UPORABE METODE GLAVNIH KOMPONENT PRI KLASIFIKACIJI TUMORJEV

EMA PLEŠKO

Fakulteta za biotehnologijo
Univerza v Ljubljani

Članek opisuje možnost uporabe metode glavnih komponent kot pomoč pri klasifikaciji tumorjev. Najprej predstavi problem diagnoze tumorjev in matematične osnove o sami metodi glavnih komponent. Sledi poglavje o uporabi metode za klasifikacijo tumorskih celicnih linij na dejanskem primeru, ki navaja tudi prednosti in slabosti analizne metode v danem kontekstu. Članek se zaključi z osnovnim primerom uporabe PCA na izbranem podatkovju, ki preveri in podkrepi učinkovitost metode.

USE OF PRINCIPAL COMPONENT ANALYSIS FOR TUMOUR CLASSIFICATION

The article describes one of the various possibilities of using principal component analysis - PCA in biotechnological context. Firstly there is a short presentation of PCA method in more general terms, followed up with presentation of a research that actually applied PCA for tumour analysis based on data collected from cancer cell lines. At the end of the article there is a simple hand-on implementation of the pca in famous iris dataset. The implementation highlights the usefulness of the PCA.

1. Uvod

Vsak tip celice ima svoj edinstven ekspresijski profil izražanja genov, ki ga lahko izkoristimo za napoved vloge celice v biološkem sistemu. Kljub temu da se normalne in rakave celice fenotipsko bistveno razlikujejo, so njuni ekspresijski profili v nekaterih delih skoraj enaki, saj je za osnovno preživetje celice potrebno izražanje enakih genov. To so geni za podvajanje DNA, celično dihanje, rast, delitev, sinteza esencialnih snovi in mnoge druge. Torej, pri analizi genske ekspresije dobimo ogromno podatkov, a so za učinkovito razlikovanje med normalnimi in rakavimi celicami ključni le nekateri [1].

Za uspešno diagnozo in zdravljenje raka je bistvenega pomena zanesljiva in natančna klasifikacija tumorjev. Trenutne metode za razvrščanje človeških rakavih tumorjev temeljijo na različnih morfoloških, kliničnih in molekularnih spremenljivkah. Vendar kljub nedavnemu napredku na tem področju še vedno obstajajo negotovosti pri diagnozi [2].

Tehnologija DNA mikromrež je omogočila spremljanje izražanja genov na genomski ravni, ki vključuje na tisoče različnih genskih produktov. Za pomoč pri diagnozi tumorjev se informacije o genski ekspresiji in produktih zbira iz vzorcev tumornih tkiv in celic. Za obravnavanje ogromne količine podatkov in luščenje koristnih informacij iz njih je pomembno, da se uporabljajo pristopi za analizo podatkov primerni za večdimenzionalne probleme. Metoda glavnih component (*angl. principal components analysis - PCA*) je močna metoda, ki se rutinsko uporablja za pridobivanje bistvenih informacij iz večdimenzionalnih podatkov [3], običajno za vizualizacijo le teh oziroma predobdelavo za nadaljnje strojno učenje.

2. Metoda glavnih komponent

Metoda glavnih komponent, v nadaljevanju PCA, je statistična metoda, ki uporablja ortogonalno transformacijo za preslikavo koreliranih podatkov v množico čim bolj linearno neodvisnih spremenljivk, ki jih imenujemo glavne komponente. Transformacija je definirana tako, da ima prva

komponenta čim večjo varianco, torej da pojasni čim več variabilnosti v podatkih kot je le mogoče. Tudi vsaka nadaljnja komponenta je definirana tako, da ima največjo možno variabilnost, pogojno, da je hkrati še ortogonalna na vse prejšnje komponente. Tako pridobljene komponente tvorijo nekorelirano ortogonalno bazo [4]. PCA je občutljiva na skalo vhodnih spremenljivk, zato je priporočjivo, da vhodne spremenljivke normaliziramo, torej vsaki spremenljivki odštejemo povprečje in jo delimo z njenim standardnim odklonom.

Definicija 1. Normalizacija slučajne spremenljivke X:

$$Z = \frac{X - mean(X)}{sd(X)}.$$

PCA je ena izmed najpreprostejših analiz, ki temelji na lastnih vrednostih. Lahko se je lotimo na dva načina [4]:

- dekompozicijo kovariančne ali korelacijske matrike na pripadajoče lastne vrednosti,
- dekompozocijo singularnih vrednostih podatkov po normalizaciji.

