■ *Pregledni znanstveni članek*

Andrej Kastrin

O nekaterih lastnostih mnogorazsežnih podatkovij

Povzetek. Število spremenljivk, s katerimi opisujemo določen predmet proučevanja, se z razvojem mnogih področij znanosti povečuje. V analizi mnogorazsežnih podatkov se srečamo s številnimi težavami, ki so med drugim povezane s slabo identifikabilnostjo modela, numerično nestabilnostjo rešitve ali preveliko prilagojenostjo modela podatkom. Preden se lotimo zahtevnejše analiza takega podatkovja, moramo poznati glavne lastnosti mnogorazsežnega prostora. V prispevku predstavimo nekatere geometrijske lastnosti mnogorazsežnega prostora. Posebej izpostavimo pojav praznega prostora, ki ga ilustriramo na primerih hiperkocke in hipersfere. Prispevek zaključimo s pregledom dodatne literature, ki bo bralcu v pomoč pri nadaljnjem študiju.

Ključne besede: statistika; strojno učenje; mnogorazsežni podatki; pojav praznega prostora.

On Some Properties of High-Dimensional Data Sets

Abstract. The extensive use of high-dimensional data to examine certain research phenomena has expanded in parallel with high-throughput technologies in various scientific fields. However, several statistical challenges arise when analysing high-dimensional datasets, such as low identifiability, numerical instability, and overfitting. Before delving into the complex analysis of high-dimensional data, a solid foundation of their inherent properties is crucial. This review aims to illustrate the geometric properties inherent in the statistical analysis of high-dimensional data by examining the behaviour of hypercubes and hyperspheres in a high-dimensional context. An overview of literature is also provided to guide the students in further study.

Key words: statistics; machine learning; high-dimensional data; empty space phenomenon.

■ **Infor Med Slov** 2023; 28(1-2): 16-23

Institucija avtorja / Author's institution: Medicinska fakulteta, Univerza v Ljubljani.

Kontaktna oseba / Contact person: doc. dr. Andrej Kastrin, Univerza v Ljubljani, Medicinska fakulteta, Vrazov trg 2, 1000 Ljubljana, Slovenija. E-pošta / E-mail: andrej.kastrin@mf.uni-lj.si.

Prispelo / Received: 27. 11. 2023. Sprejeto / Accepted: 7. 12. 2023.

Uvod

Procesiranje informacij v mnogorazsežnem prostoru je za človeka težka naloga. Če se moramo znajti v takem prostoru, je naš spoznavni aparat zelo omejen.[1] Izkaže se, da ima večina ljudi velike težave že z miselno predstavo preprostih tri- in štirirazsežnih predmetov. Nazoren primer je štirirazsežna kocka. Ko tak model kocke predstavimo ljudem in jih prosimo, naj svojo podobo kocke prenesejo na papir, bomo hitro ugotovili, da so njihove miselne predstave zelo različne. Obstajajo sicer pričevanja posameznikov (npr. igralcev računalniških igric), da lahko učinkovito miselno manipulirajo tudi v štirirazsežnem prostoru, vendar so tovrstni izsledki zelo skopi.[2,3] S preprostim besednjakom bi lahko rekli, da človek misli (le) v prostoru treh evklidskih razsežnosti, pri procesiranju informacij v več kot treh razsežnostih pa postane nemočen. Pri opisovanju podatkovnih svetov v mnogorazsežnem prostoru si zato pomagamo z računalnikom.

Z mnogorazsežnimi podatki se dandanes srečujemo na vsakem koraku. Brez posebnih zadržkov lahko rečemo, da je sodobna podatkovna analitika v veliki meri pogojena prav z obvladovanjem mnogorazsežnih podatkovij. Še pred dobrega pol stoletja se je pojem mnogorazsežnega podatkovja navezoval na podatkovno tabelo z največ štirimi ali petimi spremenljivkami,[4] medtem ko je danes podatkovje z nekaj tisoč spremenljivkami del statističnega vsakdana.[5,6] Primere mnogorazsežnih podatkovij najdemo npr. pri analizi biomedicinskih podatkov, strojnem uvrščanju besedil, analizi finančnih transakcij ali iskanju kompleksnih vzorcev v astrofizikalnih podatkih. Običajno je za taka podatkovja značilno, da število merjenih spremenljivk (močno) presega število posameznih primerov. V metodološko zahtevnejših prispevkih avtorji to radi poudarijo z neenakostjo $n \ll p$, kjer z n označimo število primerov (enot), s p pa število merjenih spremenljivk (atributov).

