SVEUČILIŠTE U RIJECI
ODJEL ZA BIOTEHNOLOGIJU
Diplomski sveučilišni studij
Medicinska kemija

Diplomski rad

Metode multivarijatne regresije za kemometrijsku analizu spektara u bliskom infracrvenom području

Rijeka, rujan 2019. Nino Požar

SVEUČILIŠTE U RIJECI
ODJEL ZA BIOTEHNOLOGIJU
Diplomski sveučilišni studij
Medicinska kemija

Diplomski rad

Metode multivarijatne regresije za kemometrijsku analizu spektara u bliskom infracrvenom području

Mentor: doc.dr.sc. Duško Čakara

Rijeka, rujan 2019.

Nino Požar

UNIVERSITY OF RIJEKA
DEPARTMENT OF BIOTECHNOLOGY
University graduate programme
Medicinal chemistry

Master's thesis

Multivariate regression methods for chemometric analysis of the near-infrared spectra

Mentor: doc. dr. sc. Duško Čakara

Rijeka, September 2019

Nino Požar

Diplomski rad je obranjen dana 16. rujna 2019. g. pred povjerenstvom:

1. izv. prof. dr. sc. Marta Žuvić
2. prof. dr. sc. Vladislav Tomišić
3. doc. dr. sc. Duško Čakara

Rad ima 107 stranica, 63 slike, 3 tablice i 65 literaturnih navoda.

Posvećeno mojoj familije ka mij' vavek vela podrška i izvor inspiracije, mudrosti i jubavi.

Zahvala

Prije svega želim se zahvaliti svom mentoru doc. dr. sc. Dušku Čakari na njegovom doprinosu u obliku ideja, rasprava i osvrta tijekom izrade ovog rada, njegovu konstruktivnu kritiku i kontinuiranu motivaciju što je doprinijelo mom osobnom i profesionalnom rastu. Također se zahvaljujem kolegi i prijatelju Dariu Matulji na svoj pomoći kroz svih 5 godina studija i izražavam ponos na sve naše zajedničke uspjehe. Zahvalu upućujem i svim profesorima i kolegama na Odjelu za biotehnologiju Sveučilišta u Rijeci.

Također se ovim putem zahvaljujem i prof. dr. sc. Rasmusu Brou i njegovom timu s Kopenhagenškog sveučilišta na ustupljenim NIRS podatcima, kao i prof. dr. sc. Miroslavu Joleru s Tehničkog fakulteta Sveučilišta u Rijeci na predlošku za izradu diplomskog rada u LaTeX-u.

Najveća zahvala ide mojoj obitelji i prijateljima koji su me podržavali tijekom ovog putovanja. Hvala vam za podršku i ohrabrenje koji su vrijedili više nego što mogu izraziti riječima.

$\sim$ *Without data you're just another person with an opinion.* $\sim$

- William Edwards Deming (1900–1993)

Ovaj rad izrađen je u Laboratoriju za fizikalnu kemiju Odjela za biotehnologiju (Sveučilište u Rijeci) te u laboratoriju za koloide, polielektrolite i međupovršine Centra za mikro i nano znanosti i tehnologije Sveučilišta u Rijeci.

U izradi rada korištena je oprema projekta Sveučilišta u Rijeci "Razvoj istraživačke infrastrukture na Kampusu Sveučilišta u Rijeci" uz potporu Europskog fonda za regionalni razvoj (EFRR) i Ministarstva znanosti i obrazovanja Republike Hrvatstke (RC.2.2.06-0001).

Sažetak

Multivarijatna analiza i regresija predstavljaju jedan od glavnih alata u analizi mjernih podataka s primjenom u kontroli kvalitete, kompresiji podataka, predviđanju varijabli u sustavima visoke kompleksnosti, a od nedavno i optimizaciji modela pomoću kojih je moguće opisati i predvidjeti ponašanje takvih sustava. Iako je matematički aparat za multivarijatnu analizu bio razvijen već početkom 20. st., puni procvat ovih metoda bilježi se tek razvojem i dostupnošću računala najnovije generacije.

Cilj diplomskog rada je istražiti primjenjivost i učinkovitost različitih metoda multivarijatne analize za dekonvoluciju spektara u bliskom infracrvenom području elektromagnetskog zračenja (NIR), izmjerenih za trokomponentnu smjesu spojeva glukoza, fruktoza i saharoza. Naglasak je na regresijskim tehnikama koje omogućuju, provjeru točnosti te pouzdanosti, kako za prilagodbu razvijenih modela, tako i za predikciju varijable odgovora. Najvažnije svojstvo tih tehnika je da primjena regresije pruža mogućnost samokonzistentne kalibracije, tj. određivanje kalibracijskih parametara optimizacijom kalibracijskog modela. Pritom, zbog podatkovne kompleksnosti NIR spektara, za primjenu svih istraženih metoda nužno je provesti i njihovu predobradu. Većina primijenjenih metoda, tj. analiza glavnih komponenti (PCA), multivarijatna linearna regresija (MLR), ridge regresija, regresija glavnih komponenti (PCR), regresija parcijalnih najmanjih kvadrata (PLS) te umjetne neuronske mreže, u prilagodbi modela sadrži korak selekcije ili redukcije podataka sadržanih u ulaznoj matrici spektara.

Istražene su i uspoređene značajke različitih metoda predobrade i prilagodbe modela te njihova učinkovitost za kvantitativnu analizu NIR spektara smjese navedenih spojeva. Kao najučinkovitija pokazala se ridge regresija, za koju ponovljena unakrsna validacija modela pokazuje *out-of-sample* pogrešku koja iznosi tek $0.63 \pm 0.08\%$. Rezultati i zaključci o istraženim metodama obrade NIR spektara primijenjivi su u praksi za brzu i efikasnu analizu smjesa kemijski sličnih spojeva u industrijskom i znanstveno-istraživačkom okružju.

Ključne riječi — **multivarijatna, kvantitativna, analiza, regresija, NIRS, kemometrija**

Abstract

Multivariate analysis and regression present one of the main tools in the analysis of measured data, with applications in quality control, data compression, forecasting of variables in highly complex systems, and more recently, model optimization which enable description and behavior prediction of such systems. Although the mathematical apparatus for multivariate analysis was developed in the early 20th century, full flourishing of these methods is recorded in parallel with the development and availability of the latest generation computers.

The aim of the present thesis is to investigate the applicability and effectiveness of different multivariate analysis methods for deconvolution of spectra in the near-infrared region of electromagnetic radiation (NIR), measured for a three-component mixture of glucose, fructose and sucrose. The emphasis is on regression techniques that allow the precision and accuracy validation both for the fitted model as well as the prediction of the response variable. The most important feature of these techniques is that regression provides the possibility of self consistent calibration, that is, determination of the calibration parameters through calibration model optimization. Due to the complexity of NIR spectra, for the application of all investigated methods, it is necessary to carry out the pre-processing of the measured data. Most of the applied regression methods, thus principal component analysis (PCA), ridge regression, principal component regression (PCR), partial least squares regression (PLS), and artificial neural networks, include variable selection or data reduction step in model fitting.

The characteristics of different methods for NIR spectra pre-processing and model fitting, and their efficiency for the quantitative analysis of the aforementioned compounds in their mixture, were investigated and compared. Ridge regression was shown to be the most effective, for which repeated cross validation showed out-of-sample error of only $0.63 \pm 0.08\%$. The results and conclusions of the investigated NIR spectra analysis methods are applicable in practice for a rapid and efficient determination of chemically similar compounds in their mixtures, both in industrial and scientific research environments.

Keywords — **multivariate, quantitative, analysis, regression, NIRS, chemometrics**

Sadržaj

Pojmovnik

ANN umjetne neuronske mreže

BIC *Bayesian information criterion*

BLUE najbolji linearni nepristrani procjenitelj

BS *best subset* selekcija

CV unakrsna validacija

dCV dvostruka unakrsna validacija

FW *forward stepwise* selekcija

GA genetički algoritmi

LOO CV pojedinačna unakrsna validacija

MLR multivarijatna linearna regresija

MSC multiplikativna korekcija signala

MSE srednja kvadratna pogreška

NIRS spektroskopija u bliskom infracrvenom području elektromagnetskog zračenja

NW Norris-Williams derivacija

OLS metoda najmanjih kvadrata

PAT procesna analitička tehnika

PCA analiza glavnih komponenti

PCR regresija glavnih komponenti

PLS regresija parcijalnih najmanjih kvadrata

Pojmovnik

rCV ponovljena unakrsna validacija

rdCV ponovljena dvostruka unakrsna validacija

RMSE korijen srednje kvadratne pogreške

RR ridge regresija

SE standardna pogreška

SG Savitzky-Golay derivacija

SNV standardna normalna varijata

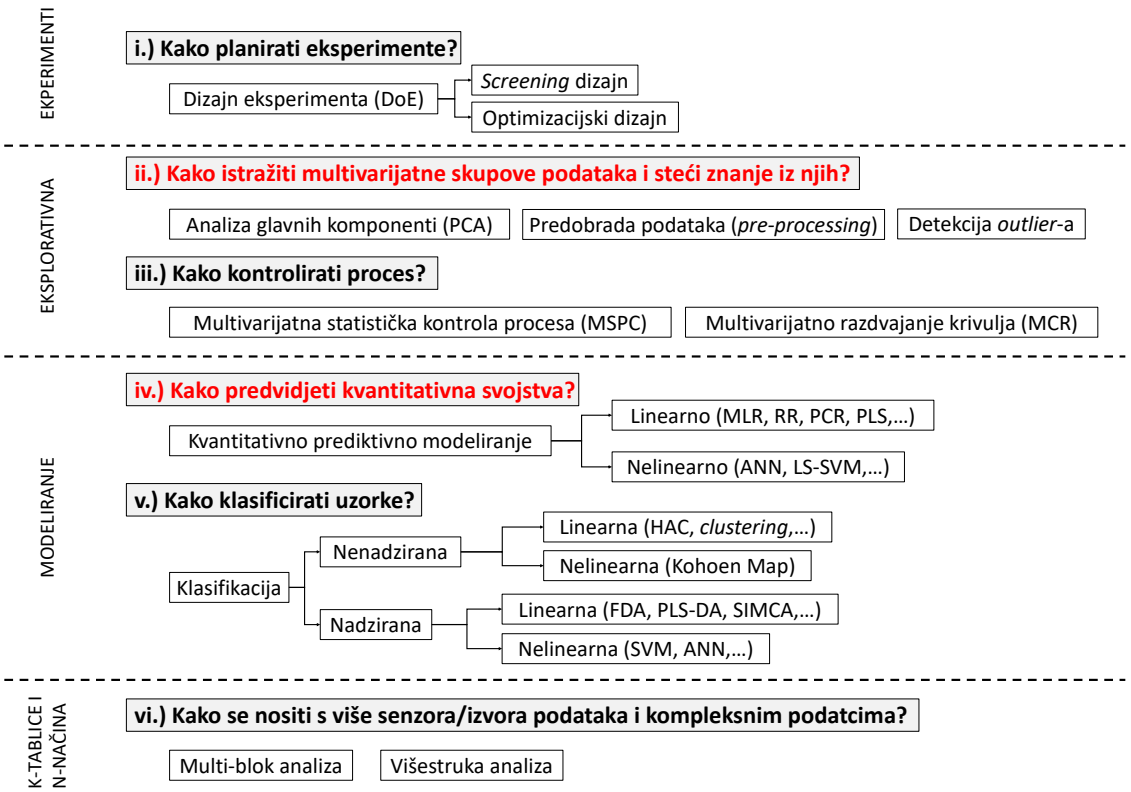
SVD dekompozicija singularnih vrijednosti

Poglavlje 1

Uvod

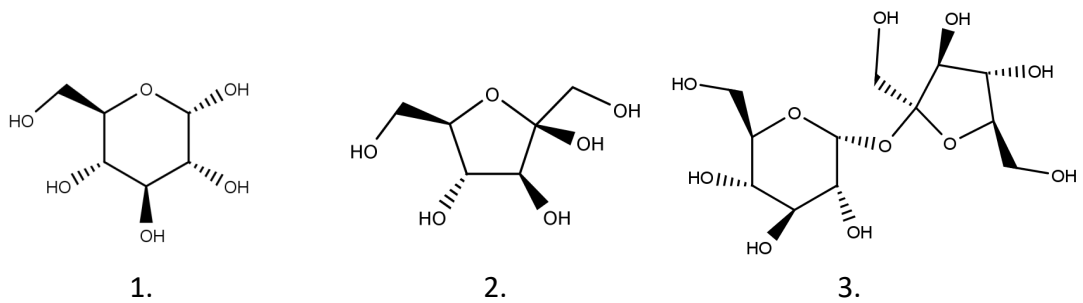
Statistika igra važnu ulogu u svim ljudskim djelatnostima, a sposobnosti primjene statističkog zaključivanja te interpretacije rezultata neizostavne su u inženjerskoj i znanstvenoj praksi. Godišnje se generira 40% novih podataka, zbog čega analize podataka postaju temelj mnogih disciplina, uz kontinuirani napredak tehnologije koja omogućava njihovu analizu. Neke od primjena su vođenje poslovanja kompanija, tzv. poslovna inteligencija, računalni vid, zatim analize tekstova, tj. obrada prirodnog jezika, ekonometrija, sportska analitika i sl. Kemometrija je kemijska disciplina koja se bavi analizama podataka, a koristi matematiku, statistiku i logiku za: a) oblikovanje ili odabir optimalnih eksperimentalnih postupaka; b) izvođenje maksimalne relevantne kemijske informacije analizom kemijskih podataka; i c) stjecanje znanja o kemijskim sustavima i procesima [1]. U kemometrijskim analizama prisutne su četiri vrste metoda prikazane u slici 1.1.

U fokusu ovog rada je analiza spektara u bliskom infracrvenom području (eng. *near-infrared spectroscopy*, NIRS) za smjesu kemijski vrlo sličnih spojeva glukoze, fruktoze i saharoze (slika 1.2). Valja obratiti pozornost na činjenicu da sva tri spoja imaju gotovo identičan atomski sastav te je saharoza dimer koji se sastoji od podjedinica glukoze i fruktoze. Cilj je odgovoriti na pitanja "Kako istražiti multivarijatne skupove podataka i steći znanje o njima?" i "Kako predvidjeti kvantitativna svojstva?". Time je fokus sužen na multivarijatne regresijske metode koje u kemometriji imaju za cilj razviti kalibracijske modele za predviđanje svojstava uzoraka na temelju mjerenih (nezavisnih, prediktorskih) varijabli [3]. U NIRSu se iz spektara



Slika 1.1 Pregled kemometrijskih metoda. Ovaj rad se bazira na pitanjima pod ii. i iv. (crveno). Preuzeto i prilagođeno iz [2].

uzoraka poznatih svojstava (standarda), pomoću multivarijatne regresije, razvijaju kalibracijski modeli koji služe za kvantitativnu analizu i predviđanje svojstava novih uzoraka iz eksperimentalno određenih spektara. Multivarijatne regresijske metode nam omogućavaju proučavanje podataka te određivanje uzoraka grupiranja i podataka koji značajno odstupaju od osnovnog skupa podataka (tzv. *outlier*-e). Kod NIR spektara u VIS/NIR području, mjerenih pri variranom sastavu smjese spojeva, broj prediktorskih varijabli značajno je veći od broja pripremljenih uzoraka, zbog čega su tehnike multivarijatne analize podataka nužne za razvoj kalibracijskih modela. Pritom su korištene sljedeće metode: multivarijatna linearna regresija, ridge regresija, regresija glavnih komponenti, regresija parcijalnih najmanjih kvadrata, i umjetne neuronske mreže.

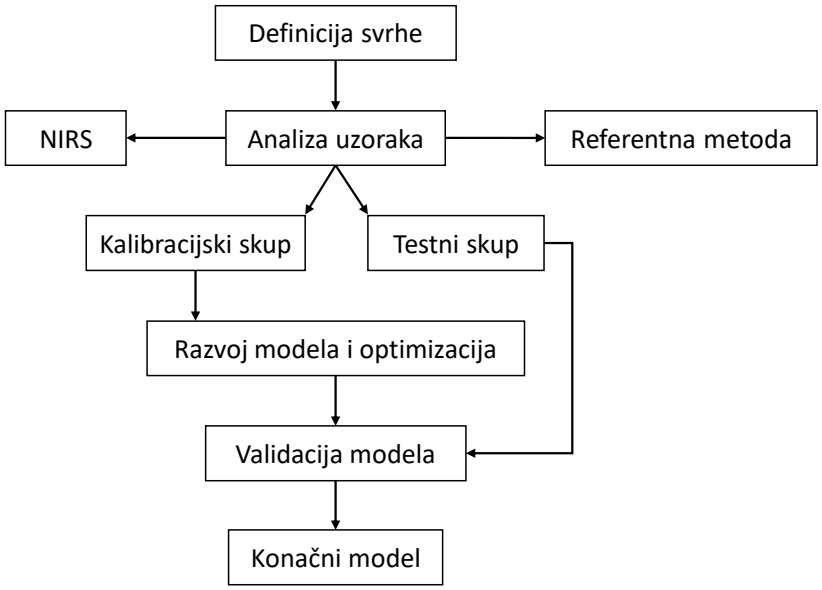


Slika 1.2 Kemijske strukture glukoze (1.), fruktoze (2.) i saharoze (3.).

Prilikom odabira i treninga najboljeg modela, najvažniji su kriteriji performanci modela: njegova točnost (određena pogreškom predikcije), interpretabilnost (zašto model nešto predvidviđa, analiza povezanosti varijabli), skalabilnost (primjenjivost za velike količine novih podataka), brzina (treninga modela i predikcije) i jednostavnost (manji broj parametara). Time se prilikom odabira optimalnog kalibracijskog modela vodi računa o četiri kompromisa: točnost u odnosu na skalabilnost, točnost u odnosu na jednostavnost, točnost u odnosu na brzinu te točnost u odnosu na interpretabilnost.

Cijeli proces razvoja kalibracijskih modela u NIRSu prikazan je u shemi 1.3. Započinje se od svrhe kalibracijskog modela, tj. postavlja se pitanje na koje želimo odgovoriti. U ovom radu to je "Možemo li iz NIRS podataka razviti model za predviđanje masenih udjela glukoze, fruktoze i saharoze u trokomponentnoj smjesi?". Potom se pripremaju uzorci koji služe za razvoj i validaciju modela (tj. kalibraciju metode) za koje se maseni udjeli mjere referentnim metodama, a također i NIR spektri. Izmjereni spektri dijele se u dva skupa, kalibracijski skup i testni. Kalibracijski služi za trening i optimizaciju modela, što uključuje i predobradu spektara. Modeli se validiraju na neovisnom testnom skupu s obzirom na prediktivne performance, na temelju kojih se odabire najbolji kao onaj s najmanjom pogreškom predikcije [4].

U industrijskim procesima, NIRS s izvedenim kalibracijskim modelom predstavlja izvrsnu analitičku tehniku, kako za *in-line* praćenje proizvodnog procesa, tako



Slika 1.3 Koraci u kalibraciji NIRS modela. Preuzeto i prilagođeno iz [4].

i kontrolu kvalitete i svojstava produkta. Kao procesna analitička tehnika (PAT), NIRS omogućava: identifikaciju procesnih parametara, praćenje procesa u realnom vremenu, dobivanje podataka koji se mogu obraditi multivarijatnim metodama, i razvoj procesne kontrolne strategije [5]. Od posebne važnosti je činjenica da je takve analize moguće provesti *in-situ* i u realnom vremenu. Prednosti takvog pristupa su nekorištenje opasnih kemikalija te smanjena mogućnost kontaminacije jer ne postoji potreba za uzimanjem alikvota iz reakcijskog sustava [6]. Zbog navedenog, *in-line* PAT se uspješno primjenjuju u različitim industrijskim granama. Npr. regulatorne agencije predlažu njihovo korištenje za praćenje proizvodnog procesa u farmaceutskoj industriji [5, 7]. Osim NIRSa, za *in-line* analize koriste se i spektroskopija u srednjem infracrvenom području (eng. *mid infrared spectroscopy*, MIR) te Ramanova spektroskopija. Svaka od tih tehnika nosi svoje prednosti i nedostatke o čemu će više riječi biti u poglavlju 3.1.

NIRS je vrlo često korištena tehnika u kvantitativnoj kemijskoj analizi, pa je razumljiva velika količina objavljene literature [4]. Među publikacijama navedenim u

[4], od posebnog interesa su npr. rad Plugge i Van Der Vlies koji su koristili NIRS za određivanje vlažnosti u lijekovima te su za razvoj kalibracijskog modela koristili multivarijatnu linearnu regresiju (MLR) pri dvije valne duljine. Također, Fortina i sur. koristili su NIRS u kombinaciji s regresijama glavnih komponenti (PCR) i parcijalnih najmanjih kvadrata (PLS) za određivanje koncentracije lijeka u formulaciji [8, 9]. Vrlo je važan i rad Jouan-Rimbaud i sur. koji su pokazali da primjena selekcije varijabli prilikom razvoja kalibracijskih modela (MLR, PCR i PLS) može poboljšati određivanje koncentracije jer se eliminiraju one varijable koje imaju nizak doprinos modelu [10]. Pritom su se genetički algoritmi pokazali kao jedna od najboljih metoda selekcije, koja kalibracijski model može poboljšati do 20%. Kod razvoja kalibracijskih modela iz NIRS podataka provodi se i predobrada spektara, od kojih multiplikativna korekcija signala najčešće daje najbolji prediktivni rezultat. Također, vrlo važan je rad Chen i sur. koji su uspoređivali modele PLS i neuronskih mreža te zaključili da neuronske mreže mogu biti korisna alternativa ukoliko konvencionalne (linearne) metode ne daju zadovoljavajući rezultat [11]. Shodno razvoju tehnologije, računalnih i statističkih metoda, broj publiciranih radova na temu razvoja kalibracijskih modela za interpretaciju NIRS podataka kontinuirano raste [4].

Poglavlje 2

Tema, cilj i zadatci

Tema diplomskog rada su regresijske metode za kemometrijsku analizu podataka izmjerenih spektroskopskim tehnikama. Cilj je prikazati i primijeniti različite metode multivarijatne regresije za određivanje masenih udjela međusobno interfrerirajućih komponenti (spojeva) iz NIR spektara izmjerenih za njihovu smjesu te ih usporediti s obzirom na točnost, brzinu, skalabilnost, jednostavnost i interpretabilnost.

ZADATCI:

1. Eksplorativna analiza spektara
2. Predobrada NIR spektara
3. Trening različitih metoda multivarijatne regresije:
    - Multivarijatna linearna regresija uz različite metode odabira prediktorskih varijabli
    - Multivarijatna linearna regresija uz regularizaciju
    - Multivarijatna linearna regresija uz redukciju dimenzionalnosti
    - Nelinearna regresija uz primjenu neuronskih mreža
4. Dijagnostika i usporedba metoda
5. Prikaz i diskusija rezultata

Poglavlje 3

Eksperimentalne tehnike i podatci

3.1 Spektroskopija u bliskom infracrvenom području

Spektroskopija u bliskom infracrvenom području (NIRS) jednostavna je, brza, nedestruktivna i neinvazivna tehnika koja omogućava analizu multikomponentnih smjesa, s visokom razinom preciznosti i točnosti [12]. Korištenje NIRSa ima velik broj prednosti u odnosu na druge PAT, što je prikazano u tablici 3.1. Iz razloga što NIRS ne zahtijeva pripremu uzoraka, a time ni korištenje opasnih kemikalija, otapala i reagensa, spada u metode prijateljske prema okolišu ("*environmental friendly*"). Nedostatak NIRSa je visoka kompleksnost spektara (kombinacijske vibracije i viši tonovi), široko preklapanje spektralnih vrpci i visok prag detekcije. Stoga su za primjenu NIRSa nužne kemometrijske metode analize podataka da bi se izvukla maksimalna relevantna informacija [13].

NIR tehnologija ima primjenu u velikom broju područja. Kvalitativne i kvantitativne NIR analize koriste se u poljoprivredi/industriji hrane, analizama polimera, petrokemiji i industriji goriva, okolišnim analizama, tekstilnoj industriji, biomedicinskim i medicinskim područjima, farmaciji i kozmetici te za NIR *imaging*. Poljoprivreda je prvo područje gdje je počelo intenzivno korištenje NIRSa i zaslužna je za plasiranje te tehnologije na tržište i početak razvoja prema današnjem standardu. Kasnije je i u drugim navednim područjima počela šira primjena, posebice

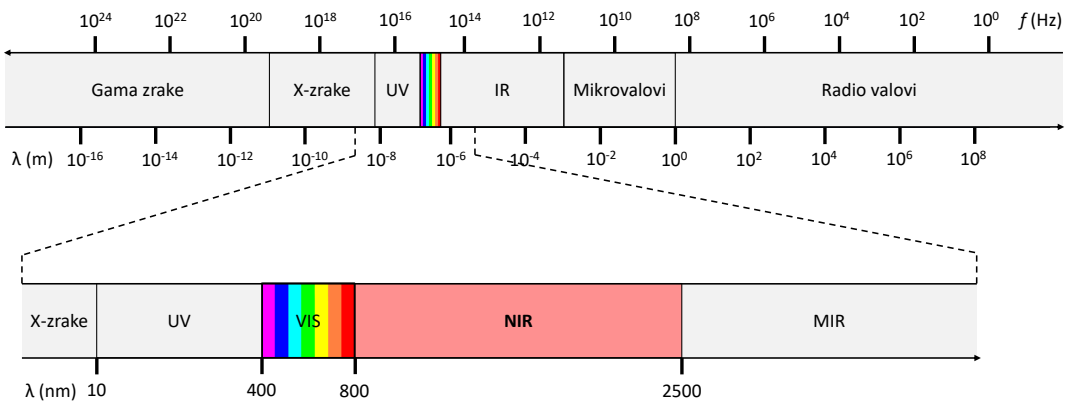
Tablica 3.1 Neke od karakteristika Raman, MIR i NIR spektroskopija. Preuzeto i prilagođeno iz [12].

	Raman	**MIR**	**NIR**
Valni broj	50-4000 cm^{-1}	200-4000 cm^{-1}	4000-12500 cm^{-1}
Veze	homonuklearne veze (npr. C-C, C=C, S-S)	polarne veze (npr. C=O, C-O, C-F)	veze koje sadrže vodik (npr. C-H, O-H, N-H, S-H)
Uzrok vrpce*	raspršenje zračenja	apsorpcija zračenja (osnovna vibracija)	apsorpcija zračenja (viši tonovi i kombinacija)
Apsorpcijske vrpce	dobro odvojene, mogu se pripisati pojedinim kemijskim skupinama	dobro odvojene, mogu se pripisati pojedinim kemijskim skupinama	serija preklapajućih vrpci
Intenzitet signala	loš	dobar	dobar
Kvantifikacija	$I \sim c$	$\log\left(\frac{I_0}{I}\right) \sim c$	$\log\left(\frac{I_0}{I}\right) \sim c$
Uvjeti pobude	promjena polarizabilnosti α	promjena diplonog momenta μ	promjena diplonog momenta μ
Selektivnost	visoka	visoka	niska, potreba za kalibracijom i kemometrijom
Interferencija	široka bazna linija fluorescencije	voda	voda, fizička svojstva (npr. veličina čestice, oblik i čvrstoća)
Veličina čestica	neovisan	ovisan	ovisan
Primjenjivost za *at-line*, *on-line*, *in-line*	dobra	loša	dobra
Priprema uzoraka	nikakva	smanjena	nikakva

* fizikalni fenomen koji uzrokuje pojavu vrpce u spektru. Radi jednostavnosti, sve vrpce nazivamo apsorpcijskim, bez obzira na način mjerenja spektara (transmisija odn. difuzna refleksija).

u farmaceutskoj i petrokemijskoj industriji (uključujući analize polimera). Novi primjeri primjene su NIR *imaging*, klinička optička tomografija te *in vivo* neinvazivne biokemijske analize [14].

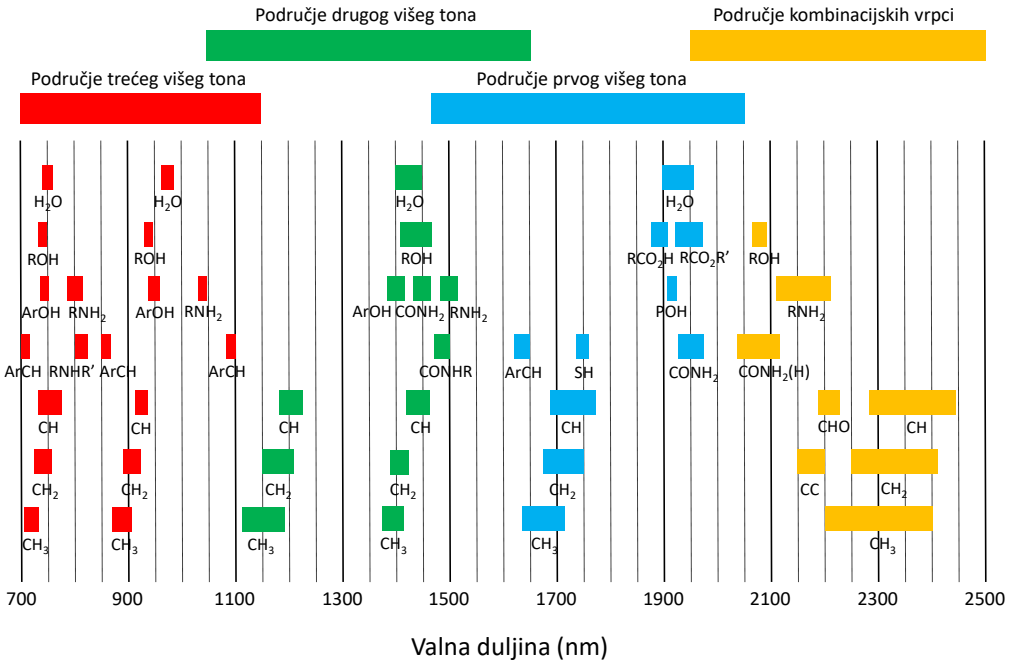
U NIRSu se koristi područje elektromagnetskog zračenja u rasponu valnih duljina 700-2500 nm (slika 3.1). U tom području najčešće se detektiraju viši tonovi vibracija veza s vodikom (C-H, N-H, O-H i S-H) [12], i to prvi viši tonovi u rasponu valnih duljina 1500-2100 nm, zatim drugi viši tonovi u rasponu 1000-1700 nm te treći viši tonovi u rasponu 700-1100 nm, dok su u rasponu 2000-2500 nm prisutne kombinacijske vrpce (slika 3.2). Također, u slici 3.2 je prikazano koje kemijske skupine apsorbiraju zračenje u specifičnim dijelovima područja NIR spektra.



Slika 3.1 Spektar bliskog infracrvenog zračenja (NIR) u kontekstu spektra elektromagnetskog zračenja.

3.1.1 Načini mjerenja NIR spektara

Specifičnost NIR tehnike je u tome što se mjerenje može provoditi na dva načina: mjerenje transmisije za analize uzoraka otopina te difuzne refleksije za analize čvrstih tvari [14]. Mjerenje transmisije provodi se na isti način kao kod UV/VIS spektroskopije: upadna zraka svjetlosti prolazi kroz uzorak u kiveti pri čemu se dio zračenja apsorbira, dio transmitira, dok se preostali dio raspršti i reflektira:



Slika 3.2 Analitičke vrpce najčešćih kemijskih skupina u organskim spojevima u NIR području. Preuzeto i prilagođeno iz [12].

$$I_0 = I_\mathrm{A} + I_\mathrm{R} + I_\mathrm{T} \tag{3.1}$$

gdje je I_0 intenzitet upadnog, I_A intenzitet apsorbiranog, I_R intenzitet reflektiranog i raspršenog, a I_T intenzitet transmitiranog zračenja. Spektrofotometrom pritom mjerimo I_0 i I_T, dok se I_R zanemaruje i nastoji svesti na minimum. Transmisijski način koristi se za analizu otopina jer je primjenjiv Beer-Lambertov zakon, koji se temelji na pretpostavci da se izmjena energije između fotona zračenja i atoma, iona ili molekula analita dešava prilikom njihovih sudara. Udio čestica analita koje se nalaze u stanjima u kojima mogu stupiti u interakciju sa zračenjem dan je Boltzmannovom raspodjelom. Stoga je I_A proporcionalan $e^{\varepsilon(\lambda)cl}$, gdje $\varepsilon(\lambda)$ označava apsorpcijski koeficijent, l duljinu puta prolaska zrake kroz uzorak, a c koncentraciju. Stoga se Beer-Lambertov zakon izražava kao:

$$A = \log\left(\frac{1}{T}\right) = -\log\left(\frac{I_\mathrm{T}}{I_0}\right) = \varepsilon c l \tag{3.2}$$

gdje se veličine A i T nazivaju apsorbanca odn. transmitanca te nemaju jedinicu, dok jedinica za ε ovisi o tome kako je izražena koncentracija. Ukoliko je c izražena kao množinska koncentracija, tada ε mora biti normiran u odnosu na množinu te ga nazivamo molarnim apsorpcijskim koeficijentom.

Za smjesu u kojoj komponente nezavisno apsorbiraju zračenje, u Beer-Lambertovom zakonu vrijedi aditivnost po njenim komponentama:

$$A = l \sum_{i=1}^{p} (\varepsilon_i c_i) \tag{3.3}$$

gdje je p broj komponenti, a ε_i i c_i su apsorpcijski koeficijent odn. koncentracija komponente i. Zbog aditivnosti vrijedi linearni odnos između apsorbance i svake od koncentracija komponenti smjese, zbog čega je za kvantitativnu analizu spektara moguće primijeniti linearne metode. Za kemometrijsku analizu, s obzirom da je mjerena veličina apsorbanca, a veličina koja se predviđa koncentracija komponente, koristi se Beer-Lambertov zakon u invertiranoj formi:

$$c = \frac{A}{\varepsilon l} \tag{3.4}$$

Drugi način predstavlja mjerenje difuzne refleksije, koje se koristi za analize čvrstih i praškaskih uzoraka. U tom slučaju ne možemo zanemariti raspršenje na površinama čestica niti ga minimizirati, već ono predstavlja mjerenu veličinu jer ima znatno jači signal od transmitiranog zračenja [14]. U tom načinu izravno ne vrijedi Beer-Lambertov zakon, već se kvantifikacijske analize provode na temeljima Kubelka-Munk teorije iz koje proizlazi linearan odnos masenog udjela i reflektance R [15]. Kubelka-Munk funkcija dana je jednadžbom:

$$f(R) = \frac{(1-R)^2}{2R} = \frac{K}{S} \tag{3.5}$$

gdje su K i S koeficijenti apsorpcije odn. raspršenja. Pritom, R je definiran kao:

$$R = \frac{I_{\mathrm{R}}}{I_{\mathrm{R0}}} \tag{3.6}$$

gdje je I_{R} intenzitet zračenja kojeg uzorak reflektira, a I_{R0} intenzitet zračenja kojega reflektira neapsorbirajući materijal. Pokazalo se da je S neovisan o valnoj duljini, a K i ε (koeficijent apsorpcije u transmisiji iz jedn. 3.2, po Beer-Lambertovom modelu) su međusobno proporcionalni, a faktor proporcionalnosti neznačajno ovisi o valnoj duljini [15]:

$$K \sim \varepsilon \tag{3.7}$$

zbog čega vertikalno pomaknuti graf logaritma spektra reflektance daje graf logaritma spektra transmitance. Stoga se Kubelka-Munk funkcija pojednostavljuje kako bi predstavljala apsorpcijski spektar:

$$f(R) = \log\left(\frac{1}{R}\right) \tag{3.8}$$

K i S za smjesu predstavljaju zbrojeve umnožaka tih varijabli za pojedine komponente K_i i S_i, i njihovih masenih udjela w_i [16]:

$$f(R) = \left(\frac{K}{S}\right)_{\mathrm{M}} = \frac{\sum_{i=1}^{p} w_i K_i}{\sum_{i=1}^{p} w_i S_i} \tag{3.9}$$

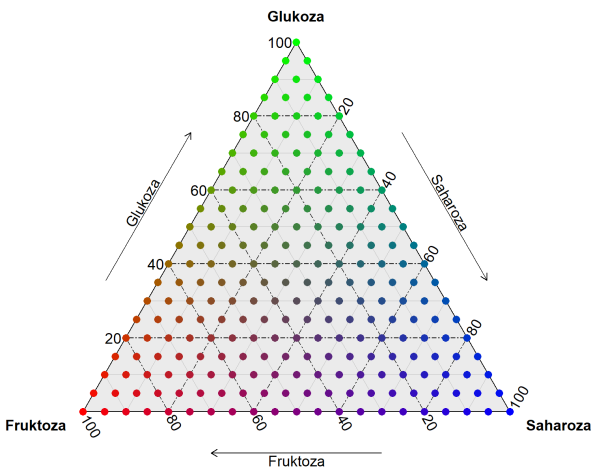
gdje M označava smjesu od p komponenti. Ukoliko je S konstantan (ne ovisi o masenim udjelima komponenti), jedn. 3.9 predstavlja linearni odnos apsorpcijskih spektara i masenih udjela:

$$f(R) = \log\left(\frac{1}{R}\right) = \frac{1}{S}\sum_{i=1}^{p} w_i K_i \tag{3.10}$$

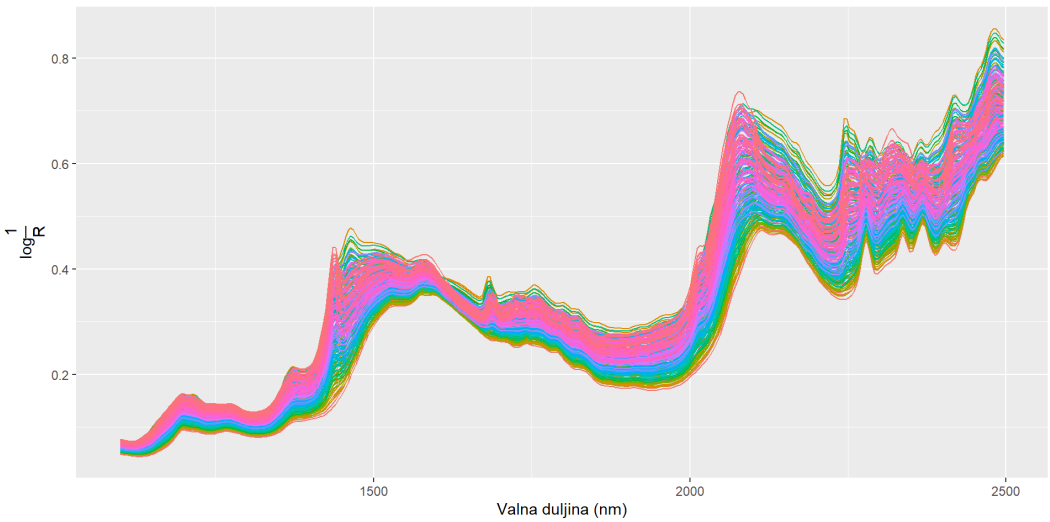
U suprotnom, ukoliko se S_i za komponente međusobno značajno razlikuju, nelinearnosti u spektrima koje su posljedica ovisnosti raspršenja o sastavu smjese (nazivnik u jedn. 3.9), uklanjaju se metodama predobrade spektara. I u ovom načinu mjerenja, za kemometrijske analize primijenjuje se invertirana jedn. 3.10 s obzirom da je cilj predvidjeti masene udjele iz mjerene reflektance.