Sledi splošnejša izplejava oziroma navodila izpeljave metode glavnih komponent. V članku je opisana samo prva točka od naštetih in naslednji način je tudi v zadnjem poglavju narejen PCA na testnem primeru.

Definicija 2. Izračun glavnih komponent [4]. Velike črke označujejo matrike in ustrezno matrično množenje:

- Naj bo X začetno podatkovje. Prvo normaliziramo vsako spremenljivko (stolpec) $X_i \in X$. Označimo normalizirano matriko podatkov z Z.
- Izračunamo kovariačno matriko, označimo jo z C:

  $$C = \frac{1}{n-1} Z^T Z,$$

  kjer je n število vrstic matrike Z.
- Izračunamo lastne vektorje in lastne vrednosti kovariančne matrike:

  $$V^{-1} C V = D,$$

  kjer V matrika lastnih vektorjev in D diagonalna matrika, ki ima po diagonali lastne vrednosti matrike C. Lastne vrednosti uredimo po velikosti in pazimo da ohranimo pripadajoče vrstni red lastnih vektorjev.

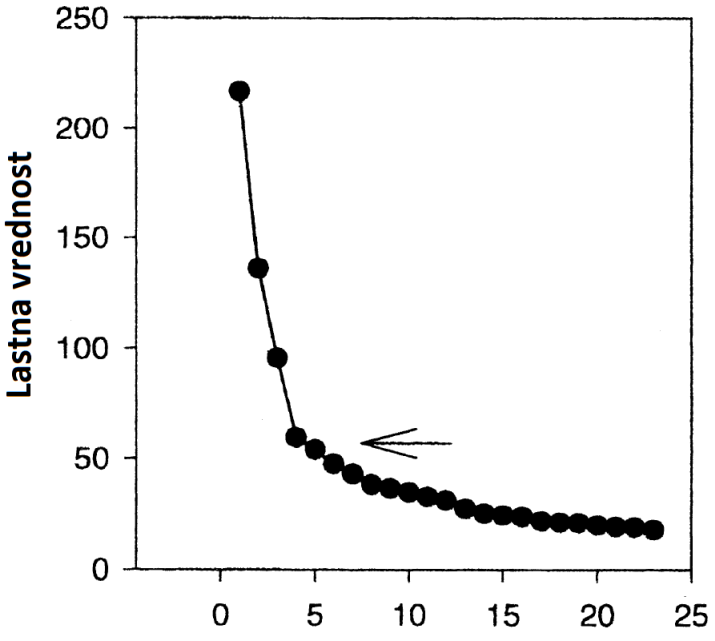
Velikost posamezne lastne vrednosti relativno na vsoto vseh nam pove koliko variabilnosti pojasni posamezna komponenta. Običajno izberemo določen odstotek variabilnosti, ki ga želimo pojasniti, na primer 90% in pogledamo koliko komponent rabimo, da pojasnimo vsaj željen odstotek variabilnosti. Začetni prostor skrčimo tako, da izbrano število največjih lastnih vektorjev zložimo v matriko in jo matrično pomnožimo z začetnimi podatki (matriko Z).

3. Uporaba PCA pri obdelavi podatkov

Tumorske celične linije, pridobljene iz ene rakave celice in gojene zunaj telesa v umetnih razmerah, se pogosto uporablajo za tako imenovane *in vitro* tumorske modele. To so modeli, ki analizirajo procese oziroma poizkuse, kateri potekajo v nadzorovanem okolju zunaj živega organizma. Nedavni podatki o genski ekspresiji stotih celičnih linij zdaj omogocajo sistematično genomsko primerjavo celičnih linij in tumorjev. S primerjavo ekspresijskih profilov tumorskih celičnih linij lahko ugotavljamo glavne skupne lastnosti, ki najverjetneje ločijo rakave celice od normalnih. Več kot takšnih lastnosti odkrijemo, boljšo in natančnejso diagnozo lahko postavimo [5].