Pravilna analiza mnogorazsežnih podatkovij je pogojena z dvema dejavnikoma. Prvič, mnogorazsežni prostor se ponaša z lastnostmi, ki so v primerjavi z eno- ali dvorazsežnim prostorom bistveno drugačne in pogosto neintuitivne.[6-9] Drugič, metod za analizo eno- in dvorazsežnih podatkovij ne moremo preprosto uporabiti nad mnogorazsežnimi podatkovnimi matrikami. Bralcu bo najbrž dobro poznana težava z linearno regresijo, kjer je v primeru $n < p$ vzorčna kovariančna matrika singularna, kar ima za posledico, da ne moremo izračunati njenega inverza.[10]

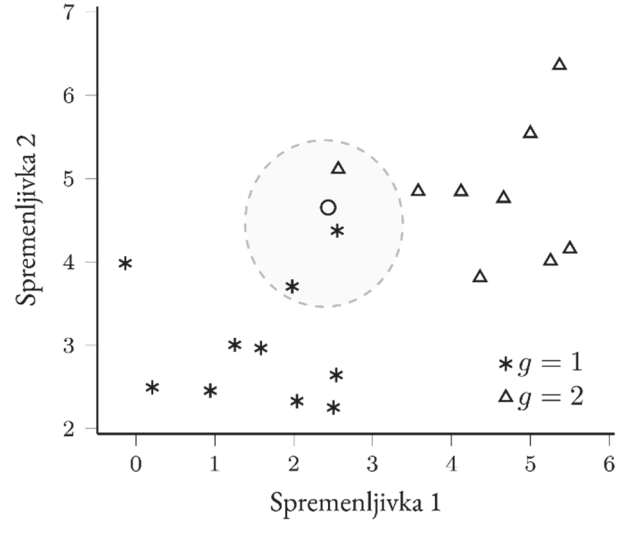
Za celovit pregled pasti, na katere naletimo v analizi mnogorazsežnega podatkovja, je potrebno predstaviti dva pojava: (i) pojav praznega prostora in (ii) pojav zgoščanja norm. Zaradi kompleksnosti tematike in omejenosti s prostorom v nadaljevanju prispevka obravnavamo le prvega. V razdelku Motivacija na primeru metode najbližjega soseda bralca najprej uvedemo v problematiko praznega prostora, ki ga nato bolj podrobno razložimo v naslednjem razdelku. V razdelku o geometrijskih lastnosti mnogorazsežnega prostora nekatere posledice praznega prostora, ki so najpomembnejše za statistično analizo, razložimo s pomočjo preproste topološke analize kocke in sfere, ki ju vložimo v mnogorazsežni prostor. Prispevek zaključimo s pregledom najpomembnejše literature, ki bo bralcu v pomoč pri nadaljnjem študiju.

Motivacija

Za boljšo predstavo obravnavajmo preprost klasifikator z metodo najbližjega soseda. Podatkovje $\mathbf{D}$ naj sestavlja n podatkovnih točk $\boldsymbol{x}_i \in \mathbb{R}^d$. Z $\mathbf{D}_i$ označimo podmnožico točk z oznako razreda g_i, tako da je $n_i = |\mathbf{D}_i|$. Napovedani razred za podatkovno točko $\boldsymbol{x}$ izračunamo kot

$$\hat{G}(x) = \arg\min g_i(K_i),$$

kjer je K_i število podatkovnih točk med K najbližjimi sosedi točke $\boldsymbol{x}$, ki so označeni z oznako razreda g_i. Situacija uvrščanja s $K = 3$ sosedi je prikazana na sliki 1.



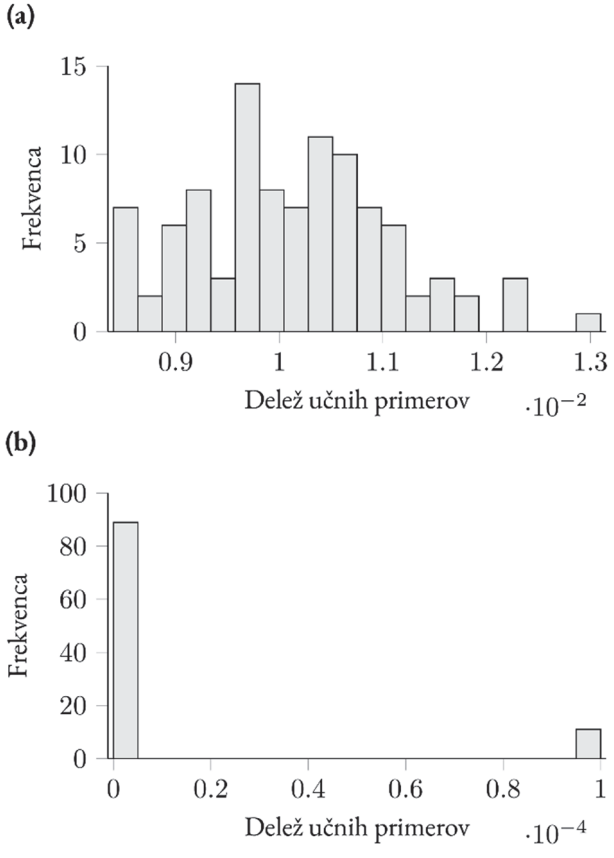
Slika 1 Metoda najbližjega soseda. Dvorazsežno podatkovje sestavlja 19 primerov, označenih z razredom 1 ali 2. Novi primer, ki ga želimo uvrstiti, ponazarja krožec. Za $K = 3$ identificiramo tri sosede znotraj senčene površine. Primer uvrstimo v večinski razred 1.