3.2 Podatci

U radu su kao eksperimentalni podatci korišteni spektri preuzeti iz publikacije Nørgaarda i suradnika [17]. Radi se o uzorcima trokomponentnih smjesa saharoze, glukoze i fruktoze u kojima je variran maseni udio svake od triju komponenti u rasponu 0% do 100% s razlikom od 5%. Pripremljeno je $n = 231$ uzoraka ($1+2+3+\ldots+20+21 = 231$) prema dijagramu prikazanom u slici 3.3. Za svaki uzorak izmjereni su spektri raspršenog zračenja $R(\lambda_i)$ u rasponu valnih duljina 1100 nm $\leq \lambda_i <$ 2500 nm uz interval $\lambda_i - \lambda_{i-1} = 4$ nm prikazani u slici 3.4. Za sva mjerenja korišten je instrument FOSS NIRSystems 6500 te način mjerenja difuzne refleksije. Izmjereni spektri sadrže $m = 350$ izmjerenih točaka $[\lambda_i, R_i]$ koje u daljnjem razvoju regresijskih modela predstavljaju podatke sadržane u matrici $\mathbf{X}^{(n \times m)}$, pri čemu retci matrice predstavljaju pojedine uzorke, a stupci prediktorske varijable, tj. $\log(\frac{1}{R})$ izmjerene pri različitim valnim duljinama. U matrici $\mathbf{Y}^{(n \times p)}$ sadržani su maseni udjeli fruktoze, glukoze i saharoze, poznati iz pripreme uzoraka, gdje su retci matrice pojedini uzorci, a stupci varijable odgovora, tj. maseni udjeli. Matrice su shematski prikazane u slici 3.5.



Slika 3.3 Maseni udjeli pojedinih komponenti u uzorcima trokomponentne smjese saharoze, glukoze i fruktoze. Kombinacije boja (crvena, zelena, plava) naglašavaju udjele pojedinih komponenti (fruktoza, glukoza, saharoza).



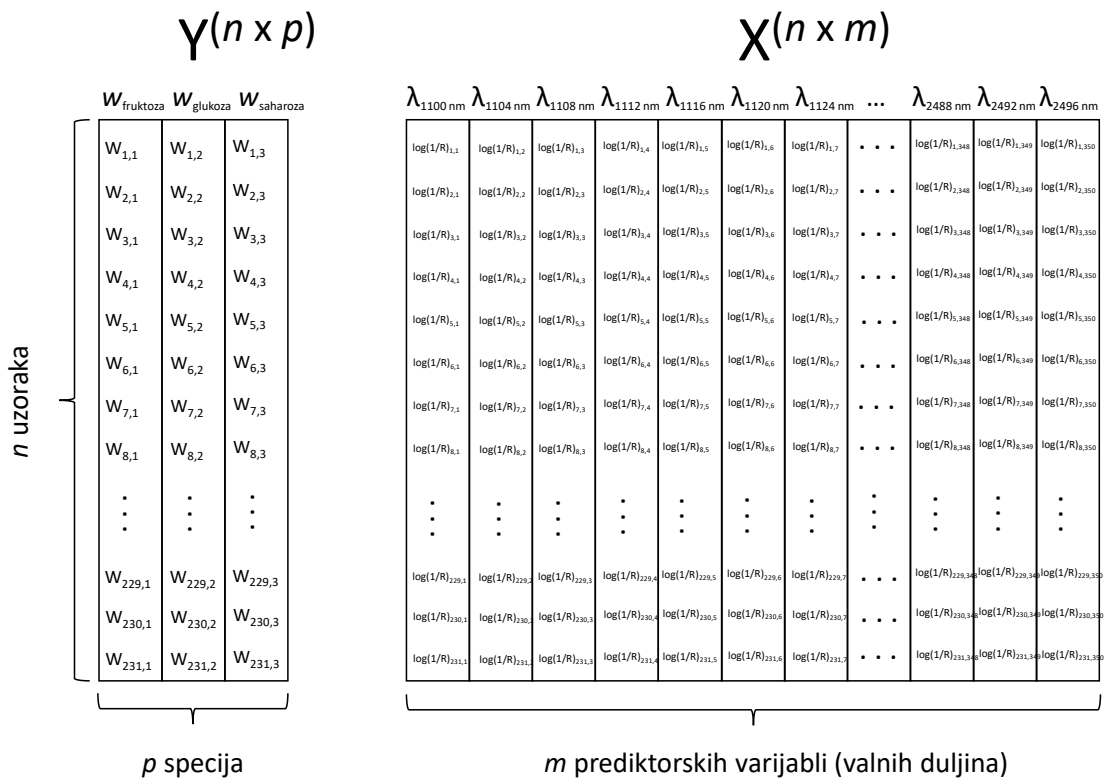
Slika 3.4 Eksperimentalni NIR spektri trokomponentnih smjesa saharoze, glukoze i fruktoze, snimljeni pomoću NIR spektrometra (FOSS NIRSystems 6500) u načinu mjerenja difuzne refleksije.

3.2.1 Notacija

Matrice su označene velikim, uspravnim i podebljanim slovima (npr. $\mathbf{X}$), vektori (stupci) malim uspravnim i (npr. $\mathbf{x}_i$), vektori (retci) malim, nakošenim i podebljanim (npr. $\boldsymbol{x}_j$), a elementi i skalari malim nakošenim (npr. $x_{j,i}$). Vektori i elementi iz iste matrice svi koriste isto slovo (npr. $\mathbf{X}$, $\mathbf{x}_i$, $\boldsymbol{x}_j$, $x_{j,i}$). Operacija transponiranja označava se s $'$, dok je matrica identiteta označena s $\mathbf{I}$.

3.3 Specifikacije računala i *software*-a

Računalne analize provedene su na stolnom računalu (CPU: Intel®Core™2 Quad CPU, Q8400 @ 2.67 GHz, RAM: 4GB, OS: Windows 10 Education 64-bit) koristeći RStudio (R *version* 3.6.1). Korišteni su R paketi: "R.matlab", "pls", "subselect", "nnet", "caret", "tidyverse", "RColorBrewer", "gridExtra", "corrplot", "leaps", "neuralnet", "knitr" i "ggtern". Sve analize su samostalno provedene te su svi R kodovi i skripte dostupni na upit autoru.



Slika 3.5 Organizacija podataka u matricama. $\mathbf{Y}$ je matrica masenih udjela fruktoze, glukoze i saharoze, dimenzija $n \times p$. $\mathbf{X}$ je matrica snimljenih spektara, tj. $\log\left(\frac{1}{R}\right)$ vrijednosti, dimenzija $n \times m$.

Poglavlje 4

Metode obrade podataka

4.1 Predobrada spektara

Prije svake analize podatke treba očistiti i pripremiti. Koraci u čišćenju i pripremi su pronalaženje (i eliminacija) pogrešnih i/ili netočnih vrijednosti u skupu, tj. detekcija *outlier*-a, imputacija vrijednosti koje nedostaju te uređivanje, kako bismo naposlijetku postigli "uredne" podatke (eng. *tidy data*). Hadley Wickham ih definira kao skup podataka za koje [18]:

1. Svaka izmjerena varijabla nalazi se u jednom stupcu tablice
2. Svako mjerenje varijable nalazi se u različitom retku
3. Svakoj "vrsti" varijable odgovara jedna tablica
4. Ukoliko postoji više tablica, one trebaju sadržavati stupac pomoću kojeg ih je moguće povezati

Na "urednim" podatcima se najprije provodi predobrada (eng. *pre-processing*) koja je sastavni dio svake analize podataka. Njome se iz NIR spektara uklanjaju smetnje koje imaju uzrok u fizikalnim fenomenima povezanim s mjerenom varijablom, s ciljem poboljšavanja prediktivnih (regresijskih ili klasifikacijskih) modela [19]. Osnovne metode predobrade spektara s obzirom na stupce su: centriranje, skaliranje i autoskaliranje spektara.

Centriranje spektara podrazumijeva oduzimanje prosječnog spektra svih uzoraka od spektra svakog pojedinog uzorka. Time se postiže da je srednja vrijednost svake prediktorske varijable u podatcima jednaka 0. Ono je nužno za multivarijatne regresijske analize s obzirom da su često bazirane na određivanju vlastitih vektora matrice kovarijanci te, ukoliko su podatci centrirani, odsječak nije potrebno računati (iznosi 0) i radi se o regresiji kroz ishodište (eng. *regression though the origin*). Centriranjem matrice, ukoliko je broj uzoraka manji od broja varijabli ($n < m$), dolazi do smanjenja ranga centrirane matrice za 1. Specifičan je slučaj matrica koncentracija ($\mathbf{Y}$) kada je ukupna koncentracija u svakom uzorku očuvana. U tom slučaju kod centrirane matrice uvijek dolazi do smanjenja ranga za 1 [20], što je slučaj s eksperimentalnim podatcima u ovom radu.

Skaliranje podrazumijeva dijeljenje prediktorskih varijabli s njihovim standardnim devijacijama pri čemu njihova varijanca poprima vrijednost 1 [13]. Koristi se u slučaju kada prediktorske varijable imaju značajno različite raspone vrijednosti, kako bi se svakoj dala jednaka težina. U praksi se prilikom predobrade NIR spektara ne koristi skaliranje $\mathbf{X}$ jer povećava težinu prediktorskih varijabli ($A/\log(\frac{1}{R})$ snimljenih pri različitim valnim duljinama) s malim rasapom vrijednosti (varijancama) koje mogu predstavljati šum [3]. Ukoliko se podatci skaliraju, najčešće je to u kombinaciji s centriranjem te tada govorimo o autoskaliranju, pri čemu sve prediktorske varijable poprimaju srednju vrijednost 0 uz varijancu 1, tj. jednaku statističku težinu.

Kod NIR spektara zbog raspršenja elektromagnetskog zračenja dolazi do pojave nelinearnosti, pomaka bazne linije i izraženog šuma u $\mathbf{X}$. Zbog toga su za predobradu NIR spektara potrebne i naprednije metode s obzirom na retke. U NIRSu se najčešće koriste multiplikativna korekcija signala (eng. *multiplicative signal correction*, MSC), korekcija putem standardne normalne varijate (eng. *standard normal variate*, SNV), Norris-Williams derivacija i Savitzky-Golay derivacija. MSC i SNV pripadaju podskupu korekcija raspršenja, dok Norris-Williams i Savitzky-Golay derivacije pripadaju podskupu korekcija spektara deriviranjem.

4.1.1 Multiplikativna korekcija signala

MSC je najkorištenija metoda korekcije NIR spektara, a predstavili su je Martens i sur. 1983. g. Sastoji se od dva koraka:

1. Određivanje korekcijskih koeficijenata:

$$\boldsymbol{x}_{\text{org}} = b_1 \boldsymbol{x}_{\text{ref}} + b_0 \mathbf{1} + \boldsymbol{e} \tag{4.1a}$$

2. Korigiranje izmjerenih spektara:

$$\boldsymbol{x}_{\text{corr}} = \frac{\boldsymbol{x}_{\text{org}} - b_0 \mathbf{1}}{b_1} \tag{4.1b}$$

gdje je $\boldsymbol{x}_{\text{org}}$ izvorni spektar uzorka, $\boldsymbol{x}_{\text{ref}}$ je referentni spektar, b_0 i b_1 su korekcijski koeficijenti izračunati linearnom regresijom u prvom koraku, $\boldsymbol{e}$ je rezidual, $\boldsymbol{x}_{\text{corr}}$ je obrađeni tj. korigirani spektar, a $\mathbf{1}$ je vektor čiji su svi elementi 1. Kao $\boldsymbol{x}_{\text{ref}}$ se najčešće koristi prosječni spektar svih izmjerenih spektara.

MSC metoda je zbog korekcije raspršenja zračenja posebice korisna u načinu mjerenja difuzne refleksije. Osim toga, njome uklanjamo i pomak bazne linije između spektara. Nedostatak je potreba za određivanjem referentnog spektra, $\boldsymbol{x}_{\text{ref}}$. Iako se kao referentni spektar najčešće koristi prosječni spektar svih izmjerenih, to nije pravilo i postoje različite mogućnosti odabira [19]. Jedna od njih je *loopy* MSC koji se bazira na ponavljanju MSC metode sve dok korigirani spektri ne konvergiraju, tj. dok Frobeniusova norma[1] razlike spektara ne poprimi vrijednost manju od 0.001. Do konvergencije najčešće dolazi nakon 2 do 3 iteracije [21]. MSC je moguće proširiti u EMSC (eng. *extended MSC*), koja osim korekcije s obzirom na referentni spektar, može uključivati dodatne korekcije valnih duljina ili poznatih spektralnih informacija; ili u ISC (eng. *Inverse Scatter Correction*), inverzijom jedn. 4.1a.

4.1.2 Korekcija putem standardne normalne varijate

SNV je druga najkorištenija metoda korekcije spektara u NIRSu, jednaka drugom koraku MSC, jedn. 4.1b. Pritom, za razliku od MSC, b_0 je prosječna vrijednost

[1] za matricu $\mathbf{A}$, $\|\mathbf{A}\|_{\text{F}} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{j,i}^2}$

$\log(\frac{1}{R})$ spektra uzorka, dok je b_1 njegova standardna devijacija. Time se pri SNV korekcijski koeficijenti ne izračunavaju iz svih spekatara, već zasebno iz pojedinačnih spektara. Prednost ove metode je izostanak korištenja referentnog spektra, dok je nedostatak izostanak optimizacije u prvom koraku, zbog čega je u korigiranim podatcima moguća prisutnost bazne linije. SNV također korigira spektre s obzirom na raspršenje i pomak bazne linije.

4.1.3 Norris-Williams derivacija

Norris-Williams derivaciju (NW) predstavili su Norris i Williams 1984. godine kao metodu deriviranja NIR spektara [22]. Sastoji se od dva koraka:

1. Ravnanje spektara uzimajući prosjek $2m + 1$ točki sljedećom jednadžbom:

$$x_{\text{smooth},i} = \frac{\sum_{j=-m}^{m} x_{\text{org},i+j}}{2m+1} \tag{4.2a}$$

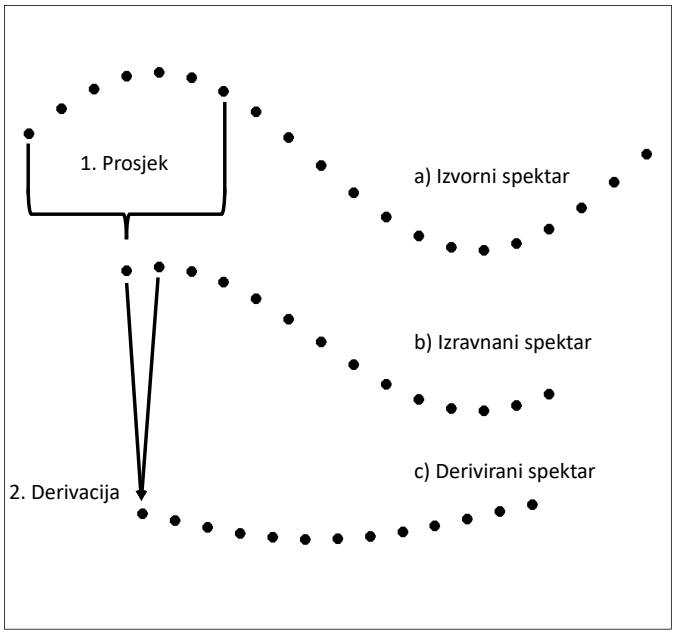
gdje je $x_{\text{smooth},i}$ vrijednost izravnanog spektra u točki i, $x_{\text{org},i+j}$ je vrijednost izvornog spektra u točki $i+j$, a $2m+1$ je broj točki kroz koje se provodi ravnanje. Ravnanje je simetrično s obzirom na točku i te je zbog toga broj točki u ravnanju neparan.

2. Korigiranje izmjerenih spektara:

$$x'_{i+\frac{1}{2}} = x_{\text{smooth},i+1} - x_{\text{smooth},i} \tag{4.2b}$$

$$x''_i = x_{\text{smooth},i+1} - 2x_{\text{smooth},i} + x_{\text{smooth},i-1} \tag{4.2c}$$

pri čemu su x' i x'' prva, odn. druga derivacija u točki. Prilikom korištenja NW nužno je odabrati broj točki za ravnanje, $2m + 1$, i stupanj derivacije. Korištenjem NW dimenzije spektara smanjuju se za $2m$ točki prilikom ravnanja i još jedna za prvu, odn. dvije za drugu derivaciju. Ravnanje spektara u prvom koraku ima ulogu da se ne smanji omjer signala i šuma u spektru te prva derivacija uklanja pomak bazne linije, dok druga uklanja baznu liniju i linearni trend u spektru [19]. Shema NW prikazana je u slici 4.1.



Slika 4.1 Princip Norris-Williams derivacije. U prvom koraku izvorni spektri ravnaju se kao prosjek kroz npr. 7 točki, nakon čega se u drugom koraku izravnani spektri deriviraju kao razlika susjednih točki. Preuzeto i prilagođeno iz [19].

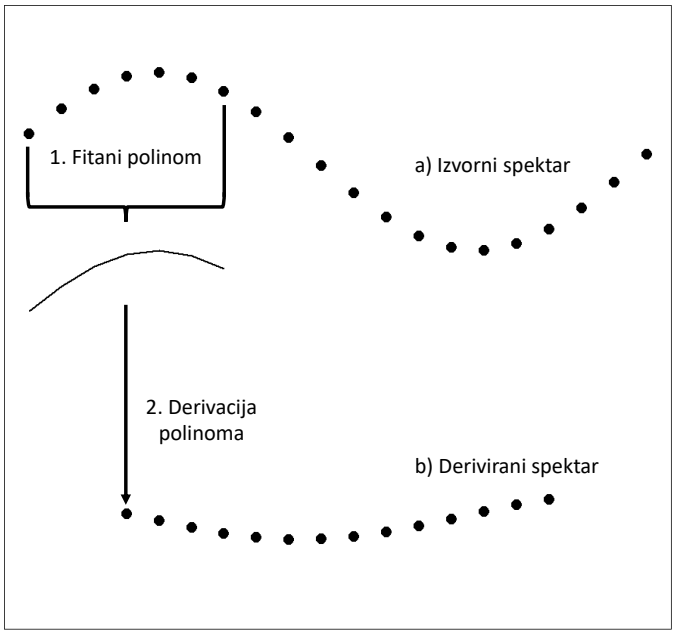
4.1.4 Savitzky-Golay derivacija

Savitzky-Golay derivaciju (SG) predstavili su Savitzky, Golay i sur. 1964. godine kao metodu numeričkog deriviranja vektora koja uključuje i korak ravnanja [23]. Ideja metode vrlo je slična NW, osim što se u prvom koraku spektri ravnaju prilagođavanjem polinoma kroz $2m + 1$ točki. Potom se polinomi deriviraju u točki i na samoj sredini točaka ravnanja, tako što se derivira funkcija polinoma ravnanja. U praksi, SG provodi se množenjem vektora u rasponu ravnanja s vektorom koeficijenata SG deriviranja i dijeljenjem sa skalarom normalizacije [24]:

$$x_{\mathrm{SG},i} = \frac{\sum_{j=-m}^{m} x_{\mathrm{org},i+j} c_{j+m+1}}{norm} \tag{4.3}$$

gdje je $x_{\mathrm{SG},i}$ vrijednost deriviranog spektra u točki i, $2m+1$ je broj točki u ravnanju, c je vektor koeficijenata SG derivacije, a *norm* je skalar normalizacije spektra. Kod SG, za razliku od NW, u konačnici se dimenzije spektara smanjuju za $2m$ točki nakon deriviranja te stupanj derivacije ne može biti veći od stupnja polinoma ravnanja. Shema SG prikazana je u slici 4.2.

Iako NW i SG imaju zajednički korak deriviranja spektara, kao i nužnost simetričnog ravnanja, zbog različitog principa ravnanja u prvom koraku, ne daju isto rješenje. Prednost SG metode u odnosu na NW je u uobičajenijem principu ravnanja polinomom i gubitku manjeg broja točaka, dok je nedostatak potreba za izračunavanjem koeficijenata deriviranja i normalizacije.



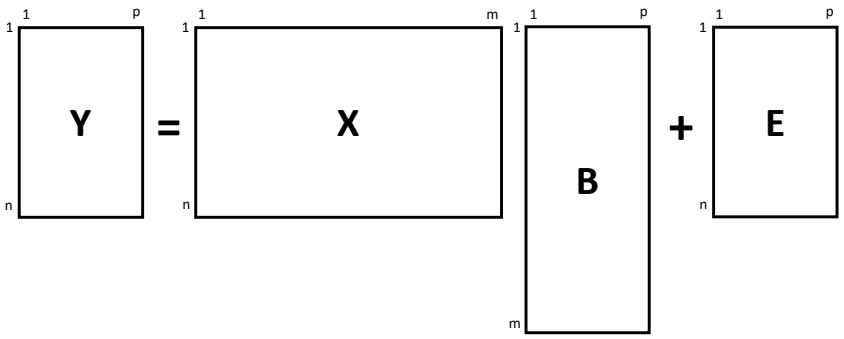
Slika 4.2 Princip Savitzky-Golay derivacije. U prvom koraku izvorni spektri ravnaju se npr. polinomom 2. stupnja kroz 7 točki, nakon čega se u drugom koraku isti polinom derivira. Preuzeto i prilagođeno iz [19].

4.2 Multivarijatna linearna regresija

Multivarijatna linearna regresija (eng. *multivariate linear regression*, MLR) proširenje je metode najmanjih kvadrata (eng. *ordinary least squares*, OLS) za slučaj s više varijabli odgovora i prediktorskih varijabli:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{4.4}$$

gdje su $\mathbf{X}$ i $\mathbf{Y}$ matrice prediktorskih varijabli odn. varijabli odgovora, $\mathbf{B}$ predstavlja matricu regresijskih koeficijenata (regresijska matrica), a $\mathbf{E}$ predstavlja matricu regresijskih reziduala. U prvom stupcu $\mathbf{X}$ sadrži dodatni vektor $\mathbf{1}$ te je dimenzija matrice $n \times (m+1)$, zbog izračunavanja odsječka. Stoga $\mathbf{B}$ ima dimenzije $(m+1) \times p$ jer prvi redak matrice predstavlja vektor odsječaka. Ukoliko su podatci prethodno centrirani, vektor $\mathbf{1}$ u $\mathbf{X}$ nije potreban. U tom slučaju $\mathbf{X}$ je dimenzija $n \times m$, a $\mathbf{B}$ dimenzija $m \times p$. Matrice MLR modela prikazane su u slici 4.3.



Slika 4.3 Matrična jednadžba multivarijatne linearne regresije. Preuzeto i prilagođeno iz [3].

Kod MLR, kao i OLS, $\mathbf{B}$ se računa minimiziranjem zbroja kvadrata reziduala, dakle minimizira se trag $\mathbf{E'E}$. Važno je naglasiti da je $\mathbf{B}_{\text{MLR}}$ moguće izračunati putem analitičkog izraza:

$$\mathbf{B}_{\mathrm{MLR}} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \tag{4.5}$$

gdje je $(\mathbf{X'X})^{-1}\mathbf{X'}$ lijevi pseudo-inverz matrice $\mathbf{X}$.

$\mathbf{B}_{\mathrm{MLR}}$ je najbolji linearni nepristrani procjenitelj (eng. *best linear unbiased estimator*, BLUE). Najbolji označava da od nepristranih procjenitelja ima najmanju varijancu, linearan jer se radi o linearnom modelu po jedn. 4.4, a nepristran je s obzirom da je očekivana vrijednost $\mathbf{B}_{\mathrm{MLR}}$ jednaka $\mathbf{B}$ ($\mathrm{E}[\mathbf{B}_{\mathrm{MLR}}] = \mathbf{B}$). Pritom, da bi $\mathbf{B}_{\mathrm{MLR}}$ bio BLUE moraju biti zadovoljeni uvjeti Gauss-Markovljevog teorema [25]:

$$\mathrm{E}[e_i] = 0 \,\forall\, i \in \{1, 2, ..., n\} \tag{4.6a}$$

$$\mathrm{Var}(e_i) = \sigma^2 < \infty \,\forall\, i \in \{1, 2, ..., n\} \tag{4.6b}$$

$$\mathrm{Cov}(e_i, e_j) = 0 \,\forall\, i \neq j \tag{4.6c}$$

gdje su e_i i e_j i-ti odn. j-ti rezidual. Uvjeti Gauss-Markovljevog teorema govore da svi reziduali imaju očekivanu vrijednost 0 i jednaku varijancu σ^2 te da su međusobno neovisni, odn. da su *IID* (*independent and identically distributed*).

$\mathbf{B}_{\mathrm{MLR}}$ je skup od p vektora regresijskih koeficijenata $\mathbf{b}_i$ koji su međusobno neovisni te se mogu izračunati provođenjem MLR zasebno za svaku varijablu odgovora $\mathbf{y}_i$ iz $\mathbf{Y}$:

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{e}_i \tag{4.7a}$$

$$\mathbf{B}_{\mathrm{MLR}} = [\mathbf{b}_1, ..., \mathbf{b}_p]; \mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_p] \tag{4.7b}$$

gdje je $\mathbf{y}_i$ i-ti vektor $\mathbf{Y}$, $\mathbf{b}_i$ regresijski vektor za $\mathbf{y}_i$, a $\mathbf{e}_i$ i-ti rezidual.

Nužni uvjeti za provođenje MLR metode su da je broj prediktorskih varijabli manji od broja uzoraka ($m < n$) te da su prediktori nekolinearni. Veći broj prediktora u odnosu na broj uzoraka u analizi dovodi do nemogućnosti predviđanja odgovora iz razloga što u tom slučaju nije moguće izračunati inverz $\mathbf{X'X}$ u jedn. 4.5. Taj problem može se nadvladati korištenjem metoda strojnog učenja, no u tom slučaju rezultat je više rješenja za $\mathbf{B}_{\mathrm{MLR}}$. $\mathbf{X'X}$ je matrica momenta koja je pozitivna semidefinitna matrica te za centriranu $\mathbf{X}$ određuje matricu kovarijanci. Kolinearnost utječe

na njen inverz na način da on postaje nestabilan te dovodi do nestabilnih regresijskih koreficijenata[2], a time i do netočnijeg modela. Problemi kolinearnosti i većeg broja prediktorskih varijabli rješavaju se selekcijom prije MLR, regularizacijom ili redukcijom dimenzionalnosti [26].

4.2.1 Selekcija varijabli

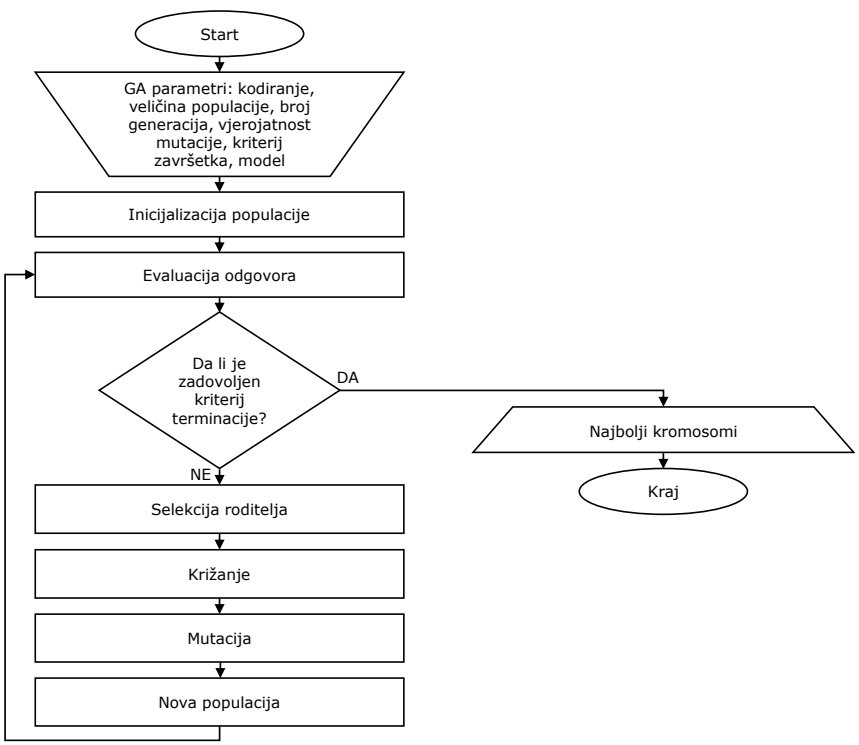
Kao što je ranije navedeno, veći broj prediktorskih varijabli od broja uzoraka te kolinearnost kod NIR podataka problem su za korištenje MLR. Zbog toga se često pristupa drugim regresijskim metodama, no moguće je koristiti i MLR uz eliminaciju prediktorskih varijabli koje: 1. ne pridonose modelu; 2. daju istu informaciju kao i neke druge varijable; 3. imaju vrlo malu povezanost s varijablama odgovora. U tom slučaju provodi se selekcija varijabli koja, ukoliko je dobro provedena, može dati model bolji i od onog izračunatog regresijskim metodama koje se baziraju redukciji varijabli ili regularizaciji [27].

Idealna metoda selekcije varijabli bila bi određivanje svih mogućih njihovih kombinacija i testiranje tih modela. To nije izvedivo jer bi primjerice za skup od 500 varijabli trebalo provest račun za $\sum_i^{500} \binom{500}{i} = 2^{500} \approx 3.27 \times 10^{150}$ kombinacija varijabli, što zahtijeva previše vremena i računalne memorije te nije moguće izvesti na osobnom računalu. Stoga su potrebni inteligentni pristupi koji će rezultirati modelom s maksimalnom prediktivnom moći u što kraćem vremenu. Neki od pristupa koji se koriste u tu svrhu su *stepwise* metode, *best subset* selekcije, selekcije bazirane na glavnim komponentama, genetički algoritmi te *lasso* regresija. Sve te metode imaju isti konačan cilj - na inteligentan način smanjiti broj kombinacija testiranih varijabli, da bi se smanjilo računalno vrijeme potrebno za izračun njihovog idealnog podskupa, a da se pritom regresijski model optimizira na temelju svih uključenih varijabli. U ovom radu korištene su selekcija varijabli genetičkim algoritmima, *forward stepwise* i *best subset* metodom.

[2]stabilnost regresijskih koeficijenata znači njihovu neovisnost o podskupu trening podataka

4.2.2 Genetički algoritmi

Genetički algoritmi (GA) poprimaju sve veću važnost u kemometriji, gdje su prve primjene bile još 70-tih godina prošlog stoljeća, dok tijekom proteklih par desetljeća raste njihova primjena u optimizaciji modela i selekciji varijabli. GA su inspirirani biološkom definicijom procesa evolucije, pri čemu populacija (podskupovi varijabli koji se modeliraju) evoluira na način da opstaju najbolji podskupovi. Upravo zbog toga prigodni su za selekciju varijabli prije regresijske obrade. Konačni cilj korištenja ove metode odabir je podskupa varijabli takvog da regresijom daje optimalan kalibracijski model [3].



Slika 4.4 Dijagram toka genetičkog algoritma (GA). Preuzeto i prilagođeno iz [28].

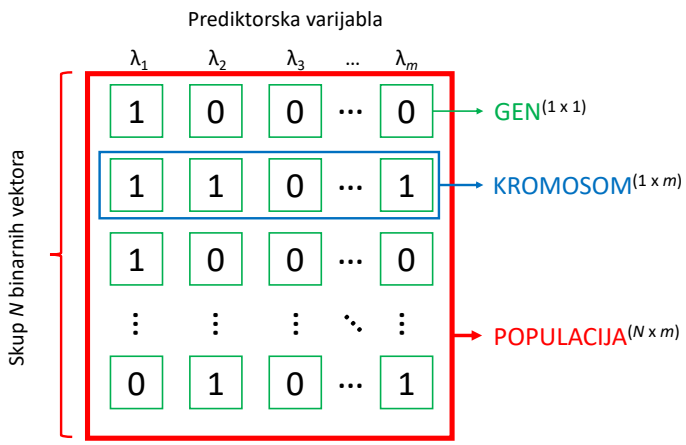
GA provode se kroz 5 koraka koji su algoritmom prikazani u slici 4.4 [28]:

1. Kodiranje varijabli

   Gen označava binarnu varijablu koja može poprimiti vrijednosti 0 ili 1. Kromosom je vektor koji se sastoji od svih gena, kojih ima onoliko koliko i prediktorskih varijabli. Množenjem prediktorskih varijabli s njima pripadajućim genima, postiže se uključivanje ili isključivanje pojedine prediktorske varijable iz kalibracijskog modela.

2. Inicijalizacija populacije

   Skup od nekoliko kromosoma predstavlja populaciju te je dizajn prikazan u slici 4.5. Na samom početku potrebno je odrediti N kromosoma kao početnu populaciju pri čemu je N između 20 i 100. Pritom se kreće s nasumično određenim kromosomima, s ograničenim maksimalnim brojem "1" gena.



Slika 4.5 Reprezentacija značenja gena, kromosoma i populacije kod genetičkih algoritama. Preuzeto i prilagođeno iz [29].