Crescenzi in soavtorji [6] so poskusili izpostaviti glavne biološke spremenljivke pri klasifikaciji tumorskih linij z uporabo PCA. Tabele s podatki o posameznih genih imajo lahko brez težav več 1000 napovednih spremenljivk, kar se izkaže za velik problem pri običajnih klasifikacijskih metodah. Zato je dobra preslikava na nižje dimenzionalni prostor zelo pomemba. Avtorji raziskave so analiziriali podatke o 1375 izvedenih genih, kar je na koncu naneslo na 1416 napovednih spremenljivk - nekateri geni so doprinesli k več kot eni napovedni spremenljivki. Cilj je bil klasificirati 60 različnih rakavih celičnih linij. Vsaka rakava celična linija je predstavljala eno vrstico v začetnih podatkih, opisano s 1416 spremenljivkami. Začetno dimenzijo opsinih spremenljivk so želeli čim bolj skrčiti, zato so se odločili za uporabo PCA. Izkazalo se je, da so lahko s samo 5 komponentami opisali kar 40% variabilnosti - kar je za avtorje raziskave pomenilo zadosten odstotek. Porazdelitev dobljenih lastnih vrednosti si lahko ogledamo na Sliki 1 in Sliki 2.

Slika 1. Porazdelitev pridobljenih lastnih vrednosti s PCA metodo urejenih po velikosti.



Vseh 60 rakavih celinčnih linij so nato predstavili s petimi glavnimi komponentami. Pridobljen podprostor so avtorji nadaljnje analizirali in poskušali klasificirati tumorje.

4. Klasifikacija tumorja

Avtorji raziskave so se klasifikacije tumorjev lotili z grupiranjem pridobljenega podprostora. Uporabili so algoritem k - centroidov (*angl.* k - *means*), kjer so izbrali $k = 6$. Algoritem k - *means*

Slika 2. Tabela pridobljenih lastni vrednosti s PCA metodo in prikazan odstotek variabilnosti.

Component number	Eigenvalue	Variance (%)	Cumulative
1	216.70	15.30	15.3
2	135.85	9.59	24.9
3	95.44	6.74	31.6
4	59.42	4.20	35.8
5	53.80	3.80	39.6
6	47.35	3.34	43.0
7	42.82	3.02	46.0
8	38.04	2.69	48.7
9	36.40	2.57	51.3
10	34.49	2.44	53.7
11	32.40	2.29	56.0
12	30.90	2.18	58.2
13	27.18	1.92	60.1
14	25.12	1.77	61.9
15	24.20	1.71	63.6

je metoda, ki poišče podobne podatke in jim dodeli enak razred glede na željeno metriko razdalje (pogosto kar evklidska). Običajno se na začetku naključno izbere k začetnih točk kot centroide, nato se za vsak učni primer poračuna razdaljo do najbližjega centroida in dodeli grupe [7]. Ko so enkrat grupe dodeljene, se poračuna povprečna razdalja točk iz iste grupe kot novi centroid (to se ponovi za vsako grupo) in nato se zopet dodeli točke novim centroidom. Ko se enkrat centroidi ne spreminjajo več, pomeni, da je metoda skonvergirala.

Avtorji so poizkusili poiskati podobne gene tudi s hiearhičnim grupiranjem, ki je pokazalo izjemno ujemanje s šestimi skupinami pridobljenimi s k *-means* algoritmom. Hierarhično grupiranje je malce drugačno, pri njem se ponovno izbere željena metrika razdalje in nato se dva najbližja učna primera združi skupaj v enega. Nato se to ponavlja dokler ne ostane samo ena grupa [8]. Rezultate se običajno prikaže z dendogramom. Tako so rezultate prikazali tudi avtorji raziskave. Vsem šestim indikatorjem so dodelili tudi smiselno biološko interpretacijo. Dendogram grupiranja in indikatorje dodeljene celičnim linijam lahko vidimo na Sliki 3.

