UNIVERZA V LJUBLJANI

FAKULTETA ZA ELEKTROTEHNIKO

Poročilo - Vaja3
Uporaba LSE, PCA in PCR metode za identifikacijo modela

Inteligentni sistemi za podporo odločanju

Martin Knap, 64180369

Mentor: prof. dr. Igor Škrjanc

Asistent: doc. dr. Dejan Dovžan

Ljubljana, 14. 4. 2019

Kazalo

Slike

Tabele

1 Definicija naloge

Pri tej vaji je naša naloga identificirati model z uporabo metod LSE, PCA ter PCR. Bolj točno gre za modeliranje procesa oksidacije amoniaka v dušikovo kislino, kjer imamo podane podatke za pretok hladilnega zraka (A_FLOW), temperaturo vode (T_H2O), koncentracijo kisline (C_ACID) ter inverzno vrednost izkoristka (I_EFF).Naloga od nas zahteva, da identificiramo enačbo s katero je možno izračunati inverzno vrednost izkoriska na podlagi ostalih podatkov.

2 Vpliv standarizacije na rezultate

Prvi del naloge od nas zahteva, da preverimo in medsebojno primerjamo modela doblejna z metodo analize glavnih komponent PCA (*angl.: Principle Component Analysis*) in metodo najmanjših kvadratov LSE. Pri tem moramo implementerati tri različne standarizacije podatkov:

(a) Standarizacija na interval [0, 1]: $X = (X - min(X))/(max(X) - min(X))$

(b) Standarizacija s standardno deviacijo in odštetim povprečjem: $X = (X - mean(X))/std(X)$

(c) Nestandarizirani podatki.

V primeru izračuna z standariziranimi podatki je po izračunu parametrov potrebna pretvorba iz implicitne v eksplicitno obliko.

2.1 LSE

Metoda najmanjših kvadratov v splošnem služi za regresijsko analizo, jedro metode pa je minimizacija vsote napak med modelom in analiziranim sistemom. V našemu primeru smo uporabili model s prostim členom r kar v matrični obliki izgleda tako, da je matriki kjer stolpci predstavlajajo izmerke posamezne merjene veličine dodan stolpec enic.

$$\underline{X} = [\underline{X_1}, \underline{X_2}, \underline{X_3}, 1] \tag{1}$$

Parametre se pri LSE izračuna po spodnji enačbi, kjer $\underline{Y}$ prestavlja stolpični vektor izhodov merjenega sistema.

$$\underline{\theta} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} \tag{2}$$

Varianca ocenjenih parametrov se določi po spodnji enačbi kjer e predstavlja razliko med modelom in izhodom sistema, N število meritev ter n število parametrov.

$$\sigma^2 = \frac{\underline{e}^T \underline{e}}{N - n} \tag{3}$$

2.2 PCA

Metoda glavnih komponent ali PCA temelji na analizi kovariančne matrike osrediščenih in normiranih podatkov. Gre za transformacijo osi koordinatnega sistema v nov ortogonalni sistem, ki rezultira v največji kovarianci podatkov glede na novo izbrane osi.

Pri PCA smo vse vhodne podatke kot tudi izhodne podatke zapisali v matriki $\underline{X} = [\underline{X_1}, \underline{X_2}, \underline{X_3}, \underline{Y}]$. Nato je sledila dekompozicija na lastne vrednosti D in lastne vektorje P (Matlab: *svd*):

$$\underline{F} = \frac{\underline{X}^T \underline{X}}{N - 1} \tag{4}$$

$$[\underline{P}, \underline{D}, \underline{P}^T] = svd(\underline{F}) \tag{5}$$

Nato smo izbrali lastni vektor $\underline{P}_1 = \underline{P}(:, i)$, ki kaže v smeri največje variance. To je lastni vektor, ki mu pripada najmanjša lastna vrednost in predstavlja enačbo (parametre) našega modela. Sedaj smo dobili enačbo v obliki:

$$\underline{P1}(\underline{X} - \underline{V}) = 0 \tag{6}$$

V zgornji enačbi $\underline{V}$ predstavlja vektor centrov oziroma srednje vrednosti obravnavane matrike podatkov $\underline{X}$.