Orodje imamo, zdaj pa ga uporabimo za simulacijo. Najprej n slučajnih vektorjev

$$\mathbf{x} = (X_1, X_1, \ldots, X_d) \in \mathbb{R}^d,$$

enakomerno porazdeljenih na intervalu [0, 1], zložimo v matriko podatkov **D**. Za vsak vektor poznamo tudi dejansko oznako razreda $g_i \in \{1, 2\}$. Naloga zahteva, da nov, neznan primer uvrstimo v ustrezni razred, z omejitvijo, da lahko pri uvrščanju uporabimo le 10 % učnih primerov v intervalu λ. Če je npr. $x = 0{,}7$, bomo pri uvrščanju upoštevali le vrednosti v intervalu [0,65; 0,75]. Zanima nas, kolikšen delež učnih primerov imamo na voljo za uvrščanje pri različnem številu razsežnosti d.

Eksperiment smo pognali 100-krat ter si beležili število učnih primerov znotraj intervala λ. Pri $d = 1$ je povprečni delež primerov enak dolžini intervala λ, tj. 0,1. Kaj pa v višjih razsežnostih? Spodaj so prikazani rezultati simulacij za $d = 2$ (slika 2a) in $d = 5$ razsežnosti (slika 2b).



Slika 2 Množica točk najbližjih sosedov. V dvorazsežnem prostoru (a) je množica zelo homogena, v prostoru s petimi razsežnostmi (b) pa že močno razpršena.

Ugotovimo lahko, da je množica točk najbližjih sosedov v $d = 2$ razsežnostih kompaktna, pri $d = 5$ razsežnostih pa že zelo difuzna. Povedano drugače, z večanjem števila razsežnosti postaja okolica posameznih podatkovnih točk vse bolj prazna.[11] Lokalnost primerov, ki je za delovanje metode najbližjih sosedov ključna, se v mnogorazsežnem prostoru izgubi, klasifikator pa odpove.

Izgubljeno lastnost lokalnosti posameznih primerov literatura poimenuje »pojav praznega prostora« (angl. *empty space phenomenon*). V angleški literaturi jo pogosto zasledimo v povezavi z nadrednico *curse of dimensionality*, kar prevajamo kot »prekletstvo dimenzionalnosti«.

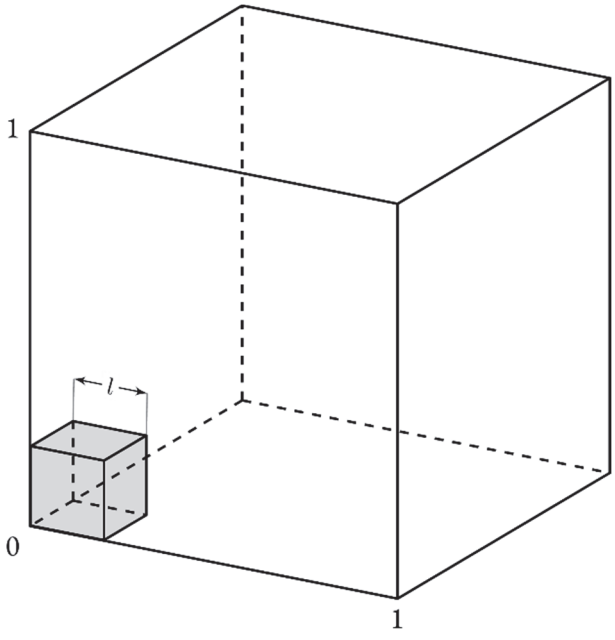
Na pojav praznega prostora v praksi pogosto naletimo v povezavi z vprašljivo identifikabilnostjo statističnega modela, numerično nestabilnostjo rešitve in prevelikim prileganjem modela podatkom. Zapomniti si velja, da metod za analizo eno- in dvorazsežnih podatkovij ne moremo preprosto uporabiti na mnogorazsežnih podatkovnih tabelah, saj problem mnogorazsežnosti prinaša s seboj mnogo statističnih težav.