3. Procjena odgovora

   Definira se kriterij po kojem se odabiru najbolji kromosomi. S obzirom da populacija određuje podskupove odabranih varijabli, odabirom najboljih kromosoma izabiru se najbolji podskupovi varijabli, tj. modeli. U našem slučaju koristi se MLR uz R^2 (koeficijent determinacije) kao kriterij kvalitete modela,

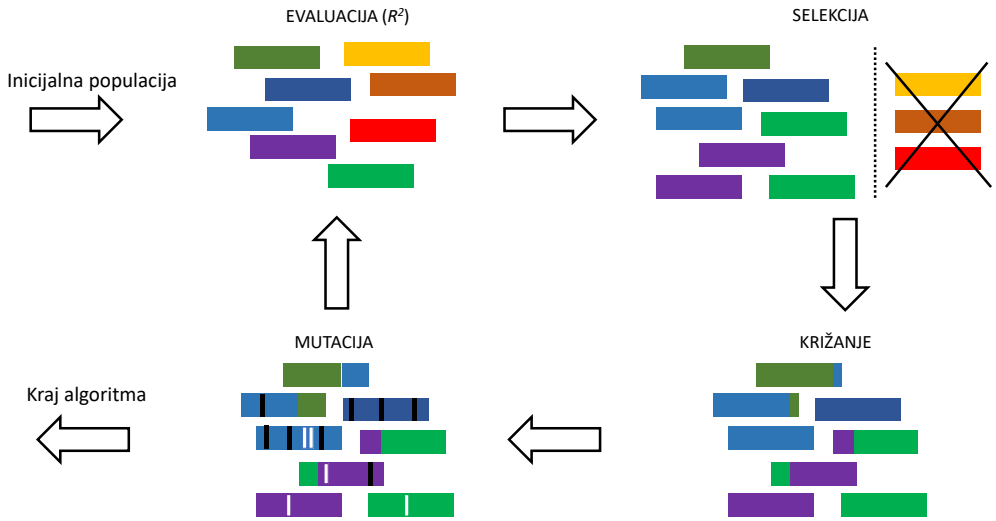
tj. najbolji kromosom je onaj za čiji model je R^2 najveći. GA su prilagođeni samo za provođenje na jednoj varijabli odgovora $\mathbf{y}$, ali s obzirom da vrijedi jedn. 4.7, GA-MLR se može provesti zasebno za svaki $\mathbf{y}$.

4. Reprodukcija

   U ovom koraku se iz postojeće populacije stvara nova populacija. Za stvaranje nove generacije koriste se 2 operatora: selekcija i križanje. Selekcija odabire N kromosoma iz postojeće populacije i kopira ih u novu na način da najbolji kromosomi imaju veću vjerojatnost odabira. Time se stvara dojam kompeticije gdje najbolji kromosomi preživljavaju, dok lošiji odumiru. Križanje je korak gdje se N novih kromosoma upari i međusobno križa. Ideja je da se kombinacijom dva "roditeljska" kromosoma stvaraju još bolji kromosomi "djece".

5. Mutacija

   Mutacijom se u svakoj generaciji malom broju kromosoma nasumično promjeni vrijednost nekih gena iz 1 u 0 i obrnuto. Nakon mutacije stvorena je nova



Slika 4.6 Operatori selekcije, križanja i mutacije kod genetičkih algoritama. Preuzeto i prilagođeno iz [30].

generacija populacije koja ponovno ulazi u analizu u korak procjene odgovora, sve dok se ne dostigne kriterij završetka (maksimalan broj populacija) te je cijelokupni proces selekcije, mutacije i križanja prikazan u slici 4.6.

4.2.3 *Best subset* selekcija

Best subset selekcija u najširem smislu je testiranje svih mogućih kombinacija varijabli i pronalaženje najbolje. Već smo napomenunli da su ovakvi izračuni često nemogući zbog velikog broja prediktorskih varijabli, zbog čega se algoritam *best subset* selekcije rastavlja na dva dijela [26]:

1. (a) Definiramo nul-model M_0 koji ne sadržava niti jedan prediktor te se radi o regresiji koja računa srednju vrijednost (eng. *mean only regression*)

   (b) Za $k = 1, 2, ..., m$:

   - Trening MLR modela koji sadrže k prediktora
   - Odabir najboljeg modela M_k s obzirom na R^2

2. Odabir najboljeg od svih $M_0, ..., M_m$ modela, onog koji ima najmanju unakrsno validiranu pogrešku predikcije, koristeći neki od kriterija: C_p, BIC ili *prilagođeni* R^2

U slučaju da je broj prediktorskih varijabli m velik, tada se k u drugom koraku limitira na neku vrijednost l, pri čemu je $l < m$. Korištenje R^2 kao kriterija kvalitete modela u prvom koraku osigurava brzinu, ali nije kvalitetno mjerilo s obzirom da računa *in-sample* pogrešku i favorizira modele s više prediktora. *Mallow's* C_p, *Bayesian information criterion* (BIC) i *prilagođeni* R^2 u drugom koraku procjenjuju *out-of-sample* pogrešku, na način da u izračunu sadržavaju termin kazne koji raste s brojem prediktorskih varijabli u modelu, tj. kažnjavaju veću kompleksnost.

Za model od k prediktora, C_p se izračunava prema jednadžbi:

$$C_p = \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + 2k\hat{\sigma}^2\right) \tag{4.8}$$

gdje je y_i stvarna vrijednost, $\hat{y_i}$ vrijednost predviđena modelom, a $\hat{\sigma}^2$ varijanca rezi-

dula (jedn. 4.6b) uobičajeno procijenjena iz punog modela koji sadržava sve prediktore. Termin $2k\hat{\sigma}^2$ predstavlja kaznu modelima s većim brojem prediktora. BIC se izračunava prema jednadžbi:

$$\text{BIC} = \frac{1}{n\hat{\sigma}^2}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \log(n)k\hat{\sigma}^2\right) \tag{4.9}$$

gdje termin $\log(n)k\hat{\sigma}^2$ predstavlja kaznu kompleksnijih modela. Prema C_p i BIC najbolji model je onaj s najmanjom vrijednošću. *Prilagođeni* R^2 se izračunava prema jednadžbi:

$$Prilagodeni\ R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-k-1)}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2/(n-1)} \tag{4.10}$$

gdje je $\bar{y_i}$ srednja vrijednost. Prema *prilagođenom* R^2, za razliku od C_p i BIC, najbolji model je onaj s navjećom vrijednošću kriterija. Za multivarijatni odgovor $\mathbf{Y}$ svi kriteriji se računaju kao prosjek kriterija izračunatih za svaku Y-varijablu $\mathbf{y}$.

4.2.4 *Stepwise* selekcija

Stepwise selekcije dijelimo na 3 oblika - *forward stepwise*, *backward stepwise* i hibridni pristup. *Forward stepwise* započinje s praznim modelom M_0, nakon čega se dodaje po jedna prediktorska varijabla u model, kao kod *best subset* selekcije. Razlika je u tome što se prilikom odabira M_k (korak 1.b) ne bira između svih mogućih kombinacija s k prediktorskih varijabli, već se na M_{k-1} dodaje još jedna što značajno ubrzava algoritam. *Backward stepwise* selekcija kreće s maksimalnim brojem prediktorskih varijabli (punim modelom), nakon čega se u svakom koraku oduzima po jedna koja mu najmanje doprinosi. Iz razloga što počinje sa svim prediktorskim varijablama, *backward stepwise* selekcija rijetko se koristi u kemometriji zbog uvjeta $n > m$. Hibridni pristup kombinacija je *forward* i *backward stepwise* selekcija na način da je polazni model prazan te kreće s dodavanjem novih prediktorskih varijabli, ali moguće je i oduzimanje [26]. U ovom radu korištena je samo *forward stepwise* selekcija.

4.3 Ridge regresija

Hoerl i Kennard predložili su 1970. godine ridge regresiju (RR) kao alternativnu metodu MLR. Za razliku od potonje, u ovoj metodi nije potrebna selekcija varijabli, s obzirom da ne podliježe uvjetima većeg broja uzoraka od prediktora i njihove nekolinearnosti [31]. Uz *lasso* i *elastic net*, ova metoda pripada grupi koje koriste tehniku regularizacije, u ovom slučaju tipa L_2 [26]. Ridge regresijom smanjuju se regresijski koeficijenti u $\mathbf{B}$ da bi se stabilizirali (vidi fusnotu 2, str. 24), čime se reducira *overfitting*. To se postiže dodatkom *tuning* parametra λ ($\lambda \geq 0$) svim dijagonalnim elementima $\mathbf{X'X}$ iz jedn. 4.5 čime se stabilizira inverz $\mathbf{X'X}$. Time je $\mathbf{B}_{\mathrm{RR}}$ određen kao:

$$\mathbf{B}_{\mathrm{RR}} = (\mathbf{X'X} + \lambda\mathbf{I})^{-1}\mathbf{X'Y} \tag{4.11}$$

S obzirom da ridge regresija nema utjecaja na odsječak potrebno je da su podatci centrirani.

Primjena ove jednostavne modifikacije u ridge regresiji omogućava da je broj prediktorskih varijabli veći od broja uzoraka, dozvoljava kolinearnost u $\mathbf{X}$ te *tuning* parametar λ omogućava optimizaciju modela. Korištenjem ridge regresije ne dolazi do redukcije varijabli, već do optimizacije regresijske matrice. Upravo zbog toga što je kod NIRSa broj varijabli gotovo uvijek veći od broja uzoraka, odlična je i jednostavna alternativa MLR.

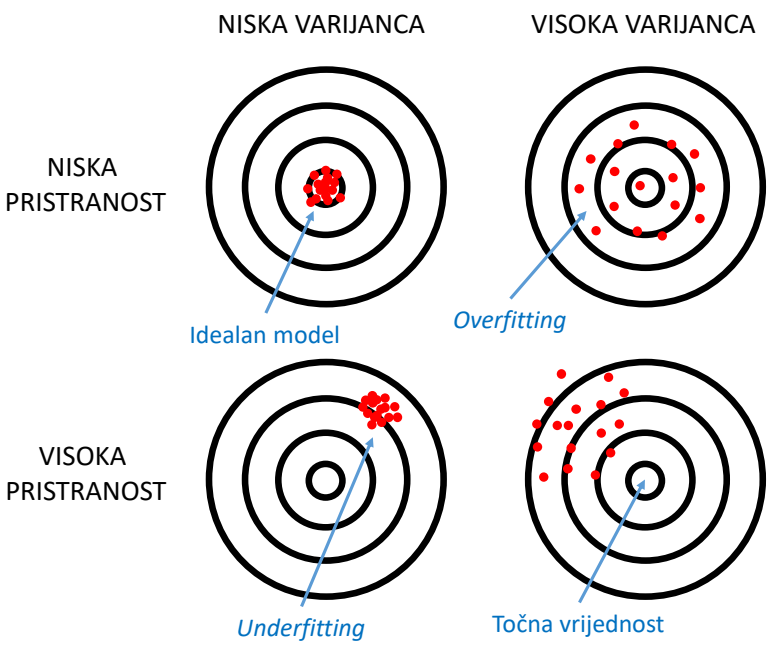
4.3.1 Kompromis između varijance i pristranosti

Nužno svojstvo prediktivnih modela u strojnom učenju je kompromis između varijance i pristranosti, pri čemu modeli s manjom pristranosti imaju veću varijancu procjene odgovora u uzorcima i obrnuto. Do utjecaja oba doprinosa dolazi iz razloga što model ne može istovremeno minimizirati i pristranost i varijancu čije je značenje prikazano u slici 4.7.

Pristranost je razlika između prosječnog predviđanja modela i točne vrijednosti koju pokušavamo predvidjeti. Problem visoke pristranosti predstavlja slučaj kada prediktivni model propušta važne odnose između prediktorskih varijabli i varijabli

odgovora (sistematska pogreška). Zbog toga prilikom regresije dolazi do *underfitting*-a pri čemu model ne uspjeva pronaći osnovni uzorak u podatcima, što se često događa kod jednostavnijih modela s manjim brojem parametara [26, 32, 33].

S druge strane, u slučaju kada je model prenaučen prilikom treninga metode, dolazi do veće varijance u predviđenoj varijabli te je model izrazito osjetljiv na promjene parametara. Taj slučaj naziva se *overfitting*-om te model nije dovoljno generaliziran s obzirom na nove podatke koji nisu prethodno bili uključeni u trening. *Overfitting* se događa kada model uči i iz šuma u podatcima, što je čest slučaj kod kompliciranijih modela s velikim brojem parametara, naročito modela dubokog učenja i klasifikacijskih/regresijskih stabala [26, 32, 33].

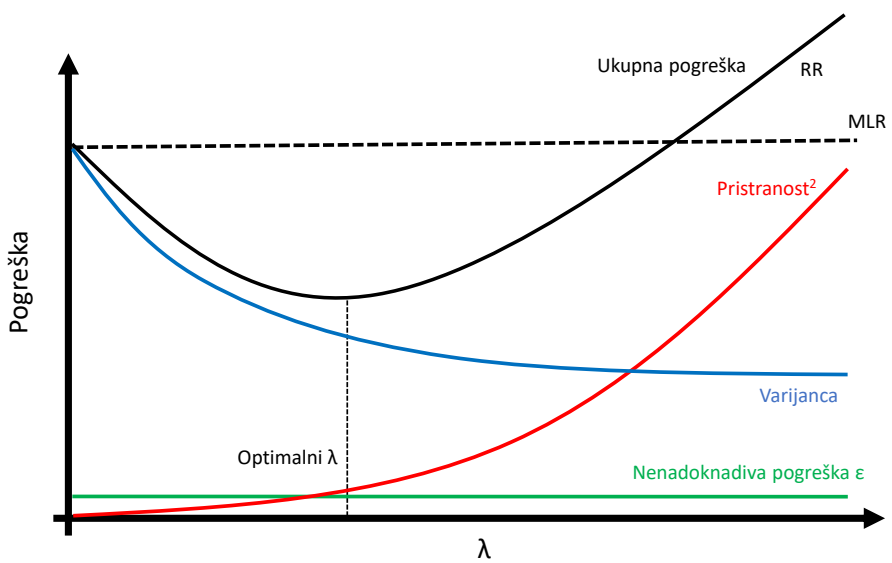


Slika 4.7 Ilustracija pristranosti i varijance. Prilagođeno iz [33].

Pogreška modela određena je kao zbroj varijance, kvadrata pristranosti i nenadoknadive pogreške ϵ, koja se ne može ukloniti [26, 32, 33]:

$$\text{Error} = \text{Var} + \text{Bias}^2 + \epsilon \tag{4.12}$$

Time kod ridge regresije smanjenjem regresijkih koeficijenata s λ dolazi do smanjenja varijance u 4.12, ali i povećanja pristranosti. Stoga λ mora biti takav da je zbroj pristranosti i varijance minimalan, tj. da je pogreška minimalna. Graf u slici 4.8 prikazuje da postoji minimum funkcije Pogreška(λ) kojim je određen optimalan λ. Varijanca ridge regresije za $\lambda = 0$ jednaka je pogrešci MLR, a pogreška ridge regresije za optimalni λ manja je od pogreške MLR. Optimalna vrijednost λ određuje se metodama ponovnog uzorkovanja (eng. *resampling*), minimizirajući pritom pogrešku predikcije [26, 32, 34].



Slika 4.8 Ovisnost kvadrata pristranosti (crveno), varijance (plavo), nenadoknadive pogreške ϵ (zeleno), ukupne pogreške ridge regresije (crno) i pogreške multivarijatne linearne regresije (iscrtano) o *tuning* parametru λ. Preuzeto i prilagođeno iz [33].

4.4 Regresija glavnih komponenti

4.4.1 Analiza glavnih komponenti

Analiza glavnih komponenti (eng. *principal component analysis*, PCA) osnovna je i široko primijenjena metoda kemometrijskih analiza. Ova metoda koristi se za pojednostavljivanje podataka, interpretaciju, vizualnu inspekciju, klasifikaciju, regresiju, pronalaženje *outlier*-a, selekciju varijabli i sl. Pearson je prvi prezentirao PCA kao metodu pronalaženja koordinata koje su najmanje udaljene od točaka u prostoru [35], na skupovima od 2 do 3 varijable zbog nemogućnosti izračuna za veći broj. Shodno razvoju tehnologije slijedio je daljnji razvoj i optimizacija PCA, a time i veća primjena u znanosti.

U PCA problem kolinearnosti između prediktorskih varijabli rješava se tako što se podatci reduciraju na manji skup od samo nekoliko latentnih varijabli koje najbolje opisuju izvorne, tj. objašnjavaju maksimalnu varijancu u podatcima (slika 4.9) [36]. To se postiže dekompozicijom $\mathbf{X}$ na matricu skorova (eng. *scores*) $\mathbf{T}$ $(n \times a)$ i matricu opterećenja (eng. *loadings*) $\mathbf{P}$ $(m \times a)$, uz minimizaciju X-reziduala sadržanih u $\mathbf{F}$:

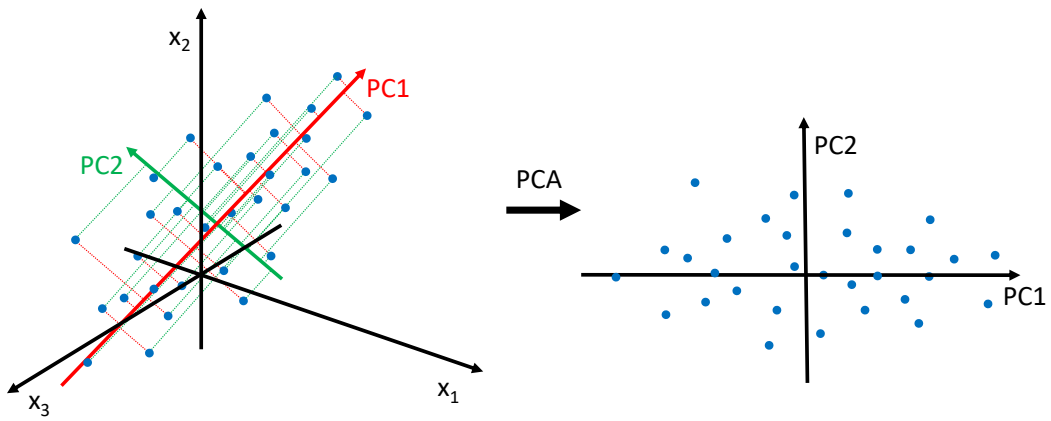
$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{F} = \sum_{i=1}^{a} \mathbf{t}_i \mathbf{p}_i{}' + \mathbf{F} \tag{4.13}$$

Vektori opterećenja $\mathbf{p}$ određuju osi novog koordinatnog sustava, dok vektori skorova $\mathbf{t}$ određuju projekcije izvornih varijabli na nove koodrinate $\mathbf{p}$. Stoga kažemo da $\mathbf{P}$ predstavlja glavne komponente, a $\mathbf{T}$ (latentne varijable) projekciju $\mathbf{X}$ na glavne komponente. Pri tom, $\mathbf{t}$ i $\mathbf{p}$ poredani su od onih koji objašnjavaju najveći udio varijance u $\mathbf{X}$ prema onima koji objašnjavaju najmanji. Desna strana jedn. 4.13 prikazuje dekompoziciju $\mathbf{X}$ kao zbroj matrica vektorskih produkata $\mathbf{t}_i\mathbf{p}_i{}'$ ranga 1, što znači da je $\mathbf{X}$ linearna kombinacija glavnih komponenti. Zbog ortonormalnosti $\mathbf{P}$ i pretpostavke da su $\mathbf{F}$ mali, jedn. 4.13 može se invertirati u:

$$\mathbf{T} = \mathbf{X}\mathbf{P} \tag{4.14}$$

Ova inverzija omogućava izračunavanje $\mathbf{T}$ za bilo koje nove skupove X-varijabli, uz poznati $\mathbf{P}$. Potonja se izračunava iz kalibracijskih X-podataka koristeći jedn. 4.13,

a potom je moguće izračunati **T** za novi skup X-podataka pomoću jedn. 4.14. PCA se naziva bilinearno modeliranje, zbog toga što vrijede jedn. 4.13 i 4.14 [37].



Slika 4.9 Princip analize glavnih komponenti. Skup od 3 varijable ($\mathbf{x}_1$, $\mathbf{x}_1$, $\mathbf{x}_3$) reducira se na skup od 2 glavne komponente (**PC1** i **PC2**) koje objašnjavaju maksimalni dio varijance u podatcima. Prilagođeno iz [37].

X se projicira na skup od a vektora opterećenja ($a < m$). Ukoliko bi bilo $a = m$, tada bi **T** jednostavno bila rotacija izvorne **X** u X-hiperprostoru, pod uvjetom da je a_{max} jednak rangu matrice **X**. Broj a je broj vektora **t** i **p** kojima je najbolje objašnjena varijanca u **X**, a preostali dio neobjašnjene varijance predstavlja šum definiran s **t** kojima je zbroj kvadrata elemenata približno jednak 0 ($\|\mathbf{t}\|^2 \approx 0$), koji nisu dio PCA modela i irelevantni su za modeliranje podataka.

PCA rješenje proizlazi iz rješenja eigenvektorske jednadžbe umnoška $\mathbf{X'X}$ koja glasi [38]:

$$(\mathbf{X'X})\mathbf{p} = \lambda\mathbf{p} \tag{4.15}$$

gdje je **p** vlastiti vektor, a λ vlastita vrijednost $\mathbf{X'X}$. $\mathbf{X'X}$ je matrica kovarijanci zbog čega **X** mora biti centrirana. Ukoliko centriranje nije provedeno, vektori opterećenja bit će pomaknuti s obzirom na podatke. Vlastiti vektori $\mathbf{X'X}$ su **p** iz jedn. 4.13, dok su korijeni vlastitih vrijednosti jednaki duljinama **t**. Eigenvektorskom jednadžbom moguće je izračunati **P** te se takav izračun naziva Jacobijeva rotacija. S obzirom da

je ta tehnika relativno spora, za računanje PCA češće se koriste druge metode poput dekompozicije singularnih vrijednosti (SVD) ili NIPALS algoritma.

Svojstva PCA

Najvažnije svojstvo PCA jesu već spomenuta ograničenja ortogonalnosti **T** i ortonormalnosti **P**. Ortogonalnost **T** znači da su svi **t** međusobno okomiti te da je točkasti produkt svakog para različitih **t** jednak 0. Točkasti produkt **t** sa samim sobom je zbroj kvadrata elemenata vektora tj. kvadrat njegove duljine ili kvadrat euklidske norme:

$$\mathbf{t}_i{}'\mathbf{t}_j = \begin{cases} 0, i \neq j \\ \|\mathbf{t}_i\|^2, i = j \end{cases} \tag{4.16}$$

Ako se jedn. 4.16 proširi na cijelu **T**, rezultat je dijagonalna matrica u kojoj su dijagonalni elementi kvadrati duljine **t**:

$$\mathbf{T'T} = \mathrm{diag}(\|\mathbf{t}_1\|^2, \|\mathbf{t}_2\|^2, ..., \|\mathbf{t}_a\|^2) \tag{4.17}$$

P je ortonormalna što podrazumijeva da su svi **p** međusobno okomiti te da je njihov točkasti produkt jednak 0, dok je točkasti produkt svakog **p** sa samim sobom 1 s obzirom da su normalizirani na duljinu 1:

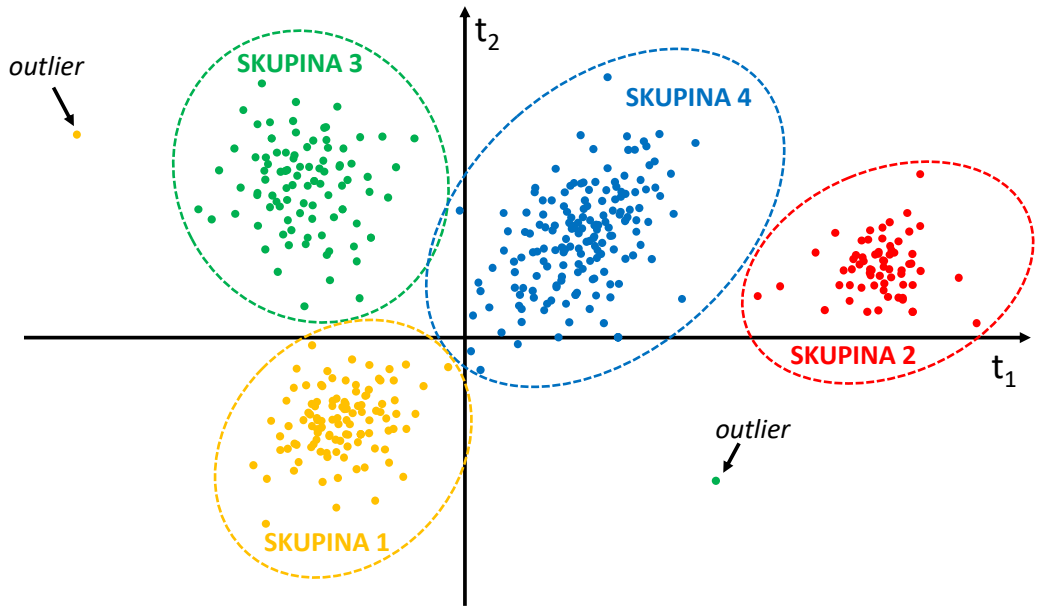
$$\mathbf{p}_i{}'\mathbf{p}_j = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases} \tag{4.18}$$

Proširenjem jedn. 4.18 na cijelu **P**, kao rezultat dobivamo matricu identiteta **I**:

$$\mathbf{P'P} = \mathbf{I} \tag{4.19}$$

Ortonormalnost **P** omogućava invertiranje jednažbe 4.13 u 4.14 zbog toga što za ortonormalne matrice vrijedi da je transponirana **P** jednaka inverzu **P** ($\mathbf{P'} = \mathbf{P}^{-1}$). Posljedica okomitosti **t**, odn. **p** je da su svi vektori u **T**, odn. **P** međusobno nekorelirani, čime je riješen problem kolinearnosti X-varijabli.

Pregled podataka može se provesti izradom grafa skorova, tzv. **t-t** grafom u kojem su prikazani parovi prvih **t** koji objašnjavaju najveći udio u ukupnoj varijanci svih podataka. Uobičajeno se prikazuju $\mathbf{t}_1$-$\mathbf{t}_2$, $\mathbf{t}_1$-$\mathbf{t}_3$ ili $\mathbf{t}_2$-$\mathbf{t}_3$ grafovi. Na njima se mogu uočiti uzorci i grupiranja u podatcima na način da točke (uzorci) koji pripadaju istom podskupu podataka imaju slične vrijednosti te se grupiraju. Uz to, mogu se uočiti *outlier*-i kao točke koje nesistematski značajno odstupaju od osnovnog skupa podataka na grafu. Primjeri grupiranja podataka i detekcije *outlier*-a **t-t** grafom prikazani su u slici 4.10.



Slika 4.10 Primjer **t-t** grafa kao rezultata PCA analize. Možemo uočiti grupiranje 4 skupine podataka i *outlier*-e kao točke koje značajno odstupaju od ostalih na grafu..

Dekompozicija singularnih vrijednosti

S obzirom na svojstva PCA, jedn. 4.13 se može riješiti pomoću dekompozicije singularnih vrijednosti (eng. *singular value decomposition*, SVD) koja dekomponira centriranu **X** na sljedeći način:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{4.20}$$

pritom su $\mathbf{U}$ i $\mathbf{V}$ ortonormalne, a $\mathbf{\Sigma}$ dijagonalna matrica. $\mathbf{P}$ iz jedn. 4.13 ekvivalentan je $\mathbf{V}$ iz jedn. 4.20, dok je $\mathbf{T}$ iz jedn. 4.13 ekvivalentan umnošku $\mathbf{U\Sigma}$ iz jedn. 4.20. Ukoliko je $\mathbf{X}$ singularna, tada su neki od dijagonalnih elemenata $\mathbf{\Sigma}$ 0, a broj ne-nul elemenata jednak je rangu $\mathbf{X}$. Vrijednosti elemenata na dijagonali predstavljaju duljinu $\mathbf{t}$ te su kvadratno proporcionalni varijanci u podatcima koju opisuje $\mathbf{t}$. Kvadrati tih vrijednosti vlastite su vrijednosti $\mathbf{X'X}$ po jedn. 4.15 [36].

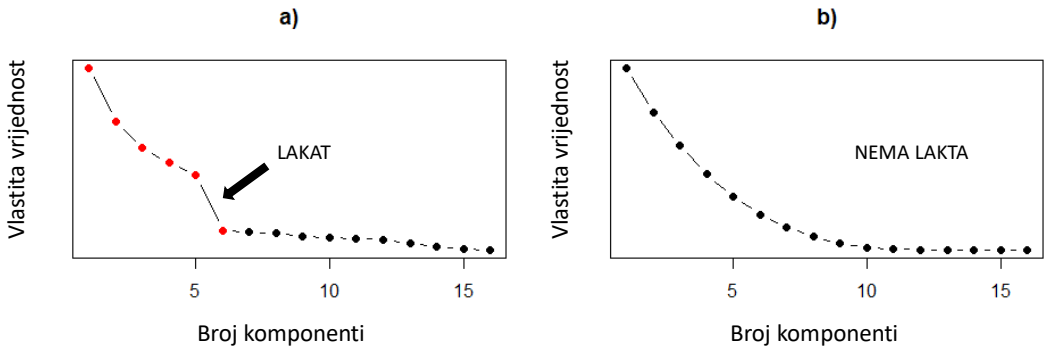
4.4.2 Određivanje broja glavnih komponenti

Kao što je prethodno navedeno, $\mathbf{X}$ se dekomponira na a glavnih komponenti koje najbolje objašnjavaju $\mathbf{X}$, no a je nepoznat. Za određivanje broja glavnih komponenti u PCA koriste se 3 kriterija: Kaiserov kriterij, *scree* test i udio objašnjene varijance [39].

Kaiserov kriterij ili vlastita vrijednost-1 kriterij najjednostavniji je pristup određivanju broja glavnih komponenti, a primjenjuje se kod modela s manje od 100 izvornih varijabli. Prednost ovog kriterija je jednostavnost, dok je nedostatak nemogućnost primjene ukoliko se vlastite vrijednosti vrlo malo razlikuju od 1 [40]. Kriterij glasi da treba zadržati sve komponente koje imaju vlastitu vrijednost veću od 1 iz jedn. 4.15, a primjenjuje se ukoliko je $\mathbf{X}$ skalirana. U tom slučaju svaka izvorna varijabla pridonosi ukupnoj varijanci podataka s 1. To znači da svaka komponenta koja ima vlastitu vrijednost veću od 1 objašnjava veći dio varijance u podatcima od izvorne X-varijable. Kaiserov kriterij može se koristiti i na neskaliranim podatcima na način da se standardizira prema zbroju vlastitih vrijednosti.

Scree test temelji se na grafičkom prikazu ovisnosti vlastitih vrijednosti o broju komponenti. Kao optimalni broj glavnih komponenti uzima se onaj nakon kojeg se u grafu uočava nagli pad, tzv. "lakat" (vidi sliku 4.11 a)) [41]. Ovaj test češće se primjenjuje ukoliko je broj varijabli veći (> 200) te je nedostatak situacija kada ne postoji nagli pad između para komponenti, kao u slici 4.11 b).

Udio objašnjene varijance je kriterij po kojem se odabiru sve glavne komponente



Slika 4.11 Odabir broja glavnih komponenti *scree* testom. a) Postoji nagli pad vlastith vrijednosti uzastopnih komponenti te *scree* test određuje 6 glavnih komponenti (crveno). b) Ne postoji nagli pad vlastith vrijednosti uzastopnih komponenti zbog čega se *scree* testom ne može odrediti optimalan broj glavnih komponenti.

koje objašnjavaju određeni udio varijance u modelu [39]. Udio objašnjene varijance za komponentu i izračunava se prema jednadžbi:

$$udio_i = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} \tag{4.21}$$

gdje je λ_i vlastita vrijednost komponente i, a m ukupan broj glavnih komponenti u podatcima, tj. varijabli. Primjerice za kriterij od 5%, udio objašnjene varijance zadržava sve komponente koje objašnjavaju min. 5% ukupne varijance prema jedn. 4.21 ($udio_i \geq 0.05$). Ovaj kriterij može se i modificirati u kriterij kumulativnog udjela varijance na način da se zbroje udjeli objašnjene varijance do komponente i:

$$udio_{cumm,i} = \frac{\sum_{k=1}^{i} \lambda_k}{\sum_{j=1}^{m} \lambda_j} \tag{4.22}$$

Po kriteriju kumulativnog udjela, primjerice za 95% zadržava se prvih i glavnih komponenti čiji je kumulativni zbroj udjela objašnjene varijance veći od 95% ($udio_{cumm,i} \geq 0.95$). Nedostatak ovog kriterija je subjektivni odabir postotka u kriteriju.

4.4.3 Regresija PCA

Regresija glavnih komponenti (eng. *principal component regression*, PCR) kombinacija je PCA i MLR [42]. Najprije se $\mathbf{X}$ reducira u $\mathbf{T}$ koja objašnjava većinu varijance u $\mathbf{X}$ prema jedn. 4.13, a potom provodi se MLR prema jedn. 4.4 gdje je $\mathbf{X}$ zamijenjena s $\mathbf{T}$:

$$\mathbf{Y} = \mathbf{T}\mathbf{B}_0 + \mathbf{E} \tag{4.23}$$

$\mathbf{B}_0$ $(a \times p)$ je u ovom slučaju regresijska matrica između $\mathbf{T}$ kao matrice prediktora i $\mathbf{Y}$ kao matrice odgovora. $\mathbf{B}_0$ se dobiva kao BLUE ovog regresijskog modela:

$$\mathbf{B}_0 = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y} \tag{4.24}$$

U ovom slučaju, s obzirom da je $\mathbf{T}$ ortogonalna, problem nestabilnosti inverza $\mathbf{T}'\mathbf{T}$ je riješen. Izostavljanjem vektora skorova koji objašnjavaju vrlo mali dio varijance u $\mathbf{X}$ riješeni su problemi kolinearnosti i većeg broja prediktorskih varijabli od uzoraka. S obzirom da vrijedi jedn. 4.14 možemo napisati i:

$$\mathbf{Y} = \mathbf{X}\mathbf{P}\mathbf{B}_0 + \mathbf{E} \tag{4.25}$$

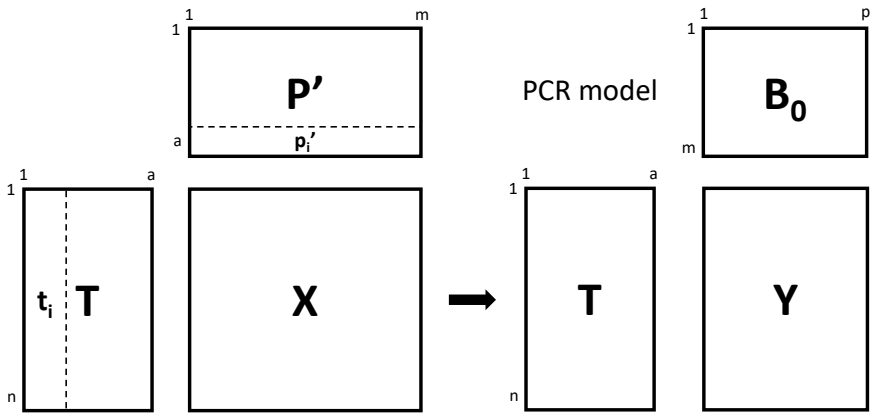
čime je definirana regresijska matrica $\mathbf{B}_{\mathrm{PCR}}$ između $\mathbf{X}$ i $\mathbf{Y}$:

$$\mathbf{B}_{\mathrm{PCR}} = \mathbf{P}\mathbf{B}_0 = \mathbf{P}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y} \tag{4.26}$$

Matrice PCR modela prikazane su u slici 4.12.

Kod regresije postoji problem određivanja optimalnog broja glavnih komponenti, tj. a je hiperparametar PCR. Za razliku od PCA, kod regresije želimo ostvariti model koji ima minimalnu prediktivnu pogrešku, neovisno o tome koliku varijancu objašnjavaju glavne komponente. Stoga se za određivanje optimalnog broja glavnih komponenti koriste metode ponovnog uzorkovanja.

