)^m = 0.03$ za $m = 2$ i 0.0052 za $m = 3$. Na panelima e) i f) uočava se da su greške procene verovatnoće sličnosti za *SampEn* praktično simetrične u odnosu na nulu.



Slika 5. Histogrami procenjenih i izračunatih verovatnoća sličnosti, i histogrami greške u proceni verovatnoće sličnosti za 100 uzastopnih vremenskih nizova i $m+1=3$. a) *ApEn*, $N = 5000$, $r = 0.3$; b) greška procene, $N = 5000$, $r = 0.3$; c) *ApEn*, $N = 3000$, $r = 0.1$; d) greška procene, $N = 3000$, $r = 0.1$; e) *ApEn*, $N = 3000$, $r = 0.4$; f) greška procene, $N = 3000$, $r = 0.4$;

of mean, s.e.m). Od prikaza standardne devijacije smo odustali jer je, za velike vrednosti greške, bila mnogostruko veća od srednje vrednosti, naročito za *SampEn*.

Greške za *SampEn* su reda veličine nekoliko procenata za malo $N=100$ i $r=0.1$, u ostalim slučajevima su bliske nuli. Za razliku od toga, greške za *ApEn* dostižu i nekoliko stotina procenata a za posmatrane uzorke nisu manje od 1%. Međutim, s.e.m. za *SampEn* za $N=100$ je veoma veliko, što potvrđuje veliki varijabilitet između partikularnih *SampEn* rezultata za kratke vremenske nizove.

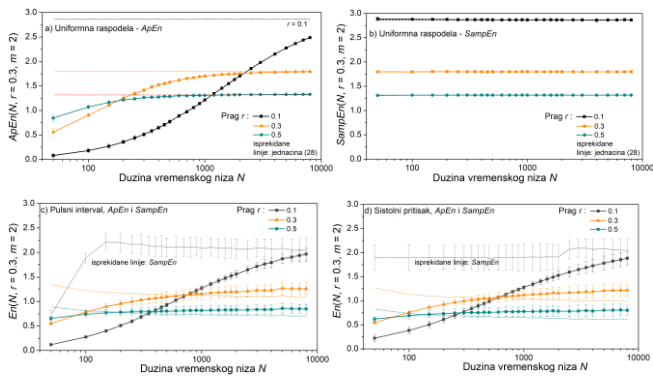
IV. PROCENE ENTROPIJE – ZAKLJUČAK

Greške u proceni entropije posledica su neadekvatnog kvantovanja verovatnoća sličnosti. Kratki nizovi imaju suviše veliki kvant, a rezultat malih vrednosti pragova je nedovoljan broj kvantnih nivoa, što poništava činjenicu da nivoi mogu da budu i vrlo fini, ako je vremenski niz dugačak.

Slika 7 pokazuje kako to utiče na entropiju. Slike 7a i 7b pokazuju procenu entropije iz uniformnih vremenskih nizova.

Sa porastom N , *ApEn* konvergira ka tačnoj vrednosti datoj jednačinom (28), dok se *SampEn* preklapa sa tačnom vrednošću i za kratke vremenske nizove.

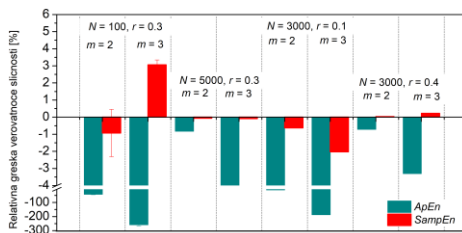
Na slikama 7c i 7d prikazana je entropija procenjena iz kardiovaskularnih vremenskih nizova mužjaka laboratorijskih Wistar pacova težine 300-350g, pod laboratorijskim uslovima (temperatura i relativna vlažnost vazduha $22 \pm 2^\circ\text{C}$ i 60–70%, uz dvanaestočasovnu smenu svetlih i zamračenih intervala). Eksperimentalne procedure su usaglašene sa Direktivom 2010/63/EU i nacionalnim pravilnikom 2009/6/RS i 2010/RS. Protokol je odobio univerzitetski etički odbor, licenca n°323-07-10519/2013-05/2. Za merenje talasnog oblika krvnog pritiska koršćena je telemetrijska oprema TL11M2-PA-C50-PX-DSI implantirana u trbušnu aortu. Radilo se o eksperimentu[24] u kojem su ispitivani vazopresinski antagonisti. Izabrali smo ovaj eksperiment jer su generisani veoma dugački vremenski nizovi, a koristili smo samo signale snimane u normalnim uslovima, bez antagonista. Iz talasnog.



Slika 7. Procene entropije u funkciji dužine niza N . a) $ApEn$ uniformnog niza; isprekidana linija je entropija po formuli (28); b) $SampEn$ uniformnog niza; c) Entropija niza pulsnih intervala laboratorijskih pacova; d) Entropija sistolnog pritiska laboratorijskih pacova.

oblika arterijskog pritiska izdvojeni su vremenski nizovi sistolnog pritiska (lokalni maksimumi) i pulsog intervala (vremenski intervali između uzastopnih maksimalnih pozitivnih promena pritiska). Artefakti i patologije su uklonjeni poluautomatski

Standardna devijacija realnih nizova je veća od devijacije simuliranih nizova sa uniformnom raspodelom, naročito za $SampEn$ (prikazana samo za $r = 0.1$, da ne opteretiti sliku). Na osnovu izloženog, sledeći korak je predložiti rešenje koje će uspeli da ponudi stabilnu procenu ovih parametara koji su postali nezaobilazni u (bio)medicinskim istraživanjima i za slučaj kratkih vremenskih nizova. Jedna od smernica je naći transformacionu funkciju vremenskog niza koja bi implementirala slično semilogaritamskim i kvazilogaritamskim funkcijama kompresije govornog signala.