Pojav praznega prostora

Zgoraj smo pokazali, da je problem analize mnogorazsežnih podatkovij neločljivo povezan s pojavom praznega prostora. Pojav je pred 60 leti prvi opisal Bellman[12] pri opisovanju problema optimizacije z metodo izčrpnega preiskovanja v produktnih prostorih. Strategija izčrpnega preiskovanja pregleda in ovrednoti vse možne rešitve v optimizacijskem prostoru, nato pa izbere zadovoljive. Pokazal je, da z linearnim povečevanjem prostora spremenljivk velikost optimizacijskega prostora raste eksponentno. To ima za posledico večjo računsko zahtevnost ter večjo verjetnost, da se optimizacija zaključi v lokalnem minimumu. Reševanje optimizacijske naloge po metodi izčrpnega preiskovanja zato že pri razmeroma majhnem številu razsežnosti preraste v neobvladljiv problem.

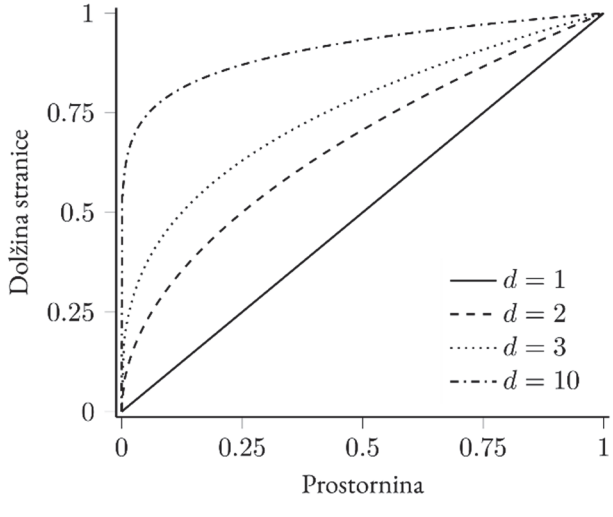
- Primer 1: Bellmanovo zakonitost ilustrirajmo s preprostim primerom. Denimo, da obravnavamo d-razsežno kartezično mrežo s korakom $\varepsilon = 1/10$. Če želimo mrežo napolniti s točkami, bomo pri $d = 10$ razsežnostih potrebovali 10^{10} točk, pri $d = 20$ razsežnostih pa se število potrebnih točk poveča že na 10^{20}; v splošnem torej potrebujemo $\mathcal{O}((1/\varepsilon)^d)$ točk. Izkaže se, da z linearnim povečevanjem prostora spremenljivk velikost prostora rešitev eksponentno raste.
- Primer 2: Imejmo d-razsežno hiperkocko, v kateri enakomerno porazdelimo podatkovne točke. Pripravimo vzorec točk iz r-tega deleža celotne

prostornine. Zanima nas dolžina stranice l (slika 3).



Slika 3 Kocko s stranico dolžine $l \leq 1$ vložimo v enotsko kocko.

Upoštevajmo, da za zvezo med dolžino stranice, številom razsežnosti in deležem zajete prostornine velja $l = r^{1/d}$. S preprostim izračunom hitro ugotovimo, da bo ob vzorčnem deležu $r = 0.01$ stranica hiperkocke pri razsežnosti $d = 1$ zavzemala 1 % celotne dolžine, pri razsežnosti $d = 10$ pa kar 63 % dolžine stranice hiperkocke. Ob vzorčnem deležu $r = 0.1$ se bo pri $d = 10$ razsežnostih dolžina stranice hiperkocke povečala na 80 %. Odnos med deležem prostornine hiperkocke in dolžino stranice je za štiri različne razsežnosti prikazan na sliki 4.



Slika 4 Odnos med deležem prostornine in dolžino stranice d-razsežne hiperkocke.

Ugotovitev, povzeta iz primerov 1 in 2, nam nalaga, da z naraščanjem števila spremenljivk v statističnem modelu zagotovimo tudi ustrezno število primerov. V nasprotnem primeru bo naš podatkovni prostor praktično prazen oziroma vsaj redek. Poznavanje pojava praznega prostora je pomembno zlasti v vsakdanji statistični praksi, saj lahko le redko zagotovimo ustrezno eksponentno rast števila primerov; večinoma imamo kljub velikemu številu spremenljivk na voljo le nekaj deset primerov.

Intuitivne predstave, ki veljajo v eno-, dvo- ali trirazsežnem prostoru, postanejo v mnogorazsežnem prostoru nepravilne. Mnogorazsežni prostor ima namreč neintuitivne geometrijske lastnosti. Predstava podatkovnih točk v večrazsežnem prostoru je lahko zato zavajajoča. Nobenih težav ne bomo imeli, če bomo želeli predstaviti podatkovje 100 enot, merjenih na dveh spremenljivkah. Iz razsevnega diagrama bomo po vsej verjetnosti lahko celo razbrali latentno strukturo podatkov (npr. skupine podatkov in odnose med spremenljivkami). Zdaj pa si predstavljajmo, da želimo predstaviti podatkovje, ki ima enako število primerov, število spremenljivk pa povečamo na 500. V razsevnem diagramu bodo podatkovne točke takega podatkovja bolj ali manj slučajno razpršene.[8] Čeprav obstaja v podatkih neka notranja struktura, bo po vsej verjetnosti iz razsevnega diagrama težko razvidna. Z večanjem števila spremenljivk namreč postajajo razdalje med posameznimi primeri v prostoru čedalje večje, kar pomeni, da se tudi najbližji primeri medsebojno zelo razlikujejo. To je glavni razlog, da se metode, ki temeljijo na lokalnosti primerov (npr. metoda najbližjega soseda, parzenova okna, Relief), slabo obnesejo pri velikem številu spremenljivk.