Kod PCR, kao i MLR, vrijedi da je rješenje jednako ukoliko se, umjesto na multivarijatnom $\mathbf{Y}$, PCR koristi na svakoj varijabli odgovora zasebno. Razlog tome je što se u prvom koraku, prilikom redukcije varijabli s PCA, koristi samo informacija



Slika 4.12 Matrični prikaz postupka za regresiju glavnih komponenti. Preuzeto i prilagođeno iz [3].

iz $\mathbf{X}$, neovisno o $\mathbf{Y}$, a drugi korak jednak je MLR za kojeg vrijedi 4.7. Zbog toga se pojedini vektori $\mathbf{B}_{\text{PCR}}$ mogu izračunati i provođenjem PCR između $\mathbf{T}$ i svake Y-varijable:

$$\mathbf{y}_i = \mathbf{T}\mathbf{b}_i + \mathbf{e}_i \tag{4.27a}$$

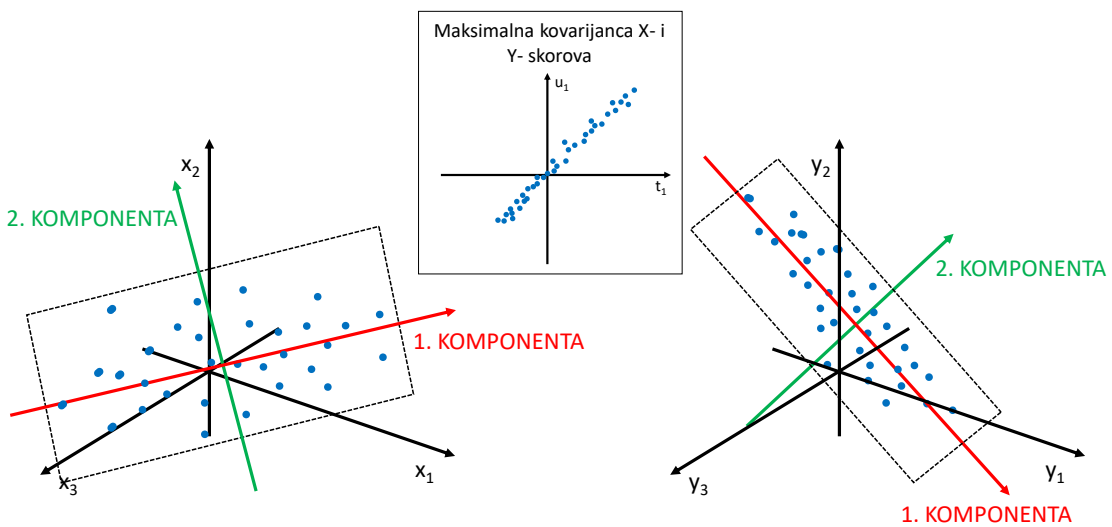
$$\mathbf{B}_{\text{PCR}} = [\mathbf{b}_1, ..., \mathbf{b}_p]; \mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_p] \tag{4.27b}$$

gdje je $\mathbf{y}_i$ i-ti vektor $\mathbf{Y}$, $\mathbf{b}_i$ regresijski vektor za $\mathbf{y}_i$, a $\mathbf{e}_i$ i-ti rezidual.

4.5 Regresija parcijalnih najmanjih kvadrata

Regresija parcijalnih najmanjih kvadrata (eng. *partial least-squares* ili *projection to latent structures*, PLS) najvažnija je regresijska metoda u kemometriji. Sam razvoj kemometrije kao kemijske discipline usko je povezan s razvitkom PLS. Ovu metodu prvi je primijenio statističar Herman Wold 70-tih godina prošlog stoljeća u ekonometrijskim analizama za modeliranje kompleksnih skupova podataka [43]. Njegov sin Svante prilagodio je i primijenio PLS u kemometrijskim analizama [44].

PLS ima za cilj razviti linearni regresijski model povezivanjem $\mathbf{X}$ i $\mathbf{Y}$ preko latentnih varijabli, određenih na način da se maksimizira kovarijanca između X- i Y-skorova (slika 4.13). Zbog toga omogućava analizu podataka s velikim brojem snažno koreliranih varijabli koje sadrže šum, uz modeliranje multivarijatnog odgovora $\mathbf{Y}$. Pri tom, PLS za izračunavanje latentnih varijabli koristi informacije iz obiju $\mathbf{X}$ i $\mathbf{Y}$, dok se kod PCR one računaju isključivo iz $\mathbf{X}$. Time se i $\mathbf{X}$ i $\mathbf{Y}$ dekomponiraju na



Slika 4.13 Princip regresije parcijalnih najmanjih kvadrata. Skup od 3 prediktorske ($\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$) i 3 varijable odgovora ($\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$) reducira se na skup od 2 PLS komponente pritom maksimizirajući kovarijancu između X- i Y-skorova ($\mathbf{t}_1$, odn. $\mathbf{u}_1$). Prilagođeno iz [45].

skorove i opterećenja, što možemo shvatiti kao PCA provedenu i na $\mathbf{X}$ i na $\mathbf{Y}$, pri čemu su njihove glavne komponente dodatno zarotirane u prostoru da bi se postigla maksimalna kovarijanca između X- i Y- skorova [46].

Baš kao i PCA, PLS pronalazi skup novih varijabli, $\mathbf{T}$, koje određuju latentne varijable modela. Stoga je dekompozicija $\mathbf{X}$ jednaka kao kod PCA:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{F}_\mathbf{X} \tag{4.28}$$

pritom svojstva $\mathbf{T}$ i $\mathbf{P}$ su različita jer $\mathbf{P}$ nije ortogonalna. Zbog toga ne vrijedi jedn. 4.14 te je za invertiranje jedn. 4.28 nužno koristiti matricu X-težina (eng. *weights*) $\mathbf{W}$:

$$\mathbf{T} = \mathbf{X}\mathbf{W}^* = \mathbf{X}\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \tag{4.29}$$

pri čemu je $\mathbf{W}$ ortonormalna, dimenzija jednakih $\mathbf{P}$ i ključna za interpretaciju PLS modela. $\mathbf{Y}$ se dekomponira na matricu Y-skorova $\mathbf{U}$ i Y-opterećenja $\mathbf{Q}$:

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F}_\mathbf{Y} \tag{4.30}$$

gdje su $\mathbf{U}$ i $\mathbf{Q}$ jednakih dimenzija kao $\mathbf{T}$ i $\mathbf{P}$, a Y-reziduali u $\mathbf{F}_\mathbf{Y}$ minimizirani. S obzirom da je cilj PLS maksimiziranje kovarijance između X- i Y- skorova, $\mathbf{T}$ odn. $\mathbf{U}$, postoji linearna povezanost njihovih vektora, što možemo prikazati kao:

$$\mathbf{U} = \mathbf{T}\mathbf{V} + \mathbf{H} \tag{4.31}$$

gdje je $\mathbf{V}$ $(a \times a)$ dijagonalna matrica regresijskih koeficijenata između parova $\mathbf{t}_i$ i $\mathbf{u}_i$, a $\mathbf{H}$ su reziduali koji nisu objašnjeni ovom relacijom. S obzirom da vrijedi jedn. 4.31, $\mathbf{Y}$ možemo izraziti kao:

$$\mathbf{Y} = \mathbf{T}\mathbf{V}\mathbf{Q}' + \mathbf{E} = \mathbf{T}\mathbf{C}' + \mathbf{E} \tag{4.32}$$

gdje je $\mathbf{C}$ matrica Y-težina, a $\mathbf{E}$ su reziduali PLS regresije koji su minimalni. Stoga konačni regresijski model za PLS izražavamo kao:

$$\mathbf{Y} = \mathbf{X}\mathbf{B}_{\mathrm{PLS}} + \mathbf{E} \tag{4.33}$$

$$\mathbf{B}_{\mathrm{PLS}} = \mathbf{W}^{*}\mathbf{C}' = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{V}\mathbf{Q}' \tag{4.34}$$

Jedn. 4.34 ne predstavlja analitičko rješenje već približno numeričko. Za provođenje PLS $\mathbf{X}$ i $\mathbf{Y}$ moraju biti centrirane te se optimalan broj PLS komponenti određuje najmanjom pogreškom predikcije koristeći metode ponovnog uzorkovanja [47]. Matrice PLS regresije prikazane su u slici 4.14.

Za razliku od PCR i MLR, rješenje PLS dobiveno zasebno za svaku Y-varijablu $\mathbf{y}$ nije jednako rješenju PLS provedenom simultano na svim Y-varijablama $\mathbf{Y}$, tj. ne vrijedi jednadžba analogna jedn. 4.7 i 4.27. Zbog toga razlikujemo dva načina provođenja: PLS1 i PLS2. PLS1 je provođenje na jednoj varijabli odgovora $\mathbf{y}$, gdje se maksimizira kovarijanca između X-skorova i $\mathbf{y}$. To znači da se u izračunu ne koriste Y-skorovi jer ih se niti ne može odrediti za jedan $\mathbf{y}$. Ukoliko imamo više varijabli odgovora, PLS1 može se provoditi zasebno na svakoj $\mathbf{y}$, gdje je konačan rezultat p PLS1 modela. PLS2 podrazumjeva simultanu predikciju više varijabli odgovora $\mathbf{Y}$, gdje se maksimizira kovarijanca X- i Y- skorova. Ovaj način češće se koristi ukoliko postoji snažna korelacija u $\mathbf{Y}$, tj. ukoliko je $\mathbf{Y}$ određen jednakim ili sličnim svojstvima, što je slučaj u ovom radu [3].

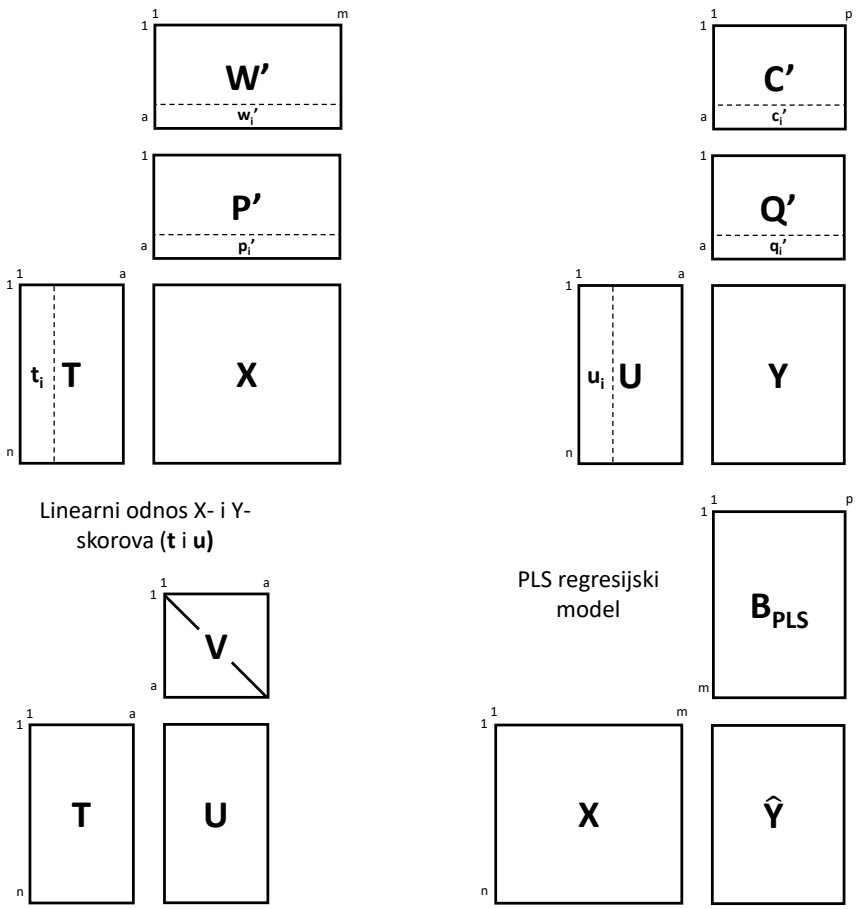
4.5.1 Svojstva PLS

Matematička i statistička svojstva PLS analizirao je Höskuldsson [48]. Dokazao je da prilikom konvergencije NIPALS algoritam se može svesti na eigenvektorski problem, na način su $\mathbf{u}$, $\mathbf{c}$, $\mathbf{t}$ i $\mathbf{w}$ ustvari vlastiti vektori s najvećim vlastitim vrijednostima tzv. *kernel* matrica $\mathbf{YY'XX'}$, $\mathbf{Y'XX'Y}$, $\mathbf{XX'YY'}$ i $\mathbf{X'YY'X}$. Stoga prilikom konvergencije rješenja PLS vrijede sljedeće 4 eigenvektorske jednadžbe:

$$\mathbf{YY'XX'u} = \lambda_1\mathbf{u} \tag{4.35}$$

$$\mathbf{Y'XX'Yc} = \lambda_2\mathbf{c} \tag{4.36}$$

$$\mathbf{YY'XX't} = \lambda_3\mathbf{t} \tag{4.37}$$



Slika 4.14 Matrični prikaz postupka za regresiju parcijalnih najmanjih kvadrata. Preuzeto i prilagođeno iz [3].

$$\mathbf{X'YY'Xw} = \lambda_4 \mathbf{w} \tag{4.38}$$

gdje su λ_1, λ_2, λ_3 i λ_4 najveće vlastite vrijednosti. Navedene jedn. omogućile su razvoj *kernel* algoritama. Höskuldsson je također dokazao da za provedbu PLS nije

potrebna deflatacija[3] $\mathbf{Y}$ nakon svake iteracije, već samo $\mathbf{X}$, s obzirom da je rješenje neovisno o deflataciji $\mathbf{Y}$ [48].

Geometrijska svojstva PLS vezana su uz ortogonalnosti $\mathbf{w}$, $\mathbf{t}$ i $\mathbf{p}$ [48]. Prvo svojstvo je da su $\mathbf{w}$ međusobno okomiti i normalizirani na duljinu 1, što znači da je $\mathbf{W}$ ortonormalna:

$$\mathbf{w}_i{}'\mathbf{w}_j = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases} \tag{4.39}$$

$$\mathbf{W}'\mathbf{W} = \mathbf{I} \tag{4.40}$$

Drugo svojstvo je da su $\mathbf{t}$ međusobno okomiti zbog čega je $\mathbf{T}$ ortogonalna:

$$\mathbf{t}_i{}'\mathbf{t}_j = \begin{cases} 0, i \neq j \\ \|\mathbf{t}_i\|^2, i = j \end{cases} \tag{4.41}$$

$$\mathbf{T}'\mathbf{T} = \mathrm{diag}(\|\mathbf{t}_1\|^2, \|\mathbf{t}_2\|^2, ..., \|\mathbf{t}_a\|^2) \tag{4.42}$$

Treće je svojstvo da su vektori $\mathbf{w}_i$ okomiti na vektore $\mathbf{p}_j$ ukoliko je $i < j$:

$$\mathbf{w}_i{}'\mathbf{p}_j = 0; i < j \tag{4.43}$$

U konačnici matematička i statistička analiza PLS metode koju je proveo Höskuldsson omogućila je razvoj novih PLS algoritama te poboljšanje postojećih [49].

4.5.2 Pretpostavke PLS

Osnovna pretpostavka PLS modela je postojanje latentnih varijabli i postizanje maksimalne kovarijance između X- i Y- skorova [50]. Na podatke utječe samo nekoliko nezavisnih latentnih varijabli. Njihov broj je nepoznat, no svakako manji od broja X-varijabli te ga je moguće odrediti metodama ponovnog uzorkovanja. Za nezavisne

[3]deflatacija označava iterativno "ljuštenje" informacije iz neke matrice, uz smanjivanje njezinog ranga. Za točan opis vidi [46].

X-varijable je broj latentih jednak broju X-varijabli. Tada su latentne samo rotacija X-varijabli te u tom slučaju PLS i MLR daju isto rješenje. U slučaju kada X-varijable nisu međusobno nezavisne, **X** ima nepotpun rang te PLS rezultira statistički robusnijim rješenjem od MLR.

Druga pretpostavka PLS modela je homogenost što znači da kroz sve iteracije postupka vrijedi isti ili sličan linearan odnos između **u** i **t**, kao i mehanizam utjecaja **X** na **Y** [50]. Homogenost se testira proučavanjem **u-t** grafova, koji dakle mogu ukazivati na prisutnost nelinearnosti u podatcima. Ukoliko je prisutna, može se pristupiti nelinearnim PLS tehnikama kao što su: transformacija podataka (INLR i GIFI-PLS), nelinearni PLS s umjetnim neuronskim mrežama (PLS-ANN) ili nelinearni PLS bez umjetnih neuronskih mreža (QPLS i SPLPLS) [51].

4.5.3 PLS algoritmi

NIPALS

NIPALS (eng. *Nonlinear Iterative PArtial Least-Squares*) algoritam koristi se i za PCA i za PLS te je jedan od prvih algoritama korištenih u kemometrijskim analizama [43]. Njime se glavne komponente računaju iterativnim procesom dok nije postignuta njihova konvergencija i maksimalna kovarijanca između X- i Y- skorova [46]. Ovaj algoritam u nekim slučajevima može biti neprikladan jer je potrebna kontinuirana deflatacija **X** i **Y**, iako su preložene modifikacije za njegovo ubrzavanje [49]. NIPALS je transparentan i točan algoritam, ali veliki nedostatak mu je sporost.

Kernel

Kernel algoritmi za izračun PLS baziraju se na *kernel* i *association* matricama te su razvijeni na temeljima Höskuldssonove matematičke i statističke analize [48]. Razlikujemo dva tipa: originalni *kernel* algoritam i tzv. *wide kernel* algoritam. *Kernel* algoritam predstavili su Lindgren i sur. kao brži i računalno manje zahtjevniji algoritam PLS regresije za skupove podataka s velikim brojem uzoraka [52], dok su *wide kernel* algoritam predstavili su Rännar i sur. kao PLS algoritam koji je brži i

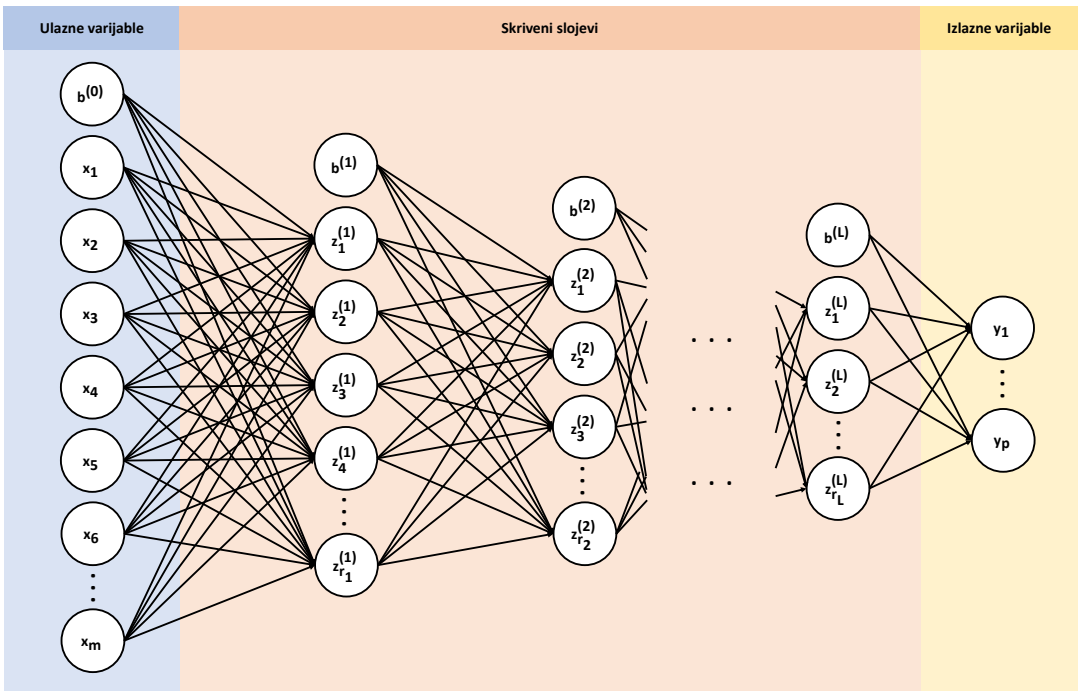
zahtijeva manje memorije od dosadašnjih za specifičan slučaj kada je broj varijabli izrazito velik, a broj uzoraka vrlo mali ($m >> n$) [53, 54].

SIMPLS

SIMPLS algoritam predstavio je de Jong 1992. g., kao PLS algoritam koji izravno računa $\mathbf{T}$ kao linearnu kombinaciju originalnih varijabli [55]. Pri tom, izbjegnuta je deflatacija matrica prisutna u NIPALS algoritmu, što pridonosi brzini i korištenju manje računalne memorije. Ovaj algoritam, za razliku od *kernel* i NIPALS algoritama, u slučaju multivarijatnog odgovora $\mathbf{Y}$ daje nešto drugačije rješenje iz razloga što uistinu maksimizira kovarijancu između X- i Y- skorova. SIMPLS je izrazito brz i točan algoritam.

4.6 Umjetne neuronske mreže

Umjetne neuronske mreže (eng. *artificial neural networks*, ANN) nelinearna su računalna metoda koja zbog svoje fleksibilnosti i prilagodbe ima široku primjenu u kemijskim analizama, uključujući prediktivne regresijske analize [56]. Ideja ANN proizlazi iz neuroznanosti, tj. iz modeliranja načina na koji su povezani neuroni u ljudskom mozgu. X- i Y-varijable povezuju se preko varijabli u nekoliko skrivenih slojeva prikazanih u slici 4.15. Skrivene varijable djeluju na principu neurona u biološkom smislu, tako što su sve varijable pojedinog sloja funkcijski povezane sa svim varijablama prethodnog i sljedećeg sloja. Stoga na njih gledamo kao na čvorove neuronskih mreža, čime se postiže imitacija prijenosa signala u ljudskom mozgu.



Slika 4.15 Shema modela dubokog učenja neuronskim mrežama.

U modelu neuronskih mreža **X** (ulazne varijable) i **Y** (izlazne varijable) povezuju se varijablama u skrivenim slojevima pomoću težina (eng. *weights*), gdje su X-

varijable povezane s r_1 skrivenih varijabli prvog sloja linearnim modelom:

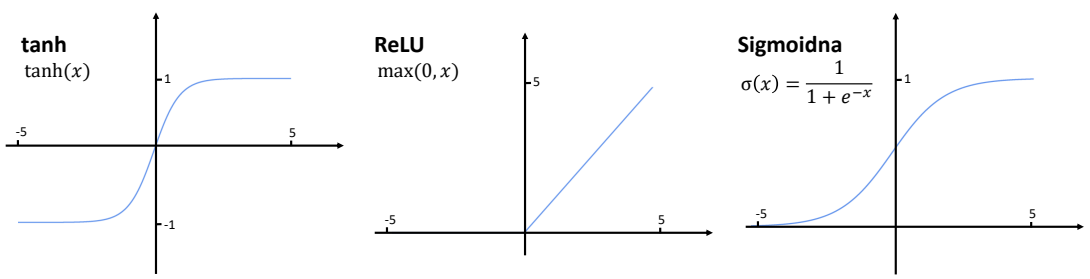
$$\mathbf{V}^{(1)} = \mathbf{X}\mathbf{W}^{(0)} = \sum_{i=0}^{m} \mathbf{x}_i \boldsymbol{w}_i^{(0)} \tag{4.44}$$

gdje je $\mathbf{W^{(0)}}$ matrica težina prvog sloja dimenzija $(m+1) \times r_1$. U $\mathbf{X}$ je dodana tzv. jedinica pomaka (eng. *bias unit*) ($\mathbf{x}_0$), tj. vektor sa svim elementima 1. Skrivene varijable se potom transformiraju nelinearnom aktivacijskom funkcijom, najčešće sigmoidnom, čime poprimaju vrijednosti u intervalu $\langle 0, 1 \rangle$:

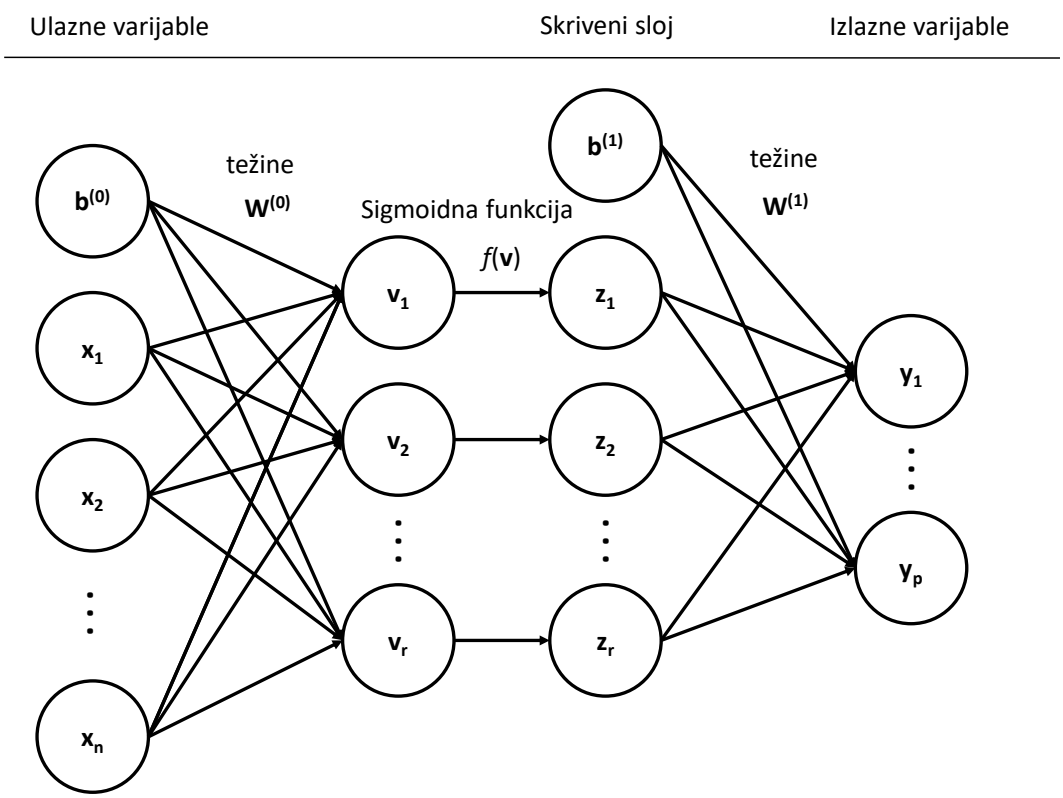
$$\mathbf{z}_1 = \sigma(\mathbf{v}_1) = \frac{1}{1 + e^{-\mathbf{v}_1}} \tag{4.45}$$

Nakon toga, skrivene varijable do L-tog sloja povezuju se na isti način prema jedn. 4.44, uz dodatak jedinice pomaka u svakom sloju, te se tranformiraju aktivacijskom funkcijom.

Pri tom, za trening modela neuronskih mreža potrebno je odabrati hiperparametre: broj skrivenih slojeva L, broj varijabli u pojedinom skrivenom sloju r_L te aktivacijsku funkciju. Osim sigmoidne, često se kao aktivacijske funkcije koriste ReLU (*rectified linear unit*) ili tanh prikazane u slici 4.16. Dodatno, modeli neuronskih mreža mogu se optimizirati hiperparametrima izbacivanja varijabli između slojeva (eng. *dropout*) i regularizacijom kako bi se izbjegao *overfitting*. Shema modela jednostavne neuronske mreže prikazana je u slici 4.17.



Slika 4.16 Aktivacijske funkcije kod neuronskih mreža: tanh, ReLU i sigmoidna. Preuzeto i prilagođeno iz [57].



Slika 4.17 Shema modela neuronske mreže s jednim skrivenim slojem.

Prilikom treninga modela neuronskih mreža izračunavaju se parametri, tj. težine u svakom sloju. Cilj neuronskih mreža je minimizirati funkciju gubitka (eng. *loss function*) $J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$, koja poprima vrijednost $\frac{1}{2}$ MSE:

$$J(\mathbf{W}^{(0,\ldots,\mathrm{L})}) = \frac{1}{2np}\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{j,i} - \hat{y}_{j,i}) \tag{4.46}$$

pri čemu $\mathbf{W}^{(0,\ldots,\mathrm{L})}$ označava sve težine modela neuronske mreže. Broj parametara može biti velik te primjerice za model od 20 X-varijabli, 3 skrivena sloja s 10 varijabli u svakom sloju te 3 Y-varijable iznosi 573. Parametri se određuju *backpropagation* algoritmom koji funkcionira u pet koraka [58, 57]:

1. Nasumično postavljanje početnih vrijednost svih težina, $\mathbf{W}^{(0,\ldots,\mathrm{L})}$.

2. Implementacija propagacije unaprijed. Za dane vrijednosti $\mathbf{W}^{(0,\ldots,\mathrm{L})}$, $\mathbf{X}$ i $\mathbf{Y}$ izračunavaju se sve vrijednosti skrivenih varijabli, predviđene vrijednosti $\hat{\mathbf{Y}}$ i $J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$.

3. Unazadna propagacija pogreške. Računaju se prve derivacije funkcije gubitka po svakom parametru w_i, $\frac{\delta}{\delta w_i} J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$. Derivacije se računaju unatrag, na način da se prve računaju one kasnijih skrivenih slojeva, a zadnje prvog sloja jer se pogreška računa unatrag, koristeći pravilo lanca za izračun derivacija.

4. Koristeći algoritam opadajućeg gradijenta (eng. *gradient descent*) i $\frac{\delta}{\delta w_i} J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$ nove vrijednosti w_i parametara se istodobno ažuriraju prema jednadžbi:

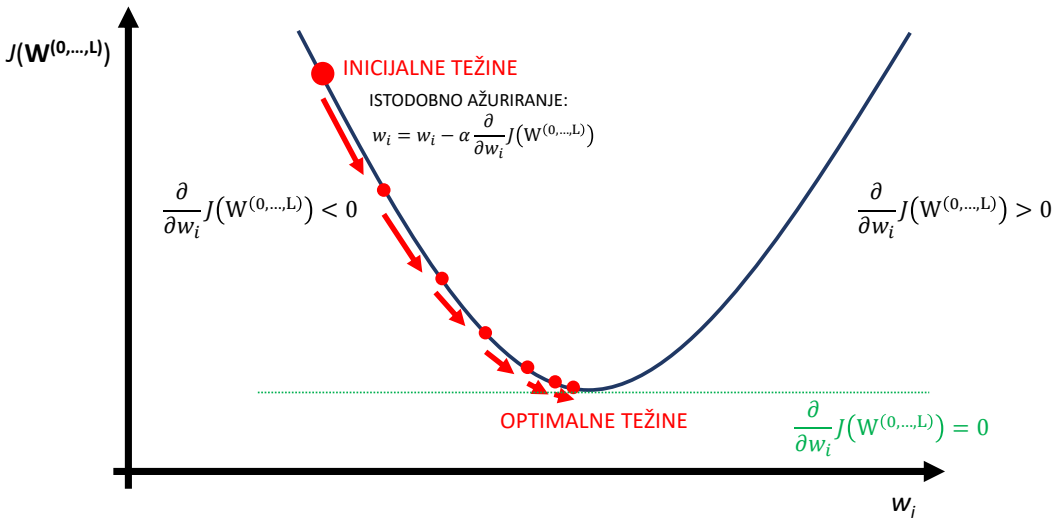
$$w_i = w_i - \alpha \frac{\delta}{\delta w_i} J(\mathbf{W}^{(0,\ldots,\mathrm{L})}) \tag{4.47}$$

   gdje je α stopa učenja (eng. *learning rate*). Broj α je hiperparametar koji, ukoliko je premali usporava konvergenciju algoritma, dok ukoliko je prevelik može dovesti do divergencije. Shema algoritma opadajućeg gradijenta prikazana je u slici 4.18.

5. Algoritam se ponavlja dok nije postignuta konvergencija, tj. minimizacija $J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$, i time određeni optimalni parametri neuronske mreže. S obzirom da $J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$ nije konveksna funkcija, postoji mogućnost da algoritam konvergira u lokalnom minimumu. Kako postoji više lokalnih minimuma, ANN model može rezultirati različitim rješenjima prilikom treninga.

Najjednostavniji oblik neuronske mreže je MLR koja je ustvari neuronska mreža bez skrivenih slojeva i aktivacijske funkcije. Logistička regresija je neuronska mreža bez skrivenih slojeva i sigmoidnom aktivacijskom funkcijom s obzirom na Y-varijable (kod klasifikacije se i izlazne varijable aktiviraju). Model s jednim skrivenim slojem naziva se *single-hidden-layer* neuronska mreža, dok modeli s više skrivenih slojeva pripadaju metodama dubokog učenja.

ANN ne zadovoljavaju nikakav zadani matematički model te mogu modelirati bilo kakvu funkcijsku ovisnost, zbog čega su izrazito korisne za nelinearno modeliranje. Kod NIRSa je eventualno prisutna nelinearnost uzrokovana kompleksnim kompleksnim biološkim, okolišnim i instrumentalnim varijacijama [59]. ANN mogu



Slika 4.18 Shema algoritma opadajućeg gradijenta za minimizaciju funkcije gubitka $J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$ i određivanje optimalnih parametara $\mathbf{W}^{(0,\ldots,\mathrm{L})}$. U svakoj iteraciji algoritma istodobno se ažuriraju svi parametri w_i, dok nije postignuta konvergencija, tj. minimalna vrijednost $J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$. Jedn. 4.47 uz odgovarajući α osigurava minimizaciju $J(\mathbf{W}^{(0,\ldots,\mathrm{L})})$, no ukoliko je α premali dolazi do spore konvergencije algortima, dok ukoliko je prevelik dolazi do divergencije.

biti spregnute s linearnim modelima. Primjer je ANN-PLS metoda koja predstavlja nelinearni kalibracijski model [51]. Nedostatak im je što zahtijevaju korištenje velikog kapaciteta računalne memorije i spore su prilikom treninga, posebice ukoliko se koristi veći broj skrivenih slojeva i jedinica u njima, pa stoga trebaju veliki skup podataka prilikom treninga, što dodatno usporava algoritam. Također, osjetljive su na nasumičnu, nesistematsku kontaminaciju, koja u puno slučaja može predstavljati isključivo šum, zbog čega dolazi do prenaučenosti na podatake [13]. Težine neuronskih mreža nemaju nikakvu interpretabilnu vrijednost, zbog čega se neuronske mreže smatraju modelima crnih kutija ("*black box*").

Od svih metoda prikazanih u ovom radu, jedino ANN su nelinearne. One su primjenjene jer, iako u teoriji vibracijski NIR spektri zadovoljavaju linearni odnos,

u stvarnosti dolazi do pojave nelinearnosti u spektrima. Nelinearnost je posljedica pomaka vibracijskih vrpci u spektrima te inter- i/ili intra- molekulskih interakcija, poput $\pi - \pi$ slaganja, vodikovih veza pa čak i djelovanja Van Der Waalsovih sila [60]. U tom slučaju nelinearne regresijske metode mogle bi biti dobra alternativa linearnim.

4.7 Performance modela

Određivanje performanci modela povezano je s određivanjem prediktivne pogreške pojedinih metoda te se kod multivarijatnih regresijskih analiza koristi u koracima optimizacije i evaluacije modela. Minimizacija prediktivne pogreške u optimizacijske svrhe koristi se s ciljem određivanja hiperparametara poput λ u ridge regresiji, broja komponenata u PCR i PLS i sl. U evaluaciji modela, minimalna pogreška predikcije koristi se za određivanje optimalne regresijske metode (modela). Metode ponovnog uzorkovanja za određivanje prediktivne pogreške su *bootstrap* i unakrsna validacija, od kojih je u ovom radu korištena isključivo unakrsna validacija.

U analizi prediktivnih modela važna je razlika između predikcije i prilagodbe, posebice prilikom određivanja hiperparametara regresije. Pogreška predikcije ukazuje koliko dobro regresijski model predviđa nove vrijednosti koje nisu uključene njegov trening (*out-of-sample*), dok pogreška prilagodbe ukazuje koliko model dobro predviđa vrijednosti koje su uključene u trening (*in-sample*). Posljedica korištenja pogreške prilagodbe umjesto predikcije prilikom određivanja broja komponenti PCR i PLS modela, je da će oni s većim brojem komponenti pokazati manju pogrešku prilagodbe. Razlog tome je što objašnjavaju veći udio varijance u podatcima, što rezultira *overfitting*-om. S druge strane, pogreška predikcije pokazat će lokalne minimume koji omogućuju određivanje optimalnog broja komponenti za predviđanje novih uzoraka jer dodatne komponente uglavnom objašnjavaju mali dio varijance koja predstavlja samo šum i smetnje. Time, za razliku od pogreške prilagodbe, ne daje preoptimističan rezultat, a pogreška prilagodbe uvijek je manja od pogreške predikcije.