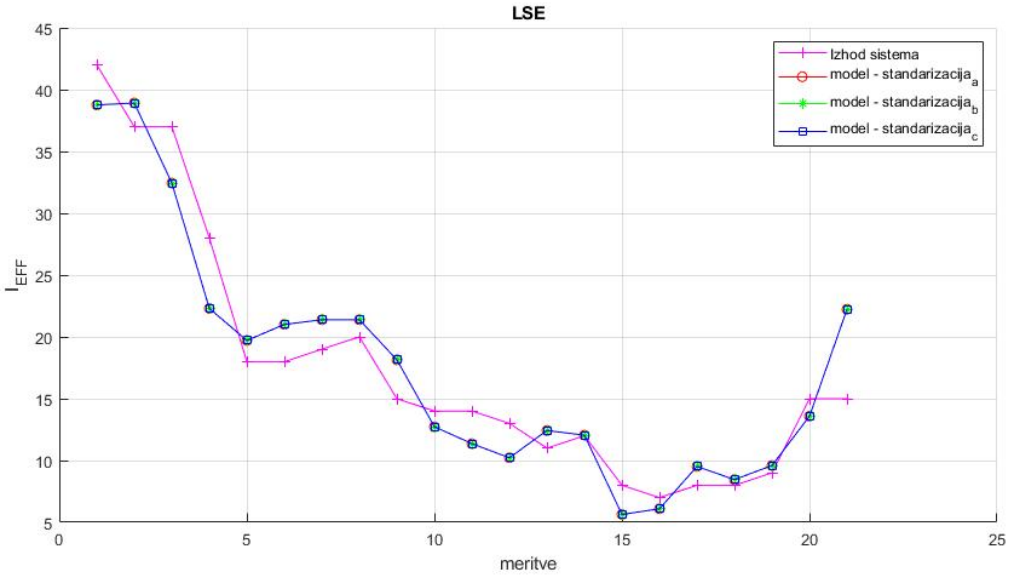
Naslednji korak predstavlja pretvorba v eksplicitno enačbo modela:

$$\underline{Y} = -(\frac{P_1}{P_4}x_1 + \frac{P_2}{P_4}x_2 + \frac{P_3}{P_4}x_3) + (\frac{P_1}{P_4}V_1 + \frac{P_2}{P_4}V_2 + \frac{P_3}{P_4}V_3) + V_4 \tag{7}$$

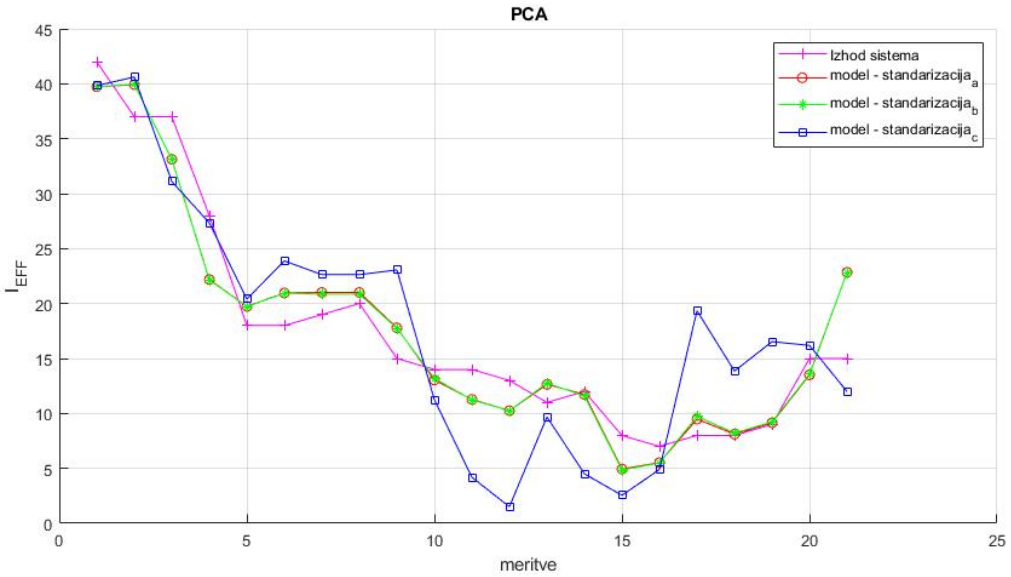
V primeru standariziranih podatkov je bila potrebna še transformacija v originalni prostor.

2.3 Rezultati

Na spodnjih slikah so prikazani izhodi modelov pri LSE in PCA metodah za vse tri vrste standarizacij.



Slika 1: Rezultati LSE analize.



Slika 2: Rezultati PCA analize.

Na spodnjih tabelah so prikazane vrednosti izračunanih parametrov z metodama LSE in PCA.