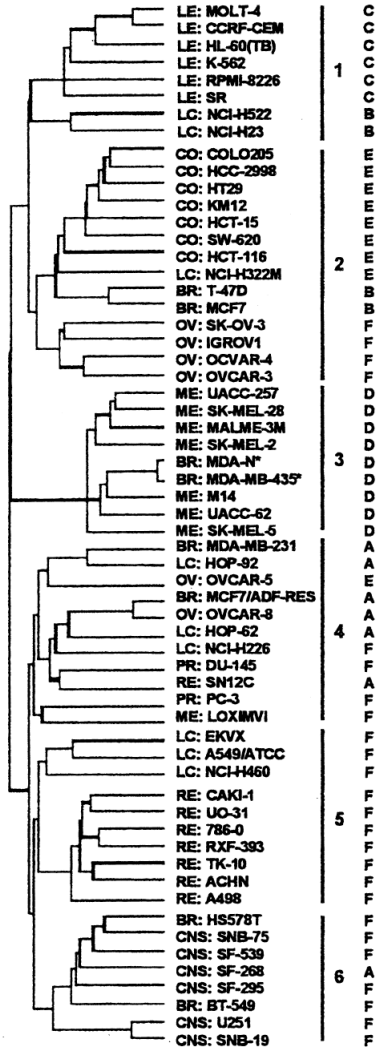
Avtorji so pokazali zanimiv pristop k analizi celičnih linij, ter našli zanimive relacije med posameznimi celičnimi linijami, ki bi jih prej zaradi obsežnosti napovednega prostora najverjetneje spregledali. Tukaj velja poudariti, da je raziskava stara že kar nekaj časa (2001) in takrat je obdelava takega števila genov predstavljala lep napredek.

5. Slabosti PCA in možnje nadgradnje

Glede na to, da je raziskava [6] iz leta 2001, je vmes prišlo kar nekaj novih metod za obdelova visoko razsežnih podatkov in tudi računska moč računalniških procesorjev je močno narasla. Po Moore-vem zakonu se računska moč računalniških procesorjev podvoji na vsake dve leti. Pojavi se vprašanje ali je PCA analiza še vedno predstavlja dovolj dober pristop.

Slabost PCA analize je, da zajame samo linearne relacije med napovednimi spremenljivkami,

Slika 3. Dendogram hierarhičnega grupiranja rakavih celičnih linij.



torej čim imamo kakršnokoli nelinearno odvisnost, ta ne bo izražena v glavnih komponentah. Sedaj je odvisno, koliko nam to zares škodi, predvsem pri analizi genov, potrebno bi bilo analizirati koliko je zares nelinearnih odvisnosti in ali so sploh pomembne.

Kot alternativa PCA analizi zadnje čase veliko dobiva na pozornosti globoko učenje (*angl. deep learning*). Predvsem avtokodirniki (*angl. autoencoders*), ki so zmožni podatke skrčiti vsaj tako dobro kot PCA, vendar so zmožni zajeti tudi nelinearno odvisnosti [9]. Slabost le teh je seveda velika časovna in računska zahtevnost, še toliko bolj, če jo primerjamo z relativno nezahtevno metodo glavnih komponent. Vse to so odprta vprašanja in možne teme nadaljnjih raziskav.

6. Preizkus metode glavnih komponent na testnem podatkovju *iris*

Cilj tega poglavja je narediti preprost preizkus, ki preveri oziroma pokaže, da metoda zares deluje in je uporabna. Za implementacijo se je uporabilo programsko okolje R in zelo znano podatkovje *iris*, katero je že vgrajeno v osnovnem programskem okolju. Podatkovje vsebuje 150 učnih primerov, vsak primer ima 4 numerične napovedne spremenljivke in oznako kateri vrsti rož pripada. Ciljna spremenljivka ima tri možne razrede.

Prvo se izračuna glavne komponente in analizira lastne vrednosti, nato pa skrči napovedni prostor na izbrano število dimenzij in oceni izgubo informacije. Programska koda je ponovljiva in dostopna v prilogi.

Analiza lastnih vrednosti:

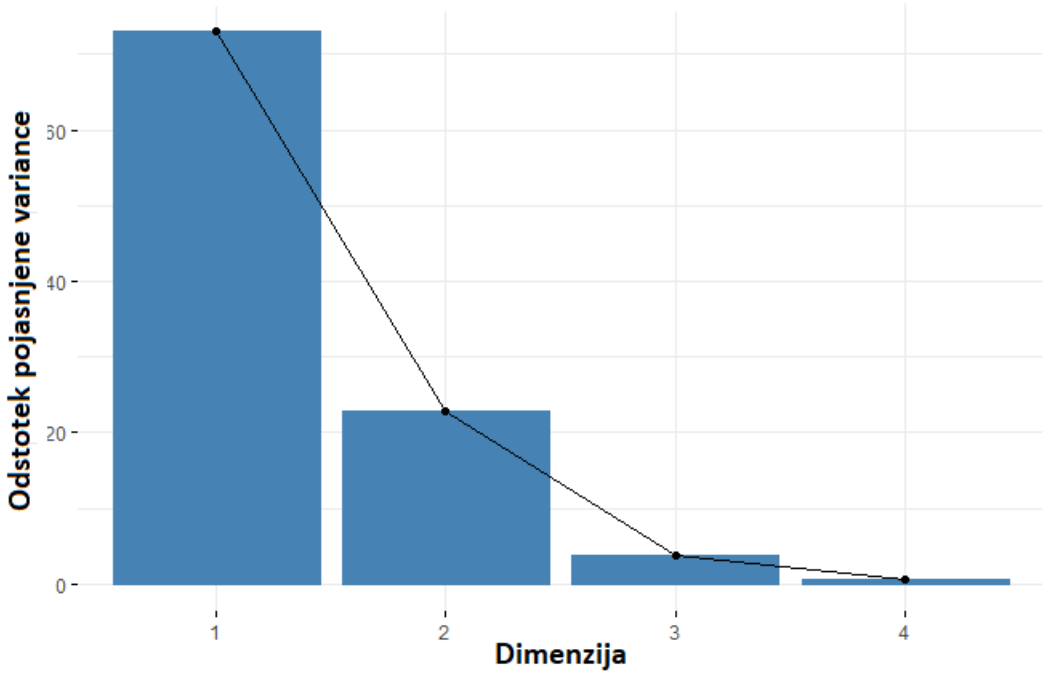
Porazdelitev lastnih vrednosti vidimo na Sliki 4, vrednosti in relativni odstotki pa so predstavljeni v Tabeli 1. Opazimo, da z dvema komponentama pojasnimo več kot 80% variabilnosti v podatkih, zato je smiselno prostor skrčiti na dve dimeziji. Prav tako sta dve dimenziji priročni za vizualno analizo.

Tabela 1. Tabela lastnih vrednosti glavnih komponent in njihova relativna velikost v odstotkih:

	V1	**V2**	**V3**	**V4**
Lastne vrednosti	**1.708**	**0.956**	**0.383**	**0.144**
Odstotek	**0.535**	**0.300**	**0.120**	**0.045**

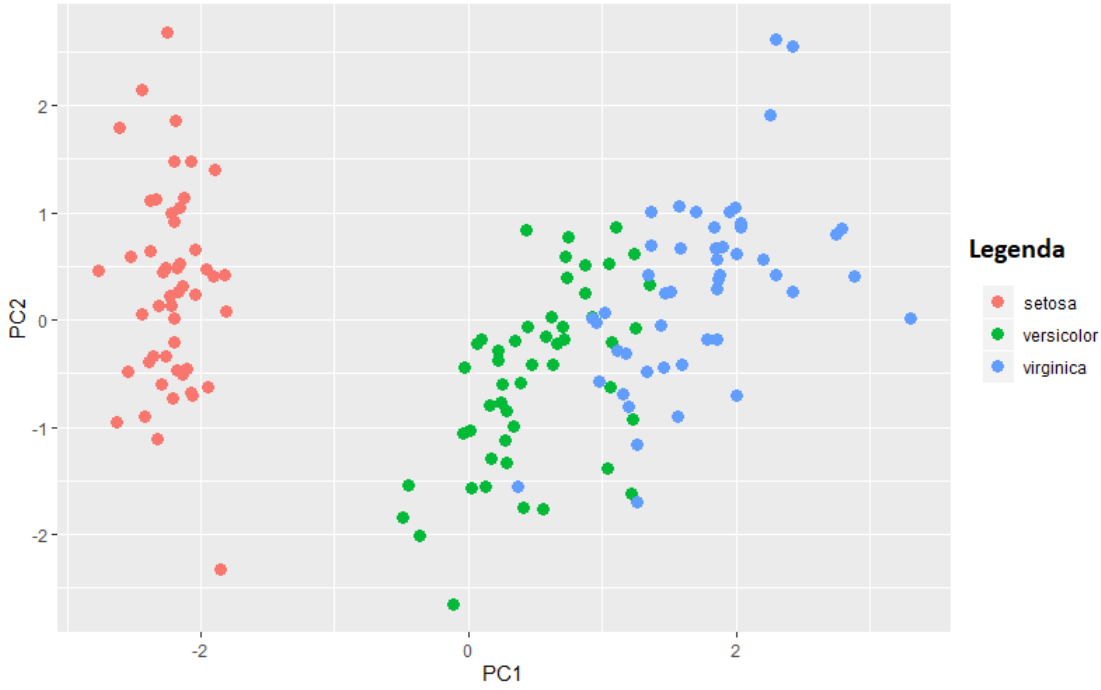
Vizualizacija:

Slika 4. Odstotek pojasnjene variance posameznih glavnih komponent:



Oglejmo si projekcijo dveh glavnih komponent na grafu. Glede na to, da imamo označene podatke, torej vemo kateri učni primeri pripadajo določeni skupini rož, ne rabimo delati grupiranja, ampak lahko samo izrišemo podatke in obarvamo učne primere po skupinah rož ter opazujemo ali smo dobili kaj smiselnega. Rezultate vidimo na sliki 5.

Slika 5. Slika dveh glavnih koponent z označenimi primeri:



Opazimo, da lahko eno skupino zelo lepo ločimo od drugih dveh (rdečo), medtem ko se zelena in modra na robovih malce prekrivata. Verjetno se tu pozna izugba 17% variabilnosti.

7. Zaključek

Metoda glavnih komponent je močno in relativno preprosto orodje za analizo večdimenzionalnih podatkov. Še posebno je uporabna pri analizi genov in zmanjšanju začetne dimenzije. V raziskavi smo videli izjemno uporabo le tega, saj so avtorji ohranili 40% variabilnosti, medtem ko so zmanjšali dimenzijo prostora iz 1416 na 5. Članek pokaže, da metoda deluje tudi na preprostih podatkovjih, kot je demonstrirano na primeru. Na koncu bi se rada zahvalila vsem, ki so mi pomagali pri razumevanju meni osebno težjih matematičnih pojmov in metod iz strojnega učenja. Hvala.

8. Priloga: programska koda R

```
set.seed(10)
data <- iris
pca <- prcomp(iris[,1:4], center = T, scale. = T)
lastne_vrednosti <- pca$sdev
relative_lastne_vrednosti <-pca$sdev / sum(pca$sdev)
reduced_space <- pca$x[,1:2]
factoextra::fviz_eig(pca)
tmp <- as.data.frame(cbind(reduced, as.character(iris[,5])))
colnames(tmp) <- c("PC1","PC2","Label")
tmp$PC1 <-as.numeric(as.character(tmp$PC1))
tmp$PC2 <- -as.numeric(as.character(tmp$PC2))
ggplot2::ggplot(data=tmp) + geom_point(aes(x=PC1, y=PC2, color=Label), lwd=3)
```

LITERATURA

[1] L. Zhang, W. Zhou, V. E. Velculescu et. al. 1997, *Gene expression profiles in normal and cancer cells*, Science. 276, 5316: 1268 - 1272

[2] S. Dudoit, J. Fridlyand, T.Speed. 2002, *Comparison of discrimination methods for the classification of tumors using gene expression data*, Journal of the American Statistical Association, 97(457): 77-87.

[3] Hua-Long Bu, Guo-Zheng Li, Xue-Qiang Zeng. 2007, *Reducing error of tumor classification by using dimension reduction with featire selection*, The first international symposium on optimization and systems biology (OSB 07). 232-241.

[4] I. Jolliffe. 2011, *Principal Component Analysis*, International Encyclopedia of Statistical Science: 1094-1096

[5] S. Domcke, R. Sinha, D. A. Levine et al. 2013, *Evaluating cell lines as tumor models by comparison of genomic profiles*, Nature communication. 4, 2126

[6] M. Crescenzi, A. Giuliani. 2001. *The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data*, FEBS Letters 507: 144-118

[7] K-means clustering, *https://en.wikipedia.org/wiki/K-means_clustering*, z dne 11.2.2020

[8] Hierarchical clustering, *https://en.wikipedia.org/wiki/Hierarchical_clustering*, z dne 11.2.2020

[9] Auteencoders, *https://en.wikipedia.org/wiki/Autoencoder*, z dne 11.2.2020