Primjena statističkog učenja na proširenu semi-empirijsku formulu mase

Bezak, Mihaela

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO- MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

Mihaela Bezak

PRIMJENA STATISTIČKOG UČENJA NA PROŠIRENU SEMI-EMPIRIJSKU FORMULU MASE

Diplomski rad

Zagreb, 2019.

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO- MATEMATIČKI FAKULTET
FIZIČKI ODSJEK

INTEGRIRANI PREDDIPLOMSKI I DIPLOMSKI SVEUČILIŠNI STUDIJ
FIZIKA; SMJER ISTRAŽIVAČKI

Mihaela Bezak

Diplomski rad

Primjena statističkog učenja na proširenu semi-empirijsku formulu mase

Voditelj diplomskog rada: prof. dr. sc., Nils Paar

Ocjena diplomskog rada: _____

Povjerenstvo: 1. _____

2. _____

3. _____

Datum polaganja: _____
Zagreb, 2019.

Zahvaljujem dr. sc. Tomislavu Marketinu na pomoći, dobroj volji, beskonačnoj susretljivosti te na predanom znanju prilikom izrade diplomskog rada. Također, zahvale i mentoru prof. dr. sc. Nils Paaru na trudu, predanosti i strpljenju.
Hvala mojim roditeljima, Mirku i Katarini, sestri Gabrijeli, čija me potpora pokretala da završim fakultet.
Na kraju, svim prijateljima, hvala na sreći i veselju koju su unijeli u moj život.

Sažetak

U nuklearnoj fizici već više od osamdeset godina u primjeni je Bethe–Weizsäcker formula mase (BW). BW formula sadrži osnovne elemente koji su dovoljni za kvantitativan opis energija vezanja. Unatoč postojanju modernih, mnogo kompleksnijih modela mase, zbog svoje jednostavnosti i intuitivne povezanosti s fizikom atomske jezgre BW formula je i dalje interesantna tema istraživanja.
Motiv ovog rada je opisivanje karakteristika BW formule uporabom podataka izmjerenih energija vezanja. Metode statističkog učenja u tome predstavljaju vrlo efikasan alat koji može uvidjeti pravilnosti te značajnost pojedinih doprinosa. Dobiva se zanimljiv rezultat da najznačajniji doprinos energiji vezanja imaju osnovni članovi iz BW formule, dok novo pridodani članovi predstavljaju korekcije manjeg doprinosa na osnovnu formulu. Dodavanje novih članova sustavu opravdano je samo ako novo pridodan član sa sobom nosi novi fizikalni sadržaj, koji predhodno nije bio uključen u modelu. Povećavanjem kompleksnosti, povećavamo varijabilnost na štetu generalizacije što je prikazano pomoću train-test pogreške.

Ključne riječi: Bethe–Weizsäcker formula mase, proširena Bethe–Weizsäcker formula mase, primjena statističkog učenja

Application of statistical learning on expanded semi-empirical mass formula

Abstract

Bethe–Weizsäcker (BW) formula has been used by nuclear physicists for more than eighty years. The BW formula contains the basic terms sufficient for a quantitative description of nuclear binding energies. Although more modern and complex mass models exist, the BW formula is still widely used and studied due to its simplicity and intuitive connection with atomic physics.

The aim of this thesis is to describe characteristics of the BW formula by using experimentally observed binding energies. Statistical learning methods are shown to be a very effective tool for identifying patterns and the weights of contributing terms. We come to conclusion that the most relevant contribution to binding energies comes from the basic terms of the BW formula, and any additional terms only give small corrections. Adding a new term to the system is only justified if it introduces physics ignored by the basic model. Adding complexity also increases variability at the expense of generalisation, as it is shown by the train-test error.

Keywords: Bethe–Weizsäcker mass formula, statistical learning, improved BW formula

Sadržaj

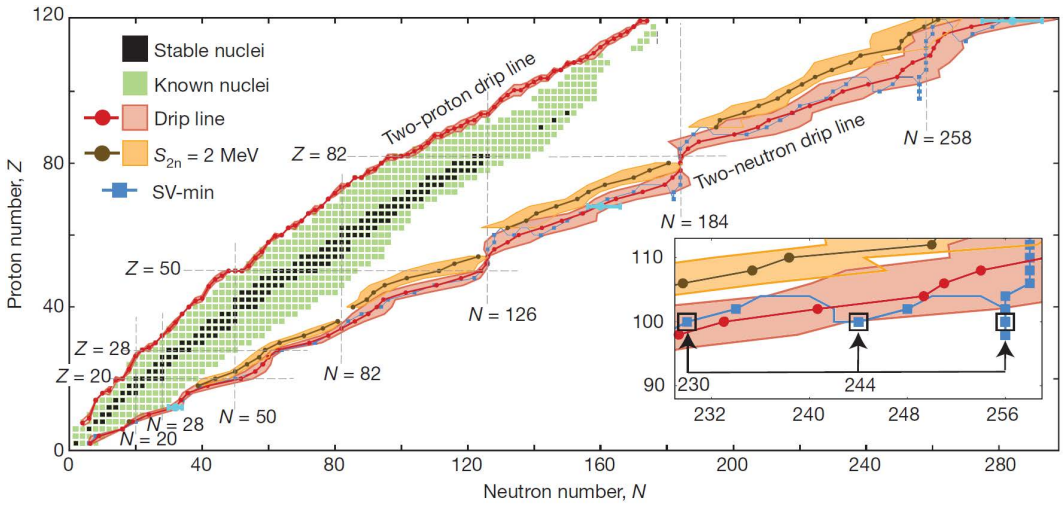
1 Uvod

Zadnjih nekoliko godina došlo je do snažnog razvoja informacijskih tehnologija vezanih uz područje znanosti o podacima. Tako se recimo, pojmovi poput strojnog učenja, umjetne inteligencije mogu čuti na svakodnevnoj razini. Taj snažan napredak discipline nije neobičan jer su nam zbog interneta dostupne velike količine podataka skladištenih u bazama. Uz dostupnost, ključnu ulogu je imao razvoj neuronskih mreža. Po uzoru na neuronski sustav, neuronske mreže su radna cjelina koja automatizirano, bez navođenja informacija, iščitava karakteristike podataka. Kako bi preciznije primjenjivali algoritme za učenje, potrebno je razumijevanje karakteristika sustava na statističkoj razini, što nas navodi prema korištenju tzv. *statističkog učenja*. Statističko učenje je razvojna cjelina strojnog učenja. Glavna razlika od strojnog učenja, koje je set metoda za automatizirano učenje iz podataka, je da statističkim učenjem formiramo odnos između podataka pomoću matematičkih jednadžbi povučenih iz funkcionalne analize i statistike. Primjena je interdisciplinarna te visoko zastupljena u područjima istraživanja poput psihologije [1] te neuroznanosti [2]. Društvene znanosti poput ekonomije koriste strojno učenje prilikom izrade marketinških planova gdje je najveći fokus na procjeni rizika. Polja društvenih znanosti poput lingvistike i antropologije svoj fokus stavljaju na prepoznavanje i simuliranje određenih jezičnih izraza što u svrhu znanstvenog istraživanja, dok ponajviše zbog uporabe u industriji [3].

Razvoj statističkog i strojnog učenja nije ostao nezamijećen od strane znanstvene zajednice iz područja fizike. Kako se skala promatranja fizikalnih problema smanjuje, ili povećava, količina podataka koja se mora obraditi postaje sve veća. Nije ništa neobično da velike eksperimentalne kolaboracije poput CERN-a godišnje proizvedu do 25 petabajtova (PB) podataka [1]. Analiza tih podataka je usko povezana s algoritmima za učenje neuronskih mreža [4]. Druga poznatija primjena statističkog učenja u okviru fizike, je ona u nuklearnoj astrofizici. Koriste se kompleksne simulacije nukleosinteze teških elemenata koje su utemeljene na kompliciranom međudjelovanju.

Na slici 1.1 prikazana je mapa nuklida, gdje su crnim kvadratićima označene stabilne jezgre, njih je svega stotinjak, točnije 288. Stabilne jezgre tvore tzv. dolinu

[1] 1 PB = 1000 TB



Slika 1.1: Mapa vezanih parno-parnih jezgara prikazanih kao funkcija Z i N. Crnim kvadratima prikazani su stabilni nuklidi, zeleno su prikazani nestabilni. Crvene linije označavaju granice vezanja [5].

stabilnosti. Zelenim kvadratićima su označene ostale, nestabilne jezgre kojih je nekoliko tisuća. Crveno su označene granice vezanja. Prikazane linije predstavljaju granicu gdje jezgra više nema dovoljno energije da spriječi ispadanje zadnjeg nukleona. Unutar granica vezanja, sveukupno nam je poznato oko 3000 jezgara, dok teorija sugerira oko 7000 jezgara [5]. Mapiranje i određivanje fizikalnih svojstava teorijski predviđenih jezgara aktualna je tema današnjih istraživanja u nuklearnoj fizici.

Za kvalitetno mapiranje jezgara potrebno je poznavanje njihovih masa tj. energija vezanja. Osim toga, poznavanje energije vezanja od fundamentalne je važnosti za proučavanje mnogih drugih nuklearnih svojstava. Svrha ovog rada je iz fizikalno jednostavne perspektive, proučavanjem isključivo samo poznatih podataka, utvrditi pravilnosti pri opisu energije vezanja. Za uočavanje pravilnosti oslonit ćemo se na metode statističkog učenja. Problematika i metode statističkog učenja teorijskim pristupom predstavljene su u drugom poglavlju. U trećem poglavlju prikazana je primjena metoda na Bethe-Weizsäckerovoj formuli. Rezultati sugeriraju na nedostatke u opisu te je stoga osnovna Bethe-Weizsäckerova formula proširena s novim članovima koji uključuju i nove parametre. U ovom radu u obzir je uzeto dodatnih šest članova te je provjeren njihov doprinos u vidu smanjenja ukupne pogreške. U nedavno objavljenom članku [5] problem izračuna energija vezanja numerički je generaliziran uvođenjem linearno nezavisnim varijablama, koje služe kao dopuna osnovnim članovim iz BW formule. Prema uzoru na članak, devet parametara iz

Bethe-Weizsäckerove je prošireno linearnim deformacijama. Ukupan model na koji je primijenjena statistička analiza sadržava 119 članova. Ostalo je otvoreno pitanje proširenja Bethe-Weizsäckerove formule sa što više kompjuterski generiranih članova koji bi pokrili sve interakcije među članovima. U svrhu toga napravljena je analiza takvog modela. Zaključno smo usporedili sve modele i metode. Prikazan je model koji najbolje optimizira kompleksnost sustava te nudi mogućnost za daljnja predviđanja.

2 Energija vezanja jezgre

Energija vezanje jezgre $E_B(N, Z)$ je energija potrebna da se svih Z protona i N neutrona izdvoje iz jezgre. Koristimo Einsteinovu relaciju $E = mc^2$ kako bi energiju vezanja povezali s masom. Dobiven izraz za energiju vezanja prikazuje razliku masa slobodnih nukleona i jezgre

$$E_B(N, Z) = Zm_pc^2 + Nm_nc^2 - m(N, Z)c^2 \tag{2.1}$$

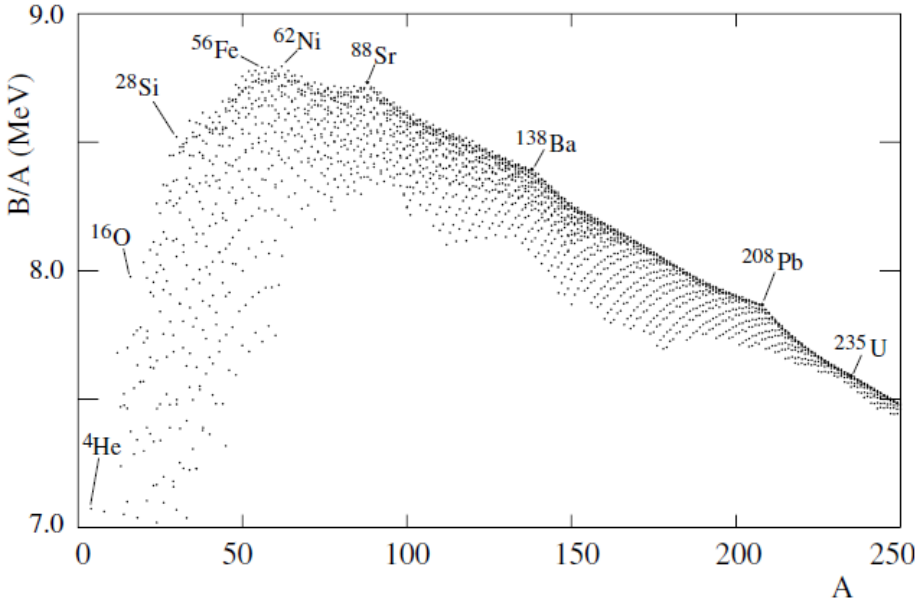
gdje je m_p masa protona, m_n masa neutrona te m(N,Z) masa jezgre.
Energija vezanja po nukleonu E_B/A prikazana je na slici 2.1 Kod lakših jezgara primjećujemo da energija vezanja po nukleonu raste s masenim brojem A. Porast energije je relativno velik u usporedbi sa susjednim nuklidima. Izdvojeno, energija vezanja ^4_2He značajnije je veća od susjednih jezgara sličnog masenog broja. Jezgra helija sadrži dva protona i dva neutrona čije nuklearno međudjelovanje je posebno jako. Navedena konfiguracija se naziva alfa grozd. Stvaranje strukture grozdova opažamo i kod ostalih jezgara koje se mogu opisati kao skup alfa čestica.
Iznad vrijednosti $A \approx 56$ energija vezanja po nukleonu počinje blago opadati. Zapišemo li energiju vezanja u njenoj prvoj aproksimaciji, gdje se promatra samo nuklearno međudjelovanje između susjednih nukleona, slijedi izraz

$$E_b \approx A \times 8 \text{ MeV}. \tag{2.2}$$

Iz izraza vidi se da prosječna energija vezanja po nukleonu iznosi

$$E_B/A = 7.7 - 8.8 \text{ MeV} \quad 12 < A < 225. \tag{2.3}$$

Trend opadanja indikacija je Coulombovog odbijanja između protona u jezgri. Energija vezanja po nukleonu grubo je podijeljena na dva područja. Područje lakih jezgara gdje energija vezanja po nukleonu raste sa masom jezgre pa je fuzija egzoterman proces, te područje teških jezgara gdje se energija oslobađa cijepanjem jezgara.

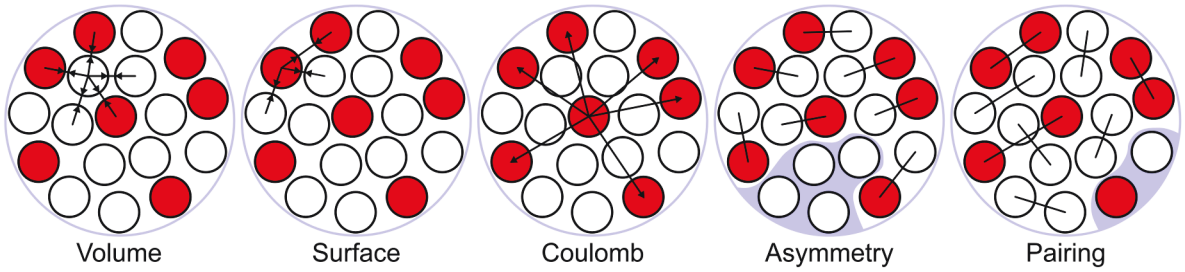


Slika 2.1: Energija vezanja po nukleonu $E_B(A, Z)/A$ kao funkcija masenog broja A [7].

2.1 Semi-empirijska formula mase

Problem izračuna energije vezanja jedan je od zahtjevnijih problema teorijske nuklearne fizike. Dva su osnovna pristupa modeliranju energije vezanja: mikroskopski, koji počiva na modeliranju nuklearnog međudjelovanja i detaljnom izračunu svojstava jezgre, i makroskopski u kojem se jezgra smatra kao makroskopski objekt. Zbog velikog opsega fizikalnih čimbenika koji se moraju uključivati u model, mikroskopski modeli postaju izrazito složeni za jezgre s atomskim brojem većim od tri.