Tabela 1: Tabela parametrov LSE.

standarizacija_a	standarizacija_b	standarizacija_c
0.7156	0.7156	0.7156
1.2953	1.2953	1.2953
-0.1521	-0.1521	-0.1521
-39.9197	-39.9197	-39.9197

Tabela 2: Tabela parametrov PCA.

standarizacija_a	standarizacija_b	standarizacija_c
0.7921	0.7988	0.2046
1.2244	1.2134	3.4608
-0.1952	-0.2159	-0.7845
-39.3239	-37.7156	-0.1573

Primerjava modelov in izhoda sistema z izračunom popvrečne vrednosti napake, standardne deviacije napake ter NRMSE napake:

Tabela 3: Napaka LSE.

	standarizacija_a	standarizacija_b	standarizacija_c
povprečna vr. napake	-5.5828e-15	1.6072e-15	7.5656e-13
std. deviacija napake	2.9902	2.9902	2.9902
NRMSE	0.2869	0.2869	0.2869

Tabela 4: Napaka PCA.

	standarizacija_a	standarizacija_b	standarizacija_c
povprečna vr. napake	-1.2688e-15	-5.9212e-16	-9.7277e-15
std. deviacija napak	3.0281	3.0343	6.0683
NRMSE	0.2905	0.2911	0.5822

Varianca ocenjenih parametrov po LSE metodi:

Tabela 5: Varianca LSE.

	standarizacija_a	standarizacija_b	standarizacija_c
σ^2	9.9350	9.9350	9.9350

2.4 Komentar

Prvo kar opazimo je da pri LSE metodi standarizacija nima vpliva na rezultat. To lahko opazimo tako iz grafa in tabele parametrov, kot tudi iz tabele napak. Povprečne napake so za vse tri standarizacije precej majhne. Relativno majhne so tudi standardne deviacije napake ter NRMSE, ki so obenem tudi enake.

Pri PCA metodi modeliranja že iz grafa opazimo, da ima standarizacija dokajšen vpliv, saj pri modelu iz nestandariziranih podatkov vidimo večja odstopanja. Slednje je moč opaziti tudi iz tabele parametrov, kjer so vrednosti precej drugačne. Povprečna napaka je sicer precej majhna, vendar je deviacija kar dvakrat večja od tiste, ki jo imata modela iz standariziranih podatkov. Povprečni napaki, standardni deviaciji napake ter NRMSE standariziranih modelov so precej blizu tistim iz LSE analize. Sklep je ta da sta modela iz standariziranih podatkov bolj konsistentna, pristranskost je pa pri vseh treh modelih približno enaka. Rezultat LSE analize v kontekstu deviacije napake ter NRMSE je v primerjavi s PCA malenkost boljši. Pri PCA sta oba načina standarizacije dala podobna rezultata.

Če hočemo pri PCA analizi pridobiti model ustrezne kvalitete je potrebno podatke standarizirati. Na ta način se poveča vpliv spremenljivkam z večjo varianco.

3 Problem kolinearnosti

V drugem delu naloge smo obstoječim podatkom dodali meritev, ki je odvisna od že obstoječe meritve.

$$\underline{X}_{dodatna} = 2 * T\_H2O + 6 + 0.1 * randn(length(T\_H2O, 1)) \tag{8}$$

Z dodano odvisno meritvijo je sedaj potrebno izračunati model po LSE ter po PCR (*angl.: Principal Component regression*) ter ju primerjati. Pri izračunu modelov so bili podatki standarizirani tako kot v prvem delu naloge.

3.1 PCR

Regresija glavnih komponent (PCR) je metoda, ki temelji na metodi glavnih komponent (PCA). Pri PCR gre za regresijo izhoda sistema na naboru neodvisnih spremenljivk (meritev) kjer se za določoanje neznanih regresijskih koeficientov uporablja PCA.

Za določanje modela tvorimo matriko meritev $\underline{X} = [\underline{X}_1, \underline{X}_2, \underline{X}_3, \underline{X}_{dodatna}]$ ter vektor izhodov sistema $\underline{Y}$. Najprej izvedemo dekompozicijo lastnih vrednosti.

$$\underline{F} = \frac{\underline{X}^T \underline{X}}{N-1} \tag{9}$$

$$[\underline{P}, \underline{D}, \underline{P}^T] = svd(\underline{F}) \tag{10}$$