Kao mjera prediktivne pogreške u ovom radu korišten je korijen srednje kvadratne pogreške predikcije (eng. *root mean squared error*, RMSE) definiran na sljedeći način:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.48}$$

pri čemu je y_i stvarna vrijednost elementa i, $\hat{y}_i$ je predviđena vrijednost elementa i pomoću modela, a n je broj predikcija. Za regresijski model na multivarijatnom odgovoru $\mathbf{Y}$, RMSE se računa pojedinačno za svaku varijablu matrice $\mathbf{Y}$ te je po-

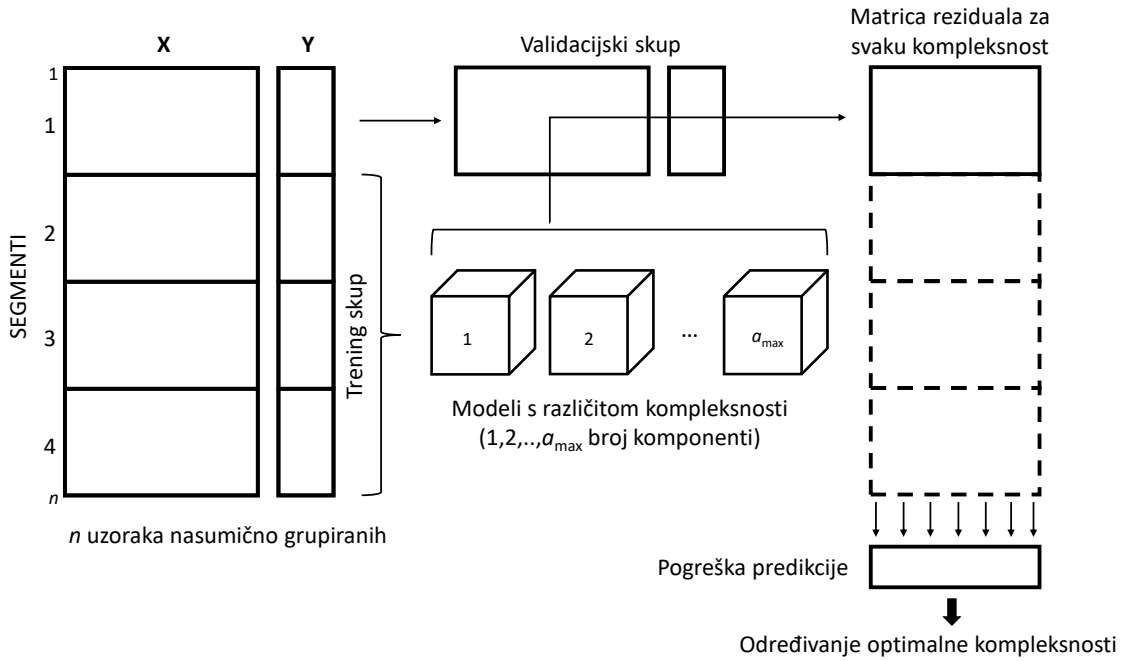
greška modela prosjek svih vrijednosti. RMSE je odabran kao najčešće korištena metoda određivanja prediktivne pogreške prilikom regresije te se izražava jedinicom jednakom izvornim Y-podatcima. Ostale mogućnosti su srednja kvadratna pogreška (eng. *mean squared error*, MSE), R^2 (koeficijent determinacije), srednja apsolutna pogreška (eng. *mean absolute error*, MAE), srednja kvadratna postotna pogreška (eng. *mean square percentage error*, MSPE), korijen srednje kvadratne postotne pogreške (eng. *root mean square percentage error*, RMSPE), srednja apsolutna postotna pogreška (eng. *mean apsolute percentage error*, MAPE), srednja kvadratna logaritamska pogreška (eng. *mean square logarithmic error*, MSLE) i korijen srednje kvadratne logaritamske pogreške (eng. *root mean square logarithmic error*, RMSLE). Uz RMSE može se izračunati i njegova standardna pogreška (SE) metodom unakrsne validacije, čime je moguće odrediti intervale pouzdanosti.

4.7.1 Unakrsna validacija

Unakrsna validacija (eng. *cross validation*, CV) najkorištenija je metoda ponovnog uzorkovanja u kemometriji i strojnom učenju da bi se ostvario veći broj predikcija za izvođenje zaključaka o kvaliteti modela [3, 27]. Koristi se za optimizaciju modela, primjerice za određivanje kompleksnosti (broja komponenti) PCR i PLS regresija, kao i optimalnog *tuning* parametra ridge regresije. Važno je naglasiti da se pomoću CV model evaluira s obzirom na prediktivnu moć (pogrešku predikcije), a ne kriterij prilagodbe modela (pogrešku prilagodbe).

Osnovna procedura CV provodi se na način kako je prikazano u slici 4.19. Skup od n uzoraka razdvaja se nasumično u k segmenata čiji broj može biti u rasponu od 2 do n, najčešće od 4 do 10. Jedan segment se izdvaja u tzv. validacijski skup, dok preostalih $k-1$ segmenata postaju tzv. trening skup. Koristeći uzorke u trening skupu razvija se regresijski model, nakon čega se njime predviđaju vrijednosti uzoraka u validacijskom skupu. Postupak se ponavlja k puta, dok svaki segment nije bio validacijski skup te se time dobivaju predviđene vrijednosti $\hat{\mathbf{Y}}_{\mathbf{CV}}$ $(n \times p)$ kojima se računaju reziduali i prediktivne pogreške. Ukoliko se CV koristi za optimizaciju modela, mogu se razviti modeli s različitom kompleksnosti (npr. brojem komponenti/*tuning* parametrom) čime je rezultat jedna $\hat{\mathbf{Y}}_{\mathbf{CV}}$ za svaku razinu kompleks-

nosti, iz koje se potom računa pogreška predikcije. Pritom, minimum iste određuje optimalnu kompleksnost.



Slika 4.19 Unakrsna validacija za određivanje optimalne kompleksnosti modela. Uzorci su nasumično segmentirani u 4 segmenta nakon čega se unakrsnom validacijom računa pogreška predikcije za svaku kompleksnost (broj komponenti), iz čega se određuje optimalna kompleksnost. Preuzeto i prilagođeno iz [3].

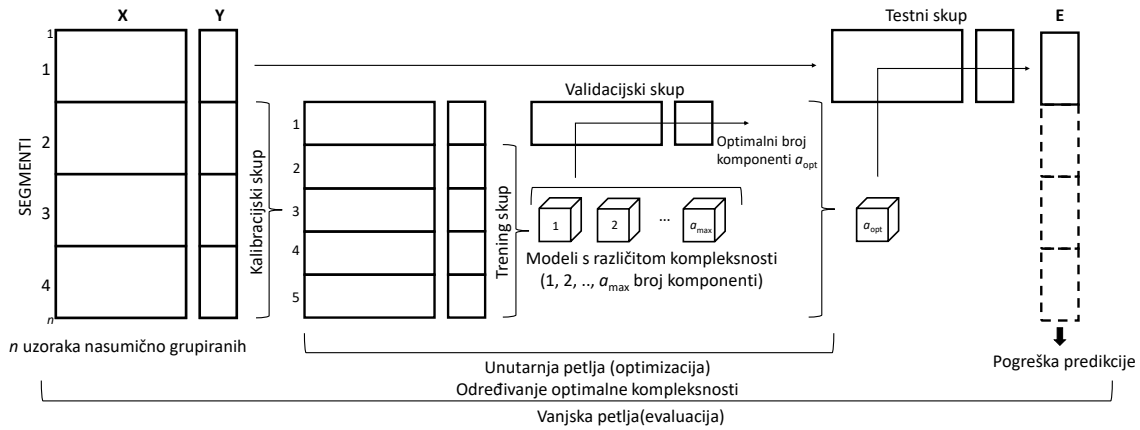
CV procedura daje n predikcija, tj. predikciju za svaki uzorak. Kako bi povećali statističku vrijednost CV, broj predikcija i smanjili utjecaj randomizacije prilikom uzorkovanja (nasumičnog razdvajanja u segmente) na rezultat, koristi se ponovljena unakrsna validacija (eng. *repeated cross validation*, rCV). Metoda rCV ponavlja CV postupak r puta, s time da se ponavlja i nasumično razdvajanje u segmenate, zbog čega svaka iteracija daje različit rezultat.

Druga mogućnost je korištenje pojedinačne unakrsne validacije (eng. *leave-one-out cross validation*, LOO CV) gdje je broj segmenata k u CV jednak n te je u validacijskom skupu prisutan samo jedan uzorak. Prednost u odnosu na uobičajenu k-struku CV je što ne koristi randomizaciju prilikom uzorkovanja koja u nekim

slučajevima može biti besmislena (npr. za mali broj uzoraka) te se jednostavnije programira i dobra je za preliminarnu evaluaciju. Specifično je korisna kod ridge regresije jer segmentiranje može imati negativan utjecaj na model zbog manje stabilnosti regresijskih koeficijenata od primjerice PLS ili PCR modela [3]. Nedostaci ove metode su što za veće skupove podataka traje znatno duže, može dati preoptimističan rezultat (posebice ako su neki uzorci slični kao kod ponavljanih mjerenja) te nema mogućnosti ponavljanja jer bi ponovljena LOO CV u svakoj iteraciji dala identičan rezultat.

Ukoliko se CV provodi s ciljem i optimizacije i evaluacije modela, tada se koristi dvostruka unakrsna validacija (eng. *double cross validation*, dCV) čiji princip je prikazan u slici 4.20. Kod te se metode optimizacija i evaluacija odvajaju u zasebne faze te razlikujemo vanjsku i unutarnju petlju, čime se izbjegava preoptimistična evaluacija modela. U vanjskoj petlji, koja ima za cilj evaluaciju modela, uzorci se nasumično segmentiraju na način da jedan segment postaje testni skup dok preostali postaju kalibracijski. U unutarnjoj petlji kalibracijski skup segmentira se u trening i validacijski s ciljem optimizacije modela. Unutarnjom petljom određuje se optimalna kompleksnost modela, kao ona za koju je pogreška predikcije najmanja. Potom se u vanjskoj petlji na kalibracijskom skupu razvija model za optimalnu kompleksnost i primjenjuje na testnom skupu. Pogreška predikcije testnog skupa koristi se za evaluaciju modela i usporedbu s drugim regresijskim modelima. Broj segmenata u vanjskoj i unutarnjoj petlji ne mora biti jednak te je češće broj segmenata vanjske petlje manji od broja segmenata unutarnje. Petlje se ponavljaju dok svi segmenti nisu uključeni kao testni, odn. validacijski skup.

Da bi se ostvario optimalan rezultat CV u smislu optimizacije i evaluacije najbolje je koristiti ponovljenu dvostruku unakrsnu validaciju (eng. *repeated double cross validation*, rdCV) koja kombinira svojstva rCV i dCV. To postiže na način da ponavlja dCV postupak r puta, čime se dobiva dovoljno veliki uzorak predikcija iz kojih možemo izvoditi zaključke o modelu. S obzirom da unutarnja petlja rdCV zbog randomizacije prilikom uzorkovanja ne mora uvijek ukazivati na istu optimalnu kompleksnost, konačni regresijski model razvija se s kompleksnošću koja je pokazala najveću frekvenciju odabira.

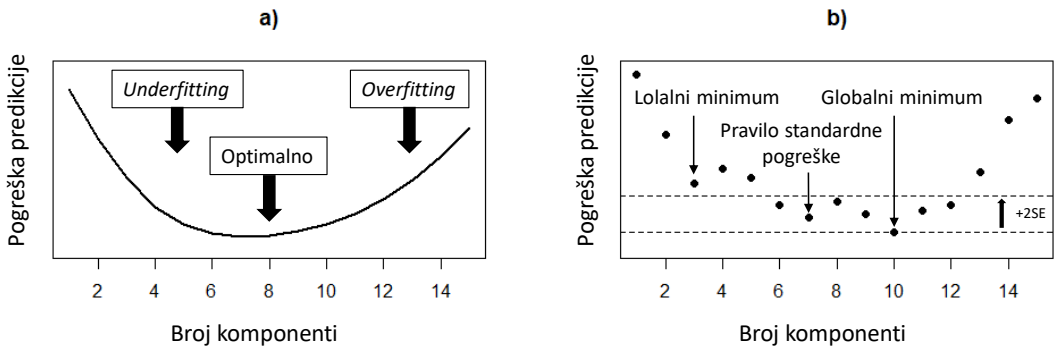


Slika 4.20 Dvostruka unakrsna validacija za optimizaciju i evaluaciju modela. Unutarnjom petljom uzorci u kalibracijskom skupu su nasumično segmentirani u 5 segmenata iz koji se određuje optimalna kompleksnost a_{opt} kao u slici 4.19. Vanjskom petljom uzorci su nasumično segmentirani u 4 segmenta na kojima je razvijen model s a_{opt} brojem komponenti te je izračunata pogreška predikcije modela. Preuzeto i prilagođeno iz [3].

4.7.2 Određivanje optimalne kompleksnosti PCR i PLS modela

Kompleksnost označava veličinu PLS ili PCR modela s obzirom na broj komponenti, analogno vrijednosti *tuning* parametra λ kod ridge regresije. Važno je odrediti model koji pokazuje minimalnu pogrešku predikcije u CV, no uz čim manju kompleksnost jer njezino povećanje smanjuje stabilnost regresijskih koeficijenata. Za broj komponenti različit od optimalnog broja dolazi do *underfitting*-a i *overfitting*-a zbog kompromisa između varijance i pristranosti (slika 4.21 a)) [3]. *Underfitting* se događa kada je broj komponenti manji od optimalnog broja te je model premalo naučen na podatcima kalibracijskog skupa, tj. ima visoku pristranost. Ukoliko je broj varijabli veći od optimalnog broja, tada je model prenaučen na kalibracijskom skupu i govorimo o *overfitting*-u, tj. o problemu visoke varijance.

Razlikujemo 3 načina određivanja optimalne kompleksnosti iz pogreške predikcije: globalni minimum, lokalni minimum i pravilo standardne pogreške koji su prikazani



Slika 4.21 Određivanje kompleksnosti regresijskih modela. a) Minimum u ovisnosti pogreške predikcije o broju komponenti određuje optimalnu kompleksnost (broj komponenti) modela. b) Rezultati optimalne kompleksnosti modela određeni globalnim minimumom, lokalnim minimumom i pravilom standardne pogreške. Preuzeto i prilagođeno iz [3].

u slici 4.21 b) [3]. Praćenje globalnog minimuma je najjednostavniji način određivanja kompleksnosti modela. To znači da je optimalan broj komponenti određen globalnim minimumom ovisnosti pogreške predikcije o broju komponenti te često rezultira *overfitting*-om. Lokalni minimum određuje optimalan broj komponenti kao prvi lokalni minimum ovisnosti pogreške predikcije o broju komponenti. Time izbjegava *overfitting*, no moguća je pojava *underfitting*-a. Pravilo standardne pogreške određuje optimalan broj komponenti kao najmanji broj komponenti za koji je pogreška predikcije manja od one u globalnom minimumu uvećane za 2 standardne pogreške (SE). To znači da je optimalan broj komponenti najmanji njihov broj čija je očekivana vrijednost pogreške predikcije unutar 95%-tnog intervala pouzdanosti vrijednosti u globalnom minimumu. Osim 2SE, mogu se koristiti 1SE ili 3SE. Ovo pravilo izbjegava *underfitting* i *overfitting*, no za određivanje SE nužni su rCV ili *bootstrap*, uz nedostatak primjene subjektivnosti prilikom odabira kriterija (1SE, 2SE ili 3SE).

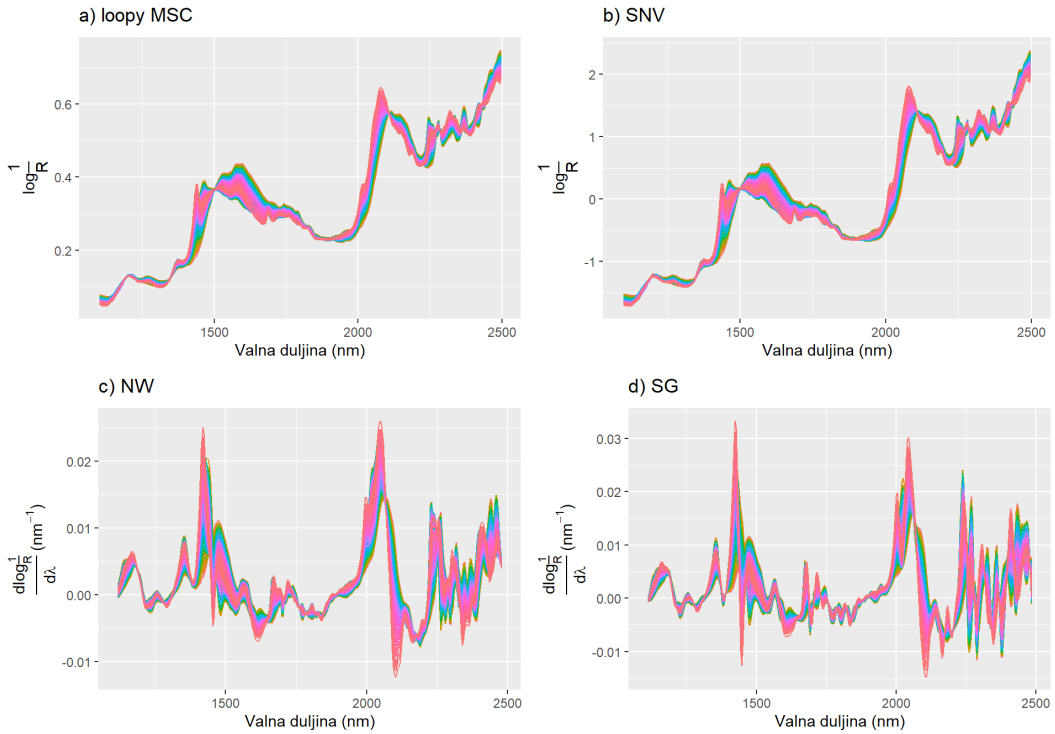
Poglavlje 5

Rezultati

5.1 Predobrada podataka

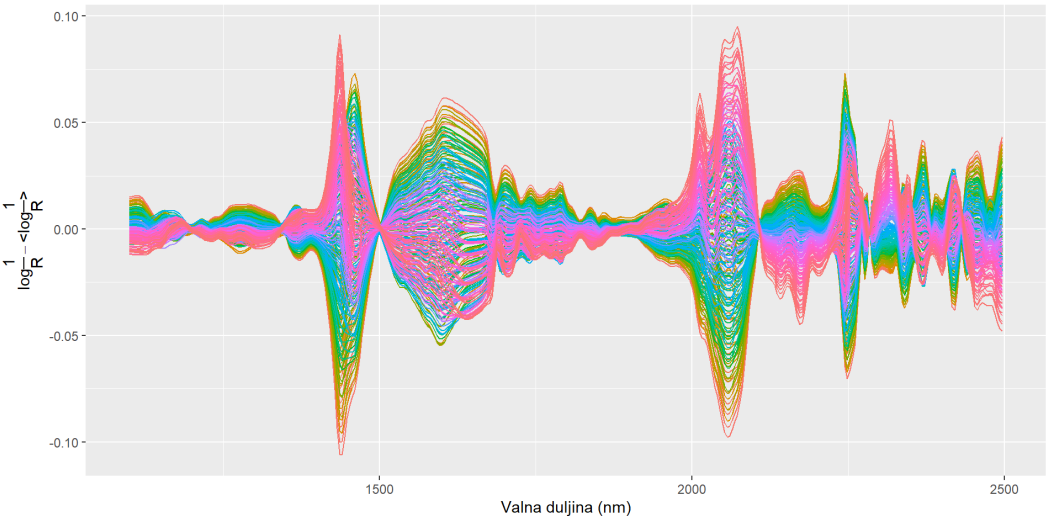
Prethodno primjeni regresijskih metoda, izvorni spektri korigirani su metodama predobrade s ciljem uklanjanja prisutnih nelinearnosti te poboljšavanja modela. Spektri su korigirani MSC, SNV, NW i SG metodama. MSC je korišten kao *loopy* MSC s prosječnim spektrom kao referentnim te je nakon 3 iteracije postignuta konvergencija. SNV je određen standardnom devijacijom i srednjom vrijednošću pojedinog spektra. NW derivacija je određena s 9 točaka ravnanja ($m = 4$) i prvom derivacijom, zbog čega broj prediktorskih varijabli nakon predobrade iznosi 341, dok je SG derivacija određena s 7 točki ravnanja ($m = 3$), polinomom prvog stupnja i prvom derivacijom, zbog čega broj prediktorskih varijabli nakon predobrade iznosi 344.

Rezultati predobrade prikazani su u slici 5.1. Metode koje korigiraju raspršenje, MSC i SNV, pokazuju sličan rezultat, osim razlike u skali $\log(\frac{1}{R})$. Obje metode uklanjaju baznu liniju u spektrima, zbog čega je varijabilnost u $\log(\frac{1}{R})$ pri svakoj valnoj duljini posljedica samo razlike u masenim udjelima komponenti smjesa. Rezultati derivacijskih metoda, NW i SG, izgledom su različiti od onih koje korigiraju raspršenje i izvornih spektara, no ova korekcija uklanja trend koji podatci pokazuju s obzirom na valne duljine. U daljnoj kemometrijskoj analizi korišteni su spektri korigirani *loopy* MSC, s obzirom da je najkorištenija metoda predobrade u NIRSu jer uklanja posljedice raspršenja i baznu liniju, nije osjetljiva na šum, ne smanjuje se

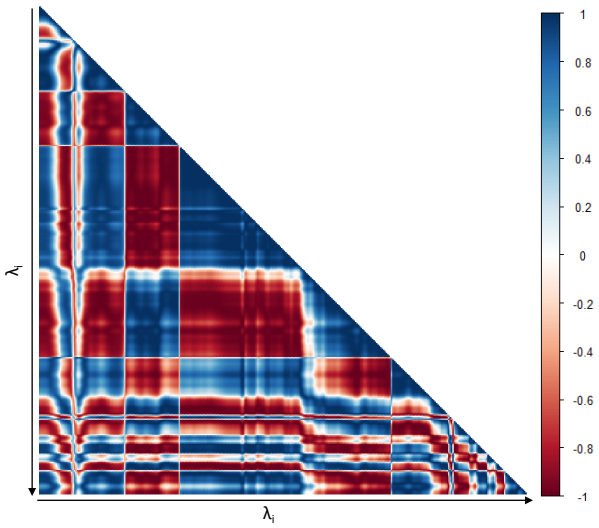


Slika 5.1 Predobrada NIR spektara trokomponenih smjesa fruktoze, glukoze i saharoze: a) *loopy* MSC gdje je prosječni spektar referentni, b) SNV koristeći standardnu devijaciju i srednje vrijednosti spektara, c) Norris-Williams derivacijom uz 9 točki ravnanja i prvu derivaciju, i d) Savitzky-Golay derivacijom uz 7 točki ravnanja, polinom prvog stupnja i prvu derivaciju.

broj varijabli te spektri ostaju najsličniji izvornim. Nakon korekcije, a prije razvoja regresijskih modela, spektri su centrirani čime je uklonjen trend u podatcima s obzirom na valne duljine, što je prikazano u slici 5.2. Spektri pokazuju visoku korelaciju između prediktorskih varijabli (slika 5.3), posljedica čega je kolinearnost varijabli te je nužno provesti selekciju ili redukciju varijabli, ili regularizaciju u fazi treniranja modela.



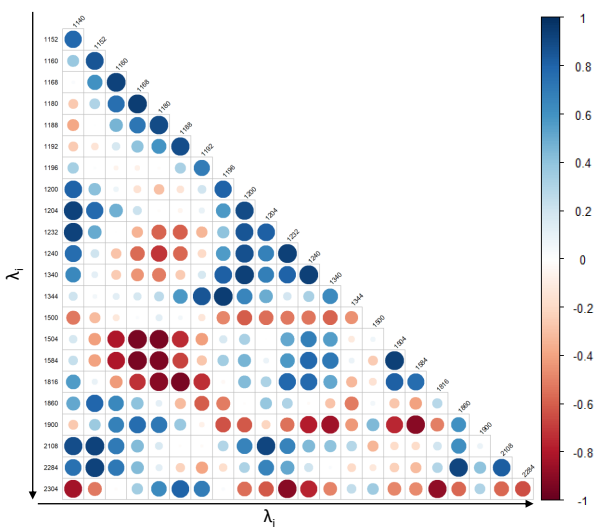
Slika 5.2 Centrirani NIR spektri trokomponenih smjesa fruktoze, glukoze i saharoze. Na spektrima je najprije provedena korekcija *loopy* MSC metodom, a potom centriranje.



Slika 5.3 Dijagram korelacija između svih prediktorskih varijabli, tj. $\log(\frac{1}{R})$ izmjerenih pri različitim valnim duljinama (λ_i). Crvena boja označava negativnu korelaciju, plava pozitivnu, dok bijela označava da ne postoji korelacija između prediktorskih varijabli.

5.2 Multivarijatna linearna regresija

U ovom radu, prvo su razvijeni MLR modeli, za čiji trening je prije potrebno provesti selekciju varijabli s obzirom da su prediktorske varijable izrazito korelirane te je njhov broj veći od broja uzoraka. Problem visoke korelacije riješen je na način da su eliminirane sve prediktorske varijable koje pokazuju korelaciju veću od 95% (slika 5.4). Time su podatci od izvornih 350 svedeni na skup od 23 prediktorske varijable, tj. $\log(\frac{1}{R})$ izmjerenih pri valnim duljinama: 1140 nm, 1152 nm, 1160 nm, 1168 nm, 1180 nm, 1180 nm, 1192 nm, 1196 nm, 1200 nm, 1204 nm, 1232 nm, 1240 nm, 1340 nm, 1344 nm, 1500 nm, 1504 nm, 1584 nm, 1816 nm, 1860 nm, 1900 nm, 2108 nm, 2284 nm i 2304 nm. Na tom skupu su dodatno prevedene selekcije genetičkim algoritmima, *best subset* te *forward stepwise* selekcija, nakon čega su rezultati modela analizirani i uspoređeni.

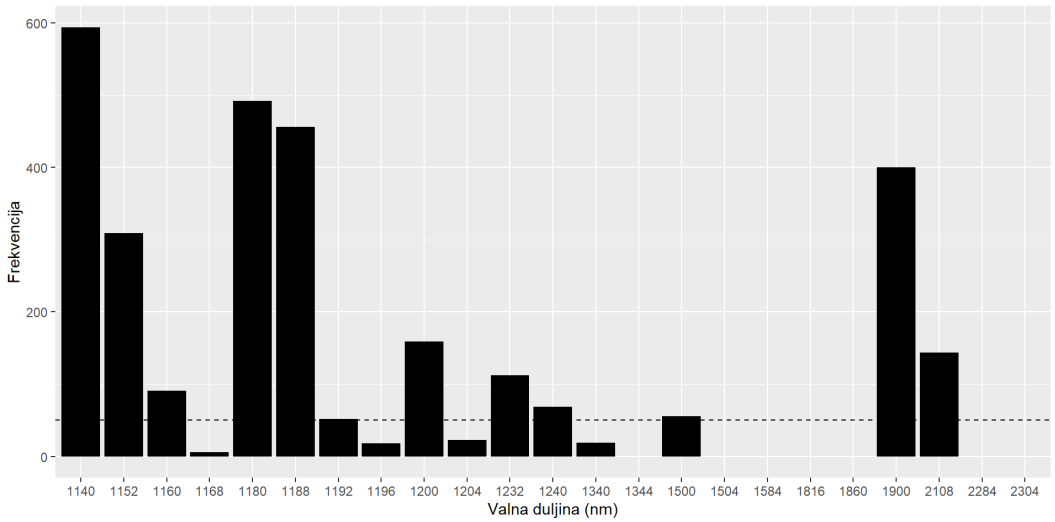


Slika 5.4 Korelacijski graf između prediktorskih varijabli, tj. $\log(\frac{1}{R})$ izmjerenih pri različitim valnim duljinama (λ_i) kojima je apsolutna korelacija manja od 95%. Crvena boja označava negativnu korelaciju, plava pozitivnu, dok bijela označava da ne postoji korelacija između prediktorskih varijabli.

5.2.1 Genetički algoritmi

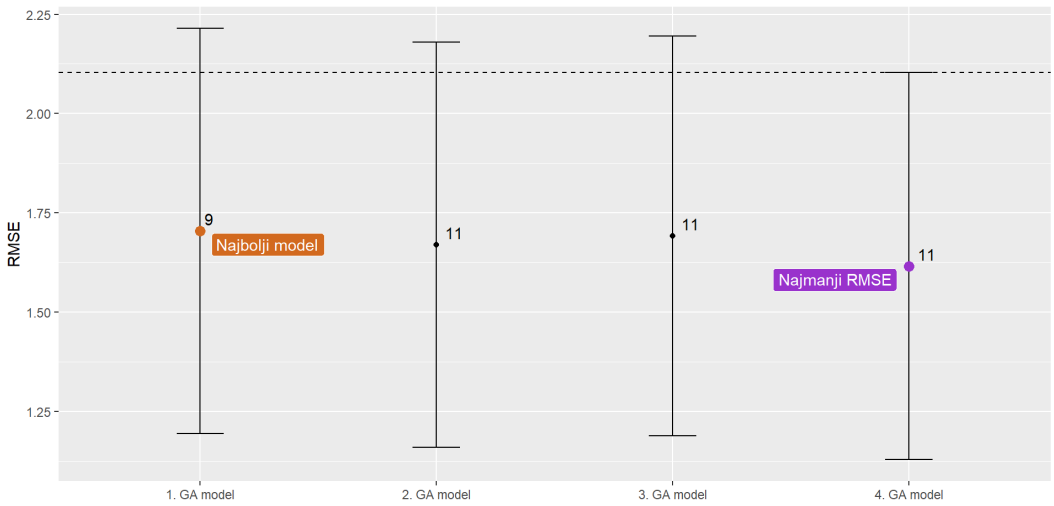
Selekcija varijabli genetičkim algoritmima u ovom radu je podijeljena u dva dijela: u prvom dijelu su eliminirane manje značajne prediktorske varijable, a u drugom se od skupa preostalih značajnijih odabiru najznačajnije [61]. GA su provedeni zasebno na svakoj varijabli odgovora $\mathbf{y_i}$ s obzirom da vrijedi 4.7.

Za eliminaciju prediktorskih varijabli koje najmanje pridonose modelu, kao kriterij brojnosti onih značajnijih, uzet je broj 15. Postupak je ponovljen 200 puta za svaku varijablu odgovora. S obzirom na različitu evoluciju populacije u svakom od 600 modela ($200 \times p = 200 \times 3 = 600$) te još uvijek značajnu korelaciju između odabranih prediktorskih varijabli, GA ne odabire uvijek isti optimalan podskup. Tim postupkom se kao rezultat prvog dijela algoritma dobiva histogram učestalosti odabira pojedine prediktorske varijable u 15-varijatni MLR model, prikazan u slici 5.5. Značajnije prediktorske varijable su one koje algoritam odabire više od 50 puta. Za obrađivane podatke su to one izmjerene pri valnim duljinama: 1140 nm, 1152 nm,



Slika 5.5 Histogram učestalosti odabira pojedine prediktorske varijable u 15-varijatni GA-MLR model, nakon 200 ponavljanja na svakoj varijabli odgovora. Iscrtana linija označava prag značajnosti prilikom odabira varijabli (frekvencija > 50).

1160 nm, 1180 nm, 1188 nm, 1192 nm, 1200 nm, 1232 nm, 1240 nm, 1500 nm, 1900 nm, 2108 nm.

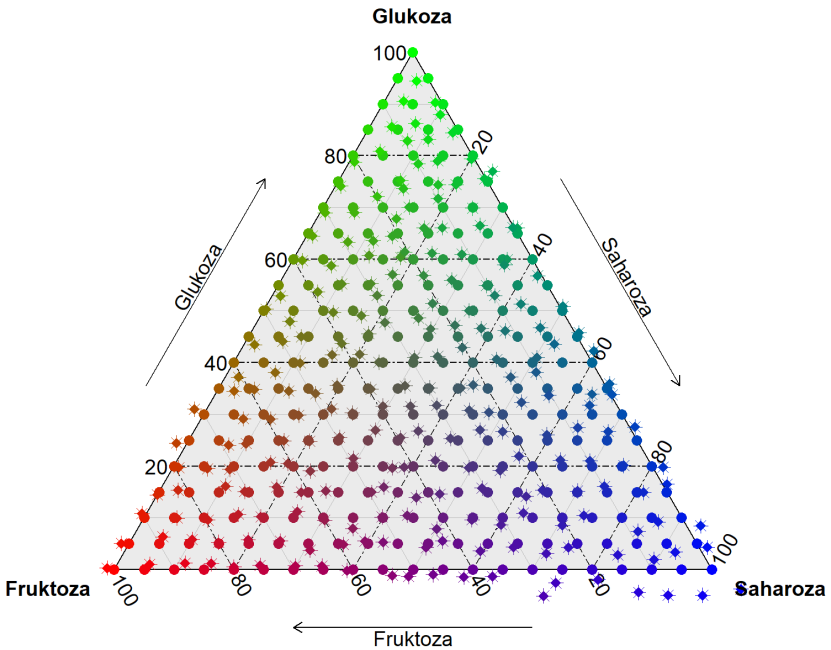


Slika 5.6 Odabir optimalnog MLR modela genetičkim algoritmima uz RMSE kao mjeru točnosti. Provedena je rCV (10-struka s 20 ponavljanja) te je optimalan model odabran pravilom standardne pogreške (2SE), onaj s 9 prediktorskih varijabli.

U drugom dijelu je na odabranih 12 prediktorskih varijabli ponovno proveden GA. Optimalni modeli su odabrani:

- provođenjem GA na svakoj varijabli odgovora zasebno
- eliminacijom prediktorskih varijabli koje najmanje pridonose modelu uz 5, 6, 7 i 8 kao kriterije brojnosti značajnijih
- grupiranjem skupa varijabli odgovora po broju značajnih prediktorskih varijabli (5, 6, 7, 8) u 4 podskupa (modela), na kojima je provedena rCV (10-struka s 20 ponavljanja) uz RMSE kao mjeru točnosti[1]
- optimalan MLR model je odabran rCV uz pravilo standardne pogreške (2SE),

[1]s obzirom da se skupovi odabranih prediktorskih varijabli djelomično preklapaju za 3 varijable odgovora, njihov ukupni broj za pojedini model može biti veći od kriterija (5, 6, 7, 8 za modele 1, 2, 3, 4 u slici 5.6)

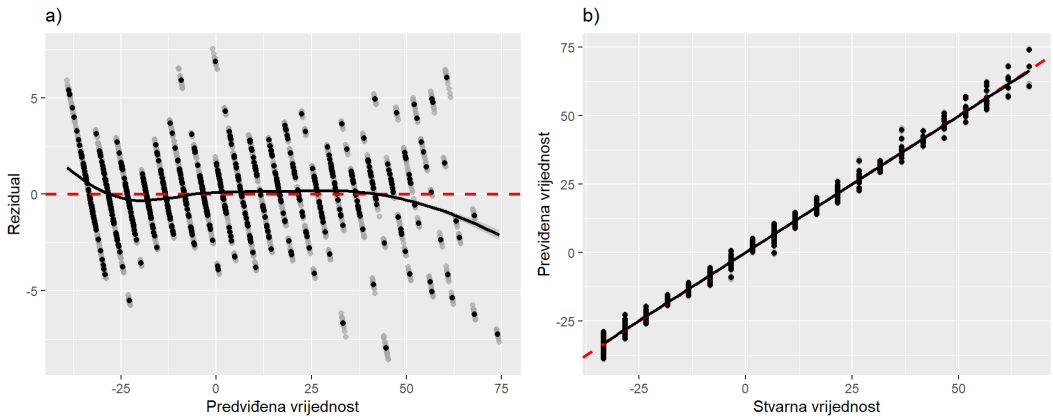


Slika 5.7 Stvarni maseni udjeli pojedinih komponenti u uzorcima trokomponentne smjese saharoze, glukoze i fruktoze (točke) te predviđeni GA-MLR modelom (zvjezdice). Kombinacije boja (crvena, zelena, plava) naglašavaju udjele pojedinih komponenti (fruktoza, glukoza, saharoza).

što je prikazano u slici 5.6

Njegova RMSE iznosi 1.71, za prediktorske varijable izmjerene pri valnim duljinama: 1140 nm, 1152 nm, 1160 nm, 1180 nm, 1188 nm, 1200 nm, 1240 nm, 1900 nm, 2108 nm.