Slika 6. Relativne greške procene verovatnoće sličnosti [%]

Literatura

- [1] J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters and Complexity*, Oxford University Press, Oxford, UK, 2006; pages XV and 67.
- [2] M. Tribus and E. C. McIrvine, "Energy and information," *Sci. Am.*, vol. 225, pp. 179–188, 1971.
- [3] C. E. Shannon, "Communications in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, 1949.
- [4] Nikolai Ivanov (translator), "A. N. Kolmogorov's and Y. G. Sinai's Papers Introducing Entropy of Dynamical Systems" Available online: <https://nikolaivivanov.files.wordpress.com/2015/05/definitionentropy2014-20151.pdf>, accessed on 4th January 2022.
- [5] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc. Natl Acad. Sci. USA*, vol. 88, pp. 2297–2301, 1991.

- [6] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Am. J. Physiol. Heart Circ. Physiol.*, vol. 278, pp. H2039–H2049, 2000.
- [7] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou, "The appropriate use of approximate entropy and sample entropy with short data sets". *Ann. Biomed. Eng.*, vol. 41, pp. 349–365, 2013.
- [8] K. H. Chon, C. G. Scully, and S. Lu, "Approximate entropy for all signals," *IEEE Eng Med Biol Mag*, vol. 28, pp. 18–23, 2009.
- [9] S. Lu, X. Chen, J. K. Kanters, I. C. Solomon, and K. H. Chon, "Automatic selection of the threshold value r for approximate entropy", *IEEE Trans Biomed Eng.*, vol. 55, pp. 1966–1972, 2009.
- [10] F. Kaffashi, R. Foglyano, C. G. Wilson, and K. A. Loparo, "The effect of time delay on approximate and sample entropy calculations", *Physica D: Nonlinear Phenomena*, vol. 237, pp. 3069–3074, 2008.
- [11] A. Porta, G. Baselli, D. Liberati, N. Montano, C. Cogliati, T. Gnecci-Ruscone, A. Malliani, and S. Cerutti, "Measuring regularity by means of a correctesconditional entropy in sympathetic outflow", *Biol. Cybern.*, vol. 78, pp. 71–78, 1998.
- [12] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of biological signals," *Physical Review E*, vol. 71, pp. 1–18, 2005.
- [13] W. Chen, Z. Wang, H. Xie, and W. Yu, "Characterization of surface EMG signal based on fuzzy entropy," *IEEE Trans. Neural Syst. and Rehabil. Eng.*, vol. 15, pp. 266–272, 2007.
- [14] A. Szczesna, "Quaternion Entropy for Analysis of Gait Data", *Entropy*, vol. 21(1), 79, 2019.
- [15] S. Behrendt, T. Dimpfl, F. J. Peter, J. David, and R. Zimmermann, "Transfer Entropy - Quantifying information flow between different time series using effective transfer entropy", *SoftwareX*, Volume 10, July–December 2019.
- [16] A. Delgado-Bonal, and A. Marshak, "Approximate Entropy and Sample Entropy: A Comprehensive Tutorial," *Entropy*, vol. 21(6), 541, 2019.
- [17] A. Boskovic, T. Loncar, O. Sarenac, N. Japundzic-Zigon, and D. Bajić, "Unbiased entropy estimates in stress: A parameter study", *Comput. Biol. Med.* vol. 42, pp. 667–679, 2012,
- [18] T. Skoric, O. Sarenac, B. Milovanovic, N. Japundzic-Zigon, and D. Bajić, "On Consistency of Cross-Approximate Entropy in Cardiovascular and Artificial Environments", *Complexity* vol. 2017, paper no. 8365685, 2017.
- [19] S. Lu, X. Chen, J.K. Kanters, I.C. Solomon, K.H. Chon, "Automatic selection of the threshold value r for approximate entropy", *IEEE Trans. Biomed. Eng.* vol. 55, no. 8, pp. 1966–1972, 2008.
- [20] K.H. Chon, C.G. Scully, S. Lu, "Approximate entropy for all signals", *IEEE Eng. Med. Biol.* . vol. 28, no. 6, pp. 18–23, 2009.
- [21] P. Castiglioni, M. Di Rienzo, How the threshold "R" influences approximate entropy analysis of heart-rate variability, *Comput. Cardiol.* . vol. 35, pp. 561–564, 2009.
- [22] M.B. Kennel, R. Brown, "Determining embedding dimension for phase-space reconstruction using geometrical reconstruction", *Phys. Rev. A Gen. Phys.* vol. 45, no. 6, pp. 3403–3411, 1992.
- [23] D. Bajić, N. Japundžić-Žigon, "On Quantization Errors in Approximate and Sample Entropy," *Entropy*, vol. 24, no. 73, 19 pages, 2022.
- [24] S. Milutinovic, D. Murphy, N. Japundzic-Zigon, "The role of central vasopressin receptors in the modulation of autonomic cardiovascular controls: A spectral analysis study", *Am. J. Physiol. Regul. Integr. no.* 291, pp. r1579–r1591, 2006.

Abstract This paper performs an error analysis of approximate and sample entropy. It shows that inconsistency is an outcome of quantization errors of matching probabilities. The analysis is performed using independent and identically uniformly distributed sample series for which the analytical ground truth can be established, and using cardiovascular signals recorded from laboratory rats.

ON APPROXIMATE ENTROPY ESTIMATION ERRORS
Dragana Bajić and Nina Japundžić-Žigon