Nato odstranimo komponento z zanemarljivim vplivom ter tvorimo matriko zadetkov $\underline{T}$:

$$\underline{P}_s = \underline{P}(:, 1:3) \tag{11}$$

$$\underline{T} = \underline{X}\underline{P}_s \tag{12}$$

Sedaj določimo paramatre modela z LSE:

$$\underline{\theta}_T = (\underline{T}^T\underline{T})^{-1}\underline{T}^T\underline{Y} \tag{13}$$

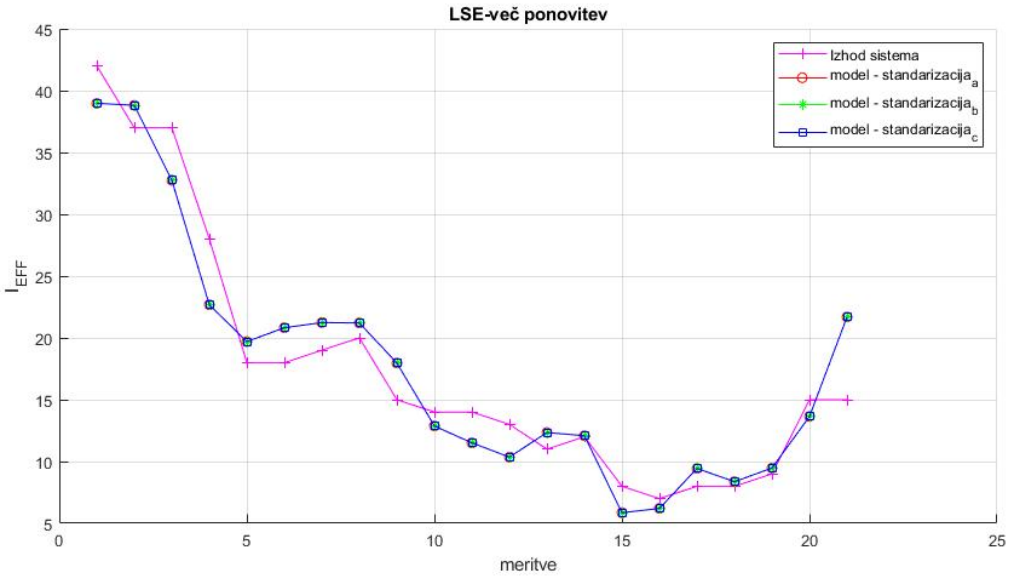
Ter jih preslikamo v originalni prostor:

$$\underline{\theta} = \underline{P}_s\underline{\theta}_T \tag{14}$$

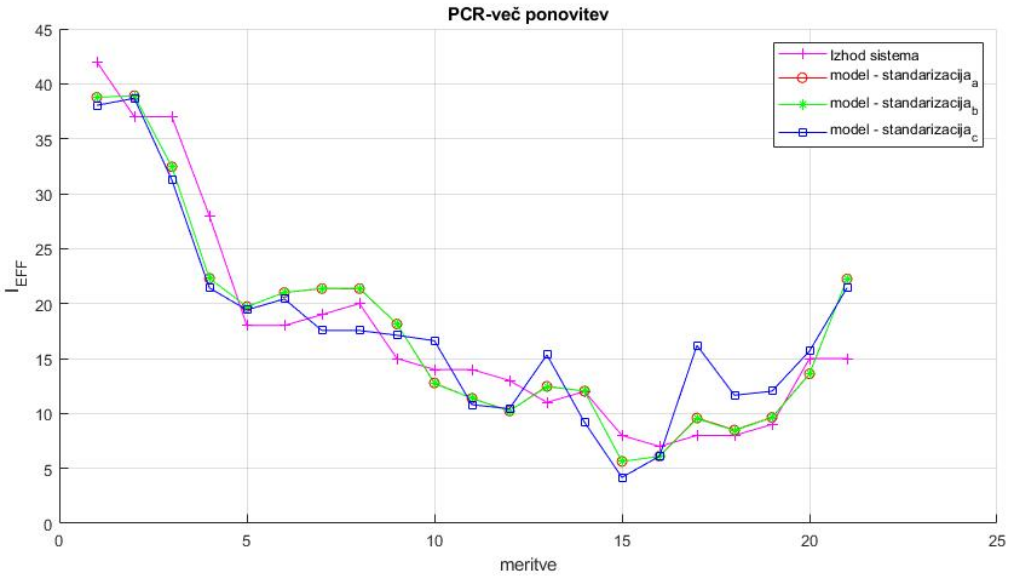
Sedaj sledi izračun modela ter v primeru standarizacije preslikava v osnovni prostor.