Nešto jednostavniji pristup problemu je da se uzmu u obzir empirijske informacije koje se primjenjuju za sveobuhvatan opis strukturnih fenomena atomskih jezgara. Takav općenit pristup daje dobro objašnjenje nuklearnih struktura te dobro opisuje nuklearne energije vezanja. Kao primjer takvog pristupa je širokoprihvaćen nuklearni model kapljice. Jezgra se promatra kao kapljica nestlačive tekućine s pretpostavkom da je nuklearna sila kratko-dosežna te jednaka za svaki nukleon. 1935. godine Carl Friedrich von Weizsäcker, vođen idejom modela kapljice objavljuje prvu semi-empirijsku formulu mase (SEMF). Bethe i Bacher 1936. godine objavljuju članak modela masa koristeći statistički pristup primijenjen na Weizsäckerovu semi-empirijsku formulu. Iz tog pristupa nastala je danas poznata *Bethe-Weizsäckerova formula mase*.



Slika 2.2: Ilustrativan prikaz doprinosa energiji vezanja atomske jezgre prikazanih u okviru modela kapljice [8].

2.2 *Bethe-Weizsäckerova formula mase*

Bethe-Weizsäckerova formula mase zasniva se na nuklearnom modelu kapljice uz dodatne članove koji opisuju kvantne efekte. Formula mase ima parametarski oblik

$$E_B = a_v A - a_s A^{2/3} - a_c \frac{Z(Z-1)}{A^{1/3}} - a_{asym} \frac{(N-Z)^2}{A} + a_p \frac{\delta(A,Z)}{A^{1/2}}. \tag{2.4}$$

Konstante a_i predstavljaju parametre koji se određuju prilagodbom na eksperimentalne podatke. Zbog lakšeg fizikalnog razumijevanja Bethe-Weizsäckerove formule masa, promotrimo svaki od članova zasebno. Na slici 2.2 ilustrativno je prikazan utjecaj pojedinačnih članova.

Volumni član

Nukleoni međudjeluju sa najbližim susjednim nukleonom stoga pretpostavljamo da je gustoća nukleona konstantna. Volumen kapljice, u našem slučaju jezgre, će tako biti proporcionalan broju nukleona tj. masenom broju A. Doprinos energiji vezanja od volumnog člana je

$$E_{vol} = a_v A. \tag{2.5}$$

Jednadžba 2.5 daje relacijski odnos između atomskog broja i radijusa jezgre $R \propto A^{1/3}$.

Površinski član

Jednadžba 2.5 opisuje energiju vezanja beskonačne nuklearne materije. Da bismo dobili opis atomske jezgre, potrebno je uključiti i površinske efekte. Uključivanje "rubnih efekata" uvodi se pomoću površinskog člana. Nukleoni koji se nalaze na

površini jezgre slabije se vežu nego oni u unutrašnjosti jezgre. Dakle, zaključujemo da mora postojati član koji je proporcionalan površini sfere. Budući da je volumen sfere radijusa R proporcionalan s masenim brojem, $R^3 \propto A$ slijedi da je član površine $R^2 \propto A^{2/3}$. Doprinos površinskog člana energiji vezanja je

$$E_{surf} = a_s A^{2/3}. \tag{2.6}$$

Ovaj doprinos smanjuje ukupnu energiju vezanja i predstavlja korekciju volumnom članu. Pozivajući se na analogiju s kapljicom tekućine, član površine povezuje se površinskom napetošću.

Coulombov član

U obzir je potrebno uzeti i odbijanje protona u jezgri koji dodatno smanjuje energiju vezanja. Ako jezgra sadrži Z protona naboja **e**, na proton će djelovati drugih Z-1 protona. Elektrostatska energija takve distribucije naboja u sferi radijusa R dana je

$$E = \frac{3}{5}\frac{e^2}{4\pi\epsilon_0}\frac{Z(Z-1)}{r} \tag{2.7}$$

gdje je r udaljenost između dva naboja. Kako je srednja udaljenost naboja u jezgri proporcionalna ukupnom radijusu sfere $R \propto A^{1/3}$ slijedi da je izraz za Coulombov član

$$E_{coul} = a_c \frac{Z(Z-1)}{A^{1/3}}. \tag{2.8}$$

Vrijedi spomenuti da se u literaturi često koristi izraz Z^2. U ovom radu koristit će se Z(Z-1) jer naglašava da elektrostatsko odbijanje postoji samo ako je prisutan više od jedan proton.

Član asimetrije

Većina lakih i srednje teških stabilnih jezgara ima jednak broj neutrona i protona. Povećanjem razlike broja protona i neutrona raste asimetrija jezgre na štetu energije vezanja. Član asimetrije također je značajan za lakše jezgre što se može vidjeti iz njegovog oblika

$$E_{asym} = a_{asym}\frac{(N-Z)^2}{A}. \tag{2.9}$$

Teorijsko obrazloženje ne leži unutar okvira klasične fizike već se radi o posljedicama kvantnog principa tj. Paulijevom principu isključivanja. Protone i neutrone opisujemo dvama različitim kvantnim stanjima (dva izospinska stanja). Uzmemo li na primjer jezgru u kojoj prevladava veći broj neutrona, neki od neutrona će morati zauzeti više energetsko stanje. Zauzimanjem viših energetskih stanja neutroni gube na energiji vezanja. Gustoća popunjavanja te broj popunjenih ukupnih nivoa ima funkcijsku ovisnost o masenom broju A.

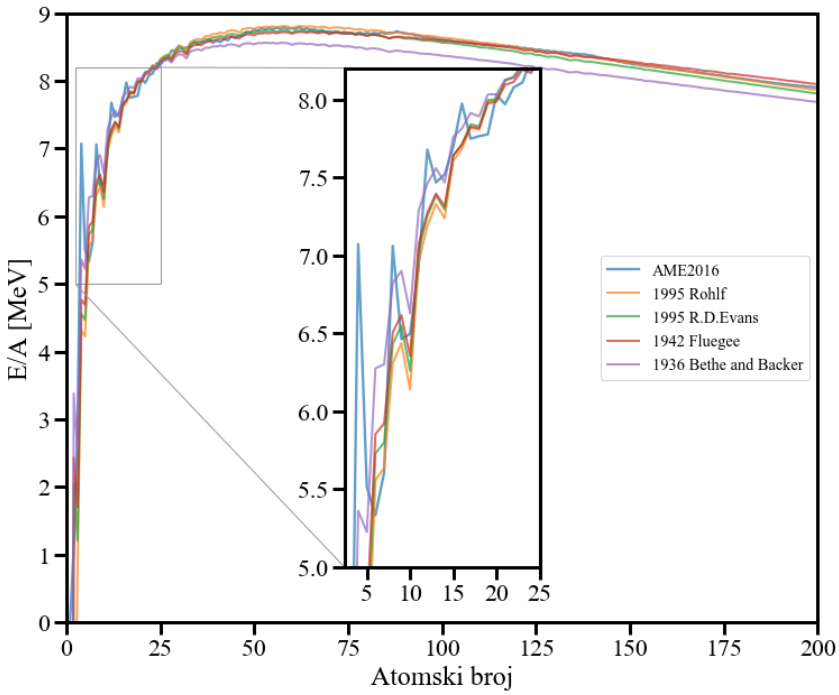
Član sparivanja

Član sparivanja dolazi od međusobnog vezanja protona s protonom te neutrona s neutronom. Jezgra s parnim brojem nukleonskih parova je stabilnija od one s neparnim brojem. Jednako kao i kod člana asimetrije teorijska pozadina člana sparivanja leži unutar okvira kvantne fizike. Član sparivanja ima oblik

$$E_p = a_p\frac{\delta(A,Z)}{A^{1/2}} \tag{2.10}$$

$$\delta(A,Z) = \begin{cases} -1 & \text{Z, N parni (A paran)} \\ 0, & \text{A neparan} \\ 1, & \text{Z, N neparni (A paran).} \end{cases} \tag{2.11}$$

Faktor $A^{-1/2}$ nije lako teorijski objasniti, određeni modeli često uzimaju i druge potencije $A^{3/4}, A^{-1}$.



Slika 2.3: Semi-empirijska formula masa za različite parametre a_i. Na povećanoj slici vidljiva su odstupanja različitih modela.

Tijekom godina korišteni su različiti BW modeli čiji su parametri dobiveni prilagodbom na različite setove podataka. Neki od značajnijih prikazani su u tablici 2.1 dok su na slici 2.3 prikazana njihova odstupanja. Unatoč širokoj primjeni BW formule primjećujemo da postoji relativno veliko odstupanja za lake jezgre, naročito za jezgre izvan doline stabilnosti [12].

Tablica 2.1: Lista koeficijenata izraženih u MeV, koeficijenti su izvađeni iz literatura [13], [14].

	a_v	a_s	a_c	a_{asym}	a_p
Bethe and Backer [1936]	13.86	13.20	0.58	19.5	-
Fluegee [1942]	14.66	15.40	0.62	20.5	$33.5/A^{3/4}$
R.D. Evans [1955]	14.30	16.75	0.68	22.6	-
Rohlf [1995]	15.75	17.80	0.71	23.7	$12/A^{1/2}$

2.3 Proširena semi-empirijska formula

Semi-empirijska formula i danas je aktualno polje istraživanja. Postoje brojni znanstveni članci koji pokušavaju opisati problematiku različitim pristupima. Grubom podjelom, članovi koji se pridodaju opisuju energiju vezanja na određenoj skali (mikro i makro modeli). U ovom će se radu uzeti članovi koji rade prepravku BW formule za cijelu energetsku skalu (makro model). Pregled članova koji su najzastupljeniji za opis makroskopskog modela dan je u literaturi [6]. BW formula je tako nadopunjena sa *proširenim površinskim*, *proširenim Coulombovim*, *Wignerovim* članom, *članom zakrivljenosti* i članom koji opisuje utjecaj ljusaka jezgre tj. *efekt ljusaka* (eng. Shell effect).

Prošireni površinski član

Dolazi zbog promjene površinske simetrije. U literaturi se često nalaze nesuglasice oko oblika ovog člana. Do nesuglasica dolazi zato što postoje naznake iz članka [6] da nelinearni efekti u jezgri ne trnu dovoljno brzo da bi bio opravdan razvoj doprinosa na linearne dijelove koji predstavljaju doprinose u najnižim redovima energije. Usprkos postojanju nesuglasica zadržan je doprinos oblika

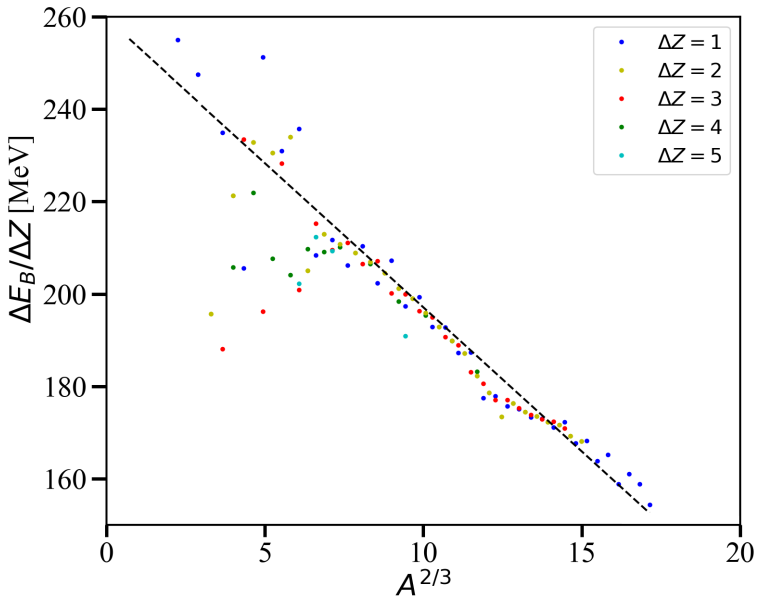
$$E_B = a_{st}\frac{(N-Z)^2}{A^{4/3}.} \tag{2.1}$$

Prošireni Coulombov član

Zrcalnim jezgrama nazivamo jezgre za koje vrijedi $A_1 = A_2 = A$, $N_1 = Z_2$ te $Z_1 = N_2$. Zbog neovisnosti nuklearne sile o naboju, energija vezanja zrcalnih jezgara razlikuje se samo u njihovim Coulombovim doprinosima energiji. Razlika između energija vezanja dviju zrcalnih jezgara je

$$\Delta E_B = a_{coul}\frac{(N^2-Z^2)}{A^{1/3}} \tag{2.2}$$

Na slici 2.1 prikazana je razlika energija zrcalnih jezgara podijeljena s razlikom naboja za 88 izmjerenih zrcalnih jezgara. Različitim bojama označene su različite vrijednosti za ΔZ čiji je iznos bio od 1 do 5. Očekivana vrijednost y odsječka je nula dok rezultati pokazuju odstupanja.



Slika 2.1: Ovisnost $\Delta E_b/\Delta Z$ o $A^{2/3}$.

Član koji popravlja ova odstupanja dolazi od kvantno mehaničke korekcije Coulombovog doprinosa te ima oblik [6]

$$E_B = a_{xc}\frac{Z^{4/3}}{A^{1/3}}. \tag{2.3}$$

Wignerov član

Član asimetrije u originalnoj BW formuli ima izospinsku pozadinu. Predložen je član koji bi dao linearnu proporcionalnost izospinu oblika

$$E_B = a_W\frac{|N-Z|}{A}. \tag{2.4}$$

Član zakrivljenosti

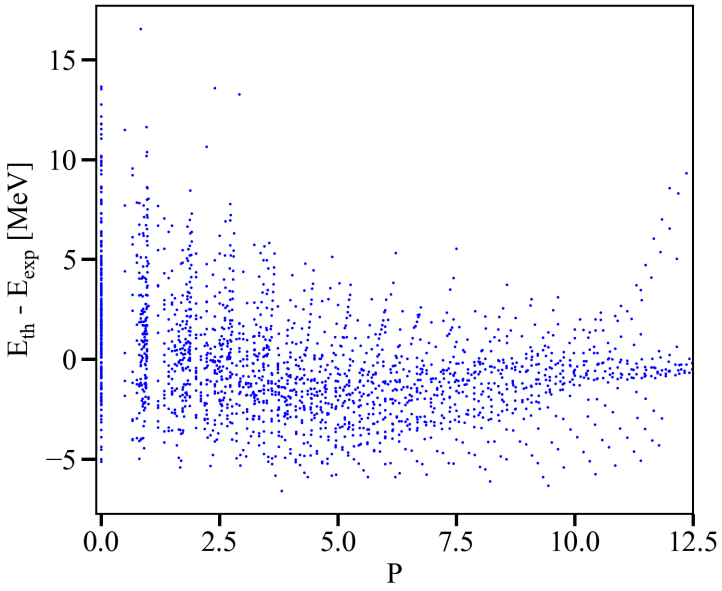
Energija vezanja za velike vrijednosti masenog broja pokazuje trend opadanja koji je proporcionalan radijusu jezgre.

$$E_B = a_R A^{1/3}. \tag{2.5}$$

Efekt ljusaka

Kako uključiti utjecaj ljusaka na energiju vezanja najviše se razlikuju unutar pojedinih modela. U ovom radu uzet će se u obzir *valentan efekt ljusaka*, koji je okarakteriziran valentnom-nukleonskom varijablom

$$P = \frac{\nu_n \nu_p}{\nu_n + \nu_p} \tag{2.6}$$



Slika 2.2: Razlika energija BW modela te podataka u ovisnosti o valentno-nukleonskoj varijabli P.

gdje su ν_p i ν_n brojevi valentnih nukleona. Nuklearni magični brojevi koji se koriste u ovom modelu su 2, 8, 20, 28, 50, 82, 126, 184 za protone i neutrone. Zamijećeno je parabolično ponašanje reziduuma BW formule između vrijednosti nuklearnih magičnih brojeva. Na slici 2.2 vidljivo je grupiranje reziduuma u ovisnosti o P te je primijećeno da dolazi do blagog zakrivljenja s povećanjem vrijednosti P. Na temelju tih opažanja sugerira se doprinos energiji vezanja

$$E_B = a_m P + a_{m1} P.^2 \tag{2.7}$$

Proširena formula

Ukupna proširena BW formula koju ćemo analizirati glasi

$$\begin{aligned} E_B =& a_v A + a_s A^{2/3} + a_c \frac{Z(Z-1)}{A^{1/3}} + \\ &+ a_{asym} \frac{(N-Z)^2}{A} + a_P \frac{\delta(N,Z)}{A^{1/2}} + a_{st} \frac{(N-Z)^2}{A^{4/3}} + a_{xc} \frac{Z^{4/3}}{A^{1/3}} + \\ &+ a_W \frac{|N-Z|}{A} + a_R A^{1/3} + a_m P + a_{m1} P^2. \end{aligned} \tag{2.8}$$

3 Statističko učenje

Statističko učenje predstavlja skup alata iz funkcionalne analize i statistike u svrhu stvaranja odnosa između ulaznih i izlaznih podataka. Pokazalo se da dobro razumijevanje odnosa između podataka omogućava preciznije generiranje algoritama za kasniju primjenu kod strojnog učenja. Širok spektar primjene strojnog učenja povlači sa sobom različite spektre podataka iz kojih model uči. Navođeni različitim problematikama, utemeljena je podjela učenja prema tipu podataka koje koristimo za učenje. Tako imamo dvije grane učenja iz podataka *nadzirano* i *nenadzirano učenje*. Pod nadziranim učenjem smatra se učenje iz označenih podataka. Izlazni podaci povezani su s poznatim ulaznim podacima. Statistički modeli kreirani ovim učenjem često služe za predviđanja ponašanja sustava. Nadzirano učenje koristi se u područjima poput medicine, biologije, astrofizike te ekonomije. Radi se o problemima kod kojih vršimo mjerenja pod utjecajem jednog ili više parametara. U nenadziranom učenju nemamo označene podatke. Cilj ovakvog modela je pronalaženje pravilnosti u strukturama podataka te njihovo razvrstanje u skupine. Ova metoda se često koristi za određivanje nepoznatih parametara u modelu.