Mjera pogreške konačnog GA-MLR je RMSE. Predviđene vrijednosti masenih udjela prikazane su u slici 5.7 te *in-sample* pogreška iznosi 1.69. *Out-of-sample* pogreška iznosi 1.70 ± 0.25 (određena 10-strukom rCV s 20 ponavljanja). U slici 5.8 prikazani su dijagnostički grafovi reziduala. Pod a) možemo uočiti da distribucija reziduala s obzirom na predviđene vrijednosti odstupa od pravca $y = 0$ i pokazuje uzorak u rezidualima, što ukazuje da primijenjeni model ne odgovara podatcima te



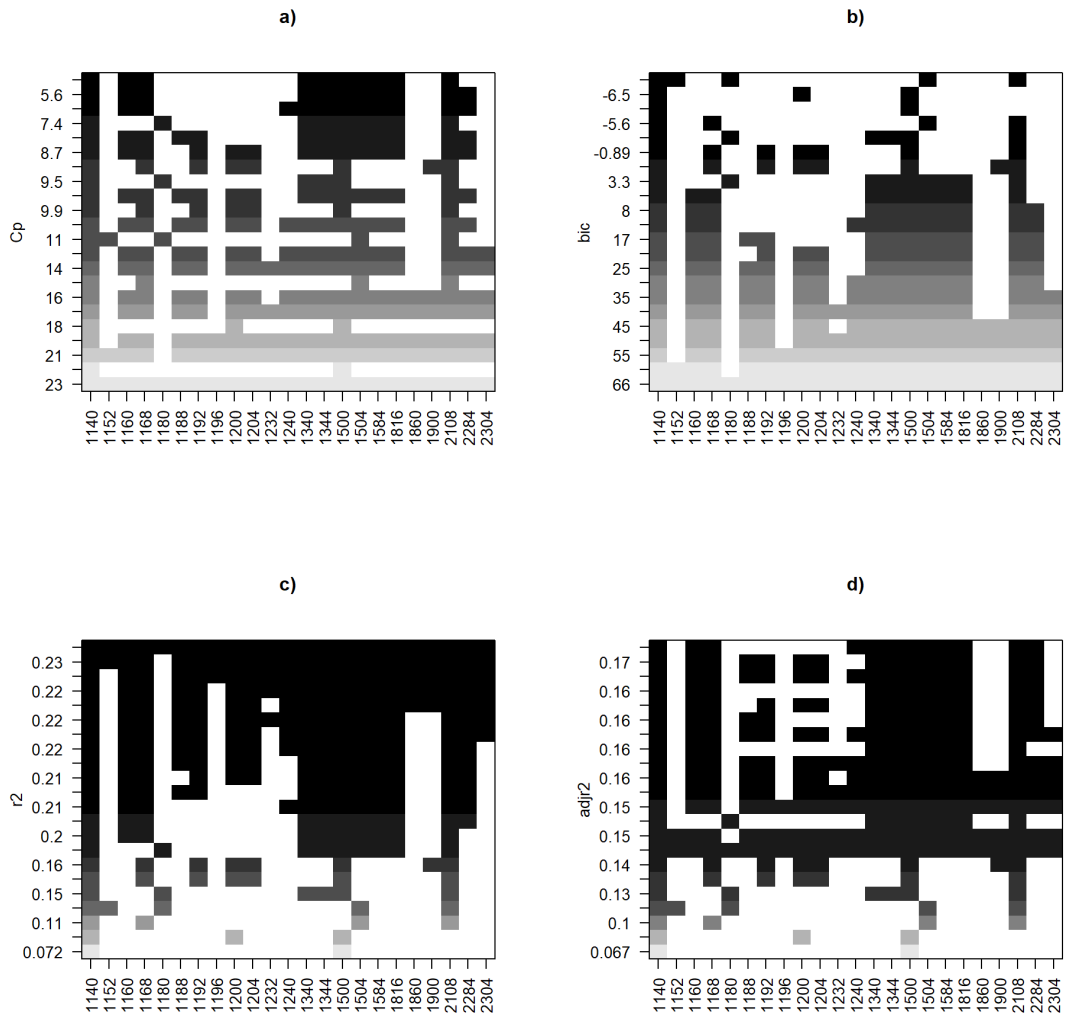
Slika 5.8 Dijagnostika reziduala izračunatih pomoću rCV (10-struka s 20 ponavljanja) za GA-MLR model. a) Ovisnost reziduala o predviđenoj vrijednosti gdje crne točke označavaju prosječne vrijednosti iz ponavljanja rCV, dok sive označavaju sve vrijednosti ponavljanja, crna puna linija njihovu *loess* krivulju[2], a crvena iscrtana pravac $y = 0$. b) Ovisnost predviđenih o stvarnim masenim udjelima gdje je crna puna linija regresijski pravac, a crvena iscrtana pravac $y = x$.

se može poboljšati. U b) možemo uočiti da postoji linearni odnos između stvarnih i predviđenih vrijednosti masenih udjela u smjesi pri čemu R^2 iznosi 0.99 ($p < 0.05$) te je visoko statistički značajan i ukazuje na činjenicu da model ima stvarnu i visoku prediktivnu moć. Normalnost distribucije reziduala testirana je Shapiro-Wilk testom koji pokazuje $p < 0.05$ ($W = 0.979$) i prihvaćamo hipotezu da reziduali nisu normalno distribuirani s vjerojatnošću većom od 95% (alternativna hipoteza).

5.2.2 *Best subset* selekcija

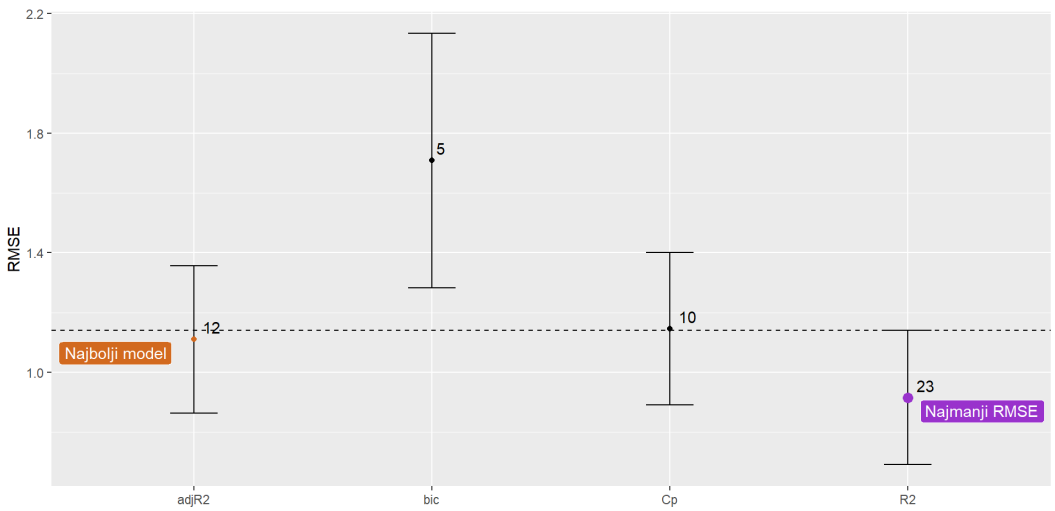
Druga korištena metoda selekcije je *best subset* provedena na podskupu od 23 prediktorske varijable koje imaju apsolutnu vrijednost korelacijskog koeficijenta manju od 95%. Tom metodom odabrani su optimalni modeli s obzirom na 4 kriterija - C_p, BIC, R^2 i *prilagođenom* R^2. U slici 5.9 prikazan je odabir prediktorskih varijabli

[2] *loess* (eng. *local regression*) je generalizacija pomičnog prosjeka i polinomne regresije razvijena za izravnavanje raspršenih grafova. Za više informacija vidi [26].



Slika 5.9 Evaluacija modela koji uključuju *best subset* selekciju testiranih pomoću 4 kriterija - a) C_p, b) BIC, c) R^2, i d) *prilagođenom* R^2. U grafovima je prikazana vrijednost pojedinog kriterija (naglašena gradijentom boje) za svaku odabranu prediktorsku varijablu. Najbolji modeli nalaze se u samom vrhu grafa.

prema pojedinom kriteriju te njihov optimalni broj iznosi: 10 prema C_p, 5 prema BIC, 12 prema *prilagođenom* R^2 te 23 prema R^2. Na svakom od tih modela prove-

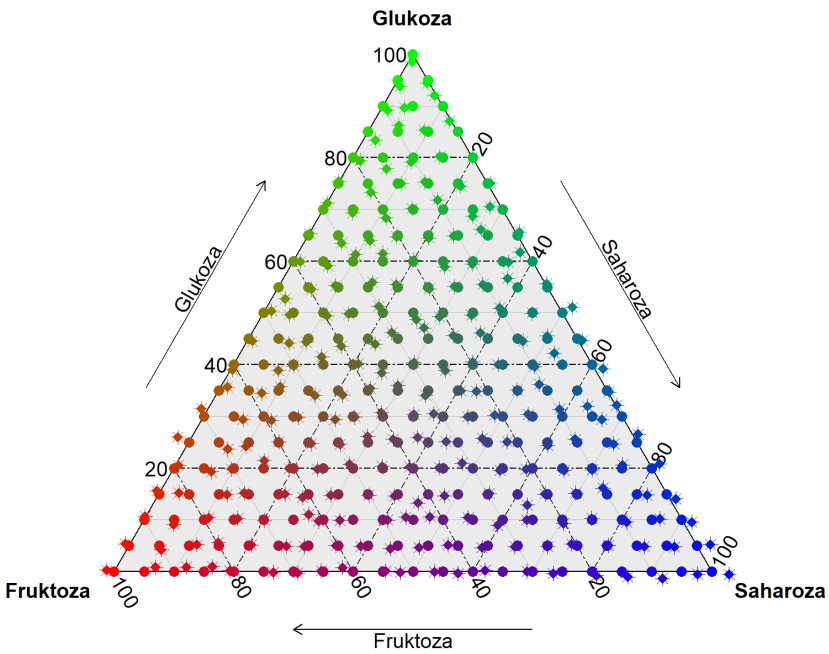


Slika 5.10 Odabir optimalnog modela *best subset* selekcijom uz RMSE kao mjeru točnosti. Provedena je rCV (10-struka s 20 ponavljanja) te je optimalni model odabran pravilom standardne pogreške (2SE), onaj s 12 prediktorskih varijabli (prema *prilagođenom* R^2 kriteriju).

dena je rCV (10-struka s 20 ponavljanja) te je koristeći RMSE kao mjeru pogreške i pravilo standardne pogreške (2SE), određen optimalni MLR model (slika 5.10). To je onaj prema *prilagođenom* R^2 kriteriju, s 12 prediktorskih varijabli, izmjerenih pri valnim duljinama: 1140 nm, 1160 nm, 1168 nm, 1340 nm, 1344 nm, 1500 nm, 1504 nm, 1584 nm, 1816 nm i 2108 nm.

Mjera pogreške konačnog BS-MLR modela je RMSE. Predviđene vrijednosti masenih udjela prikazane su u slici 5.11 te *in-sample* pogreška iznosi 1.06. *Out-of-sample* pogreška iznosi 1.12 ± 0.12 (određena 10-strukom rCV s 20 ponavljanja). U slici 5.12 prikazani su dijagnostički grafovi reziduala. Pod a) možemo uočiti da distribucija reziduala s obzirom na predviđene vrijednosti odstupa od pravca $y = 0$ i pokazuje uzorak u rezidualima, što ukazuje da primijenjeni model ne odgovara podatcima te se može poboljšati. U b) možemo uočiti da postoji linearni odnos između stvarnih i predviđenih vrijednosti masenih udjela u smjesi pri čemu R^2 iznosi 0.99 ($p < 0.05$) te je visoko statistički značajan i ukazuje na činjenicu da model ima stvarnu i visoku prediktivnu moć. Normalnost distribucije reziduala testirana je Shapiro-Wilk testom

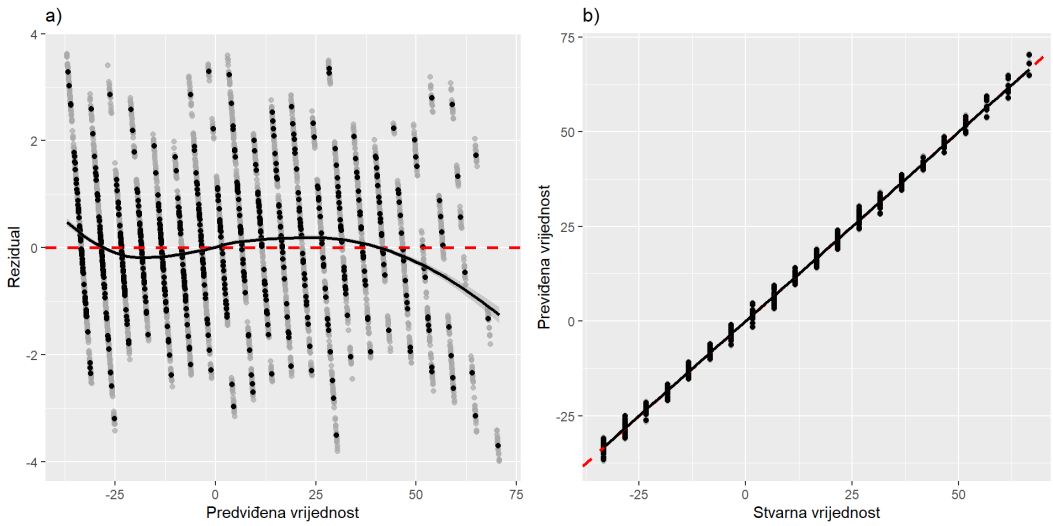
koji pokazuje $p = 0.11$ ($W = 0.996$) i s 95%-tnom vjerojatnošću možemo tvrditi da su reziduali normalno distribuirani (nul-hipoteza).



Slika 5.11 Stvarni maseni udjeli pojedinih komponenti u uzorcima trokomponentne smjese saharoze, glukoze i fruktoze (točke) te predviđeni BS-MLR modelom (zvjezdice). Kombinacije boja (crvena, zelena, plava) naglašavaju udjele pojedinih komponenti (fruktoza, glukoza, saharoza).

5.2.3 *Forward-stepwise* selekcija

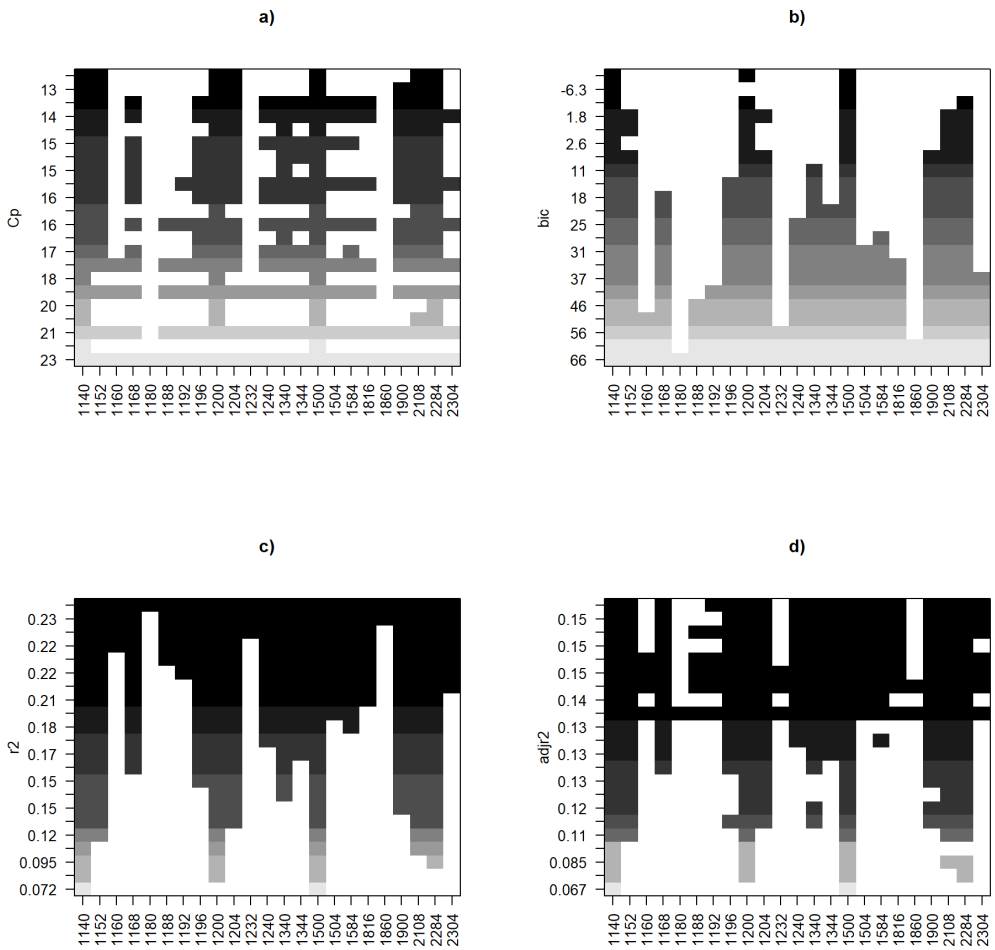
Treća korištena metoda selekcije je *forward stepwise* provedena na podskupu od 23 prediktorske varijable koje imaju apsolutnu vrijednost korelacijskog koeficijenta manju od 95%. Tom metodom su, kao i kod *best subset* selekcije, odabrani optimalni modeli s obzirom na 4 kriterija - C_p, BIC, R^2 i *prilagođeni* R^2. U slici 5.13 prikazan je odabir prediktorskih varijabli prema pojedinom kriteriju te njihov optimalni broj iznosi: 7 prema C_p, 3 prema BIC, 18 prema *prilagođenom* R^2 te 23 prema R^2. Na



Slika 5.12 Dijagnostika reziduala izračunatih pomoću rCV (10-struka s 20 ponavljanja) za BS-MLR model. a) Ovisnost reziduala o predviđenoj vrijednosti gdje crne točke označavaju prosječne vrijednosti iz ponavljanja rCV, dok sive označavaju sve vrijednosti ponavljanja, crna puna linija njihovu *loess* krivulju (vidi fusnotu 2, str. 67), a crvena iscrtana pravac $y = 0$. b) Ovisnost predviđenih o stvarnim masenim udjelima gdje je crna puna linija regresijski pravac, a crvena iscrtana pravac $y = x$.

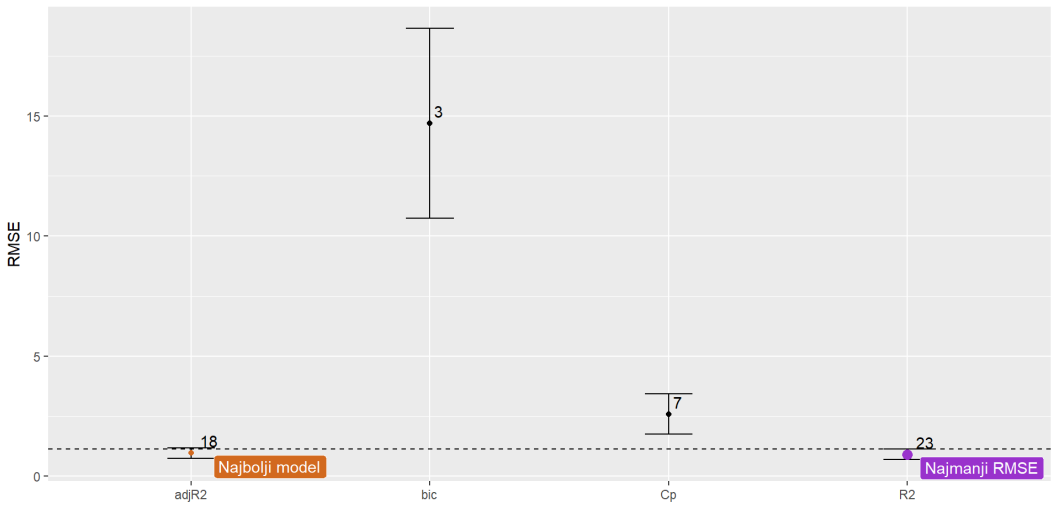
svakom od tih modela provedena je rCV (10-struka s 20 ponavljanja) te je koristeći RMSE kao mjeru pogreške i pravilo standardne pogreške (2SE) određen optimalni MLR model (slika 5.14). To je onaj prema *prilagođenom* R^2 kriteriju, s 12 prediktorskih varijabli izmjerenih pri valnim duljinama: 1140 nm, 1152 nm, 1168 nm, 1192 nm, 1196 nm, 1200 nm, 1204 nm, 1240 nm, 1340 nm, 1344 nm, 1500 nm, 1504 nm, 1584 nm, 1816 nm, 1900 nm, 2108 nm, 2284 nm, 2304 nm.

Mjera pogreške konačnog FW-MLR modela je RMSE. Predviđene vrijednosti masenih udjela prikazane su u slici 5.15 te *in-sample* pogreška iznosi 0.90. *Out-of-sample* pogreška iznosi 0.97 ± 0.11 (određena 10-strukom rCV s 20 ponavljanja). U slici 5.16 prikazani su dijagnostički grafovi reziduala. Pod a) možemo uočiti da distribucija reziduala s obzirom na predviđene vrijednosti odstupa od pravca $y = 0$ i pokazuje uzorak u rezidualima, što ukazuje da primijenjeni model ne odgovara podatcima te se može poboljšati. U b) možemo uočiti da postoji linearni odnos



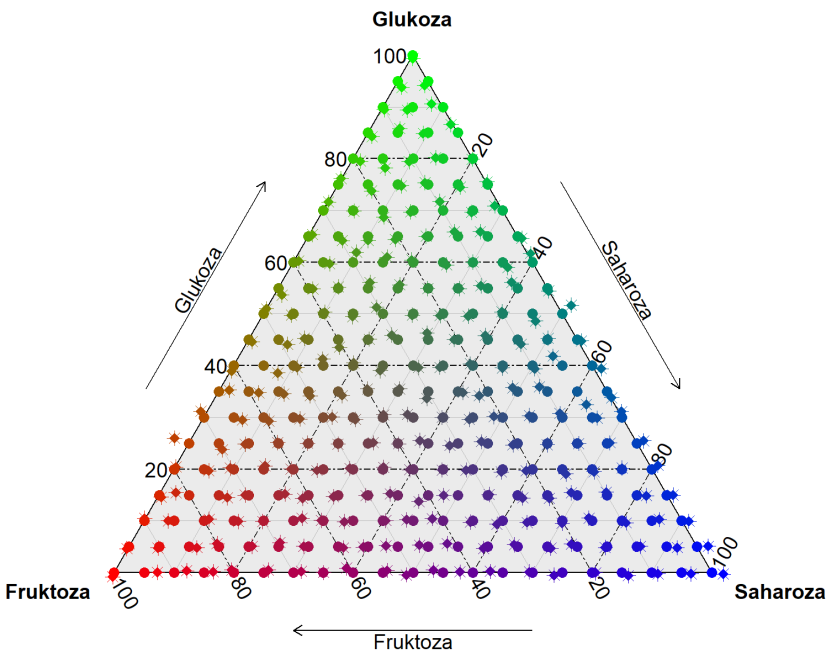
Slika 5.13 Evaluacija modela koji uključuju *forward stepwise* selekciju testiranih pomoću 4 kriterija - a) C_p, b) BIC, c) R^2, i d) *prilagođenom* R^2. U grafovima je prikazana vrijednost pojedinog kriterija (naglašena gradijentom boje) za svaku odabranu prediktorsku varijablu. Najbolji modeli nalaze se u samom vrhu grafa.

između stvarnih i predviđenih vrijednosti masenih udjela u smjesi pri čemu R^2 iznosi 0.99 ($p < 0.05$) te je visoko statistički značajan i ukazuje na činjenicu da model ima stvarnu i visoku prediktivnu moć. Normalnost distribucije reziduala testirana je Shapiro-Wilk testom koji pokazuje $p = 0.22$ ($W = 0.997$) i s 95%-tnom vjerojatnošću

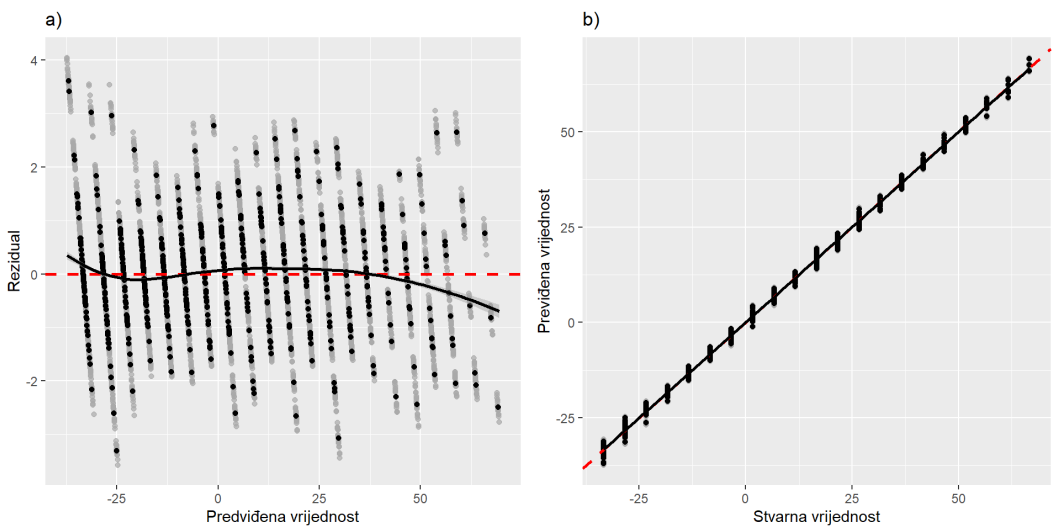


Slika 5.14 Odabir optimalnog modela *forwad stepwise* selekcijom uz RMSE kao mjeru točnosti. Provedena je rCV (10-struka s 20 ponavljanja) te je optimalan model odabran pravilom standardne pogreške (2SE), onaj s 18 prediktorskih varijabli (prema *prilagođenom* R^2 kriteriju).

možemo tvrditi da su reziduali normalno distribuirani (nul-hipoteza).



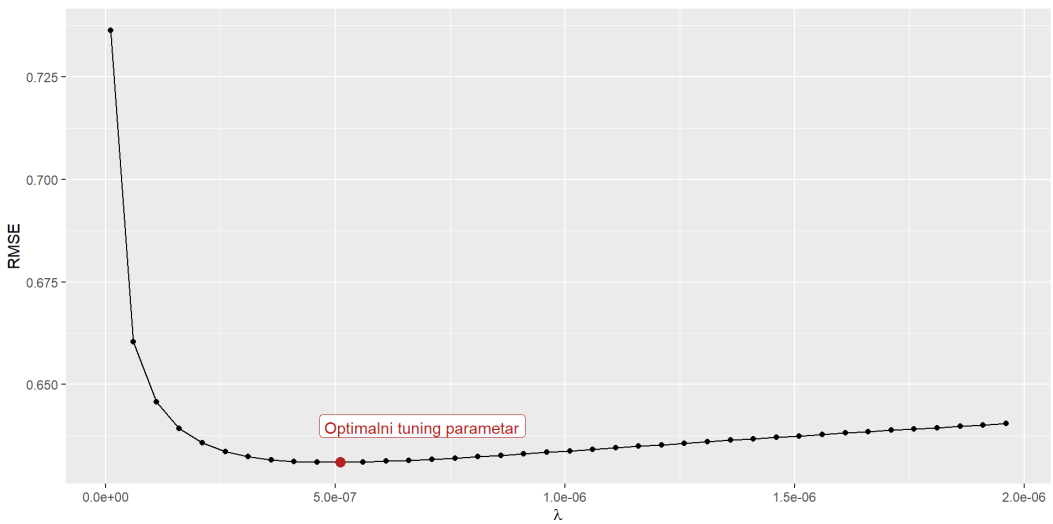
Slika 5.15 Stvarni maseni udjeli pojedinih komponenti u uzorcima trokomponentne smjese saharoze, glukoze i fruktoze (točke) te predviđeni FW-MLR modelom (zvjezdice). Kombinacije boja (crvena, zelena, plava) naglašavaju udjele pojedinih komponenti (fruktoza, glukoza, saharoza).



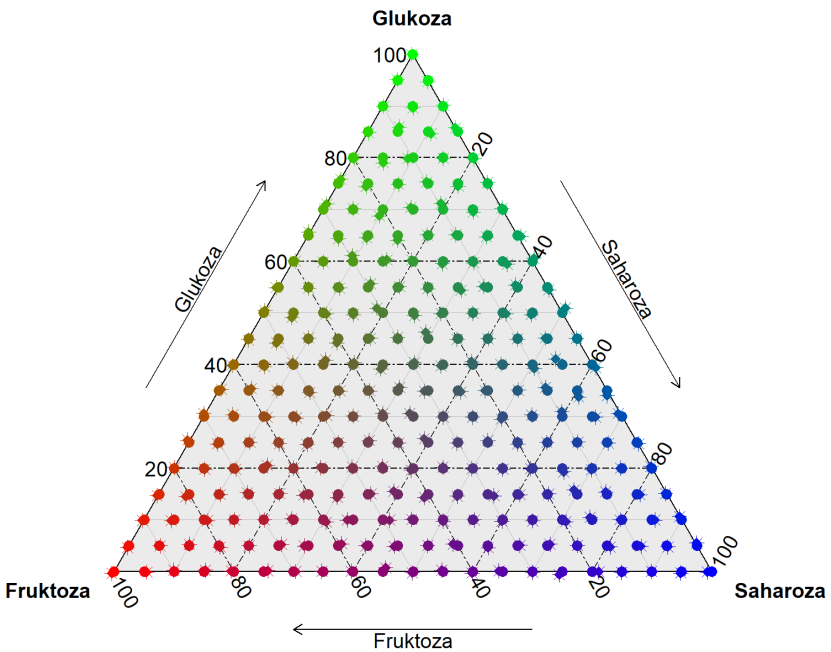
Slika 5.16 Dijagnostika reziduala izračunatih pomoću rCV (10-struka s 20 ponavljanja) za FW-MLR model. a) Ovisnost reziduala o predviđenoj vrijednosti gdje crne točke označavaju prosječne vrijednosti iz ponavljanja rCV, dok sive označavaju sve vrijednosti ponavljanja, crna puna linija njihovu *loess* krivulju (vidi fusnotu 2, str. 67), a crvena iscrtana pravac $y = 0$. b) Ovisnost predviđenih o stvarnim masenim udjelima gdje je crna puna linija regresijski pravac, a crvena iscrtana pravac $y = x$.

5.3 Ridge regresija

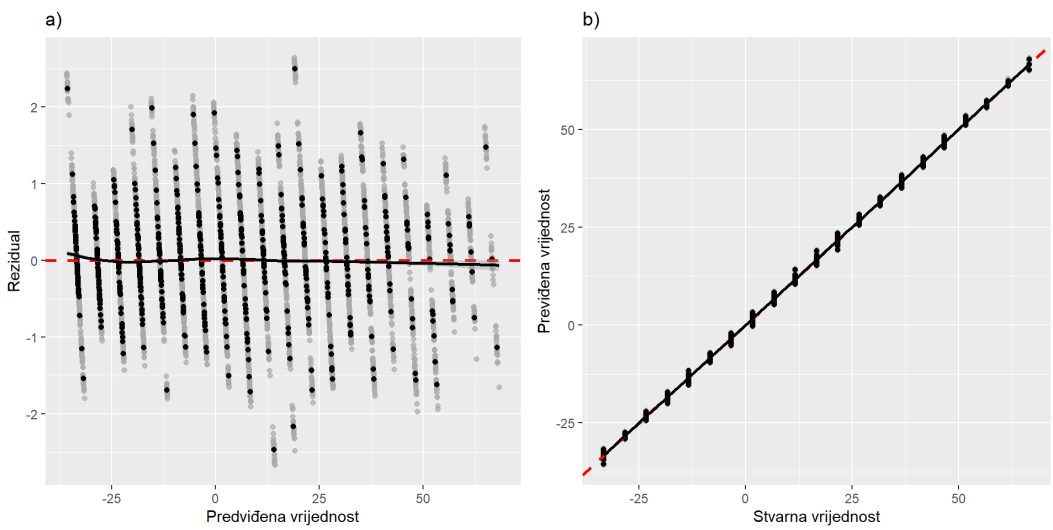
Prilikom treninga modela ridge regresije potrebno je odrediti *tuning* parametar λ, za što je provedena optimizacija pomoću LOO CV. Rezultat, odnosno ovisnost RMSE o λ prikazan je u slici 5.17 te optimalna vrijednost λ iznosi 5×10^{-7}. Predviđene vrijednosti masenih udjela prikazane su u slici 5.18 te *in-sample* pogreška iznosi 0.43. Dodatno je za izračunavanje *out-of-sample* pogreške provedena rCV (10-struka s 20 ponavljanja) s obzirom da LOO CV može dovesti do *overfitting*-a i preoptimistične vrijednosti pogreške te iznosi 0.63 ± 0.08. U slici 5.19 prikazani su dijagnostički grafovi reziduala. Pod a) možemo uočiti da distribucija reziduala s obzirom na predviđene vrijednosti ne odstupa od pravca $y = 0$ (simetrično je distribuirana oko pravca), teži vrijednostima oko 0 i ne pokazuje nikakve značajne uzorke, što ukazuje da primijenjeni model odgovara podatcima. U b) možemo uočiti da postoji linearni odnos između stvarnih i predviđenih vrijednosti masenih udjela u smjesi pri čemu R^2 iznosi 0.99 ($p < 0.05$) te je visoko statistički značajan i time ukazuje na činjenicu da model ima stvarnu i visoku prediktivnu moć.



Slika 5.17 RMSE kao funkcija *tuning* parametra λ, određena LOO CV. Crvena točka označava minimalni RMSE koji odgovara optimalnom modelu ridge regresije ($\lambda = 5 \times 10^{-7}$).



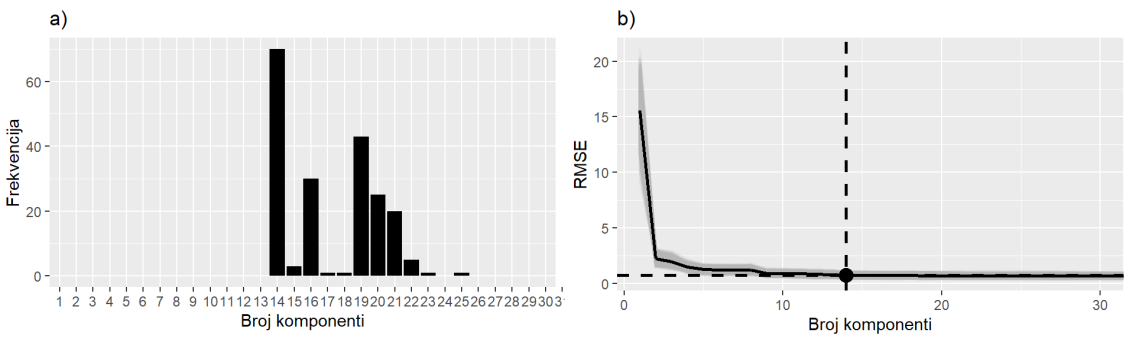
Slika 5.18 Stvarni maseni udjeli pojedinih komponenti u uzorcima trokomponentne smjese saharoze, glukoze i fruktoze (točke) te predviđeni modelom ridge regresije s $\lambda = 5 \times 10^{-7}$ (zvjezdice). Kombinacije boja (crvena, zelena, plava) naglašavaju udjele pojedinih komponenti (fruktoza, glukoza, saharoza).



Slika 5.19 Dijagnostika reziduala izračunatih pomoću rCV (10-struka s 20 ponavljanja) za model ridge regresije s $\lambda = 5 \times 10^{-7}$. a) Ovisnost reziduala o predviđenoj vrijednosti gdje crne točke označavaju prosječne vrijednosti iz ponavljanja rCV, dok sive označavaju sve vrijednosti ponavljanja, crna puna linija njihovu *loess* krivulju (vidi fusnotu 2, str. 67), a crvena iscrtana pravac $y = 0$. b) Ovisnost predviđenih o stvarnim masenim udjelima gdje je crna puna linija regresijski pravac, a crvena iscrtana pravac $y = x$.