Geometrijske lastnosti mnogorazsežnega prostora

V tem razdelku si bomo ogledali nekatere geometrijske lastnosti mnogorazsežnega evklidskega prostora. Podrobneje bomo obravnavali (i) hiperkocko, (ii) hipersfero, (iii) razmerje med prostorninama hipersfere in hiperkocke ter (iv) prostornino tanke lupine.

Topološka analiza je za študij mnogorazsežnega prostora zelo primerna in nam bo nekoliko olajšala njegovo razumevanje. Pri pregledu smo se zgledovali po enem od novejših učbenikov s področja statističnega učenja,[13] več matematičnih podrobnosti pa bo bralec našel v starejših monografijah.[8,14]

Hiperkocka

Minimalno in maksimalno vrednost spremenljivke X_j iz podatkovne matrike **D** zapišemo kot

$$\min(X_j) = \min_i \{x_{ij}\}$$

in

$$\max(X_j) = \max_i \{x_{ij}\}.$$

Podatkovni hiperprostor **D** si lahko predstavljamo v prispodobi d-razsežnega hiperpravokotnika, ki je definiran s predpisom

$$R_d = \prod_{j=1}^{d} \left[\min(X_j), \max(X_j) \right] = \left\{ x = (x_1, x_2, \ldots, x_d)^T \right\},$$

kjer je $x_j \in [\min(X_j), \max(X_j)]$, za $j = 1,\ldots, d$. Predpostavimo še, da smo surove vrednosti spremenljivk predhodno pretvorili v odklonske vrednosti, tako da je vektor njihovih aritmetičnih sredin enak $\boldsymbol{\mu} = \mathbf{0}$. Največjo absolutno vrednost v podatkovni matriki **D** definirajmo s predpisom

$$m = \max_{j=1}^{d} \max_{i=1}^{n} \left\{ \left| x_{ij} \right| \right\}.$$

Podatkovni hiperprostor lahko zdaj obravnavamo kot hiperkocko s središčem v točki **0** in dolžino stranice $l = 2m$. Formalno bomo to zapisali kot

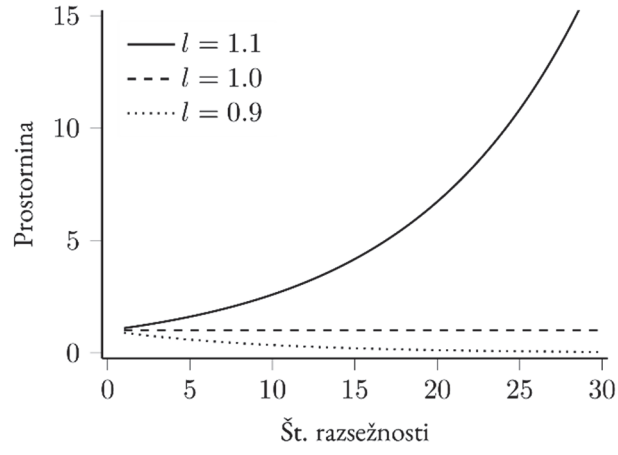
$$H_d(l) = \left\{ x = (x_1, x_2, \ldots, x_d)^T \,\middle|\, \forall i, x_i \in \left[-\frac{l}{2}, \frac{l}{2} \right] \right\}.$$

Prostornino hiperkocke s stranico dolžine l izračunamo po obrazcu

$$V\left(H_d(l)\right) = l^d.$$

Če je $l = 1$, je prostornina hiperkocke neodvisna od števila razsežnosti. Prostornina bo v tem primeru vedno enaka $V(H_d(1)) = 1$. Če je $l > 1$, bo prostornina z naraščanjem števila razsežnosti divergirala k neskončnosti, pri $l < 1$ pa konvergirala k vrednosti nič.

- Primer 3: Odnos med številom razsežnosti podatkovnega prostora in prostornino hiperkocke je za tri različne dolžine stranice prikazan na sliki 5.



Slika 5 Odnos med številom razsežnosti in prostornino hiperkocke za različne dolžine stranice.