3.1 Opis problema i terminologija

Pretpostavimo da imamo *n* mjerenja s *odgovorom*[2] (eng. *output*) *Y* za *p* različitih *prediktora*[3] (eng. *input*), $X = (X_1, ..., X_p)$. Prediktore ćemo u ostatku rada uglavno označavati simbolom *X*, njegove komponente označata ćemo sa X_j. Opservacije ćemo označavati malim slovima. Postoji odnos između odgovora i prediktora koji možemo zapisati kao

$$Y = f(X) + \epsilon \tag{3.1}$$

gdje je f nepoznata funkcija od $(X_1, ..., X_p)$, a ϵ je slučajan član pogreške. Pogreška je ovdje neovisna o prediktoru X te se traži da njezina srednja vrijednost iščezava. Cilj statističkog učenja je prilagođavanje funkcije f(X). Prilagođavanje će ovisiti o prirodi odgovora Y. Grubom podjelom odgovor Y dijeli se na *kvantitativan* i *kvalitativan*. Ako je odgovor kvantitativnog tipa koristi se *regresija*, dok se za kvalitativni odgovor koristi *klasifikacija*. Neovisno o tipu odgovora, predviđanja modela zapisujemo u

[2]U literaturi često se koriste nazivi *izlazne* ili *zavisne* varijable.
[3]Često korišteni nazivi su *ulazne* ili *nezavisne* varijable.

obliku

$$\hat{Y} = \hat{f}(X) \tag{3.2}$$

gdje je $\hat{f}$ prilagodba f, a $\hat{Y}$ predstavlja rezultirajući odgovor za Y. Preciznost predviđanja $\hat{Y}$ ovisi o reducibilnoj i ireducibilnoj pogrešci. Reducibilna pogreška smanjuje se s povećavanjem preciznosti predviđanja, dok ireducibilna pogreška uvijek postoji u mjerenju što se vidi iz jednadžbe (2.1). Njena vrijednost je neovisna o prediktorima te ne može biti smanjena s boljom prilagodbom za funkciju f. Veće vrijednosti ϵ mogu biti naznaka da u sustavu postoje skrivene varijable.

3.2 Linearne metode za regresiju

Razumijevanje linearnih metoda od velike je praktičnosti za kasnije nelinearne metode. Radi se o metodama koje se koriste kod nadziranog učenja za predviđanje kvantitativnog odgovora Y. Skup takvih metoda naziva se linearna regresija.

3.2.1 Linearna regresija i metoda najmanjih kvadrata

Predikcijska funkcija za linearan model zapisuje se kao

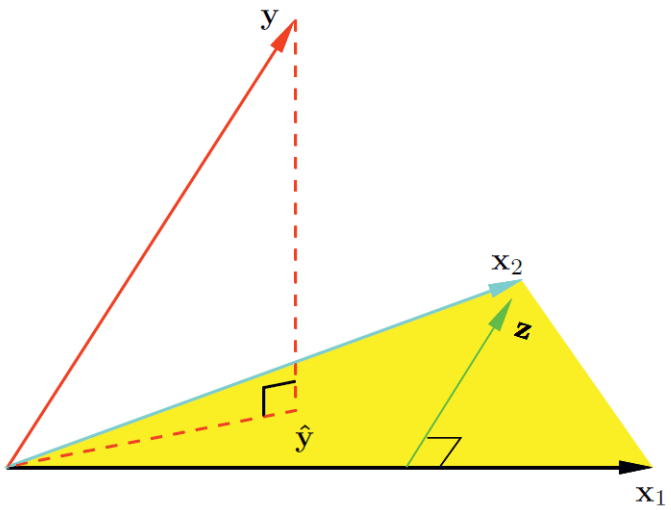
$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{3.3}$$

gdje su β_j [4] nepoznati parametri (koeficijenti). Nevezano za prirodu prediktora X_j model je linearan u parametrima.
Iz tipičnog seta podataka oblika $(x_1, y_1) \dots (x_N, y_N)$ određuju se parametri β_j. Vektor $x_i = (x_1^{(i)}, \dots, x_p^{(i)})$ predstavlja svojstvo mjerenja za i-ti slučaj. Najpoznatija metoda prilagodbe je *metoda najmanjih kvadrata*. Koeficijenti $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ biraju se tako da se minimizira suma kvadrata eng. *Residual Sum of Squares*, RSS

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 . \end{aligned} \tag{3.4}$$

[4] $\beta_j = a_j$, a_j je uobičajena oznaka za parametare prilagodbe kad se nalaze kontekstu energije vezanja.

Kako bi lakše minimizirali vrijednost RSS jednadžbu (3.4) zapisujemo pomoću N $\times$



Slika 3.1: Prikaz N-dimenzionalne geometrije za metodu najmanjeg kvadrata s dva prediktora. Vektor $\hat{\mathbf{y}}$ predstavlja ortogonalnu projekciju na površinu razapetu s ulaznim vektorima $\mathbf{x}_1$ i $\mathbf{x}_2$ [9].

(p+1) matrice $\mathbf{X}$ te ulaznog N-dim. vektora $\mathbf{y}$ [9]

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta). \tag{3.5}$$

Jednadžbu (3.5) deriviramo kako bi pronašli njen minimum

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \tag{3.6}$$

nadalje zahtijevamo da

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0. \tag{3.7}$$

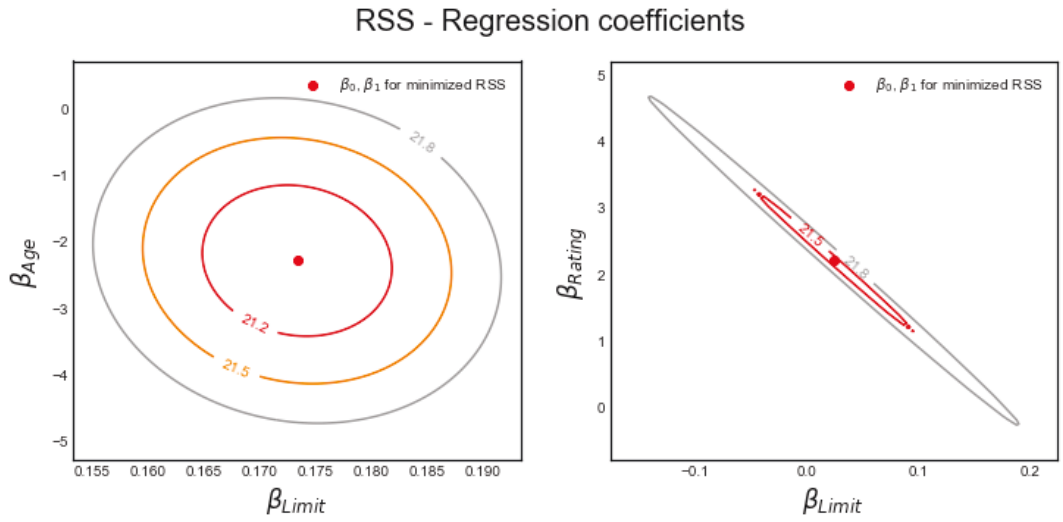
Iz jednadžbe (3.7) slijedi da je

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.8}$$

Uvrštavanjem u početnu funkciju dobivamo [9]:

$$\hat{\mathbf{y}} = \mathbf{X}\beta = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}.^T \tag{3.9}$$

Ako je $\mathbf{X}^T\mathbf{X}$ singularna, koeficijenti metode najmanjih kvadrata $\hat{\beta}$ nisu jedinstveni. Do singularnosti dolazi ako se pojave kolinearni ulazni vektori ($x_i \propto x_j$ za $i \neq j$).



Slika 3.2: ”Contour plot” za različite vrijednosti RSS prikazanih preko funkcijske ovisnosti o koeficijentima predikcije β. Crvenom točkom prikazana je najmanja vrijednost za RSS [10].

Kolinearnost vektora vodi uskoj povezanosti između parametara koje predviđamo. Razlučivanje doprinosa parametara važno nam je kako bismo uočili njihove pojedinačne doprinose odgovoru Y.

Slika 3.2 prikazuje ”contour plot” odabranih parametra β iz modela. Lijevo, parametri nisu kolinearni, različite vrijednosti parametra daju različite vrijednosti RSS-a. Minimum RSS je dobro definiran, dok to ne vrijedi za slučaj prikazan na desnoj slici. Ovdje mala promjena vrijednosti podataka uzrokuje promjenu u paru vrijednosti parametara $(\beta_{Limit}, \beta_{Rating})$, dok RSS ostaje približno jednak.

3.3 Selekcija podskupa i regularizacije

Kod metode najmanjih kvadrata dolazi do problema, ako je broj opservacija n približno jednak broju varijabla p dolazi do *preprilagođenosti* (eng. overfiting). To je situacija u kojoj model preslikava slučajni šum u funkcijsku ovisnost, model je izbubio mogućnost prilagodbe na novi skup podataka. Ako je $p > n$, u modelu ne postoji jedinstven set koeficijent te metoda nije uopće primjenjiva. Problem se može riješiti smanjenjem broja parametara ili njihovim ograničavanjem. U ovom poglavlju predstavit ćemo tri pristupa koja omogućavaju uporabu metode najmanjih kvadrata, a to su *sažimanje parametara, selekcija podskupa* te *metoda redukcije dimenzionalnosti*.

3.3.1 Sažimanje parametara

Uvijek možemo uzeti model koji će sadržavati svih p prediktora tako da koeficijente prilagodbe sažmemo ili regulariziramo. Provođenjem ovog postupka koeficijente smanjujemo prema nuli što rezultira smanjenjem varijance sistema. Dvije najpoznatije tehnike provođenja ove metode su *LASSO* i *hrbat-regresija*.

Hrbat-regresija

Hrbat-regresija[5] (eng. ridge regression) jednako kao i metoda najmanjeg kvadrata teži smanjenju rezidualne sume kvadrata RSS, samo je veličina koju želimo minimizirati nešto drugačija. Drugim riječima, koeficijenti prilagodbe hrbat regresije $\hat{\beta}^R$ su vrijednosti koje minimiziraju

$$RSS + \alpha \sum_{j=1}^{p} \beta_j^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \alpha \sum_{j=1}^{p} \beta_j^2 \tag{3.10}$$

gdje je $\alpha \geqslant 0$ *parametar podešavanja (eng. tuning parameter)*. Veličina $\alpha \sum_{j=1}^{p} \beta_j^2$ eng. *shrinkage penalty* ima efekt smanjenja koeficijenata prilagodbe k nuli. Parametar podešavanja α služi kao kontrola na relativan učinak smanjenja. Ako je $\alpha = 0$ član $\alpha \sum_{j=1}^{p} \beta_j^2$ nema nikakvog utjecaja te dolazi do redukcije jednadžbe 3.10 na jednadžbu 3.4. Ukoliko $\alpha \to \infty$ utjecaj smanjenja je velik te $\hat{\beta}^R \to 0$. Usporedno s metodom najmanjih kvadrata, hrbat regresija ne sadrži jedinstven skup rješenja za koeficijente prilagodbe β. Članovi $\hat{\beta}^R$ su određeni za svaku vrijednost α. Važno je napomenuti da je član β_0 izostavljen iz "penalty" člana te je time izbjegnuta ovisnost odgovora Y o izboru ishodišta. Konačna rješenja za koeficijente prilagodbe možemo dobiti ako zapišemo (3.10.) pomoću **X, y**

$$RSS = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \alpha\beta^T\beta. \tag{3.11}$$

Iz jednadžbe (3.11) slijedi da su rješenja

$$\hat{\beta}^R = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.12}$$

[5]Kako je znanost o podacaima relativno novo polje istraživanja, ne postoje službeni prijevodi engleskih naziva metoda. Prijevod je preuzet iz literature [15].

gdje je $\mathbf{I}$ $p \times p$ jedinična matrica. Iz jednadžbe (3.12) možemo vidjeti jedno pogodno svojstvo hrbat metode, problem je prestao biti nesingularan zato što je $\mathbf{X}^T\mathbf{X}$ dodana pozitivna konstanta.

LASSO regresija

LASSO (eng. Least Absolute Shrinkage and Selection Operator) je metoda sažimanja parametara. Promotrimo li hrbat-regresiju, naš model će uvijek sadržavati svih p prediktora. Koeficijenti prilagodbe nikad neće biti točno nula već će samo težiti k nuli. Ovo naizgled ne predstavlja nikakvo ograničenje na preciznost samog modela već otežava njegovu interpretaciju. Uzmimo kao primjer model koji sadrži poveći broj prediktora *p*. Nekolicina parametara ima značajniji doprinos u odnosu na druge parametre u ukupnom odgovoru Y. Pristupimo li ovakvom modelu s hrbat regresijom morat ćemo interpretirati važnost svih p prediktora jer će oni u konačnom modelu biti prisutni. Povećanjem varijable α reducirat ćemo vrijednosti konačnih koeficijenata no oni neće biti isključeni iz konačnog rezultata. Stoga pristupimo problemu na malo drugačiji način te jednadžbu (3.4) proširujemo sa sumom norme

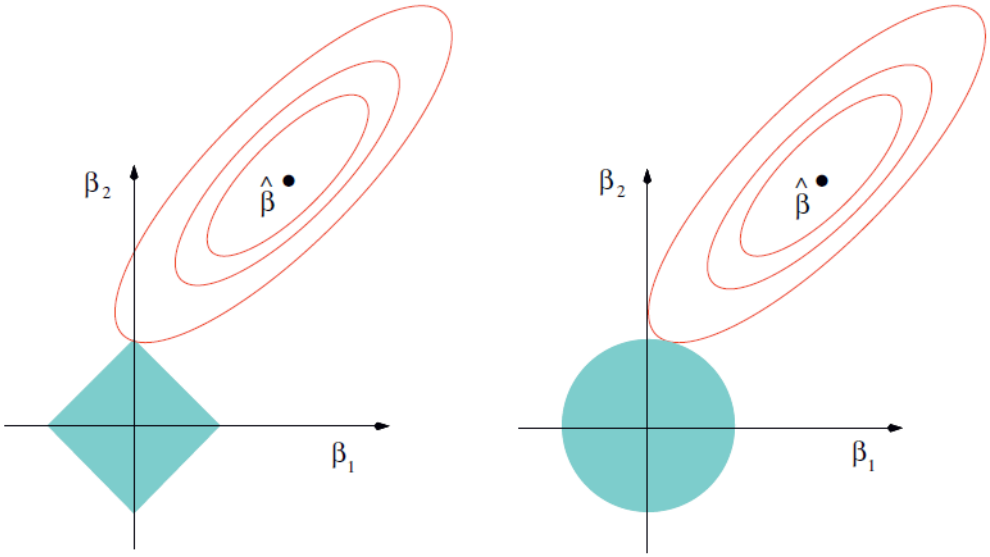
$$RSS + \alpha \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \tag{3.13}$$

što je definicija LASSO prilagodbe. Koeficijenti $\hat{\beta}^L$ minimiziraju veličinu (3.13.) te ih nazivamo *LASSO koeficijenti*. Matematički zapisano kao izraz

$$\hat{\beta}^L = \min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.14}$$

Naizgled manja promjena, u odnosu na hrbat-regresiju, čini rješenje nelnearnim u y_i te stoga zapis u zatvorenoj formi nije moguć. Zbog preglednijeg razumijevanja metoda, ilustriramo model prikazan na slici 3.3. Radi se od jednostavnom modelu s dva prediktora te sukladno tome imamo i dva regresijska koeficijenta β_1 i β_2.
Rješenje dobiveno metodom najmanjih kvadrata prikazano je sa $\hat{\beta}$. Plavom bojom prikazana su rješenja LASSO i hrbat-regresije s ograničenjem, tj.