3.2 Rezultati

Ker je dodana odvisna meritev naključna je bil postopek modeliranja opravljen 100-krat, rezultati pa so povprečeni. Na spodnjih slikah vidimo izhode modelov za LSE ter za PCR analizo.



Slika 3: Rezultati LSE analize.



Slika 4: Rezultati PCR analize.

Na spodnjih tabelah so prikazane vrednosti izračunanih parametrov za metodi LSE in PCR.

Tabela 6: Tabela parametrov LSE.

standarizacija_a	standarizacija_b	standarizacija_c
0.7219	0.7219	0.7219
-0.8658	-0.8658	-0.8658
-0.1513	-0.1513	-0.1513
1.0725	1.0725	1.0725
-46.4674	-46.4674	-46.4674

Tabela 7: Tabela parametrov PCR.

standarizacija_a	standarizacija_b	standarizacija_c
0.4105	0.4449	0.8266
0.3716	0.4032	0.2213
-0.0896	-0.0946	-0.6476
23.3148	22.2398	0.3930

Primerjava modelov in izhoda sistema z izračunom popvrečne vrednosti napake, standardne deviacije napake ter NRMSE napake:

Tabela 8: Napaka LSE.

	standarizacija_a	standarizacija_b	standarizacija_c
povprečna vr. napake	-9.0167e-12	-8.9274e-12	-3.0763e-11
std. deviacija napak	2.7955	2.7922	2.7922
NRMSE	0.2682	0.2679	0.2679

Tabela 9: Napaka PCR.

	standarizacija_a	standarizacija_b	standarizacija_c
povprečna vr. napake	-0.0075	2.5377e-16	-0.1482
std. deviacija napak	2.9921	2.9920	3.9573
NRMSE	0.2871	0.2871	0.3800

Glavne komponente analize PCR pri standarizaciji a) in b):

Tabela 10: Lastni vektorji PCR - standarizacija a).

	P1	P2	P3
A_FLOW	-0.4027	0.2575	-0.8783
T_H2O	-0.4633	0.4145	0.3425
C_ACID	-0.6431	-0.7624	0.0712
dodatna	-0.4578	0.4250	0.3260

Tabela 11: Lastni vektorji PCR - standarizacija b).

	P1	P2	P3
A_FLOW	-0.5219	-0.0115	-0.8529
T_H2O	-0.5508	0.2903	0.3408
C_ACID	-0.3472	-0.9105	0.2247
dodatna	-0.5511	0.2943	0.3255

Varianca ocenjenih parametrov po LSE metodi:

Tabela 12: Varianca LSE.

	standarizacija_a	standarizacija_b	standarizacija_c
σ^2	9.3709	9.3805	9.3805

3.3 Komentar

Pri LSE analizi ponovno vidimo da standarizacija nima vpliva. Povprečne napake so v vseh treh primerih večje od tistih iz LSE analize, ki je izvedena brez dodane meritve, a vendar še vedno zelo majhne. Standarde deviacije ter NRMSE so pa malenkost manjše. Ocenjena varianca parametrov je malenkost manjša v primeru dodane kolinearne meritve.

Pri PCR vidimo da standarizacija b) da daleč najbolj konsistenten model. Konsistenca modela z standarizacijo a) je precej slabša. Pristranskost teh dveh standarizacij je precej podobna. Pri nestandariziranemu naboru podatkov opazimo zelo slabo konsistenco ter večjo pristranskost.

Če primerjamo med sabo LSE in PCR vidimo da s PCR dobimo bolj konsistenten model, če uporabimo standarizacijo b), pri LSE so pa vsi modeli manj pristranski.

Medsebojno odvisnost spremenljivk oziroma kolinearnost je moč opaziti iz tabel lastnih vektorjev. Vidimo namreč da imata odvisni spremenljivki podobno vrednost v lastnih vektorjih.

4 Zaključek

Pri tej vaji smo se seznanili s tremi metodami modeliranja in jih medsebojno primerjali. Sklep vaje je da, je LSE metoda precej robustna ter natančna v primeru nekolinearnosti vhodnih podatkov. Pri PCA metodi je potrebno zagotoviti standarizacijo podatkov na enega od dveh načinov. Drugi način standarizacije (*zscore*) vrne malenkost boljše rezultate. PCR modliranje se izkaže za ustrezno v primeru

kolinearnosti podatkov, če podatke standariziramo tako, da jim odštejemo srednjo vrednost ter jih delimo s standardno deviacijo.

Literatura

[1] Igor Škrjanc *Inteligentni sistemi za podporo odločcanju.* Ljubljana, 2016