Hipersfera

Podobno kot zgoraj predpostavimo, da spremenljivke nastopajo v odklonski obliki, tako da je $\boldsymbol{\mu} = \mathbf{0}$. Razdaljo med središčem podatkovnega hiperprostora **D** in najbolj oddaljeno podatkovno točko definirajmo s predpisom

$$r = \max_i \left\{ \left\| x_i \right\| \right\}.$$

Podatkovni hiperprostor lahko zdaj predstavimo kot d-razsežno hiperkroglo s središčem v točki **0** ter polmerom r, tako da je

$$B_d(r) = \left\{ x \,\middle|\, \|x\| \le r \right\}.$$

Površino hiperkrogle B_d ponazarja hipersfera S_d. Hipersfero sestavljajo vse podatkovne točke, ki so od izhodišča **0** oddaljene natanko za r:

$$S_d(r) = \left\{ x \,\middle|\, \|x\| = r \right\}.$$

Prostornino hipersfere v nižjih razsežnostih znamo enostavno izračunati s pomočjo znanih obrazcev, npr.

$$V\left(S_1(r)\right) = 2r,$$
$$V\left(S_2(r)\right) = \pi r^2,$$
$$V\left(S_3(r)\right) = \frac{4}{3}\pi r^3.$$

Splošen obrazec za izračun prostornine d-razsežne hipersfere je

$$V\left(S_d(r)\right) = \left(\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} \right) r^d,$$

kjer je

$$\Gamma\left(\frac{d}{2}+1\right)=\begin{cases}\left(d/2\right)! & \text{če } d \text{ sodo}\\ \sqrt{\pi}\left(\frac{d!!}{2^{(d+1)/2}}\right) & \text{če } d \text{ liho}\end{cases}.$$

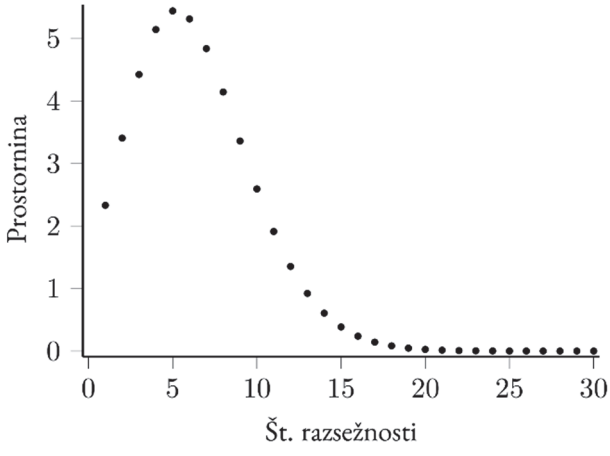
Zgoraj je z Γ označena funkcija gama, dvojna fakulteta (d!!) pa je definirana s predpisom

$$d!!=\begin{cases}1 & \text{če } d=0 \text{ ali } d=1\\ d(d-2)!! & \text{če } d\geq 2\end{cases}.$$

S povečevanjem števila razsežnosti prostornina hipersfere najprej narašča, nato pa začne padati in se približuje vrednosti nič. Za enotsko hipersfero zato velja

$$\lim_{d\to\infty} V\left(S_d(1)\right)=\lim_{d\to\infty}\frac{\pi^{d/2}}{\Gamma\left(d/2+1\right)}=0.$$

- Primer 4: Na sliki 6 je predstavljen odnos med številom razsežnosti in prostornino enotske hipersfere. Prostornina sfere najprej narašča in doseže največjo prostornino pri $d = 5$, kjer znaša $V(S_5(1)) = 5{,}26$. Prostornina se nato začne zmanjševati in pri $d = 30$ doseže zanemarljivo vrednost.



Slika 6 Odnos med številom razsežnosti in prostornino hipersfere.

Razmerje med prostorninama hipersfere in hiperkocke

Podatkovni prostor zopet omejimo s hiperkocko H, na enak način, kot smo to naredili v razdelku o hiperkocki. Vanjo postavimo karseda veliko hipersfero S. Polmer hipersfere označimo z r, stranico hiperkocke pa z $2r$. Obravnavajmo razmerje med prostorninama obeh teles. Za začetek primerjajmo obe prostornini v dveh in treh razsežnostih. V prvem primeru znaša razmerje

$$\frac{V\left(S_2(r)\right)}{V\left(H_2(2r)\right)}=\frac{\pi r^2}{4r^2}=\frac{\pi}{4}=78.5\%,$$

kar pomeni, da krožnica omejuje $\pi / 4$ površine kvadrata, v katerega je vrisana. V treh razsežnostih znaša razmerje