$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \qquad \text{za} \qquad \sum_{j=1}^{p} |\beta_j| \leqslant s \tag{3.15}$$



Slika 3.3: Ilustracija LASSO (lijevo) i hrbat-regresije (desno). Crnom točkom prikazana je najmanja vrijednost RSS-a dobivena metodom njamanjih kvadrata. Crveno su prikazane funkcijske ovisnosti koeficijenata predikcije β za određenu vrijednost RSS-a. Oragničenja hrbat, $\beta_1^2 + \beta_2^2 \leqslant s$, te LASSO, $|\beta_1| + |\beta_2| \leqslant s$ su prikazana plavom bojom [10].

te

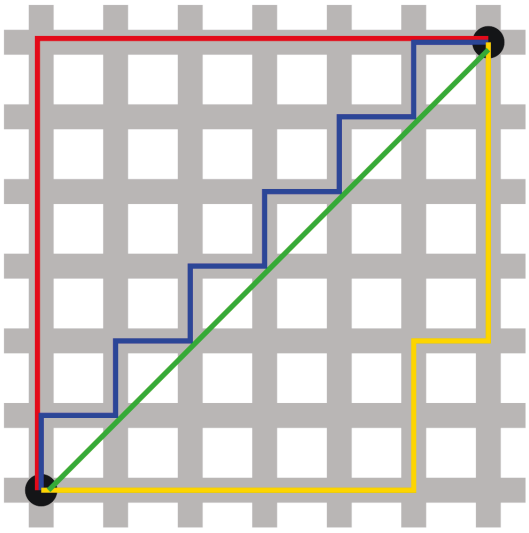
$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{za} \quad \sum_{j=1}^{p} \beta_j^2 \leqslant s. \tag{3.16}$$

Crveno označene elipse centrirane oko $\hat{\beta}$, predstavljaju područja konstantne vrijednosti RSS, sve točke dane elipse predstavljaju parove vrijednosti $\beta_{1,2}$ koji imaju jednak RSS. Ako je s dovoljno velik rješenja za regularizacije sadržavat će $\hat{\beta}$, što znači da će se regularizacije svesti na metodu najmanjih kvadrata. Iz jednadžbi 3.15 i 3.16 vidimo da su rješenja LASSO i hrbat-regresije one točke za koje elipse prve dotaknu regularizacijski ograničeno područje. Ograničenje hrbat-regresije je kružnog oblika, posljedica odabira $\beta_1^2 + \beta_2^2 \leqslant s$, te sugerira da točka dodira generalno neće biti na osima. U primjeni to znači da koeficijenti neće biti jednaki nuli. LASSO ograničenje ima kvadratičan oblik, $|\beta_1| + |\beta_2| \leqslant s$, što znači će točka dodira biti pretežito na osima. Na slici je prikazano da je točka dodira $(0, |\beta_2|)$ te rezultirajući model sadrži samo prilagodbu β_2.

Važna razlika između regularizacija očitava se i njihovim algoritmima koji služe za određivanje parametara. Algoritam pronalazaka vrijednosti parametara je problem pronalaska norme vektora **X**

$$\mathbf{X} = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Za vrijednost $p = 1$ imamo L1 normu tzv. *taxicab normu*[6] što je osnova algoritma LASSO regularizacije. "Taxicab" norma opisana je pomoću "taxicab" geometrije. U "taxicab" geometriji funkcija udaljenosti zamijenjena je novom metrikom u kojoj je udaljenost dviju točaka suma razlike apsolutnih vrijednosti njihovih kartezijskih kordinata.



Slika 3.4: Udaljenost dviju točaka promotrena u Euklidskoj i texicab geometrijom. Crvenom, plavom i žutom bojom označene su minimalne udaljenosti između označenih točaka u "taxicab" geometriji. Minimalna udaljenost točaka u Euklidskom prostoru označena je zelenom bojom.

[6]U nekim literaturama pod nazivom Manhattan norma/udaljenost.

3.3.2 Odabir podskupa varijabli

Pod odabirom podskupa varijabli zapravo smatramo odabir podskupa prediktora x_i koji će biti uključeni u model. Odabir se vrši na temelju izračunatog RSS-a, koji sadrži informaciju o varijabilnosti našeg podskupa na odgovor y. U interesu nam je odabrati najmanji podskup koji pokazuje najveći utjecaj na varijablu odgovora.
Postoje generalno dva razloga zbog kojih se odlučujemo na metode odabira podskupa varijabli. Prvi razlog je smanjenje parametarskog prostora. Smanjenjem parametarskog prostora dobivamo na interpretabilnosti modela. Žrtvujući preciznost modela podskup od k<p prediktora može sadržavati i samo nekoliko članova koji će nam davati širu sliku modela. Drugi razlog je samo izvršenje metode najmanjih kvadrata. Metoda najmanjih kvadrata obično daje model visoke varijance i niske pristranosti. Modeli visoke varijance, ukoliko ne znamo očekivan rezultat, daju kriva predviđanja o čemu će biti više riječi u poglavlju 3.4.
Postoje tri metode odabira podskupa: *odabir najboljeg podskupa*, *odabir unaprijed* te *odabir unazad*.

Odabir najboljeg podskupa

Metoda odabira najboljeg podskupa za svaki k<p, vraća podskup veličine k, koji ima najmanju predikcijsku pogrešku. Metoda najboljeg podskupa postaje nepraktična za velike p. Ukupan broj modela sa *p* prediktora je 2^p. Za određen *k* provjeravamo sve mogućnosti kojih je

$$\binom{p}{k}.$$

Iz čega je vidljivo da za $p > 20$ ova metoda postaje tehnički i vremenski neisplativa. Metoda ima i prednosti, a to je da podskup veličine k+1 ne mora sadržavati najbolji podskup varijabli iz podskupa veličine *k*.

Odabir unaprijed i unazad

Metoda odabira unaprijed (eng. Forward Stepwise Selection) te odabir unazad (eng. Backward Stepwise Selection) alternativan su izbor ukoliko sustav sadrži veći broj prediktora p. Metode, u odnosu na metodu najboljeg podskupa, pružaju suboptimalno rješenje. Kod odabira unaprijed počinjemo sa slobodnim članom. Modelu, u svakom novom koraku iteracije, pridodajemo po jednu nezavisnu varijablu tako da

svaka varijabla koju dodamo, od svih preostalih varijabli, najviše smanji pogrešku. Model veličine k+1 sadrži sve varijable koje su bile prisutne u podskupu veličine k. Princip metode odabira unazad jednak je kao i kod metode odabira unaprijed. Metode se razlikuju jedino u početnom podskupu, pa metoda unazad počinje od punog modela. Ukupan broj modela je

$$\sum_{k=0}^{p-1}(p-k) = \frac{p(p+1)}{2} \tag{3.17}$$

što predstavlja značajnu razliku u odnosu na metodu najboljeg podskupa gdje je za $p = 20$ potrebno prilagoditi $2^p = 1,048,576$ modela. Metoda unaprijed ili unazad zahtijeva prilagodbu od samo 211 modela.

3.3.3 Metode redukcije dimenzionalnosti

Do sada sve metode koje smo naveli koriste puni skup orginalnih prediktora, $X_1, \ldots, X_p$. Metode koje transformiraju početne prediktore te na transformiranim prediktorima upotrebljavaju metodu najmanjih kvadrata se nazivaju *metode redukcije dimenzionalnosti*.

Neka za $M < p$, $Z_1, Z_2, \ldots, Z_M$ predstavljaju linearnu kombinaciju prediktora p. Tako da

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{3.18}$$

gdje su $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$ za $m = 1, \ldots, M$ konstante. Korištenjem metode najmanjeng kvadrata možemo fitati linearni regresijski model oblika

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i \tag{3.19}$$

gdje su θ_i novi regresijski koeficijenti. Odgovarajućim odabirom $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$ imamo bolji model od metode najmanjih kvadrata. Naziv dimenzionalna redukcija dolazi od činjenice da model koji sadrži p prediktora te p+1 koeficijenata prilagodbe β svodi na sličan problmem sa M prediktora sa M+1 koeficijentom. Možemo zapaziti da

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij} \tag{3.20}$$

gdje je

$$\beta_j = \sum_{j=1}^{M} \theta_m \phi_{jm}. \tag{3.21}$$

Za probleme gdje je p relativno velik u odnosu na *n*, odabir $M << p$ može znatno reducirati varijancu fitanih koeficijenata. Sve metode redukcije dimenzionalnosti rade u dva koraka. Prvi korak je dobivanje $Z_1, \ldots, Z_M$ transformacijama, dok nakon njega slijedi prilagodba modela uporabom M prediktora. Odabir $Z_1, \ldots, Z_M$ ekvivalentan je odabiru ϕ_{jm}. Postoje različite metode odabira. Jedna od najzastupljenijih metoda je *analiza principalnih komponenata* (eng. *principal components analysis, PCA*).

3.3.4 Analiza principalnih komponenata

PCA je ortogonalna linearna transformacija podataka u novi koridinatni sustav čije su osi definirane projekcijama podataka duž osi najveće varijance. Prva principalna komponenta, x os novog koordinatnog sustava, predstavlja os duž koje prediktori imaju najveću varijancu. Neka je $\mathbf{X}$ $n \times p$ matrica podataka. PCA je spektralna dekompozicija matrice $\mathbf{X}^T\mathbf{X}$, gdje kao rezultat dobivamo svojstvene vektore w_i te svojstvene vrijednosti λ_i. Ključ PCA je poredak svojstvenih vrijednosti λ_i po vrijednosti od najveće prema najmanjoj. Jasnije je, zapiše li se dekompoziciju matrice $\mathbf{X}$ kao

$$\mathbf{T} = \mathbf{XW} \tag{3.22}$$

gdje je $\mathbf{W}$ $p \times p$ matrica svojstenih vektora posloženih tako da prvi stupac matrice odgovara najvećoj vrijednosti λ. Prvi stupac matrice $\mathbf{W}$ je tražena prva principalna komponenta.
Odabirom samo prvih $m < p$ principalnih komponenata smanjujemo dimenziju matrice W. Transformaciju zapisujemo

$$T_m = XW_m \tag{3.23}$$

gdje je dimenzija matice T_m sada $n \times m$.

3.4 Ocjena i odabir modela

Cilj statističkog učenja nije samo pronaći što bolju aproksimaciju $\hat{f}(x)$ funkcije f(x) već stvoriti kvalitetan model koji će predvidjeti ponašanje sustava na novom skupu podataka. Ovdje govorimo o sposobnosti generalizacije modela. Provjera kvalitete modela obično se vrši tako da se početni podaci razdijele na dva podskupa. Skup podataka za trening, na kojima se vrše selekcije te regularizacije i skup za testiranje koji nam služi kao ocjena našeg treniranog modela.
U slučaju regresijskih metoda često se koristi eng. *mean squared error* (MSE), definiran sa

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \tag{3.24}$$

gdje je $\hat{f}(x_i)$ predviđanje koje $\hat{f}$ daje za i-to opažanje. MSE vrijednost se smanjuje kako razlika između predviđanja i prave vrijednosti opada.

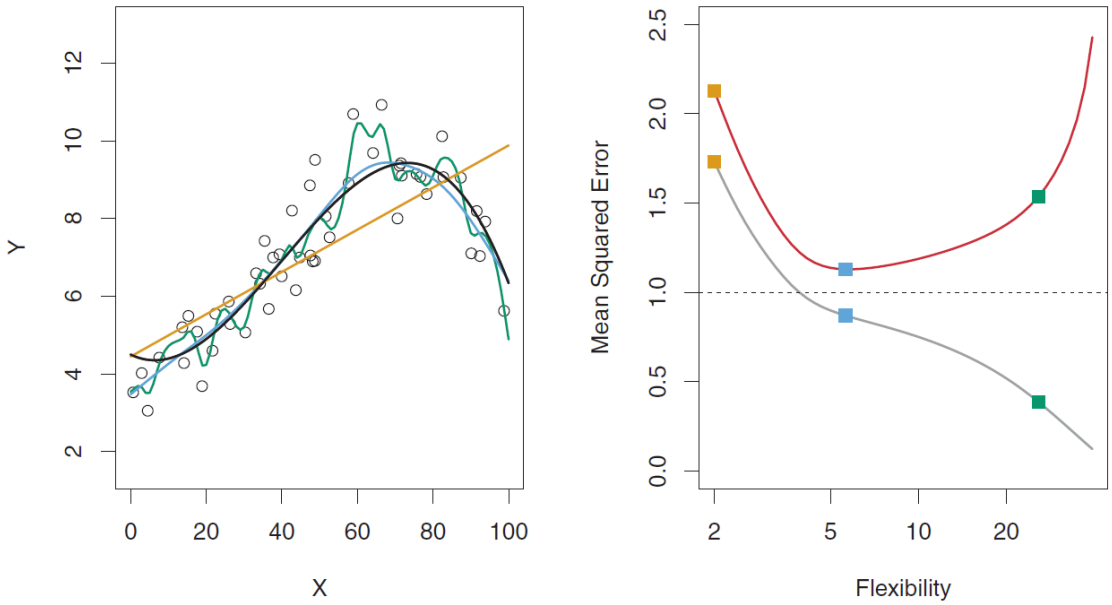
MSE definiran u jednadžbi 3.24 koristi se za izračun pogreške na skupu za testiranje, jednako kao i na skupu podataka za treniranje. Za bolju generalizaciju modela, interes nam je smanjiti MSE izračunatog na skupu za testiranje. Ako je priroda podataka takva da nam nije dostupan skup za testiranje, predviđanja se rade u korist smanjenja MSE vrijednosti na cjelokupnom skupu podataka. Porastom kompleksnosti, trening MSE pada, no to nije nužno i slučaj za test MSE. Trenutak kada trening daje najmanji MSE, a test set ima najveću vrijednost MSE kažemo da je došlo do preprilagođenosti podataka što je prikazano na slici 3.5. U praksi se često koristi i *root mean squared error* (RMS) koji je definiran kao drugi korijen MSE. Kvaliteta modela često se provjerava s koeficijentom determinacije R^2 (eng. coefficient of determination, R squared). R^2 definiran je sa

$$R^2 = \frac{RSS}{SS_{TOT}} \tag{3.25}$$

gdje je RSS definiran sa 3.4, a SS_{TOT} je ukupna suma kvadrata (eng. total sum of squares)

$$SS_{TOT} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{3.26}$$

gdje je $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ srednja vrijednost.
Vrijednost koje R^2 može poprimiti su unutar intervala $0 < R^2 < 1$. R^2 je često izražen



Slika 3.5: Crnom bojom su prikazani podaci generirani pomoću funkcije f, narančastom bojom prikazana je linearna regresijska linija, dok su plavom i zelenom prikazane regresije pomoću više polinoma. Desno: Sivom bojom je prikazan trening MSE, test MSE je prikazan crvenom bojom. Primjećuje se opadanje vrijednost trening MSE i test MSE do vrijednosti označene plavim kvadratićem koja predstavlja prilagodbu na lijevoj slici označenu plavom bojom. Nakon te vrijednosti dolazi do preprilagođenosti i model gubi na kvaliteti generalizacije (porast vrijednosti test MSE-a.) [10]

pomoću veličine VIF (eng. Variance Inflation Factor) koja je dana sa

$$\text{VIF} = \frac{1}{1 - R^2}. \tag{3.27}$$

VIF služi za provjeru kolinearnosti uz opću smjernicu da je za vrijednost VIF>10 kolinearnost koeficijenata velika.

U odabiru modela, mogu nam pomoći veličine AIC (eng. Akaike information criterion), BIC (eng. Bayesian information criterion) te podešen R^2 (eng. adjusted R^2).