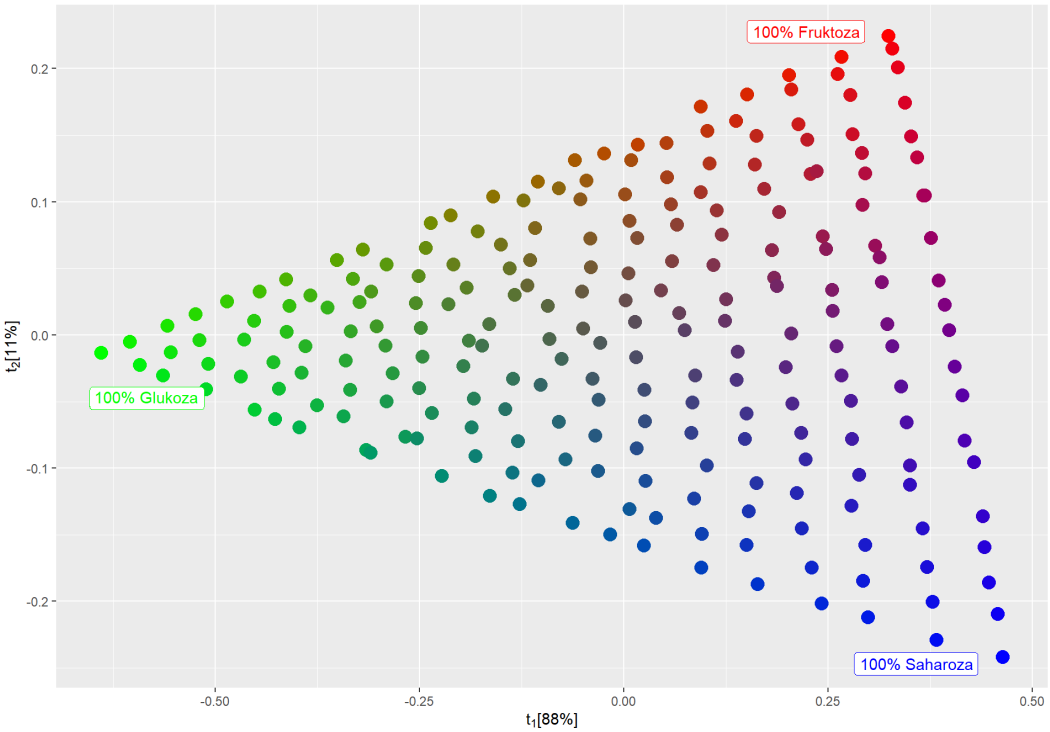
5.4 Regresija glavnih komponenti

S obzirom da je kod PCR potrebno i odrediti optimalan broj glavnih komponenti, potrebno je primijeniti rdCV kao metodu ponovnog uzorkovanja. Korištena je rdCV s 10-strukom unutarnjom petljom (određivanje optimalnog broja glavnih komponenti uz pravilo standardne pogreške (2SE)) i 10-strukom vanjskom petljom (određivanje *out-of-sample* pogreške), uz 20 ponavljanja. Testirani su modeli do max. 50 glavnih komponenti te su rezultati prikazani u slici 5.20, gdje pod a) možemo uočiti da je u unutarnjoj petlji rdCV model s 14 glavnih komponenti najčešće određen kao optimalan. Pod b) prikazana je ovisnost RMSE modela o broju glavnih komponenti s padajućim trendom te nakon 14 glavnih komponenti RMSE prestaje ovisiti o njihovom broju.



Slika 5.20 Određivanje optimalnog broja glavnih komponenti za PCR pomoću rdCV (10-struka unutarnja petlja, 10-struka vanjska petlja s 20 ponavljanja). a) Histogram odabranog broja komponenti unutarnje petlje. b) RMSE kao funkcija broja komponenti unutarnje petlje. Sivo područje označava preklopljene linije za ponavljanja unutarnje petlje, dok crna linija predstavlja srednju vrijednost svih ponavljanja. Optimalan broj glavnih komponenti odabran je pravilom standardne pogreške (2SE) i označen točkom (14 glavnih komponenti).

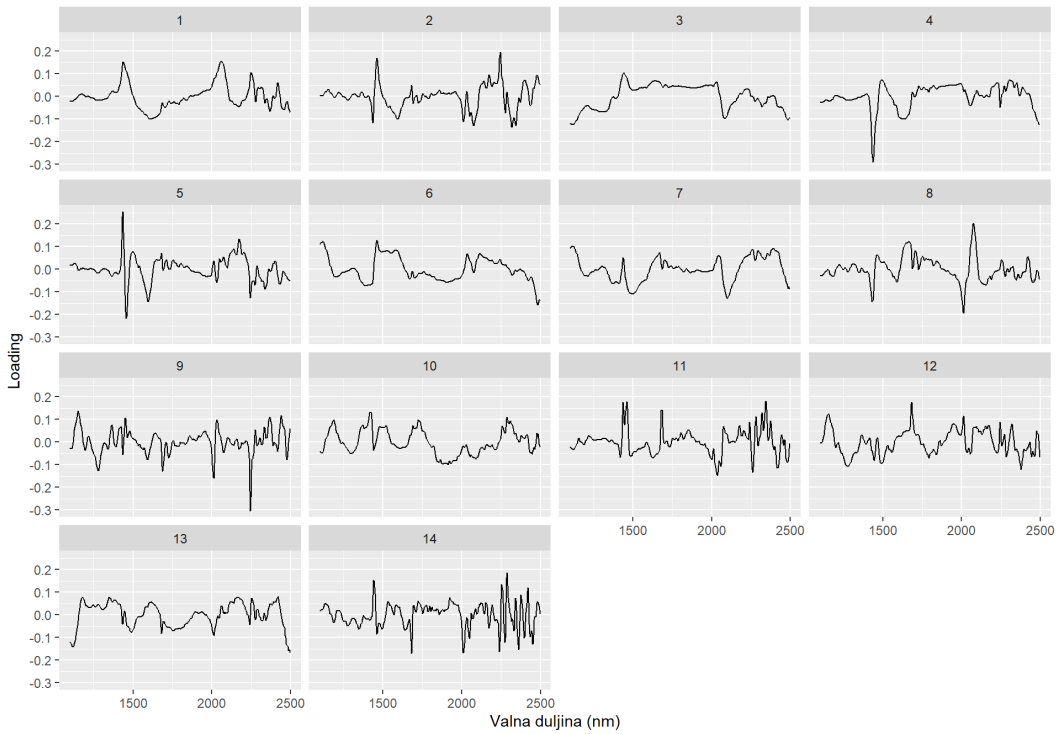
Slika 5.21 prikazuje prva dva vektora skorova. Možemo uočiti razdvajanje uzoraka u tri smjera, nalik na njihov dizajn s obzirom na masene udjele fruktoze, glukoze i saharoze u smjesi. Letimičnim pregledom ovog grafa mogli bismo, čak i bez znanja o dizajnu uzoraka, pretpostaviti da se radi o trokomponentnim smjesama. Slika



Slika 5.21 Skorovi $\mathbf{t_1}$ i $\mathbf{t_2}$ kao rezultat PCR analize NIR spektara trokomponentne smjese glukoze, fruktoze i saharoze.

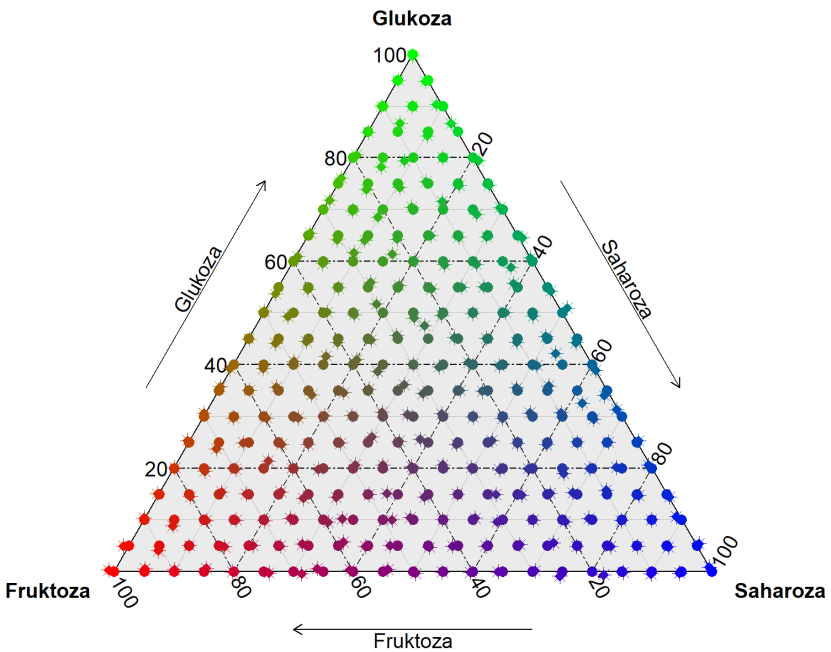
5.22 prikazuje opterećenja za svaku od 14 glavnih komponenti uključenih u PCR model. Najvažnija su prva dva vektora opterećenja (prve dvije glavne komponente) s obzirom da objašnjavaju većinu ($> 99\%$) varijance u podatcima, dok preostali imaju manji doprinos. Dva najznačajnija vektora opterećenja pokazuju da su najvažnije prediktorske varijable u rasponu valnih duljina 1400-1700 nm te 2000-2500 nm.

Predviđene vrijednosti masenih udjela prikazane su u slici 5.23 te *in-sample* pogreška iznosi 0.69. *Out-of-sample* pogreška iznosi 0.72 ± 0.09. U slici 5.24 prikazani su dijagnostički grafovi reziduala. Pod a) možemo uočiti da distribucija reziduala s obzirom na predviđene vrijednosti ne odstupa od pravca $y = 0$ (simetrično je distribuirana oko pravca), teži vrijednostima oko 0 i ne pokazuje nikakve značajne uzorke, što ukazuje da primijenjeni model odgovara podatcima. U b) možemo uočiti da postoji linearni odnos između stvarnih i predviđenih vrijednosti masenih udjela u smjesi

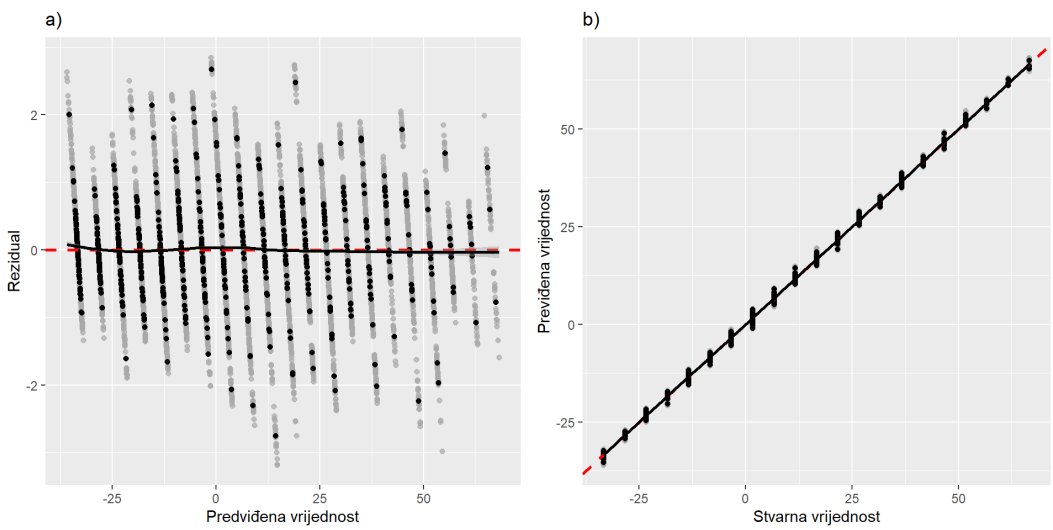


Slika 5.22 PCR opterećenja. Ovisnost PCR opterećenja o valnoj duljini za svaku od 14 glavnih komponenti.

pri čemu R^2 iznosi 0.99 ($p < 0.05$) te je visoko statistički značajan i time ukazuje na činjenicu da model ima stvarnu i visoku prediktivnu moć.



Slika 5.23 Stvarni maseni udjeli pojedinih komponenti u uzorcima trokomponentne smjese saharoze, glukoze i fruktoze (točke) te predviđeni PCR modelom s 14 glavnih komponenti (zvjezdice). Kombinacije boja (crvena, zelena, plava) naglašavaju udjele pojedinih komponenti (fruktoza, glukoza, saharoza).



Slika 5.24 Dijagnostika reziduala izračunatih pomoću vanjske petlje rdCV (10-struka s 20 ponavljanja) za PCR model. a) Ovisnost reziduala o predviđenoj vrijednosti gdje crne točke označavaju prosječne vrijednosti iz ponavljanja rdCV, dok sive označavaju sve vrijednosti ponavljanja, crna puna linija njihovu *loess* krivulju (vidi fusnotu 2, str. 67), a crvena iscrtana pravac $y = 0$. b) Ovisnost predviđenih o stvarnim masenim udjelima gdje je crna puna linija regresijski pravac, a crvena iscrtana pravac $y = x$.

5.5 Regresija parcijalnih najmanjih kvadrata

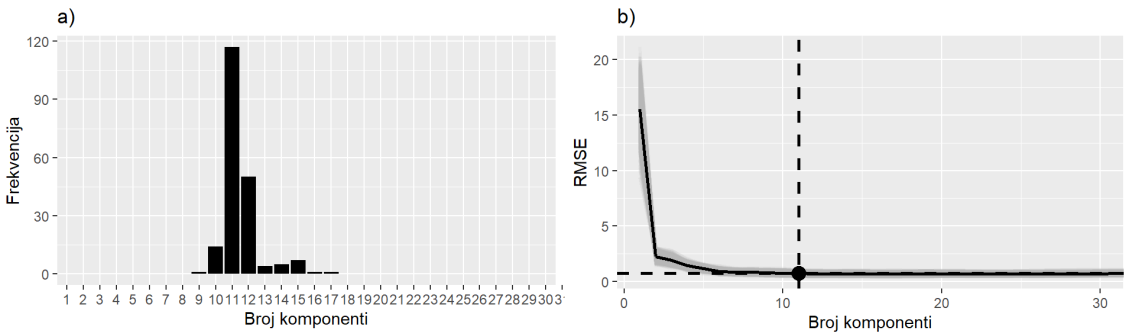
Prilikom PLS najprije su uspoređeni različiti algortimi - *kernel*, *wide kernel*, NIPALS i SIMPLS, s obzirom na *out-of-sample* pogrešku, optimalan broj PLS komponenti i brzinu. Korištena je rdCV s 10-strukom unutarnjom petljom (određivanje optimalnog broja PLS komponenti uz pravilo standardne pogreške (2SE)) i 10-strukom vanjskom petljom (određivanje *out-of-sample* pogreške), uz 20 ponavljanja. Rezultati usporedbe prikazani su u tablici 5.1. Svi algoritmi pokazuju slične rezultate s obzirom na *out-of-sample* pogrešku (0.749) i optimalan broj PLS komponenti (11). No, algoritmi se razlikuju s obzirom na brzinu gdje je *kernel* algoritam najbrži s brzinom od 0.025 s za trening modela, zbog čega je korišten u daljnjoj analizi.

Tablica 5.1 Usporedba PLS algoritama - *kernel*, *wide kernel*, NIPALS i SIMPLS. Algoritmi su uspoređeni koristeći rdCV, s obzirom na *out-of-sample* pogrešku (RMSE), optimalan broj PLS komponenti (a) i brzinu izraženu kao komputacijsko vrijeme t.

	kernel	***wide kernel***	**NIPALS**	**SIMPLS**
RMSE	0.749	0.749	0.749	0.751
a	11	11	11	11
t/s	0.025	0.954	0.211	0.036

Nakon odabira algoritma, sljedeći korak u razvoju PLS modela je određivanje optimalnog broja komponenti. Kao i kod PCR, korištena je rdCV s 10-strukom unutarnjom petljom i 10-strukom vanjskom petljom, uz 20 ponavljanja. Testirani su modeli do max. 50 PLS komponenti te su rezultati prikazani u slici 5.25 gdje pod a) možemo uočiti da je u unutarnjoj petlji rdCV model s 11 PLS komponenti najčešće određen kao optimalan. U slici b) prikazana je ovisnost RMSE modela o broju PLS komponenti s padajućim trendom te nakon 11 RMSE prestaje ovisiti o njihovom broju.

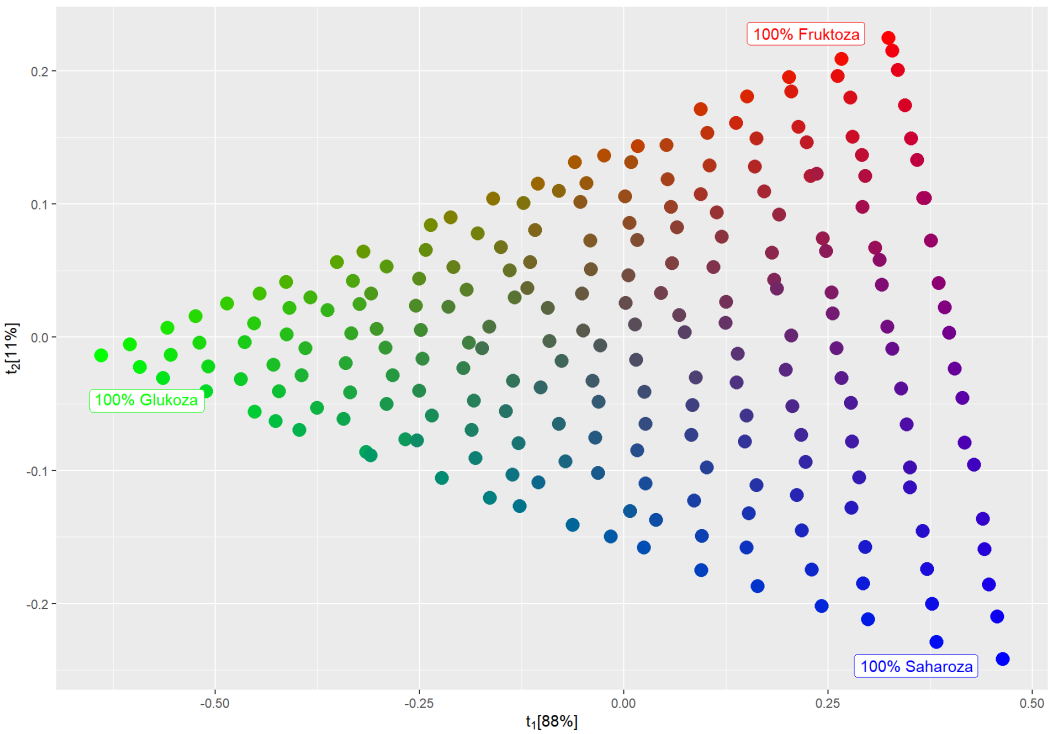
Slika 5.26 prikazuje međusobnu ovisnost prvih dvaju vektora skorova. Podjednako kao i kod PCR, možemo uočiti razdvajanje uzoraka u tri smjera, nalik na njihov dizajn s obzirom na masene udjele. Letimičnim pregledom i ovog grafa mogli bismo, čak i bez znanja o dizajnu uzoraka, pretpostaviti da se radi o trokomponentnim smje-



Slika 5.25 Određivanje optimalnog broja PLS komponenti pomoću rdCV (10-struka unutarnja petlja, 10-struka vanjska petlja s 20 ponavljanja). a) Histogram odabranog broja komponenti unutarnje petlje. b) RMSE kao funkcija broja komponenti unutarnje petlje. Sivo područje označava preklopljene linije za ponavljanja unutarnje petlje, dok crna linija predstavlja srednju vrijednost svih ponavljanja. Optimalan broj PLS komponenti odabran je pravilom standardne pogreške (2SE) i označen točkom (11 PLS komponenti).

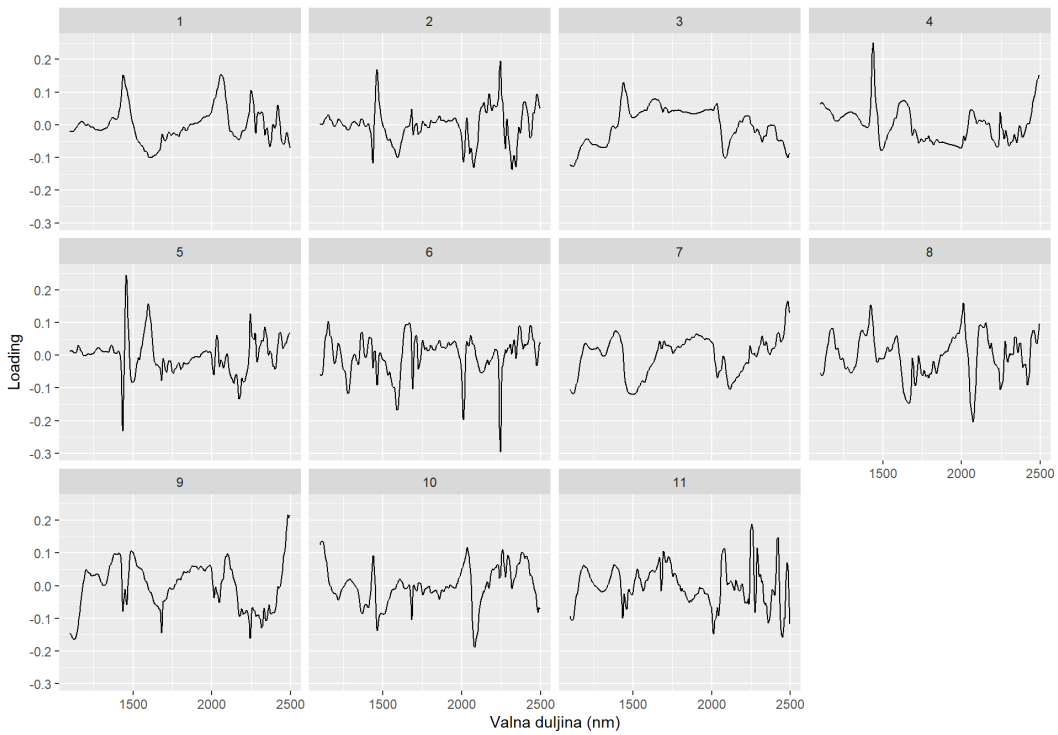
sama. Slika 5.27 prikazuje opterećenja za svaku od 11 PLS komponenti uključenih u model. I u ovom slučaju, najvažnija su prva dva vektora opterećenja (prve dvije PLS komponente) s obzirom da objašnjavajau većinu ($> 99\%$) varijance u podatcima. Kao i kod PCR, dvije najznačajnije komponente pokazuju da su najvažnije prediktorske varijable u rasponu valnih duljina 1400-1700 nm te 2000-2500 nm. U slici 5.28 prikazan je **u-t** graf za sve odabrane komponente. Možemo uočiti da su prva dva para **u** i **t** (koji objašnjavaju $> 99\%$ varijance u podatcima) u linearnom odnosu s izrazito visokom korelacijom. Daljnje komponente pokazuju sve manju korelaciju, što tumačimo time da one objašnjavaju mali udjel nelinearnosti koja je prisutna u podatcima, no ipak pridonose modelu.

Predviđene vrijednosti masenih udjela prikazane su u slici 5.29 te *in-sample* pogreška iznosi 0.66. *Out-of-sample* pogreška iznosi 0.72 ± 0.10. U slici 5.30 prikazani su dijagnostički grafovi reziduala. Pod a) možemo uočiti da distribucija reziduala s obzirom na predviđene vrijednosti ne odstupa od pravca $y = 0$ (simetrično je distribuirana oko pravca), teži vrijednostima oko 0 i ne pokazuje nikakve značajne uzorke, što ukazuje da primijenjeni model odgovara podatcima. U b) možemo uočiti da pos-

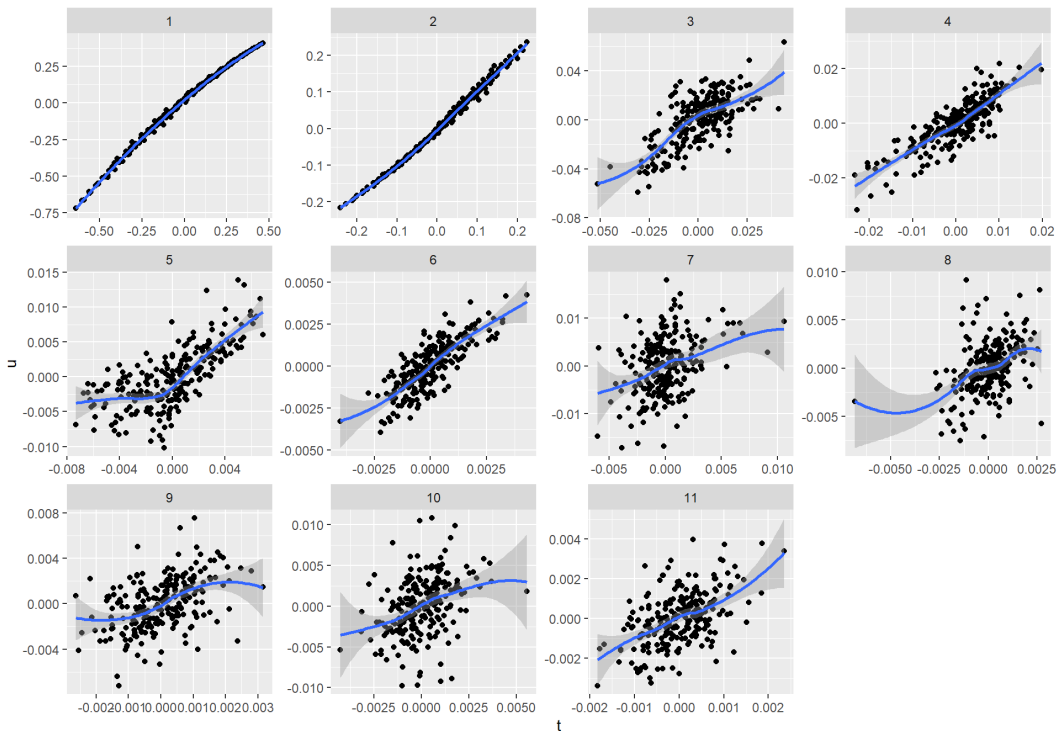


Slika 5.26 Skorovi $\mathbf{t_1}$ i $\mathbf{t_2}$ kao rezultat PLS analize NIR spektara trokomponentne smjese glukoze, fruktoze i saharoze.

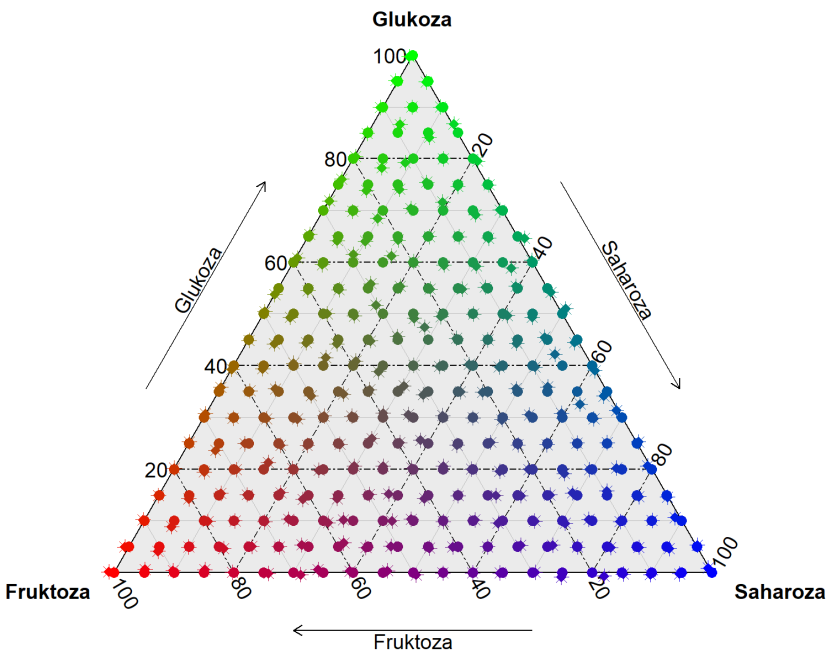
toji linearni odnos između stvarnih i predviđenih vrijednosti masenih udjela u smjesi pri čemu R^2 iznosi 0.99 ($p < 0.05$) te je visoko statistički značajan i time ukazuje na činjenicu da model ima stvarnu i visoku prediktivnu moć.



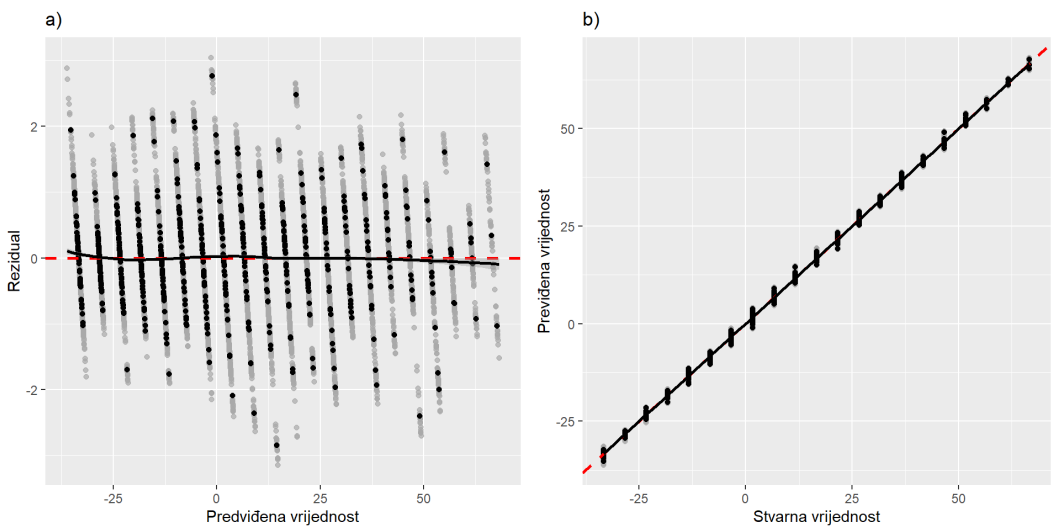
Slika 5.27 PLS opterećenja. Ovisnost PLS opterećenja o valnoj duljini za svaku od 11 PLS komponenti.



Slika 5.28 Međusobna ovisnost PLS skorova **u** i **t** za 11 komponenti uključenih u model.



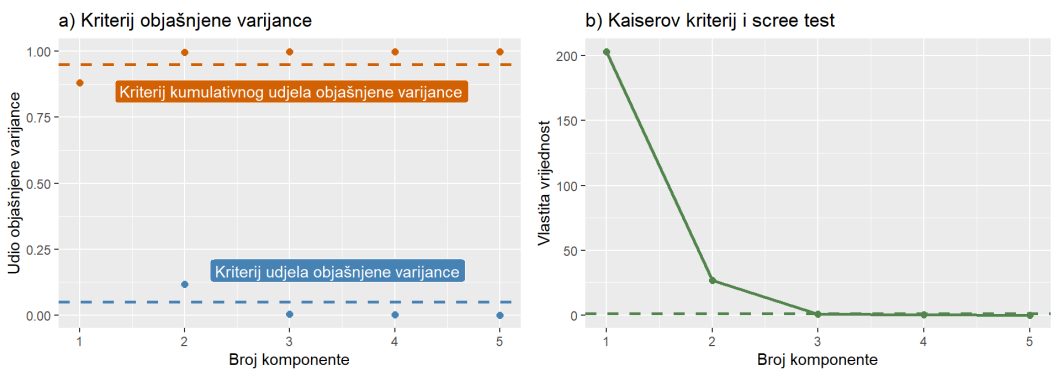
Slika 5.29 Stvarni maseni udjeli pojedinih komponenti u uzorcima trokomponentne smjese saharoze, glukoze i fruktoze (točke) te predviđeni PLS modelom s 11 glavnih komponenti (zvjezdice). Kombinacije boja (crvena, zelena, plava) naglašavaju udjele pojedinih komponenti (fruktoza, glukoza, saharoza).



Slika 5.30 Dijagnostika reziduala izračunatih pomoću vanjske petlje rdCV (10-struka s 20 ponavljanja) za PLS model. a) Ovisnost reziduala o predviđenoj vrijednosti gdje crne točke označavaju prosječne vrijednosti iz ponavljanja rdCV, dok sive označavaju sve vrijednosti ponavljanja, crna puna linija njihovu *loess* krivulju (vidi fusnotu 2, str. 67), a crvena iscrtana pravac $y = 0$. b) Ovisnost predviđenih o stvarnim masenim udjelima gdje je crna puna linija regresijski pravac, a crvena iscrtana pravac $y = x$.

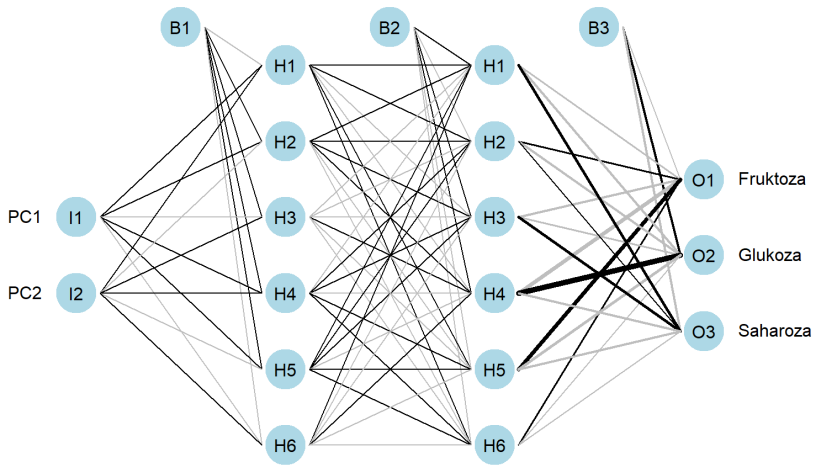
5.6 Umjetne neuronske mreže

Prilikom razvoja modela neuronskih mreža korištena je nasumična podjela podataka na trening skup od 80% podataka i testni skup od 20% koji služi za određivanje *out-of-sample* pogreške. Prethodno treningu neuronske mreže, provedena je PCA na trening skupu podataka. Odabir glavnih komponenti prikazan je u slici 5.31 te po svim kriterijima zadržavamo samo prve dvije, čime je skup od 350 prediktorskih varijabli reduciran na samo dvije glavne komponente kojima je objašnjeno više od 99% varijance u podatcima. Utrenirana neuronska mreža sadrži dva skrivena sloja (model dubokog učenja) od po 7 varijabli u svakom sloju, uključujući i jedinicu pomaka, prikazana u slici 5.32. Kao aktivacijska funkcija u svakom sloju odabrana je sigmoidna te su za optimizaciju funkcije gubitka korišteni algoritmi opadajućeg gradijenta i *resilient backpropagation*.



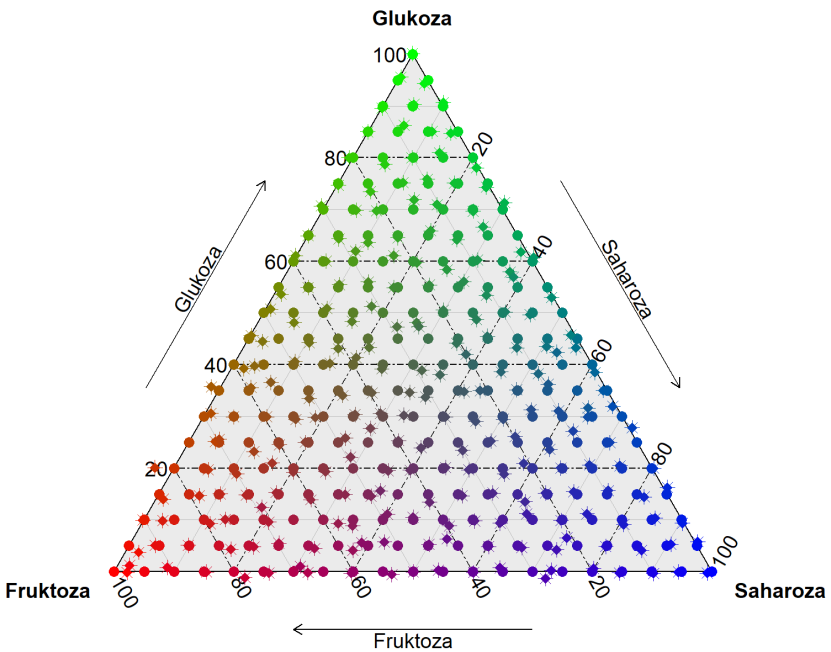
Slika 5.31 Odabir broja glavnih komponenti za razvoj modela neuronskih mreža. a) Odabir kriterijem objašnjene varijance. Prikazana je ovisnost udjela objašnjene varijance o broju glavnih komponenti. Plavom bojom je označen odabir po udjelu kojeg objašnjava pojedina komponenta te iscrtana linija označava kriterij 0.05, dok je narančasto označen odabir po kumulativnom udjelu kojeg objašnjavaju glavne komponente te iscrtana linija označava kriterij 0.95. b) Odabir Kaiserovim kriterijem i *scree* testom. Iscrtana linija označava kriterij vlastita vrijednost = 1.