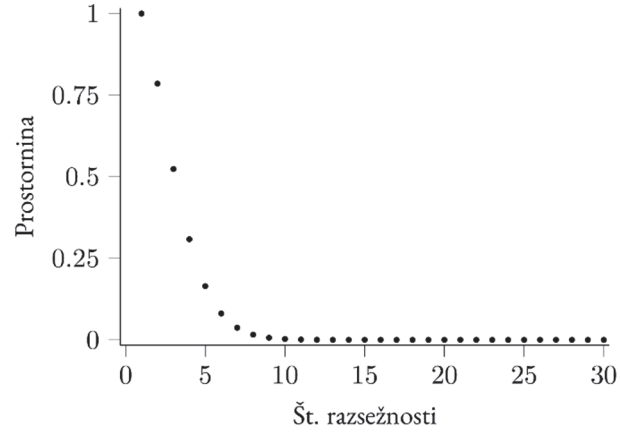
$$\frac{V}{V}=\frac{4/3\,\pi r^3}{8\pi^3}=\frac{\pi}{6}=52.4\%,$$

kar je le še $\pi / 6$ prostornine kocke. V splošnem s povečevanjem števila razsežnosti d velja

$$\lim_{d\to\infty}\frac{V\left(S_d(r)\right)}{V\left(H_d(2r)\right)}=0,$$

kar pomeni, da je asimptotična prostornina hiperkocke zgoščena ob robovih prostora, medtem ko je središče prazno.

- Primer 5: Na sliki 7 je prikazano razmerje med prostorninama enotske hipersfere in hiperkocke za različne razsežnosti prostora. Pri razsežnosti $d = 2$ znaša razmerje $\pi / 4$, kar pomeni, da hipersfera (ki je v tem primeru krog) obsega skoraj celotno prostornino (v tem primeru ploščino) kvadrata. Z naraščanjem števila razsežnosti se razmerje hitro približuje vrednosti nič ter pri $d = 10$ doseže zanemarljivo vrednost.



Slika 7 Razmerje med prostorninama hipersfere in hiperkocke za različno število razsežnosti.

Prostornina tanke lupine

Obravnavajmo še prostornino tanke lupine debeline ε, ki jo omejujeta notranja hipersfera s polmerom r ter zunanja hipersfera s polmerom $r + ε$.

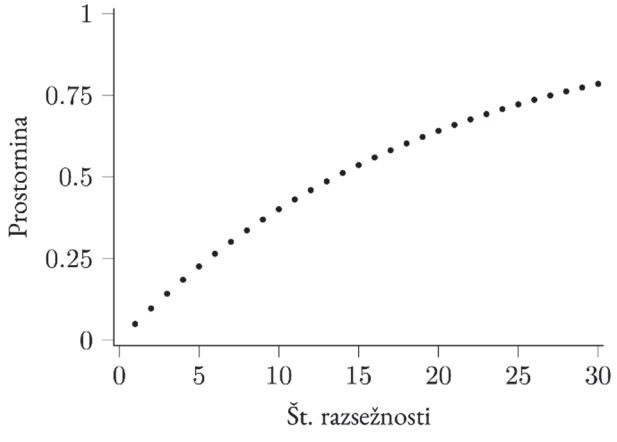
Prostornino tanke lupine $S_d(r, ε)$ izračunamo kot razliko prostornin obeh hipersfer po obrazcu

$$V\left(S_d(r,\grave{o})\right) = V\left(S_d(r)\right) - V\left(S_d(r-\grave{o})\right),$$

razmerje med prostorninama tanke lupine in zunanje sfere pa po obrazcu

$$\frac{V\left(S_d(r,\grave{o})\right)}{V\left(S_d(r)\right)} = 1 - \left(1 - \frac{\grave{o}}{r}\right)^d.$$

- Primer 6: V tanki lupini razmerje obeh prostornin narašča eksponentno z večanjem razsežnosti. Polmer fiksirajmo na $r = 1$, debelino lupine pa na $\varepsilon = 0{,}01$. V dveh razsežnostih je prostornina tanke lupine enaka $1 - 0{,}99^2 \approx 2\,\%$. V treh razsežnostih se delež prostornine poveča na $1 - 0{,}99^3 \approx 3\,\%$. Pri $d = 30$ pa prostornina lupine naraste kar na $1 - 0{,}99^{30} \approx 26\,\%$. Zaradi boljše nazornosti je na sliki 8 prikazano razmerje med dvema sferama s parametroma $r = 1$ in $\varepsilon = 0{,}05$.



Slika 8 Odnos med številom razsežnosti in prostornino tanke lupine.

Ko število razsežnosti d narašča prek vseh mej, velja

$$\lim_{d\to\infty} \frac{V\left(S_d(r,\grave{o})\right)}{V\left(S_d(r)\right)} = 1.$$

Izkaže se, da se s povečevanjem števila razsežnosti prostornina hipersfere zgošča v tanki lupini. Večji del prostornine zato najdemo v okolici površine hipersfere (znotraj ε), medtem ko je središče hipersfere prazno. Z drugimi besedami, če so podatkovne točke v d-razsežnem prostoru porazdeljene enakomerno, se bo večina točk zgostila ob robovih tega prostora.