Neka je *k* broj prilagođenih parametara u modelu te $\hat{L}$ maksimalna vrijednost funkcije vjerojatnosti za model. AIC je dan jednadžbom

$$\text{AIC} = 2k - 2ln(\hat{L}). \tag{3.28}$$

Kriterij BIC je dan

$$\text{BIC} = ln(n)k - 2ln(\hat{L}). \tag{3.29}$$

3.4.1 Pristranost, varijanca i kompleksnost

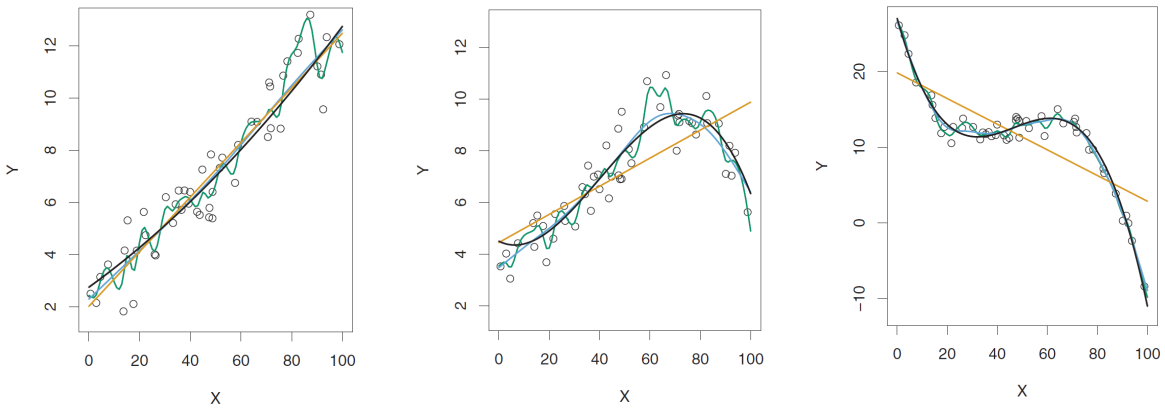
Može se pokazati da se za danu vrijednost x_0 MSE raspisuje na tri fundamentalna doprinosa [9]

$$E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon) \tag{3.30}$$

gdje je $\text{Var}(\hat{f}(x_0))$ varijabilnost funkcije $\hat{f}(x_0)$, $[\text{Bias}(\hat{f}(x_0))]^2$ je kvadrat pristranosti (eng. bias), a $\text{Var}(\epsilon)$ je varijabilnost pogreške ϵ. Varijabilnost nam govori koliko se varijabla $\hat{f}$ mijenja sa promjenom x. Promotrimo prijašnji primjer prikazan na slici 3.5, narančasto označen kvadratić točka je niske varijabilnosti, dok je zeleno označena točka visoke varijabilnosti. Visoka varijabilnost odgovara fleksibilnijoj funkciji predviđanja $\hat{f}$.

Pristranost u modelu se javlja zbog aproksimacije koje se radi. Pretpostavka linearnog odnosa između Y i X_i za $i = 1, \ldots, p$ će imati tako utjecaja na predviđanje $\hat{f}$.

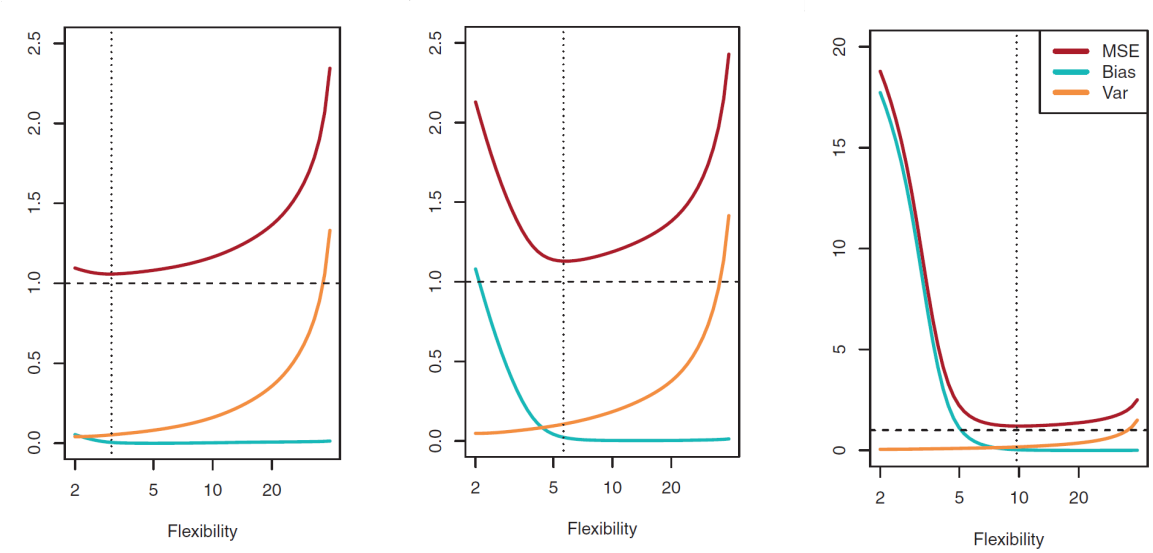
Vrijedi generalno pravilo, koristimo li fleksibilnije modele, varijabilnost će rasti dok će pristranost opadati. Modelu prikazanom na slici 3.5 pridodat ćemo još dva generirana skupa podataka. Jedan koji opisuje više linearno ponašanje podataka i drugi koji je udaljeniji od linearnog ponašanja. Sva tri skupa, poredana po linearnosti generiranih podataka prikazana su na slici 3.6. Promotrimo vrijednosti testne pogreške,



Slika 3.6: Prikaz generiranih podataka i njihovih točnih vrijednosti (crna boja) od linearnijeg do najmanje linearnog ponašanja. Predikcijski modeli prikazani su naračastom, plavom i zelenom bojom ovisno o fleksibilnosti predikcije [10].

prikazane na slikama 3.7. U sva tri slučaja varijabilnost raste u korist pristranosti sa povećanjem fleksibilnosti modela. Ukoliko model ima veću tendenciju ka linearnosti početni iznos pristranosti će biti manji. Manje linearni modeli će imati znatno

veće iznose pristranosti koje će značajno padati sa povećanjem fleksibilnosti modela. Za odabir optimalnog modela potrebno je pronaći *ravnotežu pristranosti i varijabilnosti* (eng. bias-variance trade off). Pronalazak optimalne ravnoteže varijabilnosti i



Slika 3.7: Vrijednost testnog MSE rastavljenog na doprinose od varijabilnosti i pristranosti prikazano od linearnijeg do najmanje linearnog ponašanja generiranih podataka. Isprekidanim linijama prikazani su položaji ravnoteže varijabilnosti i pristranosti za sva tri slučaja [10].

pristranosti najvažniji je zadatak koji se treba riješiti za postizanje optimalnog opisa podataka pomoću statističkog učenja. Na slici 3.7. ravnoteža varijabilnosti i pristranosti prikazana je vertikalnom linijom.

3.4.2 Unakrsna provjera

U poglavlju ranije opisana je razlika i važnost između test i trening pogreške, no u realnosti često ne možemo skup podataka razdijeliti na podatke za trening i test. Jedna od metoda koja direktno procjenjuje testnu pogrešku je *unakrsna provjera* (eng. cross-validation, CV). Kod CV metode ne dolazi do razdvajanja podataka na skupove, već se svako mjerenje ”reciklira”. Podatci se dijele na *k* jednakih dijelova (eng. k-folds), gdje *k-1* dio koristimo za treniranje,a jedan podskup podataka se ostavlja kao test. Neka je κ indeksna funkcija $\kappa : \{1, \ldots, N\} \to 1, \ldots, K$ koja za svaki indeks i točke vraća indeks podskupa u kojem se nalazi, $\kappa(i) = k$. Neka je $\hat{f}^{-k}$ model treniran na svim podskupovima osim k-tom. Procjena greške predviđanja je

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa}(x_i)). \tag{3.31}$$

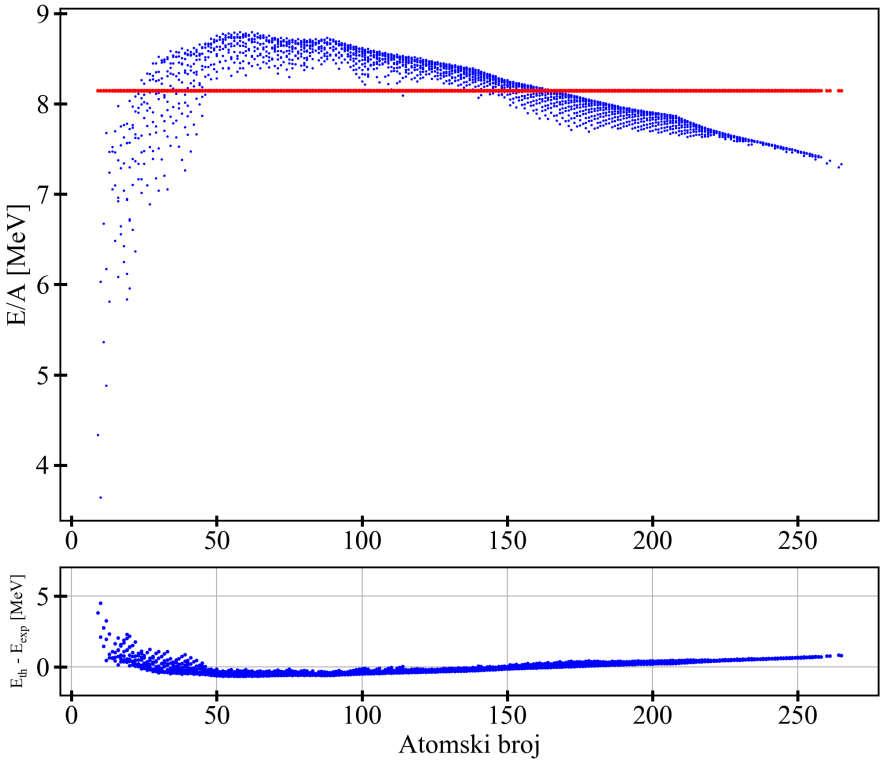
Ostaje otvoreno pitanje odabira K particija. Za vrijednost $K = N$, skupovi za trening su međusobno slični te rezultantni model ima visoku varijabilnost. S malim vrijednostima K metoda može imati pristranost. K biramo ovisno o setu podataka imajući na umu opisana ograničenja. Literatura [9] sugerira odabir K između 5-20.

4 Statistička analiza Bethe-Weizsäckerove formule

Cilj statističke analize je prikupiti što više informacija iz podataka kako bi mogli doći do zaključaka koji bi šire opisivali fiziku semi-empirijske formule. Krećemo od analize članova iz Bethe-Weizsäckerove formule danom izrazom 2.4. Konstante a_i predstavljaju parametre koje je potrebno odrediti na temelju eksperimentalnih podataka. Podaci koje upotrebaljvamo preuzeti su sa stranica Atomic Mass Data Centra [11]. Eksperimentalno je određena energija vezanja za 2497 jezgara. Podaci su dobiveni različitim metodama te su u sklopu Atomic Mass Evaluation 2016 (AME2016) prikupljeni u standardiziranom formatu.

Analizi semi-empirijske formule pristupamo tako da krećemo od prilagodbe na konstantan član, te postepeno dodajemo ostale članove iz osnovnog modela. Prije primjene prilagodbe, podatke dijelimo sa masenim brojem kako bi dobili energiju vezanja po nukleonu. Statistička analiza je točnija ukoliko se prilagodba radi unutar energija koje su istog reda veličine.

Prvi član koji pridodajemo je volumni doprinos $E_B/A = a_{vol}$.

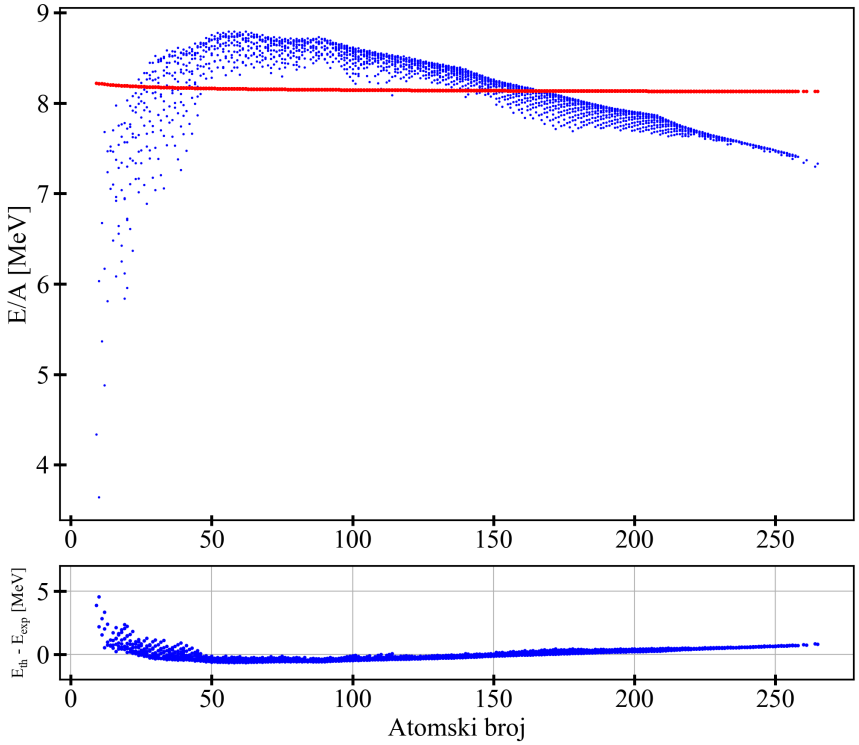


Slika 4.1: Energija vezanja po nukleonu u ovisnosti o masenom broju. Plavom bojom su prikazani podaci, crvenom bojom prikazana je prilagodba za konstantan volumni član koji iznosi $a_{vol} = 8.145(9)$ MeV-a.

Očekujemo da vodeći član energije vezanja bude linearan s obzirom na maseni broj A. Dobiveni rezultati su prikazani na slici 4.1 te tablici 4.1 za N=1.

Prilagodba odabire vrijednost parametra a_{vol} na način da opiše najviše jezgara, što rezultira prosječnom vrijednošću energije vezanja po nukleonu od skoro 8 MeV.

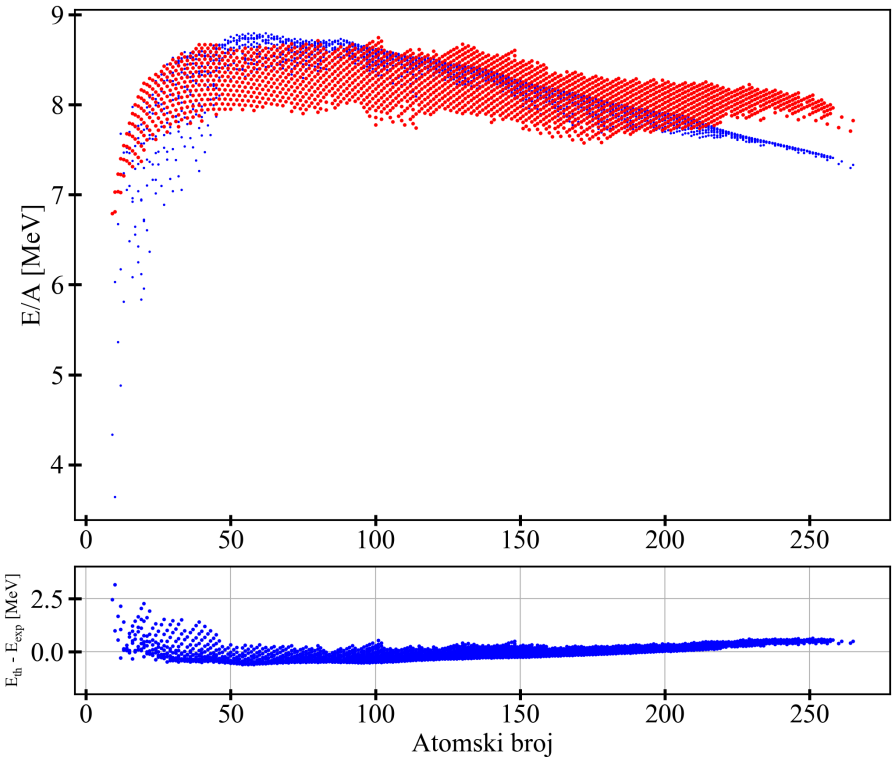
Slijedeći član iz BW formule je doprinos površine, teorijski očekujemo da će doći do smanjenja energije vezanja. Na slici 4.2 prikazana je prilagodba $E_B/A = a_{vol} + a_{surf}A^{-1/3}$.



Slika 4.2: Energija vezanja po nukleonu u ovisnosti o masenom broju. Plavom bojom su prikazani podaci, crvenom bojom prikazana je prilagodba za volumni i površinski doprinos koji iznose $a_{vol} = 8.08(4)$ MeV, $a_{surf} = 0.3(2)$ MeV.

Drugi član prilagodbe opisuje smanjenje energije vezanja zbog površinske napetosti, no u ovom modelu je pozitivan, što nije teorijski očekivano. Volumni član prilagodbe je konstantan tj. linearan s obzirom na maseni broj, dok drugi član ne može opisati energije vezanja na visokim i niskim A. Da bi se spustila suma kvadrata odstupanja prilagodba opisuje ponašanje jezgara kojih ima brojno više.

Na slici 4.3 prikazana je prilagodba $E_B/A = a_{vol} + a_{surf}A^{-1/3} + a_c\frac{Z(Z-1)}{A^{4/3}}$.