Predviđene vrijednosti masenih udjela prikazane su u slici 5.33 te *in-sample* pogreška iznosi 1.09. *Out-of-sample* pogreška iznosi 1.40. U slici 5.34 prikazani su

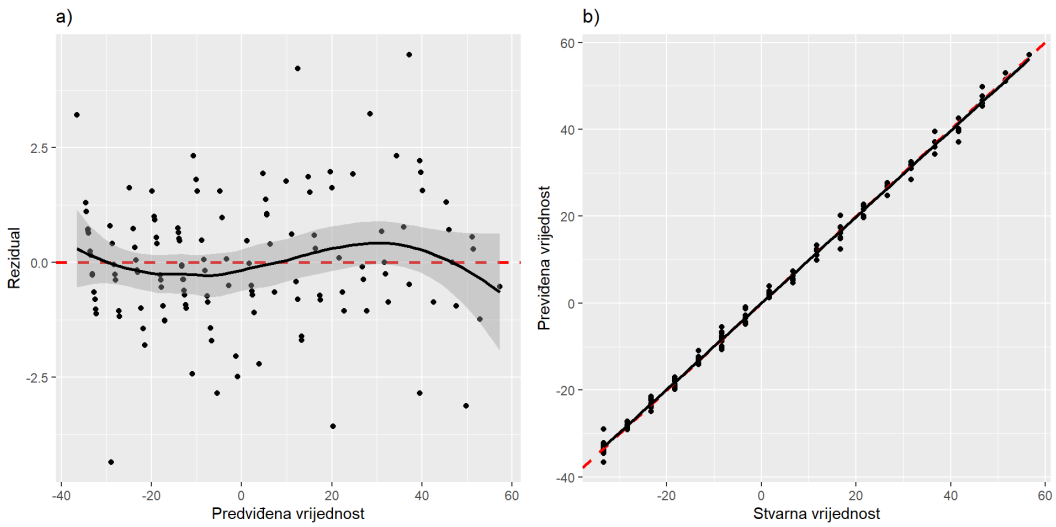


Slika 5.32 Dizajn neuronske mreže koja sadrži 2 ulazne varijable, 2 skrivena sloja od po 7 varijabli (uključujući jedinicu pomaka) i sigmoidnu aktivacijsku funkciju. Debljina linija označava aktivaciju (vrijednost težine) pojedine varijable u sloju.

dijagnostički grafovi reziduala za neuronske mreže. Pod a) možemo uočiti da distribucija reziduala s obzirom na predviđene vrijednosti odstupa od pravca $y = 0$ te pokazuje uzorak u rezidualima, što ukazuje da primijenjeni model ne odgovara podatcima te se može poboljšati. U b) možemo uočiti da postoji linearni odnos između stvarnih i predviđenih vrijednosti masenih udjela u smjesi pri čemu R^2 iznosi 0.99 ($p < 0.05$) te je visoko statistički značajan i time ukazuje na činjenicu da model ima stvarnu i visoku prediktivnu moć.



Slika 5.33 Stvarni maseni udjeli pojedinih komponenti u uzorcima trokomponentne smjese saharoze, glukoze i fruktoze (točke) te predviđeni modelom neuronskih mreža s 2 skrivena sloja od po 7 skrivenih varijabli i sigmoidnu aktivacijsku funkciju (zvjezdice). Kombinacije boja (crvena, zelena, plava) naglašavaju udjele pojedinih komponenti (fruktoza, glukoza, saharoza).



Slika 5.34 Dijagnostika reziduala za model neuronskih mreža izračunatih koristeći 80%-20% razdvajanje podataka. a) Ovisnost reziduala o predviđenoj vrijednosti gdje crna puna linija označava njihovu *loess* krivulju (vidi fusnotu 2, str. 67), a crvena iscrtana pravac $y = 0$. b) Ovisnost predviđenih o stvarnim masenim udjelima gdje je crna puna linija regresijski pravac, a crvena iscrtana pravac $y = x$.

Poglavlje 6

Rasprava

6.1 Usporedba i dijagnostika regresijskih metoda

Svi prediktivni modeli i njihove performance sažeti su u tablici 6.1. Iako su svi pokazali značajnu prediktivnu moć, kao najtočnijim se pokazao model dobiven ridge regresijiom, što se može prepoznati iz najmanje *out-of-sample* RMSE, koja iznosi 0.63. PCR i PLS modeli daju neznatno manje točna rješenja uz RMSE = 0.72 za obje metode. Uz to što je ridge regresija najtočnija, također je vrlo brza, jednostavna i skalabilna te je zbog toga najbolja za primjenu u industrijskim procesima gdje je važno odrediti masene udjele komponenti u smjesi iz NIRS podataka. Nedostatak ridge regresije je smanjena interpretabilnost zbog koje nije moguće odrediti povezanost prediktorske varijable i varijable odgovora, tj. izvoditi asocijacijske zaključke. PCR i PLS modeli imaju veću interpretabilnost jer omogućavaju eksploraciju podataka grafovima skorova i opterećenja, iako glavne komponente nemaju stvarno značenje. PLS i PCR su, kao i ridge regresija, relativno brze, jednostavne i skalabilne metode. Između njih, PLS model je jednostavniji jer zahtijeva manji broj komponenti za trening optimalnog modela ($a = 11$) od PCR ($a = 14$), što je posljedica korištenja informacije iz obju $\mathbf{X}$ i $\mathbf{Y}$. MLR metode sa selekcijom varijabli pokazuju nešto manju točnost (RMSE = 1.70 za genetičke algortime, RMSE = 1.12 za *best subset* selekciju i RMSE = 0.97 za *forward stepwise* selekciju), a sam proces selekcije varijabli usporava algoritam za trening metode. Time model čini kompliciranijim, iako je sama

Tablica 6.1 Rezultati svih utreniranih regresijskih modela. Hiperparametara: λ - faktor smanjenja ridge regresije, a - broj glavnih komponenti, L - broj skrivenih slojeva, r_1 i r_2 - broj varijabli u prvom, odn. drugom skrivenom sloju.

Metoda	***In-sample* RMSE**	***Out-of-sample* RMSE**	**Broj prediktora**	**Hiperparametri**
GA-MLR	1.688	1.702	9	/
BS-MLR	1.064	1.115	12	/
FW-MLR	0.897	0.971	18	/
RR	0.433	0.631	350	$\lambda = 5 \times 10^{-7}$
PCR	0.686	0.722	350	$a = 14$
PLS	0.657	0.719	350	$a = 11$
ANN	1.094	1.404	350 (a=2)	L = 2 $r_1 = 6$, $r_2 = 6$

regresija direktna (neiterativna), a time najbrža i najjednostavnija. Unatoč tome, MLR je izrazito interpretabilna te omogućuje uspostavu algebarskog odnosa između $\log(\frac{1}{R})$ i masenog udjela komponente u smjesi, neovisno o valnoj duljini i izvođenje asocijacijskih zaključaka. Za razliku od linearnih metoda, ANN su pokazale manju točnost od ridge regresije, PCR i PLS. Zbog većeg broja parametara i kompliciranijeg dizajna, one predstavljaju sporiju, kompliciraniju i manje interpretabilnu metodu od linearnih.

U radu su uspoređeni različiti PLS algoritmi (*kernel*, *wide kernel*, NIPALS i SIMPLS). Pokazano je da *kernel*, *wide kernel* i NIPALS algoritmi daju jednako rješenje ($\mathbf{B}_{\mathrm{PLS}}$) uz jednaki RMSE = 0.749, dok SIMPLS rezultira neznatno različitm rješenjem uz RMSE = 0.751. Taj rezultat je očekivan s obzirom na različit koncept

SIMPLS algoritma za multivarijatni odgovor ($\mathbf{Y}$) [62]. Svi algoritmi razlikuju se s obzirom na brzinu te se kao najbrži ističu *kernel* (0.025 s za trening) i SIMPLS (0.036 s za trening) algoritmi. Zbog iterativnog pristupa određivanja glavnih komponenti NIPALS algoritam (0.211 s za trening) je sporiji, što su pokazale i prijašnje slične studije [62], dok je *wide kernel* algoritam (0.954 s za trening) sporiji od svih ostalih algoritama. Razlog tome je što je *wide kernel* specifično brži za skupove podataka s vrlo malim brojem uzoraka (10-15) [53, 54].

Kod primjene MLR modela, pokazano je da je selekcija varijabli pomoću *best subset* ili *forward stepwise* metoda korisna za postizanje veće točnosti ukoliko je broj varijabli veći od broja uzoraka. Naspram njih, genetički algoritmi su se pokazali neuspješnom alternativom s manjom točnošću. Sve metode imaju prediktivnu moć, no unatoč tome sva tri MLR modela mogu biti poboljšana, što se može prepoznati iz slika 5.8, 5.12 i 5.16 pod a). Iz svega navedenog možemo zaključiti da je selekcijom varijabli izgubljen važan dio informacije u podatcima za razliku od ostalih metoda.

Primjena neuronskih mreža rezultira s RMSE = 1.40, iz čega zaključujemo da one u ovom slučaju predstavljaju najmanje točnu metodu. S obzirom da je linearna regresija u stvari podskup neuronskih mreža (neuronska mreža bez skrivenih slojeva i aktivacijske funkcije), razlog veće RMSE vrijednosti leži u tome što se radi o vrlo malom broju uzoraka (231), a velikom broju parametara regresije (81), zbog čega se događa *overfitting*. U prilog tome ide i značajno manja vrijednost *in-sample* pogreške ANN modela [26]. *Overfitting* predstavlja problem visoke varijance u predviđanju, a moguće ga je nadvladati primjenom regularizacije u modelu, redukcijom varijabli, izbacivanjem varijabli između slojeva ili snimanjem dodatnih uzoraka.

U ovom radu prezentirana je uspješna kalibracija za određivanje masenih udjela u uzorcima iz NIR spektara. Postavlja se pitanje mogućnosti primjene kalibracijskih modela na podatke izmjerene drugim NIRS uređajima zbog različite točnosti i preciznosti uređaja, tj. šuma koji se uz signal pojavljuje u spektru. Smith i sur. bave se tim pitanjem te se pokazalo da, iako postoje razlike u spektrima, što utječe na rezultate, postoje algortimi kojima se dodatno korigiraju spektri, poput Kennard-Stone algoritma [63]. Stoga bi za razvoj modela primjenjivih na više NIRS uređaja najprije trebalo provesti navedenu korekciju spektara te potom razviti kalibracijski model koji bi imao jednake perfomance predviđanja za sve uređaje.

6.2 Eksplorativna analiza podataka

Proučavanje rezultata dobivenih PCR i PLS metodama nudi mogućnost izvođenja zaključaka o samim podatcima. Grafovi opterećenja pokazali su da su najvažnije prediktorske varijable u rasponu valnih duljina 1400-1700 nm te 2000-2500 nm. Ti rasponi odgovaraju područjima prvog i drugog višeg tona te kombinacijskih vrpci prema slici 3.2. Posebno važnom ističe se područje oko 1400 nm koja predstavlja više tonove ROH, CH i CH_2 skupina, kombinacijska vrpca ROH skupine oko 2100 nm i kombinacijske vrpce CH i CH_2 skupina 2200-2500 nm. Ovo zapažanje je razumljivo jer su upravo te skupine prisutne u proučavanim spojevima (slika 1.2). Dodatno, pregledom **u-t** grafova kod PLS može se jasno uočiti da isključivo linearni odnos između $\log(\frac{1}{R})$ i masenih udjela, koji proizlazi iz NIR teorije, nije dovoljan za postizanje optimalnog PLS modela, već je potreban i mali doprinos nelinearnosti koje objašnjavaju glavne komponente nakon prve dvije (objašnjavaju $\approx 0.36\%$ varijance u podatcima). Uz to, vizualnom inspekcijom podataka sadržanih u **t-t** grafovima mogu se izvoditi zaključci o dizajnu uzoraka, čak i bez *a priori* znanja o masenim udjelima u $\mathbf{Y}$. Time je pokazano da PLS i PCA mogu biti izrazito vrijedan alat u eksplorativnoj analizi podataka [3, 17].

6.3 Predobrada NIRS podataka

U fazi predobrade podataka prikazani su (slika 5.1) i uspoređeni rezultati dviju metoda korekcije raspršenja - MSC i SNV, te dviju metoda derivacije spektara - NW i SG. Uočene su značajne razlike u rezultatima korekcija raspršenja i derivacija pri čemu su spektri predobrađeni metodama korekcije raspršenja zadržali sličan izgled, za razliku od spektara predobrađenih metodama derivacije. Dosadašnja istraživanja nisu uspjela utvrditi najbolju metodu predobrade te se smatra da odabir optimalne ovisi od slučaja do slučaja [19]. U većini slučajeva sve spomenute metode rezultiraju sličnom kvalitetom rezultata, no metode derivacije mogu prouzročiti nepouzdanost rezultata ukoliko su loše podešeni parametri (broj točki ravnanja, stupanj derivacije i sl.).

6.4 Budući rad

S obzirom da je u ovom radu prediktivna analiza izvedena samo na podatcima predobrađenim MSC, u budućim analizama preporučljivo bi bilo proučiti prediktivne perfomance preostalih metoda. Time bi se eventualno moglo poboljšati točnost, iako je upitno koliko je daljnja studija vrijedna s obzirom na vjerojatno malu mogućnost poboljšanja [64]. Nadalje, buduća istraživanja moguća su primjenom drugih regresijskih metoda, kao i primjenom drugih metoda selekcije varijabli, koje nisu korištene u ovom radu. U tom slučaju bi valjalo posebno naglasiti mogućnost primjene *lasso* regresije koja bi bila zanimljiva jer ta metoda ujedno spada i u regularizacijske (L_1 regularizacija) i selekcijske metode. Također, u radu su korištene primarno linearne regresijske metode, dok bi bila moguća i primjena drugih skupina metoda poput regresijskih stabala, uz *bagging* i *boosting*, i k-najbližih susjeda (eng. *k-nearest neighbors*, k-NN). Potonje, kao i ANN, zahtjavaju veliku količinu podataka s obzirom da ne zadovoljavaju nikakvu funkcijsku ovisnost te uključuju veliki broj parametara i hiperparametara. U literaturi se kao posebno zanimljiva metoda navodi stroj potpornih vektora (eng. *support vector machine*, SVM): "SVM predstavljaju najvažniji razvoj u kemometriji nakon (kronološki) PLS i ANN" [65]. Stoga bi bilo vrijedno istražiti primjenu i ove regresijske metode u kemometriji. Na koncu, svi razvijeni modeli mogli bi se spojiti u ansambl s ciljem razvoja još boljeg meta-modela. Pri razvoju takvog modela trebalo bi u obzir uzeti i pitanje koliko poboljšanje točnosti je moguće postići u odnosu na vrijeme koje bi bilo potrebno uložiti u razvoj takvih modela te kopromise između točnosti, jednostavnosti, skalabilnosti, brzine i interpretabilnosti.

Poglavlje 7

Zaključak

U ovom diplomskom radu predstavljene su i analizirane različite metode multivarijatne regresije za analizu NIR spektara, s ciljem određivanja njihovih prediktivnih performaci od važnosti za primjenu NIRSa u farmaceutskim i kemijskim pogonima. Koristeći NIRS i regresijske metode prezentirane u ovom radu moguće je odrediti masene udjele glukoze, fruktoze i saharoze u smjesi s pogreškom manjom od 1% (0.63 ± 0.08 za ridge regresiju uz *loopy* MSC predobradu). Iz razloga što se ridge regresija pokazala najtočnijom i vrlo brzom, upravo je ona preporučljiva za primjenu u industrijskim *in-line* analizama. Prethodno regresijskoj analizi, NIR spektri su podvrgnuti predobradi koja je uključivala *loopy* MSC i centriranje. Time je u radu proveden i prezentiran cjelokupni postupak razvoja prediktivnog modela za određivanje masenih udjela iz NIRS podataka.

Statističke metode prezentirane u ovom radu olakšavaju interpretaciju i analizu NIRS podataka. U tom vidu posebno se ističu PCA i PLS gdje se iz grafova skorova i opterećenja mogu izvesti zaključci o kemijskom sastavu uzoraka, tj. funkcionalnih skupina prisutnih u spojevima koji sačinjavaju smjesu. Kemometrijskim analizama kao što su PCA i PLS moguće je relativno kompleksne spektroskopske podatke svesti na znatno jednostavniji oblik, bez gubitka informacije.

Općenito možemo zaključiti da kontinuirani razvoj kemometrije omogućuje uspješnu primjenu NIRSa kao *in-line* analitičke tehnike u kemijskim i farmaceutskim analizama. Prednosti poput nedestruktivne prirode, brzine, mogućnosti kontrole u

realnom vremenu i potvrđenost od strane regulatornih tijela, promoviraju njegovu primjenu kao standardne tehnike u farmaceutskoj industriji. NIRS omogućava optimizaciju proizvodnog procesa u svim koracima te se očekuje da će njegova uloga u budućnosti biti još značajnija.

Bibliografija

[1] D. L. Massart, B. G. Vandeginste, L. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke, and C. K. Mann, Handbook of chemometrics and qualimetrics: Part A, ser. Data handling in science and technology. Elsevier Science, 1998, vol. 20A. , putem Interneta, https://books.google.hr/books/about/Handbook_of_Chemometrics_and_Qualimetric.html?id=jF0QhuxXeIwC&redir_esc=y

[2] S. Roussel, S. Preys, F. Chauchard, and J. Lallemand, "Multivariate data analysis (chemometrics)," in Process analytical technology for the food industry. Springer, 2014, ch. 2, pp. 7–59. , putem Interneta, https://books.google.hr/books?id=beosBQAAQBAJ&pg=PA7&hl=hr&source=gbs_toc_r&cad=4#v=onepage&q&f=false

[3] K. Varmuza and P. Filzmoser, Introduction to multivariate statistical analysis in chemometrics. CRC press, 2016.

[4] J. Luypaert, D. Massart, and Y. Vander Heyden, "Near-infrared spectroscopy applications in pharmaceutical analysis," Talanta, vol. 72, no. 3, pp. 865–883, 2007.

[5] T. Jednačak and P. Novak, "Procesne analitičke tehnike temeljene na vibracijskoj spektroskopiji in-line i primjena u industriji," Kemija u industriji, vol. 62, no. 3-4, pp. 71–80, 2013.

[6] W. P. Findlay and D. E. Bugay, "Utilization of Fourier transform-Raman spectroscopy for the study of pharmaceutical crystal forms," Journal of pharmaceutical and biomedical analysis, vol. 16, no. 6, pp. 921–930, 1998.

[7] X. Y. Lawrence, R. A. Lionberger, A. S. Raw, R. D'Costa, H. Wu, and A. S. Hussain, "Applications of process analytical technology to crystallization processes," Advanced Drug Delivery Reviews, vol. 56, no. 3, pp. 349–369, 2004.

[8] W. Plugge and C. Van der Vlies, "Near-infrared spectroscopy as an alternative to assess compliance of ampicillin trihydrate with compendial specifications," Journal of pharmaceutical and biomedical analysis, vol. 11, no. 6, pp. 435–442, 1993.

[9] M. Forina, M. Casolino, and C. De la Pezuela Martínez, "Multivariate calibration: applications to pharmaceutical analysis," Journal of pharmaceutical and biomedical analysis, vol. 18, no. 1-2, pp. 21–33, 1998.

[10] D. Jouan-Rimbaud, M. Khots, D. Massart, I. Last, and K. Prebble, "Calibration line adjustment to facilitate the use of synthetic calibration samples in near-infrared spectrometric analysis of pharmaceutical production samples," Analytica chimica acta, vol. 315, no. 3, pp. 257–266, 1995.

[11] Y. Chen, S. S. Thosar, R. A. Forbess, M. S. Kemper, R. L. Rubinovitz, and A. J. Shukla, "Prediction of drug content and hardness of intact tablets using artificial neural network and near-infrared spectroscopy," Drug development and industrial pharmacy, vol. 27, no. 7, pp. 623–631, 2001.

[12] NIR Spectroscopy-A guide to near-infrared spectroscopic analysis of industrial manufaturing processes. Metrohm, AG, 2013.

[13] H. Heise and R. Winzen, "Chemometrics in Near-Infrared Spectroscopy," in Near-Infrared Spectroscopy: Principles, Instruments, Applications, H. Siesler and Y. Ozaki, Eds. Wiley Online Library, 2001, ch. 7, pp. 125–162. , putem Interneta, https://books.google.hr/books?id=U7vqrf2YqmcC&printsec=frontcover&hl=hr&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

[14] C. Pasquini, "Near infrared spectroscopy: fundamentals, practical aspects and analytical applications," Journal of the Brazilian chemical society, vol. 14, no. 2, pp. 198–219, 2003.

[15] J. Torrent and V. Barrón, "Diffuse reflectance spectroscopy," Methods of soil analysis. Part 5, vol. 5, pp. 367–387, 2008.

[16] D. Duncan, "The colour of pigment mixtures," Proceedings of the Physical Society, vol. 52, no. 3, p. 390, 1940.

[17] L. Nørgaard, R. Bro, and S. B. Engelsen, "Principal component analysis and near infrared spectroscopy," a FOSS white paper, http://www. foss. de/industry-solution/chemical-analysis/papers, 2012.

[18] J. Leek, The Elements of Data Analytic Style. Leanpub, 2015.

[19] Å. Rinnan, F. Van Den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," TrAC Trends in Analytical Chemistry, vol. 28, no. 10, pp. 1201–1222, 2009.

[20] R. J. Pell, M. B. Seasholtz, and B. R. Kowalski, "The relationship of closure, mean centering and matrix rank interpretation," Journal of chemometrics, vol. 6, no. 1, pp. 57–62, 1992.

[21] W. Windig, J. Shaver, and R. Bro, "Loopy MSC: a simple way to improve multiplicative scatter correction," Applied spectroscopy, vol. 62, no. 10, pp. 1153–1159, 2008.

[22] K. Norris and P. Williams, "Optimization of mathematical treatments of raw near-infrared signal in the," Cereal Chem, vol. 61, no. 2, pp. 158–165, 1984.

[23] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures." Analytical chemistry, vol. 36, no. 8, pp. 1627–1639, 1964.

[24] P. A. Gorry, "General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method," Analytical Chemistry, vol. 62, no. 6, pp. 570–573, 1990.

[25] J. M. Wooldridge, Introductory econometrics: A modern approach. Nelson Education, 2015.

[26] G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning. Springer, 2013, vol. 112.

[27] K. Varmuza, P. Filzmoser, and M. Dehmer, "Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS," Computational and structural biotechnology journal, vol. 5, no. 6, 2013.

[28] A. Niazi and R. Leardi, "Genetic algorithms in chemometrics," Journal of Chemometrics, vol. 26, no. 6, pp. 345–351, 2012.

[29] V. Mallawaarachchi, "Introduction to genetic algorithms-including example code," Towards Data Science, 2017. , putem Interneta, https://towardsdatascience.com/introduction-to-genetic-algorithms-includingexample-code-e396e98d8bf3

[30] S. Jain, "Introduction to genetic algorithm & their application in data science," Analytics Vadhya, 2017. , putem Interneta, https://www.analyticsvidhya.com/blog/2017/07/introduction-to-genetic-algorithm

[31] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55–67, 1970.

[32] J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning. Springer series in statistics New York, 2001, vol. 1, no. 10.

[33] F. Scott, "Understanding the bias-variance tradeoff," An essay by Scott Fortmann-Roe, 2012. , putem Interneta, http://scott.fortmann-roe.com/docs/BiasVariance.html

[34] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," Technometrics, vol. 21, no. 2, pp. 215–223, 1979.

[35] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, pp. 559–572, 1901.

[36] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and intelligent laboratory systems, vol. 2, no. 1-3, pp. 37–52, 1987.

[37] H. Abdi and L. J. Williams, "Principal component analysis," Wiley interdisciplinary reviews: computational statistics, vol. 2, no. 4, pp. 433–459, 2010.

[38] H. Anton and C. Rorres, Elementary Linear Algebra. John Wiley & Sons, 2013.

[39] N. O'Rourke, L. Hatcher, and E. J. Stepanski, A step-by-step approach to using SAS for univariate & multivariate statistics. SAS institute, 2005. , putem Interneta, https://books.google.hr/books/about/A_Step_by_Step_Approach_to_Using_SAS_for.html?id=pfUfzykTZ1AC&redir_esc=y

[40] H. F. Kaiser, "The application of electronic computers to factor analysis," Educational and psychological measurement, vol. 20, no. 1, pp. 141–151, 1960.

[41] R. B. Cattell, "The scree test for the number of factors," Multivariate behavioral research, vol. 1, no. 2, pp. 245–276, 1966.

[42] W. F. Massy, "Principal components regression in exploratory statistical research," Journal of the American Statistical Association, vol. 60, no. 309, pp. 234–256, 1965.

[43] H. Wold, "Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach," Journal of Applied Probability, vol. 12, no. S1, pp. 117–142, 1975.

[44] S. Wold, A. Ruhe, H. Wold, and W. Dunn, III, "The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses," SIAM Journal on Scientific and Statistical Computing, vol. 5, no. 3, pp. 735–743, 1984.

[45] K. Dunn, Process improvement using data, 2010.

[46] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," Analytica chimica acta, vol. 185, pp. 1–17, 1986.

[47] I. N. Wakeling and J. J. Morris, "A test of significance for partial least squares regression," Journal of Chemometrics, vol. 7, no. 4, pp. 291–304, 1993.

[48] A. Höskuldsson, "PLS regression methods," Journal of chemometrics, vol. 2, no. 3, pp. 211–228, 1988.

[49] B. S. Dayal and J. F. MacGregor, "Improved PLS algorithms," Journal of Chemometrics: A Journal of the Chemometrics Society, vol. 11, no. 1, pp. 73–85, 1997.

[50] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," Chemometrics and intelligent laboratory systems, vol. 58, no. 2, pp. 109–130, 2001.

[51] M. Mörtsell and M. Gulliksson, An overview of some non-linear techniques in chemometrics. Mitthögskolan, FSCN, 2001.

[52] F. Lindgren, P. Geladi, and S. Wold, "The kernel algorithm for PLS," Journal of Chemometrics, vol. 7, no. 1, pp. 45–59, 1993.

[53] S. Rännar, F. Lindgren, P. Geladi, and S. Wold, "A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm," Journal of Chemometrics, vol. 8, no. 2, pp. 111–125, 1994.

[54] S. Rännar, P. Geladi, F. Lindgren, and S. Wold, "A PLS kernel algorithm for data sets with many variables and few objects. Part II: Cross-validation, missing data and examples," Journal of Chemometrics, vol. 9, no. 6, pp. 459–470, 1995.

[55] S. De Jong, "SIMPLS: an alternative approach to partial least squares regression," Chemometrics and intelligent laboratory systems, vol. 18, no. 3, pp. 251–263, 1993.

[56] F. Marini, R. Bucci, A. Magrì, and A. Magrì, "Artificial neural networks in chemometrics: History, examples and perspectives," Microchemical journal, vol. 88, no. 2, pp. 178–185, 2008.

[57] M. A. Nielsen, Neural networks and deep learning. Determination press San Francisco, CA, USA, 2015, vol. 25.

[58] D. E. Rumelhart, G. E. Hinton, R. J. Williams et al., "Learning representations by back-propagating errors," Cognitive modeling, vol. 5, no. 3, p. 1, 1988.

[59] L. Norgaard, M. Lagerholm, and M. Westerhaus, "Artificial neural networks and near infrared spectroscopy-a case study on protein content in whole wheat grain," Foss White Paper http://www. foss. dk/campaign/-/media/242657904D734CE9B0652C3D885776AE. ashx, 2013.

[60] R. M. Balabin and E. I. Lomakina, "Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data," Analyst, vol. 136, no. 8, pp. 1703–1712, 2011.

[61] H. Yoshida, R. Leardi, K. Funatsu, and K. Varmuza, "Feature selection by genetic algorithms for mass spectral classifiers," Analytica Chimica Acta, vol. 446, no. 1-2, pp. 483–492, 2001.

[62] J. P. A. Martins, R. F. Teofilo, and M. M. Ferreira, "Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets," Journal of Chemometrics, vol. 24, no. 6, pp. 320–332, 2010.

[63] M. R. Smith, R. D. Jee, and A. C. Moffat, "The transfer between instruments of a reflectance near-infrared assay for paracetamol in intact tablets," Analyst, vol. 127, no. 12, pp. 1682–1692, 2002.

[64] C. B. Zachariassen, J. Larsen, F. van den Berg, and S. B. Engelsen, "Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in low methoxylated amidated pectin production," Chemometrics and Intelligent Laboratory Systems, vol. 76, no. 2, pp. 149–161, 2005.

[65] O. Ivanciuc, "Applications of support vector machines in chemistry," Reviews in computational chemistry, vol. 23, p. 291, 2007.

europass Životopis

OSOBNE INFORMACIJE

Požar Nino

Rupa 70, 51214 Šapjane (Hrvatska)

(+385) 98 180 7909

pozar.nino@gmail.com

Spol Muško | Datum rođenja 23/05/1994 | Državljanstvo hrvatsko

RADNO ISKUSTVO

23/07/2019–danas

Data scientist/machine learning engineer

Adacta d.o.o., Zagreb (Hrvatska)

Rad na projektu u suradnji s Raiffeisenbank Hrvatska. Tema projekta je predviđanje ponašanja klijenata banke.

06/2019–07/2019

Predavač (vanjski suradnik) na kolegiju: MK103 Kemometrija

Odjel za biotehnologiju, Sveučilište u Rijeci, Rijeka (Hrvatska)

Održavanje predavanja i seminara u sklopu kolegija MK103 Kemometrija na temu multivarijatne regresije - diplomski studiji "Istraživanje i razvoj lijekova" te "Medicinska kemija".

2017–danas

Suradnik na istraživačkom projektu

Centar za mikro- i nanoznanosti i tehnologije, Sveučilište u Rijeci, Rijeka (Hrvatska)

- izrada diplomskog rada u Laboratoriju za koloide, polielektrolite i međupovršine te Laboratoriju za fizikalnu kemiju pod mentorstvom doc. dr. sc. Duška Čakare

- naziv rada: "Metode multivarijatne regresije za kemometrijsku analizu spektara u bliskom infracrvenom području"

02/2018–02/2018

Demonstrator na kolegiju: BIL302 Fizikalna kemija

Odjel za biotehnologiju, Sveučilište u Rijeci, Rijeka (Hrvatska)

11/2016–11/2016

Demonstrator na kolegiju: BIL302 Fizikalna kemija

Odjel za biotehnologiju, Sveučilište u Rijeci, Rijeka (Hrvatska)

27/06/2016–08/07/2016

Stručna praksa

Jadran galenski laboratorij, Rijeka (Hrvatska)

- Stručna praksa na Odjelu proizvodnje (mentor: Vladimir Maleš)

OBRAZOVANJE I OSPOSOBLJAVANJE

2016–danas

Diplomski sveučilišni studij "Medicinska kemija", Rijeka (Hrvatska)

Sveučilište u Rijeci - Odjel za biotehnologiju

2013–2016

Sveučilišni prvostupnik biotehnologije i istraživanja lijekova (univ. bacc. biotech. et pharm. inv.)

Preddiplomski sveučilišni studij "Biotehnologija i istraživanje lijekova", Rijeka (Hrvatska)

Sveučilište u Rijeci - Odjel za biotehnologiju

- pohvala CUM LAUDE

2009–2013

Gimnazija Andrije Mohorovičića Rijeka, Rijeka (Hrvatska)

- prirodoslovno matematička gimnazija

OSOBNE VJEŠTINE

Materinski jezik: hrvatski

Strani jezici

	RAZUMIJEVANJE		GOVOR		PISANJE
	Slušanje	Čitanje	Govorna interakcija	Govorna produkcija	
engleski	C1	C1	C1	C1	C1

Stupnjevi: A1 i A2: Početnik - B1 i B2: Samostalni korisnik - C1 i C2: Iskusni korisnik
Zajednički europski referentni okvir za jezike

Komunikacijske vještine: Dobre komunikacijske vještine stečene kroz rad sa studentima tijekom predavanja i demonstratura, prezentacijom seminara u sklopu studija, prezentacijom radova na znanstvenim skupovima i natjecanjima te polaskom tečajeva/radionica.

Organizacijske / rukovoditeljske vještine: Dobre voditeljske vještine stečene vođenjem 7 iznimno uspješnih timova na case study i data science natjecanjima.

Digitalne vještine

SAMOPROCJENA				
Obrada informacija	Komunikacija	Stvaranje sadržaja	Sigurnost	Rješavanje problema
Iskusni korisnik	Iskusni korisnik	Iskusni korisnik	Iskusni korisnik	Iskusni korisnik

Digitalne vještine - Tablica za samoprocjenu

- odlično vladanje alatima MS Office
- programi za obradu podataka: R (RStudio), SQL, Python, Qlik
- dobro vladanje računalnim (bio)kemijskim programima

Vozačka dozvola: B

DODATNE INFORMACIJE

Priznanja i nagrade

- 1. nagrada eSTUDENT Mozgalo 2019

  zadatak: Client behavior prediction: A Machine Learning Challenge (zadatak zadali Adacta i Raiffeisenbank Hrvatska)
- 1. nagrada Case Study Realizator 2018 (JGL d.d.)

  tema: Penetracija (rast prodaje) brenda Vizol S kroz inovaciju
- 1. nagrada JGL Case Study Competition 2017

  tema: Rješavanje stvarnih problema iz područja registracije lijekova
- 3. nagrada Case Study Realizator 2017 (JU Priroda)

  tema: Razvoj programske djelatnosti Centra za posjetitelje s oporavilištem za bjeloglave supove u mjestu Beli na otoku Cresu
- eSTUDENT Case Study Competition 2019 (L'Oréal Adria d.o.o.) - finale

  tema: Pokretanje online trgovine brenda za njegu kože

- eSTUDENT Case Study Competition 2018 (GSK d.o.o. Croatia) - finale
  tema: Multi-Channel Marketing kampanja za Duac
- nagrada za najaktivniji tim Case Study Realizator 2017

MOOC

Data Science Specialization by Johns Hopkins University on Coursera

- courses: Data Scientist's Toolbox, R Programming, Getting and Cleaning Data, Exploratory Data Analysis, Reproducible Research, Statistical Inference, Regression Models, Practical Machine Learning, Developing Data Products, Data Science Capstone

Strategic Business Analytics by ESSEC Business School on Coursera

- courses: Foundations of strategic business analytics, Foundations of marketing analytics, Case studies in business analytics with ACCENTURE, Capstone: Create Value from Open Data

Executive Data Science by Johns Hopkins University on Coursera

- courses: A Crash Course in Data Science, Building a Data Science Team, Managing Data Analysis, Data Science in Real Life, Executive Data Science Capstone

Machine Learning course by Stanford University on Coursera

Stipendije

- Stipendija Općine Matulji temeljem akademskog uspjeha- 2014./15., 2015./16., 2016./17.
- Sveučilišna stipendija za izvrsnost- 2017./18., 2018./19.

Kongresi i konferencije

5. Simpozij studenata kemičara - Zagreb

- poster "Ispitivanje samonakupljanja amfifilnog derivata rodamina B u vodenoj otopini"
  autori: D. Matulja, **N. Požar**, N. Malatesti, D. Čakara

32nd European Colloid and Interface Society Conference (ECIS 2018) - Ljubljana

- poster "Partial least squares regression for fitting the dissociation constants from the UV-Vis spectra without calibration"
  autori: **N. Požar**, D. Matulja, D. Čakara
- poster "Self-assembly of weakly acidic photoactive dyes in aqueous solution at varied pH and ionic strength"
  autori: D. Matulja, **N. Požar**, N. Malatesti, D. Čakara

2. studentski kongres Okolišnog zdravlja - Rijeka

- poster i prezentacija "Analiza utjecaja zagađenja zraka na zdravlje ljudi u Bosni i Hercegovini"
  autori: K. Pavlović, **N. Požar**, J. J. Castillo, S. Sanchez, K. Džepina
  1. nagrada za najbolji studentski rad

16th Conference of the International Association of Colloid and Interface Scientists (IACIS 2018) - Rotterdam

- poster "Multivariate regression methods for fitting the dissociation constatnts from the UV/VIS spectra of self-assembled dyes"
  autori: **N. Požar**, D. Matulja, N. Malatesti, D. Čakara

Natjecanja

- eSTUDENT Mozgalo 2019
- eSTUDENT Case Study Competition 2019
- Case Study Realizator 2018
- eSTUDENT Case Study Competition 2018

- Case Study Realizator 2017
- JGL Case Study Competition 2017
- KK Cedevita Case Study Competition 2017

Članstva Udruga studenata biotehnologije (USBRI)- član

Volontiranje
- Sudjelovanje u aktivnostima otvorenog dana Odjela za biotehnologiju (2015., 2017.)
- Znanost i umjetnost na ulici (2016.)
- Student mentor (2016./17.)