Priporočena literatura za nadaljnji študij

Pojav praznega prostora in z njim povezane težave na kratko predstavijo mnogi učbeniki multivariatne statistične analize in statističnega učenja. Na enostaven način je pojav razložen v knjigi *An Introduction to Statistical Learning*,[15] z nekoliko več matematizacije pa tudi v sestrski *The Elements of Statistical Learning*.[16] Bralca opozarjamo, da slednja – v drugi, razširjeni izdaji – vsebuje tudi zelo lepo berljivo samostojno poglavje o analizi mnogorazsežnih podatkovij. Zahtevnejši bralec lahko poseže po Bishopovi klasiki *Pattern Recognition and Machine Learning*[17] ali prvih dveh (imenovanih *Book 0* in *Book 1*) Murphyjevih učbenikih iz serije *Probabilistic Machine Learning*.[18,19]

Nekatere najpomembnejše geometrijske lastnosti mnogorazsežnih podatkovij so posebej obravnavane v monografiji *Data Mining and Machine Learning*,[13] pa tudi v starejši *Multivariate Density Estimation*.[8] Bralcu, ki ga bo tematika posebej zanimala, priporočamo *A Course in Geometry of N Dimensions*.[14]

Zaključek

V prispevku smo obravnavali problematiko mnogorazsežnega podatkovja v analizi podatkov. Namerno smo izpostavili le prvo od dveh ključnih lastnosti, tj. pojav praznega prostora, ki smo ga ilustrirali z obnašanjem hiperkocke in hipersfere v mnogorazsežnem prostoru. Drugo lastnost, t. i. pojav zgoščanja norm, smo prihranili za objavo v prihodnosti.

Pregled lastnosti s tem nikakor ni izčrpen, je pa dovolj temeljit, da bo bralec laže krmaril med Scilo in Karibdo mnogorazsežnih podatkov.

Zahvala

Prispevek je nastal ob finančni podpori Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (J5-2552). Hvala izr. prof. dr. Roku Blagusu za pripombe in nasvete, odgovornemu uredniku, prof. dr. Gaju Vidmarju, pa za potrpežljivost ob pripravi prispevka.

Reference

1. Kellert SH: Space perception and the fourth dimension. *Man World* 1994; 27(2): 161–180.
2. Poincare H: *Mathematics and science: last essays*. Whitefish 2008: Kessinger.
3. Rucker R: *The fourth dimension: A guided tour of the higher universe*. Boston 1996: Houghton Mifflin.

4. Rao CR: The utilization of multiple measurements in problems of biological classification. *J R Stat Soc Ser B Methodol* 1948; 10(2): 159–203.
5. Guyon I, Elisseeff A: An introduction to variable and feature selection. *J Mach Learn Res* 2003; 3: 1157–1182.
6. Verleysen M: Learning high-dimensional data. In: Ablameyko S, Goras L, Gori M, Piuri V (eds.). *Limitations and future trends in neural computation*. Amsterdam 2003: IOS Press; 141-162.
7. Lee JA, Verleysen M: *Nonlinear dimensionality reduction*. New York 2007: Springer.
8. Scott DW: *Multivariate density estimation: theory, practice, and visualization*. Hoboken 1992: Wiley.
9. Verleysen M, François D: *The curse of dimensionality in data mining and time series prediction*. New York 2005: Springer.
10. Kirk M: *Thoughtful machine learning with Python: a test-driven approach*. Boston 2017: O'Reilly.
11. Beyer K, Goldstein J, Ramakrishnan R, Shaft U: When is "Nearest neighbor" meaningful? In: Beeri C, Buneman P (eds.). *Database Theory – ICDT'99: 7th International Conference; 1999 Jan 10-12; Jerusalem*. Berlin 1999: Springer; 217–235.
12. Bellman RE: *Adaptive control processes: a guided tour*. Princeton 1961: Princeton University Press.
13. Zaki MJ, Meira W: *Data mining and machine learning: fundamental concepts and algorithms*. Cambridge, UK 2020: Cambridge University Press.
14. Kendall MG: *A course in geometry of N dimensions*. London 2018: Forgotten Books.
15. James G, Witten D, Hastie T, Tibshirani R: *An introduction to statistical learning: with applications in R*. New York 2013: Springer.
16. Hastie T, Tibshirani R, Friedman J: *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York 2016: Springer.
17. Bishop CM: *Pattern recognition and machine learning*. New York 2006: Springer.
18. Murphy KP: *Machine learning: a probabilistic perspective*. Cambridge, MA 2013: MIT Press.
19. Murphy KP: *Probabilistic machine learning: An introduction*. Cambridge, MA 2022: MIT Press.