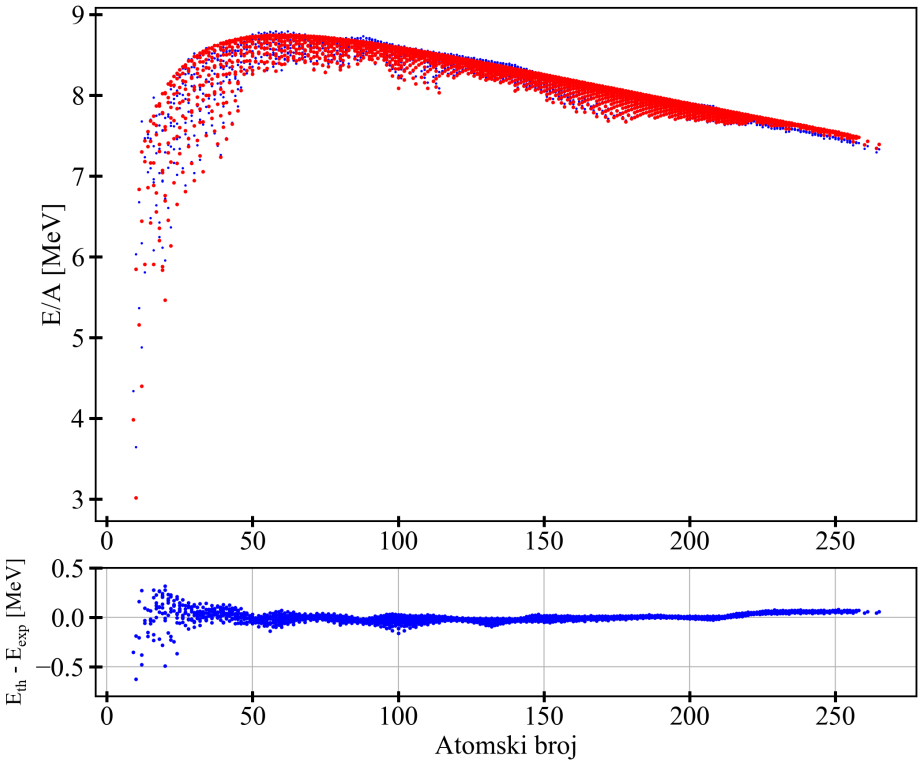
Slika 4.3: Energija vezanja po nukleonu u ovisnosti o masenom broju. Plavom bojom su prikazani podaci, crvenom bojom prikazana je prilagodba za volumni, površinski i Coulombov doprinos. Iznosi parametara zapisani su u tablici 4.1.

Dodavanje Coulombovog člana energiji vezanja nije doprinjelo značajnijem smanjenju pogreške modela, ali je predznak a_{surf} popravljen. Funkcija prilagodbe više nije jednoznačna. Model sugerira da za jezgre sačinjene samo od neutrona postižu najvišu energiju vezanja, što nije fizikalna realnost. Dodavanjem člana asimetrije prilagodbi otklanjamo taj problem, što je prikazano na slici 4.4.

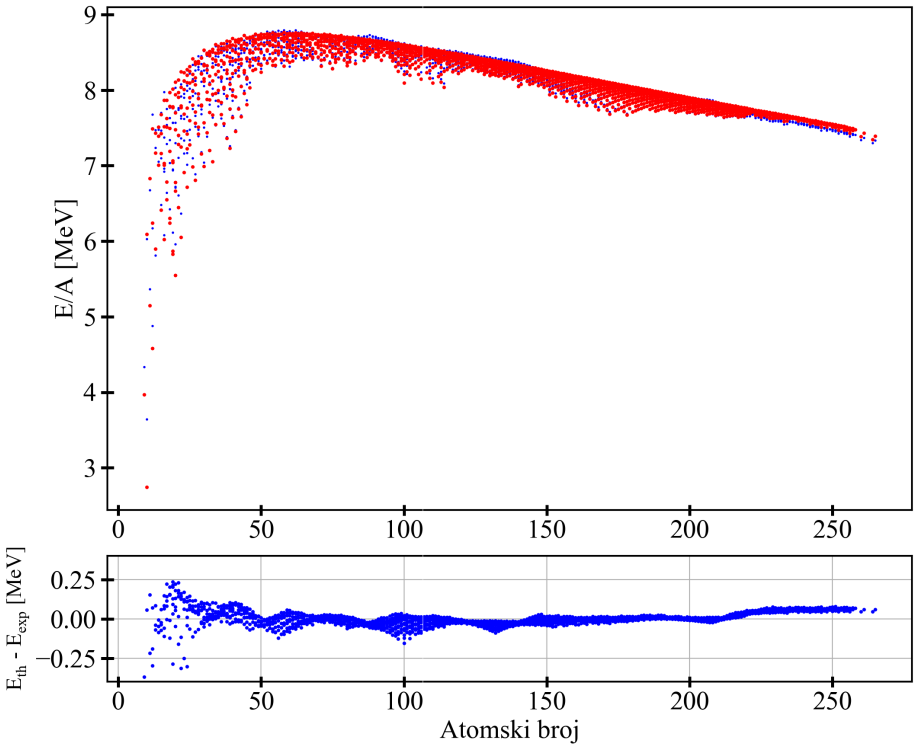
Model upotpunjujemo članom sparivanja, slika 4.5. Provjereni su i drugi potencijski oblici člana sparivanja, $A^{3/4}$, $A^{1/2}$ i A^{-1}. Za dani set podataka najmanja pogreška je bila sa uporabom člana sparivanja sa potencijskom ovisnošću $A^{-1/2}$.

Tablica 4.1: Lista koeficijenata izraženih u MeV-ima, N označava broj parametara u modelu dok RMS odgovara pogrešci energije vezanja po nukleonu za 2497 poznatih jezgara iz [11].

N	a_{vol}	a_{surf}	a_{coul}	a_{asym}	a_{pair}	$RMS_{nukl.}$ [MeV]
1	8.145(9)	-	-	-	-	0.437
2	8.08(4)	0.3(2)	-	-	-	0.436
3	11.8(1)	-9.2(3)	-0.39(1)	-	-	0.354
4	15.01(2)	-15.99(5)	-0.658(2)	-20.52(7)	-	0.051
5	15.03(2)	-16.06(5)	-0.659(2)	-20.54(6)	8.1(4)	0.048

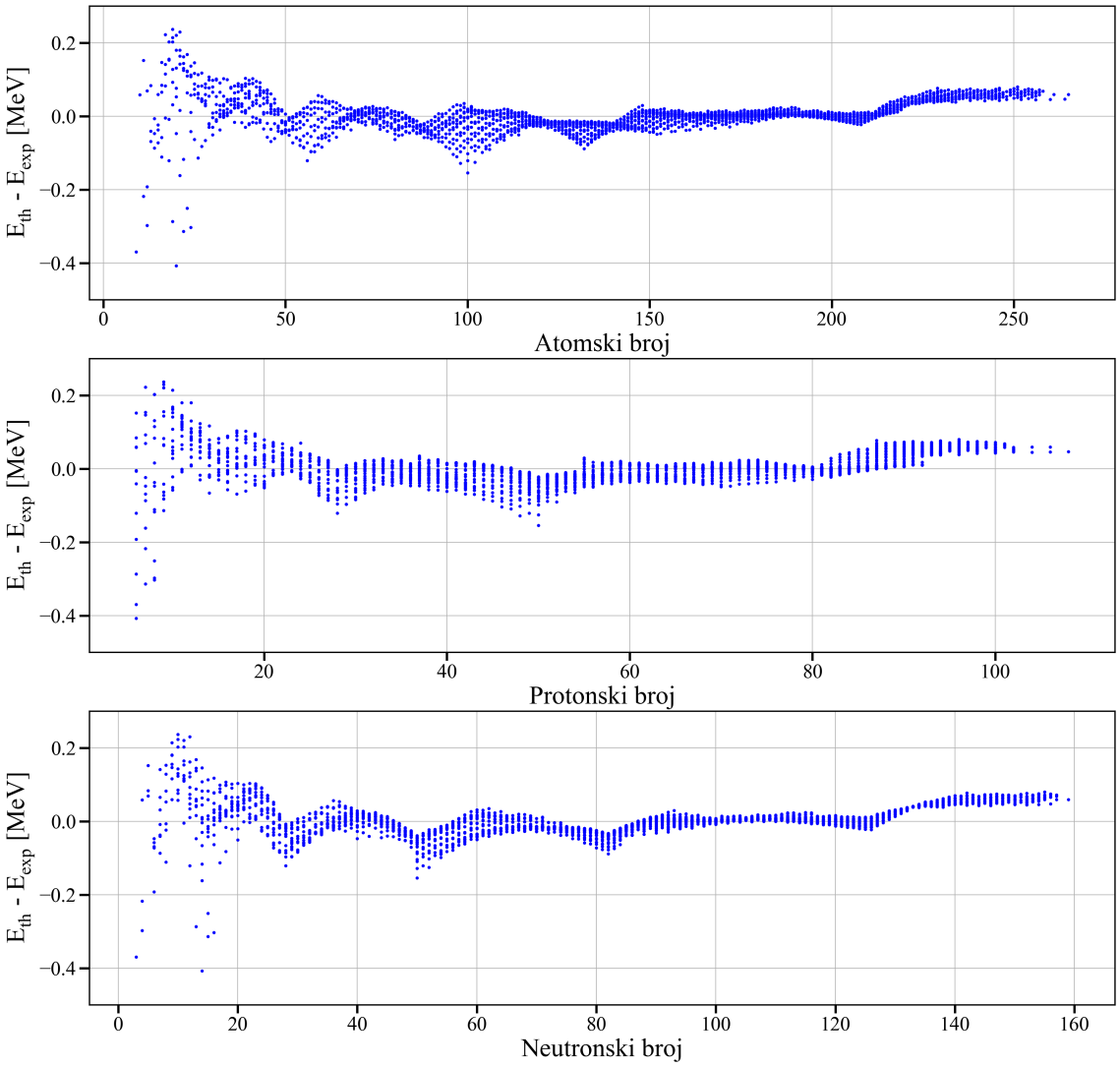


Slika 4.4: Energija vezanja po nukleonu u ovisnosti o masenom broju. Plavom bojom su prikazani podaci, crvenom bojom prikazana je prilagodba za volumni, površinski, Coulombov i asimetrijski doprinos. Iznosi parametara su zapisani u tablici 4.1. za N=4.



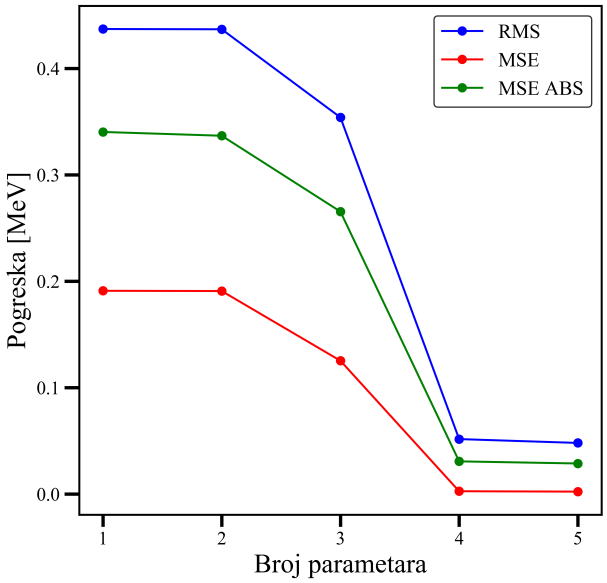
Slika 4.5: Energija vezanja po nukleonu u ovisnosti o masenom broju. Plavom bojom su prikazani podaci, crvenom bojom prikazana je prilagodba za sve doprinose iz BW formule. Iznosi parametara su zapisani u tablici 4.1. za N=5.

Slika 4.6 prikazuje kako se odsupanja predviđanja od eksperimentalnih podataka periodički pojavljuju za magične nuklearne brojeve. Ova opažanja pokazatelj su nepotpune teorijske pozadine koju nudi BW formula.

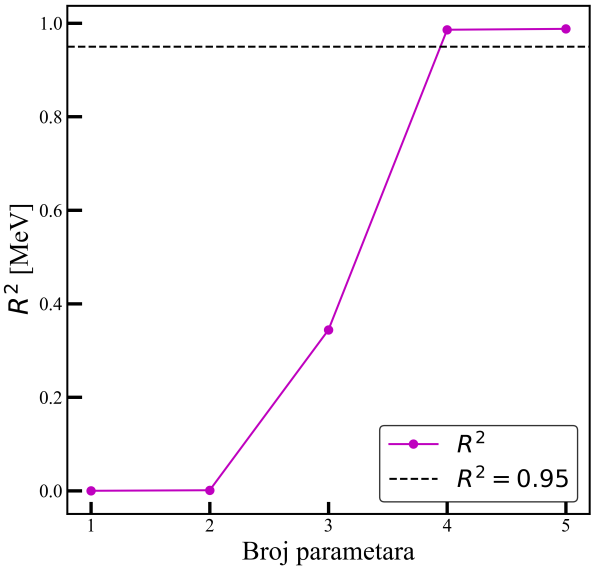


Slika 4.6: Razlika očekivanih i predviđenih energija u ovisnosti o masenom, protonskom te neutronskom broju za model u kojem su uključeni svi članovi BW formule.

Zapažen je nagli pad u vrijednosti pogrešaka pri dodavanju četvrte komponente što se vidi na slici 4.7. Pogreška energije vezanja po nukleonu se nije pretežito promjenila sa dodavanjem površinskog člana.



Slika 4.7: Ovisnost RMS, MSE i apsolutne vrijednosti MSE pogreške o broju prediktora uključenih u prilagodbu. Prikazani rezultati su pogreške energija vezanja po nukleonu.



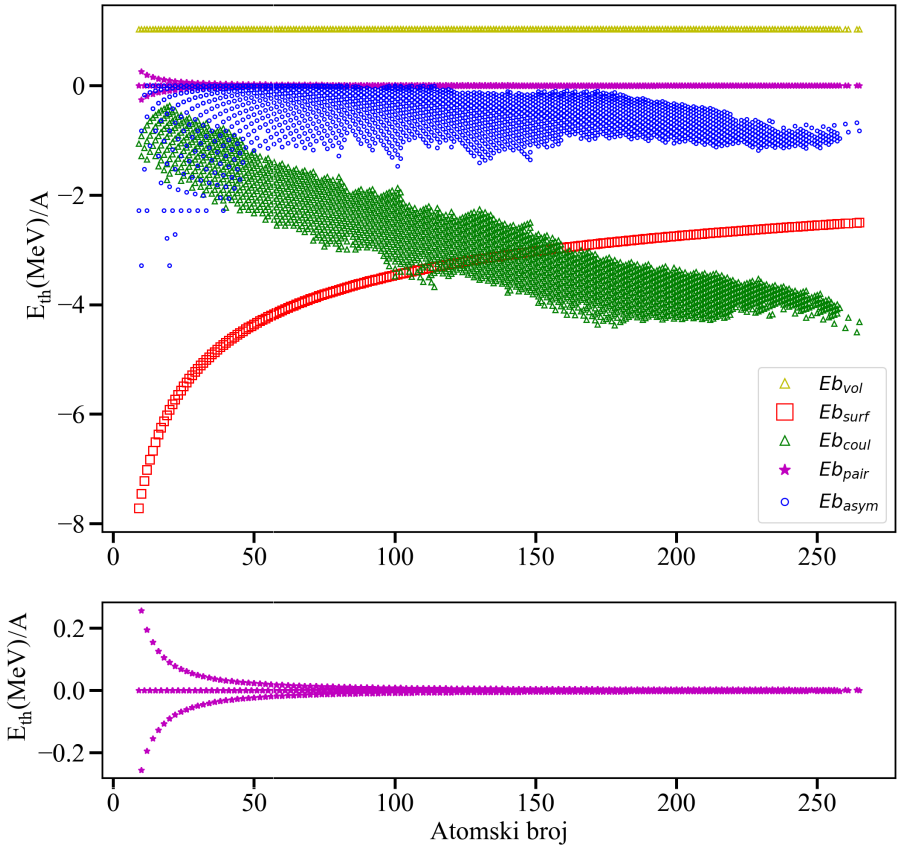
Slika 4.8: R^2 u ovisnosti o broju parametara za energiju vezanja po nukleonu.

Vrijednost R^2 je prikazana na slici 4.8. Energija vezanja po nukleonu ima malu varijabilnost za parametre koji doprinose volumni i površinski dio u ukupnoj energiji. Izračunati su korelacijski odnosi između parametara i ukupni doprinosi energiji vezanja što je prikazano u tablici 4.9. Parametri volumena, površine i Coulomba visoko su međusobno korelirani. Parametar asimetrije nešto je manje koreliran s ostalima te stoga doprinosi novu informaciju u sustav što rezultira padom vrijednosti pogreške. Parametar sparivanja negativno je koreliran s doprinosima volumena, površine te asimetrije dok je pozitivno koreliran s Coulombovim doprinosom. Ukupni doprinosi energiji vezanja i njihove korelacije odgovaraju teorijskim očekivanjima.

Na slici 4.10 prikazani su doprinosi energiji vezanja za pet prediktora. Vidljivo je smanjenje energije vezanja s porastom masenog broja. Coulombov član doprinosi najvećem smanjenju energije što je i teorijski očekivano. Volumni, konstantan član, jedini pozitivno doprinosi energiji vezanja. Površinski doprinos veći je za lakše jezgre te s porastom masenog broja njegov doprinos opada.

Slika 4.9: Tablica korelacija dobivenih energija uporabom MNK.

	Eb_vol	Eb_surf	Eb_coul	Eb_asym	Eb_pair
Eb_vol	1	-0.889	-0.908	-0.222	-0.0346
Eb_surf	-0.889	1	0.885	0.0711	0.0642
Eb_coul	-0.908	0.885	1	-0.14	0.0337
Eb_asym	-0.222	0.0711	-0.14	1	0.00795
Eb_pair	-0.0346	0.0642	0.0337	0.00795	1

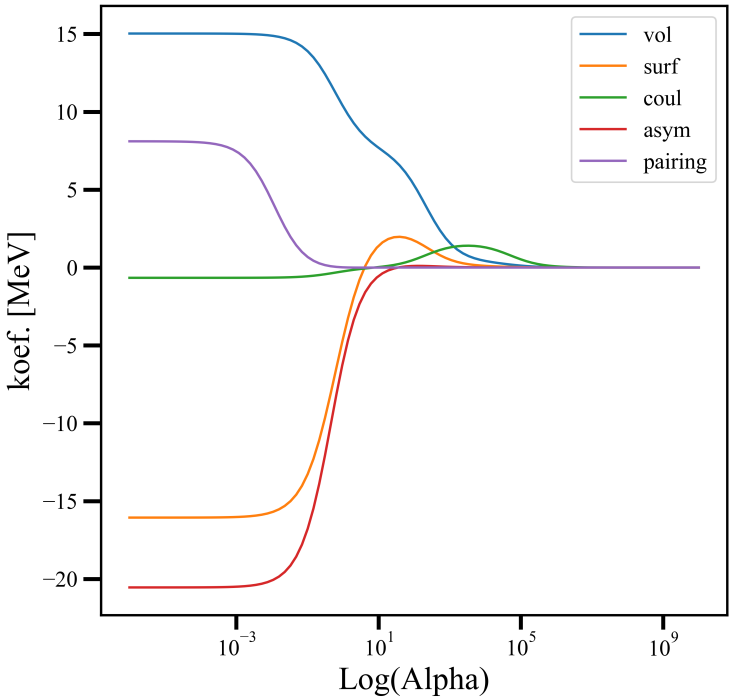


Slika 4.10: Prikaz doprinosa energiji vezanja za pojedine članove za jednostavan model od pet prediktora. Na slici ispod je preglednije prikazan član sparivanja. Zbog lakšeg prikaza, volumni doprinos je spušten na nižu vrijednost. Korišteni koeficijenti se nalaze u tablici 4.1.

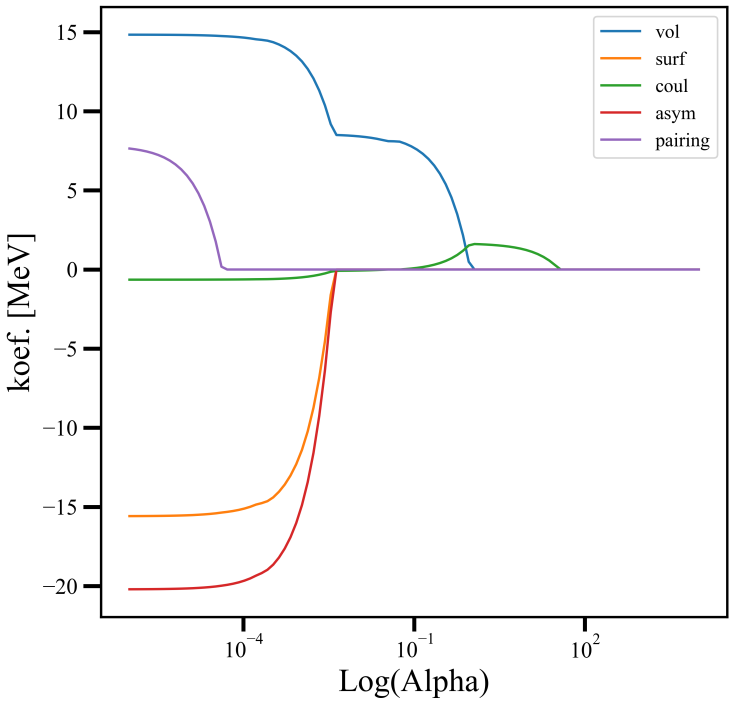
4.1 LASSO i hrbat-regresija

Hrbat-regresija i LASSO obično se koriste kako bi se opisao model koji sadrži više parametara prilagodbe te je potrebna regularizacija parametara kako bi se izbjegla preprilagođenost. Druga pogodnost regularizacijskih metoda je da daju pregled odnosa parametara prilagodbe. Za svaku vrijednost parametra regularizacije α koeficijenti, tj. parametri će poprimiti "novu" vrijednost. Prikaz ovisnosti koeficijenata o parametru regularizacije se nalazi na slikama 4.11 i 4.12.

Vidljiva je korelacija između parametara. Povećavanjem parametra alpha dolazi do suženja vrijednosti parametara, najdulje u modelu odsatje parametar koji doprinosi Coulombov član energije. Parametar čija vrijednost najprije počinje opadati je doprinos sparivanja. Situacija za LASSO regularizaciju vrlo je slična, no postoji jasna razlika, a to je da parametri postaju u potpunosti isključeni iz modela.

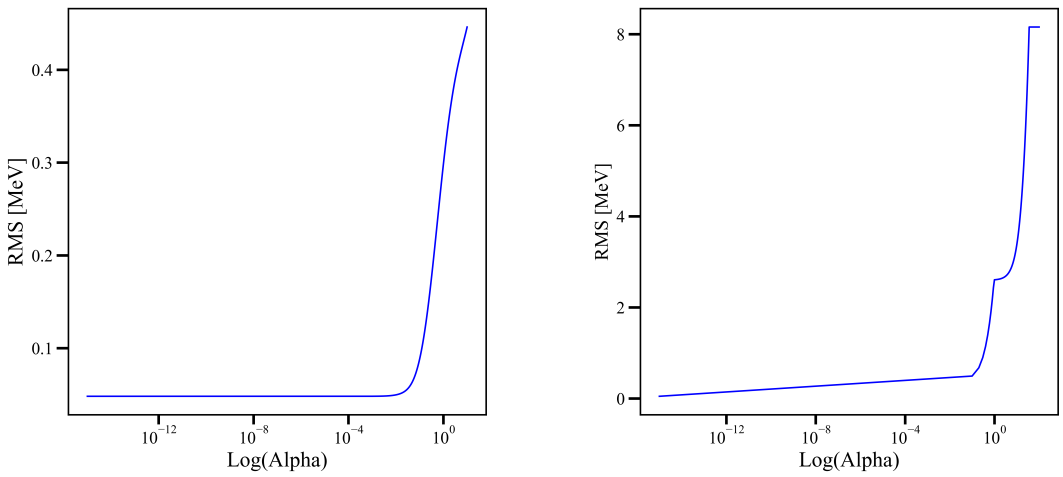


Slika 4.11: Koeficijenti u ovisnosti o parametru regularizacije alpha za hrbat-regresiju.



Slika 4.12: Koeficijenti u ovisnosti o parametru regularizacije alpha za LASSO.

LASSO regularizacija relativno rano iz modela isključi sve prediktore što se vidi u naglom porastu pogreške. Kako se RMS pogreška mijenja u ovisnosti o parametru regularizacije alpha prikazano je na slici 4.13. Za istu vrijednost $\alpha \approx 1$ vrijednost pogreške LASSO regularizacije je 1.8 MeV-a, dok hrbat-regresija daje 0.3 MeV-a. Unakrsnom provjerom obje regularizacijske metode odabiru parametre prilagodbe koji po iznosu odgovaraju parametrima dobivenim metodom najmanjih kvadrata.

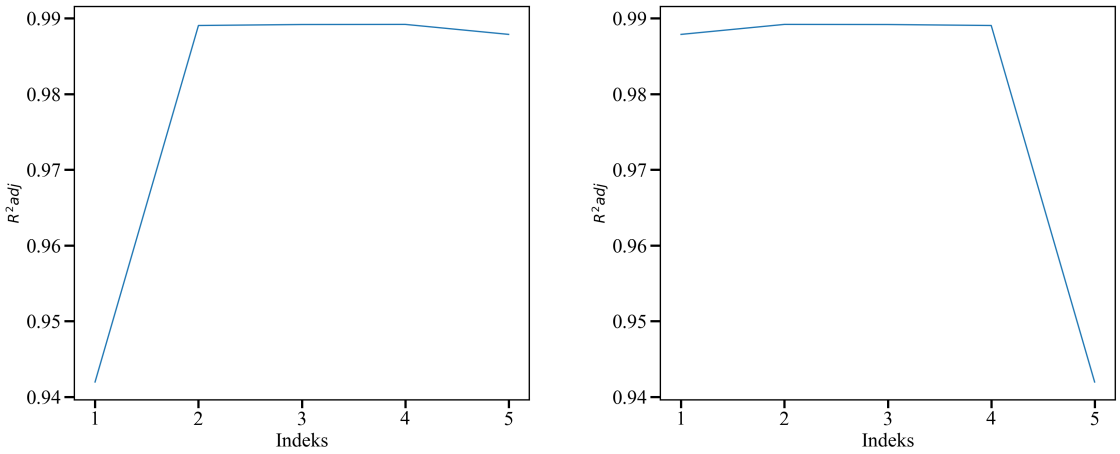


(a) Hrbat-regresija (b) LASSO

Slika 4.13: RMS u ovisnosti o parametru regularizacije alpha.

4.2 Selekcija unaprijed i unazad

Napravljena je selekcija podskupa metodom unaprijed i unazad. Uključivanje tj. izbacivanje prediktora u sustav određeno je pomoću tri kriterija: podešen R^2, BIC i AIC. Podešen R^2 daje iste rezultate za selekciju unaprijed i unazad. Volumni doprinos, kao konstantan član, smanjuje vrijednost R^2. Podskup koji ovaj model odabire, sastoji se od 4 prediktora. Doprinos površinskog člana ima najveću vrijednost podešenog R^2, kako se ostali prediktori uključuju u model prikazano je pomoću slike 4.16.

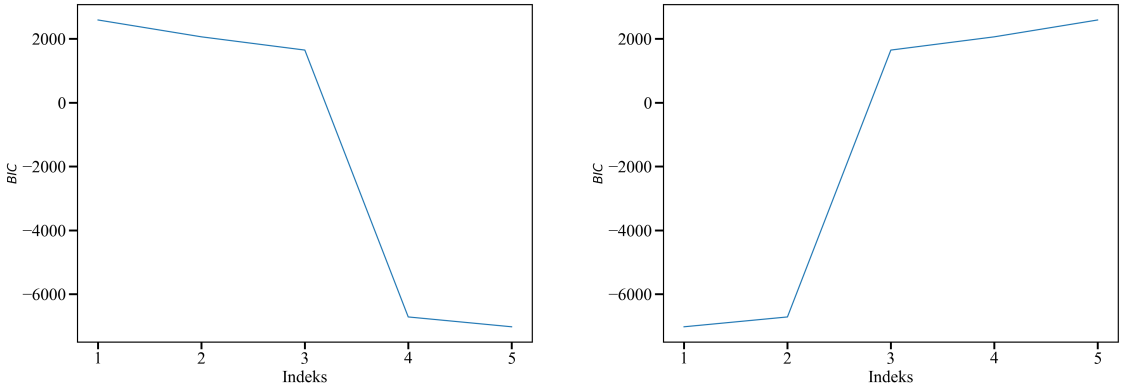


Slika 4.14: Ovisnost vrijednost podešenog R^2 o broju prediktora u modelu. Lijevo je prikazano za selekciju unaprijed, desno se nalazi selekcija unazad.

Slika 4.15: Tablica vrijednosti podešenog R^2 za selekciju unaprijed i selekciju unazad.

	predictors	rsquared_adj
1	[surf]	0.941971
2	[coul, surf]	0.989056
3	[asym, coul, surf]	0.989187
4	[pair, asym, coul, surf]	0.989201
5	[vol, pair, asym, coul, surf]	0.987883

	predictors	rsquared_adj
1	[vol, surf, coul, asym, pair]	0.987883
2	[surf, coul, asym, pair]	0.989201
3	[surf, coul, asym]	0.989187
4	[surf, coul]	0.989056
5	[surf]	0.941971



Slika 4.16: Vrijednost BIC ovisno o broju prediktora u modelu. Prediktori u modelu označeni su indeksom, sukladno indeksu iz tablice. Lijevo je prikazano za selekciju unaprijed, desno se nalazi selekcija unazad.

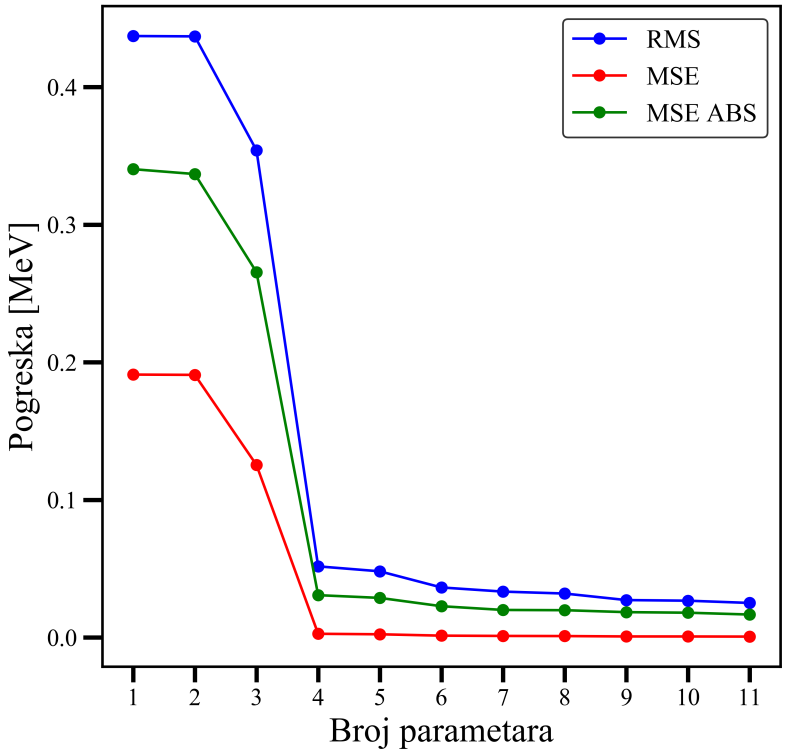
Slika 4.17: Tablica vrijednosti za BIC kriterij. Lijevo je prikazana selekcija unaprijed, desno selekcija unazad.

	bic	predictors
1	2590.727598	[vol]
2	2062.692682	[asym, vol]
3	1647.775336	[coul, asym, vol]
4	-6712.449373	[surf, coul, asym, vol]
5	-7020.791441	[pair, surf, coul, asym, vol]

	bic	predictors
1	-7020.791441	[vol, surf, coul, asym, pair]
2	-6712.449373	[vol, surf, coul, asym]
3	1647.775336	[vol, coul, asym]
4	2062.692682	[vol, asym]
5	2590.727598	[vol]

Odabir podskupova pomoću AIC kriterija daje iste podskupove kao i BIC kriterij. Razlika BIC i AIC kriterija je član 2k tj. ln(n)k. Za k=5 odstupanje AIC i BIC kriterija je unutar istog reda veličine, stoga su isti rezultati odabira podskupa očekivani.

Nakon analize BW formule s N=5 članova, modelu pridodajemo proširenja predložena u poglavlju 2.3. Proširena BW formula dana je izrazom 2.8. Sveukupno imamo 11 parametara čije su vrijednosti zapisane u tablici 4.2. Na slici 4.18 vidljivo je da do znatnog pada vrijednosti pogreške dolazi s uključivanjem člana asimetrije u model. Vrijednosti pogreške energije vezanja po nukleonu za uključenih $N = 11$ prediktora iznosi $RMS_{nukl.} = 0.025$ MeV-a.Radi se o smanjenju od 48 % od početne vrijednosti za $N = 5$ prediktora. Pogreška energije vezanja je $RMS(N = 11) = 2.256$ MeV-a, ostale pogreške energije vezanja po nukleonu te koeficijenti su prikazani u tablici 4.2.



Slika 4.18: Ovisnost RMS-a o broju prediktora u modelu proširene semi-empirijske formule

Tablica 4.2: Lista koeficijenata izraženih u MeV-ima, N označava broj parametara u modelu, RMS je izračunat za energiju vezanja po nukleonu za 2497 poznatih jezgara iz [11].

N	a_{vol}	a_{surf}	a_{coul}	a_{asym}	a_{pair}	a_{st}	a_{xc}	a_W	a_R	a_m	a_{m1}	RMS[MeV]
6	15.42(2)	-17.33(5)	-0.699(1)	-26.7(2)	8.9(3)	22.7(6)	-	-	-	-	-	0.036
7	15.35(2)	-17.98(5)	-0.697(2)	-25.5(2)	9.0(3)	21.4(5)	0.80(4)	-	-	-	-	0.033
8	15.19(2)	-17.53(6)	-0.686(2)	-26.6(2)	9.2(2)	31.4(9)	0.87(4)	-17(1)	-	-	-	0.031
9	16.68(5)	-26.1(3)	-0.759(3)	-32.4(2)	9.2(3)	54(1)	1.15(3)	-42(1)	13.1(5)	-	-	0.027
10	16.73(5)	-26.4(3)	-0.759(2)	-32.4(2)	9.2(2)	55(1)	1.15(3)	-42(1)	13.5(5)	-0.04(2)	-	0.026
11	16.69(5)	-26.1(3)	-0.752(3)	-32.0(2)	9.2(2)	53(1)	1.08(3)	-40(1)	13.0(4)	-0.096(1)	0.100(6)	0.025

4.3 Generalizirano proširena semi-empirijska formula

Idea za proširivanje semi-empirijske formule dolazi iz članka Numerical Generalization of the Bethe-Weizsäcker Mass Formula [5]. Autori članka predlažu nelinearan odnos parametara prilagodbe kao rješenje. Pridodana proširenja opisuju deformacije koje je teško jednoznačno predvidjeti za cijeli spektar jezgara. Model sadrži dvjestotinjak nelinearnih parametara prilagodbe koji daju pogrešku energije vezanja od 1.11 do 1.565 MeV-a ovisno o tome jesu li u modelu prisutni izotopi vodika ili ne. Uvođenje nelinearnih odnosa za izračun parametara je tehnički zahtjevan postupak, stoga će se u ovom radu promotriti samo linearno proširenje BW formule. Polazna točka nam je proširena BW formula iz jednadžbe 2.8 s jedanaest članova. Svi doprinosi osim efekta ljusaka proširuju se s predloženim parametrima

$$v_1 = \frac{Z}{A}, \quad v_2 = \frac{N}{A}, \quad v_3 = \frac{N-Z}{A}, \quad v_4 = \frac{Z}{N+1} \tag{4.1}$$

te njihovim potencijama

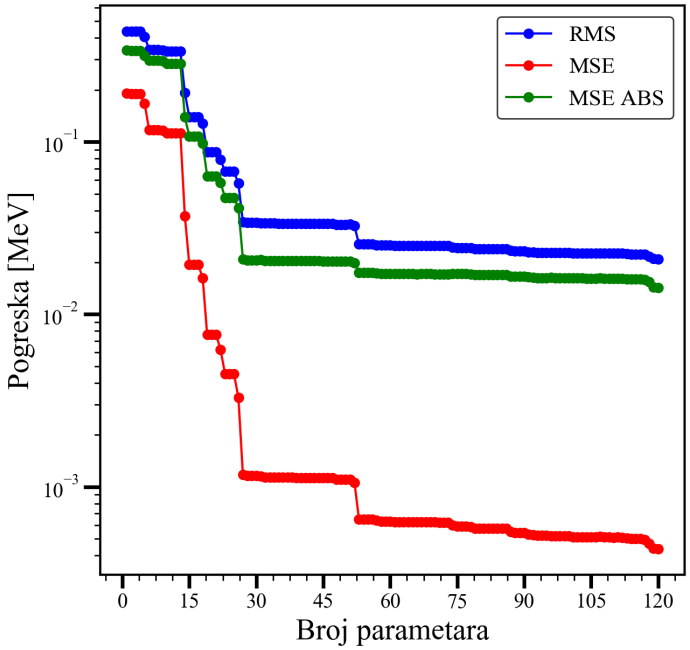
$$v_5 = \left(\frac{Z}{A}\right)^2, \quad v_6 = \left(\frac{N}{A}\right)^2, \quad v_7 = \left(\frac{N-Z}{A}\right)^2, \quad v_8 = \left(\frac{Z}{N+1}\right)^2 \tag{4.2}$$

$$v_9 = \left(\frac{Z}{A}\right)^3, \quad v_{10} = \left(\frac{N}{A}\right)^3, \quad v_{11} = \left(\frac{N-Z}{A}\right)^3, \quad v_{12} = \left(\frac{Z}{N+1}\right)^3. \tag{4.3}$$

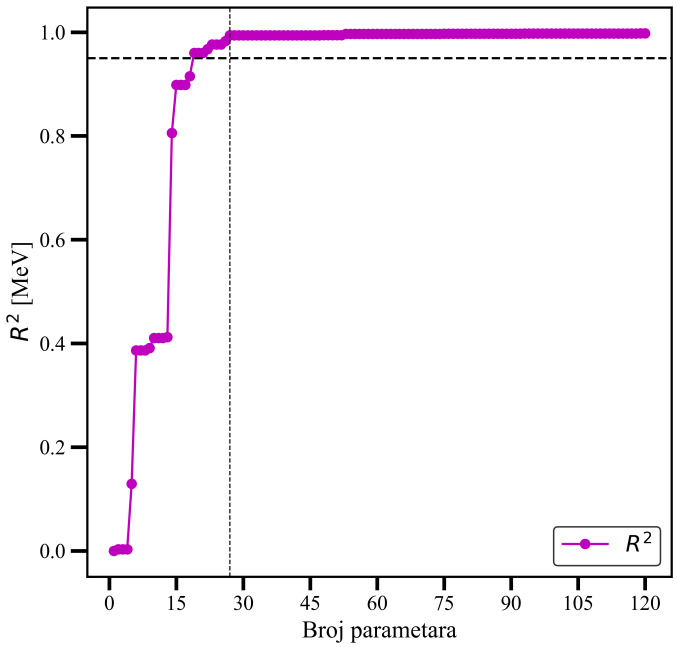
Sveukupno, model sadrži jedanaest članova *k= vol, surf, coul, asym, pairing, coul exc, wigner, curvature, surf sym*, 108 proširenja oblika

$$a_k^1 v_1 k + a_k^2 v_2 k + \cdots + a_k^{12} v_{12} k \tag{4.4}$$

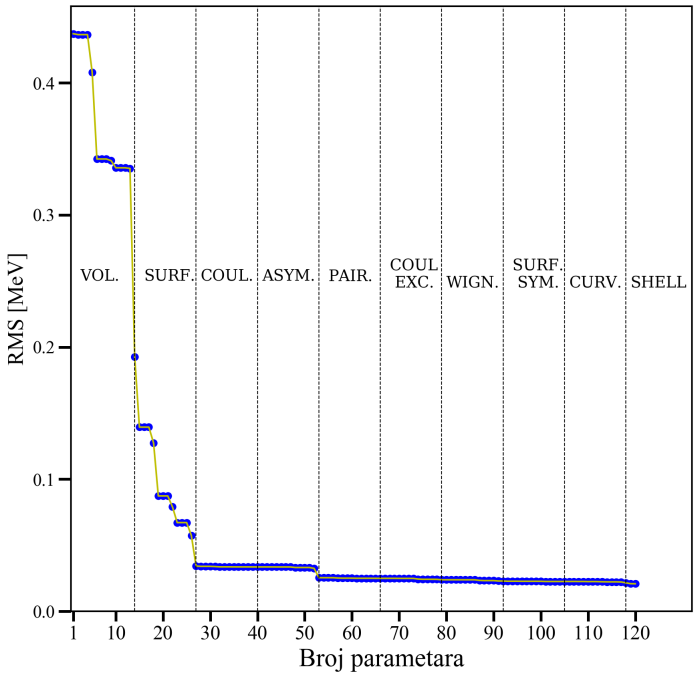
što zajedno daje 119 a_k^j parediktora za $j = 1, \ldots, 12$. Na model primijenimo metodu najmanjih kvadrata. Vrijednosti pogrešaka i R^2 prikazane su na slikama 4.19 i 4.20. Pogreška energije vezanja po nukleonu iznosi $RMS_{nukl.}(N = 119) = 0.0201$ MeV-a što je 58% bolje od rezultata dobivenog u modelu s neproširenom BW formule. Pogreška energije vezanja dobivena MNK iznosi $RMS(N = 119) = 1.726$ MeV-a.



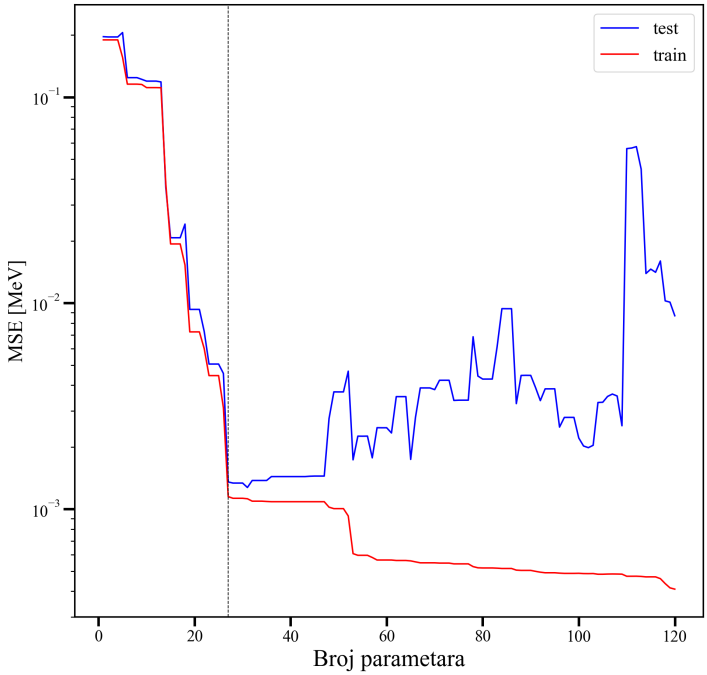
Slika 4.19: Pogreška energije vezanja po nukleonu u ovisnosti o broju prediktora. Pogreška je prikazana na logaritamskoj skali.



Slika 4.20: Vrijednost R^2 ovisno o broju prediktora. Horizontalnom linijom je prikazana vrijednost nakon koje dolazi do preprilagođenosti. Vertikalno je prikazana vrijednost $R^2 = 0.98$.



Slika 4.21: RMS pogreška energije vezanja po nukleonu u ovisnosti o broju prediktora u modelu. Vertikalnim linijama napravljena je podjela na jedanaest dijelova. Najmanja pogreška $RMS_{nukl.}(N = 119) = 0.0201$ MeV-a.



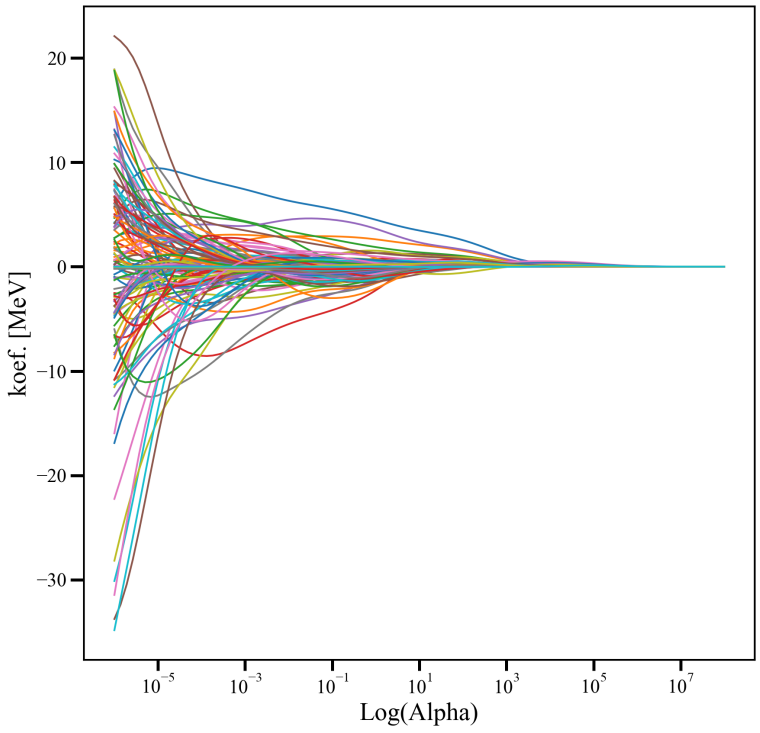
Slika 4.22: Vrijednost MSE pogreške energije vezanja po nukleonu u ovisnosti o broju parametara. Plavom bojom prikazana je pogreška testnog skupa, crvenom je pogreška treniranog skupa podataka.

Zašto je došlo do relativno malog odstupanja između pogrešaka u modelima sa $N = 119$ i $N = 11$ članova jasno je prikazano na slici 4.21. Vrijednost pogreške

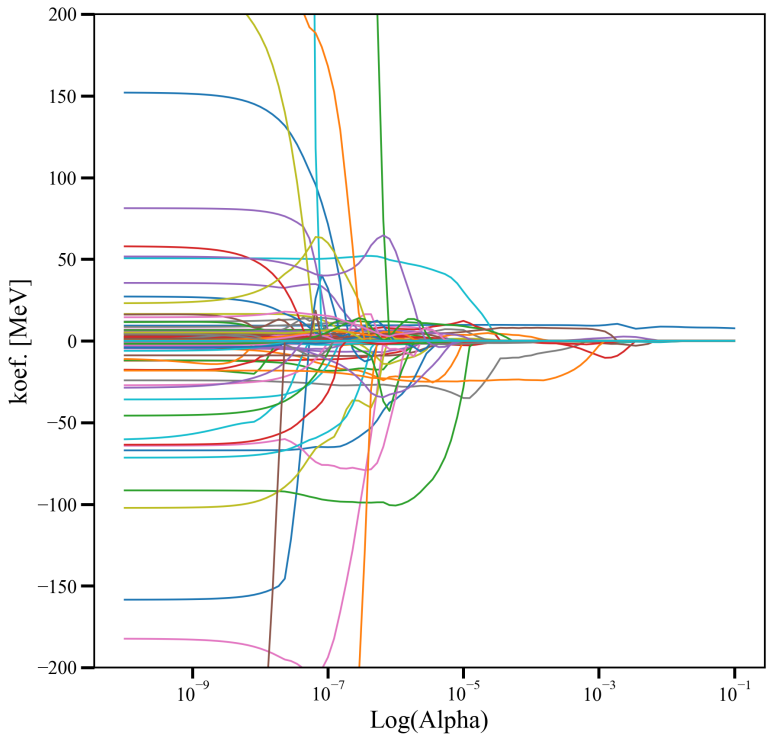
pada kod uključivanja vodećeg člana proširenja. U modelu ne dolazi do značajnijih promjena kod pridodavanja članova koji opisuju deformacije. Moramo imati na umu da metoda najmanjih kvadrata daje vrlo fleksibilan model tj. model visoke varijabilnosti. Metoda najmanjih kvadrata snažan je statistički alat za predviđanja ako imamo relativno malen broj prediktora, kod većeg broja prediktora se često javlja preprilagođenost. Provjera pretreniranosti seta se prikazuje pomoću podjele podataka na trening i test. U našem slučaju, test set sadrži 490 nasumično odabranih jezgara. Iz slike 4.22 vidljiva je preprilagođenost modela nakon uključivanja 28. člana. Nedostaci ovog modela su: nemogućnost generalizacije i visoka kompleksnost.

Hrbat-regresija i LASSO

Kako bi izbjegli eskaliranje parametara u ekstremne vrijednosti, model prilagodili za buduća predviđanja primjenjujemo hrbat-regresiju i LASSO. Rezultati regularizacija prikazani su na slikama 4.23 i 4.24.

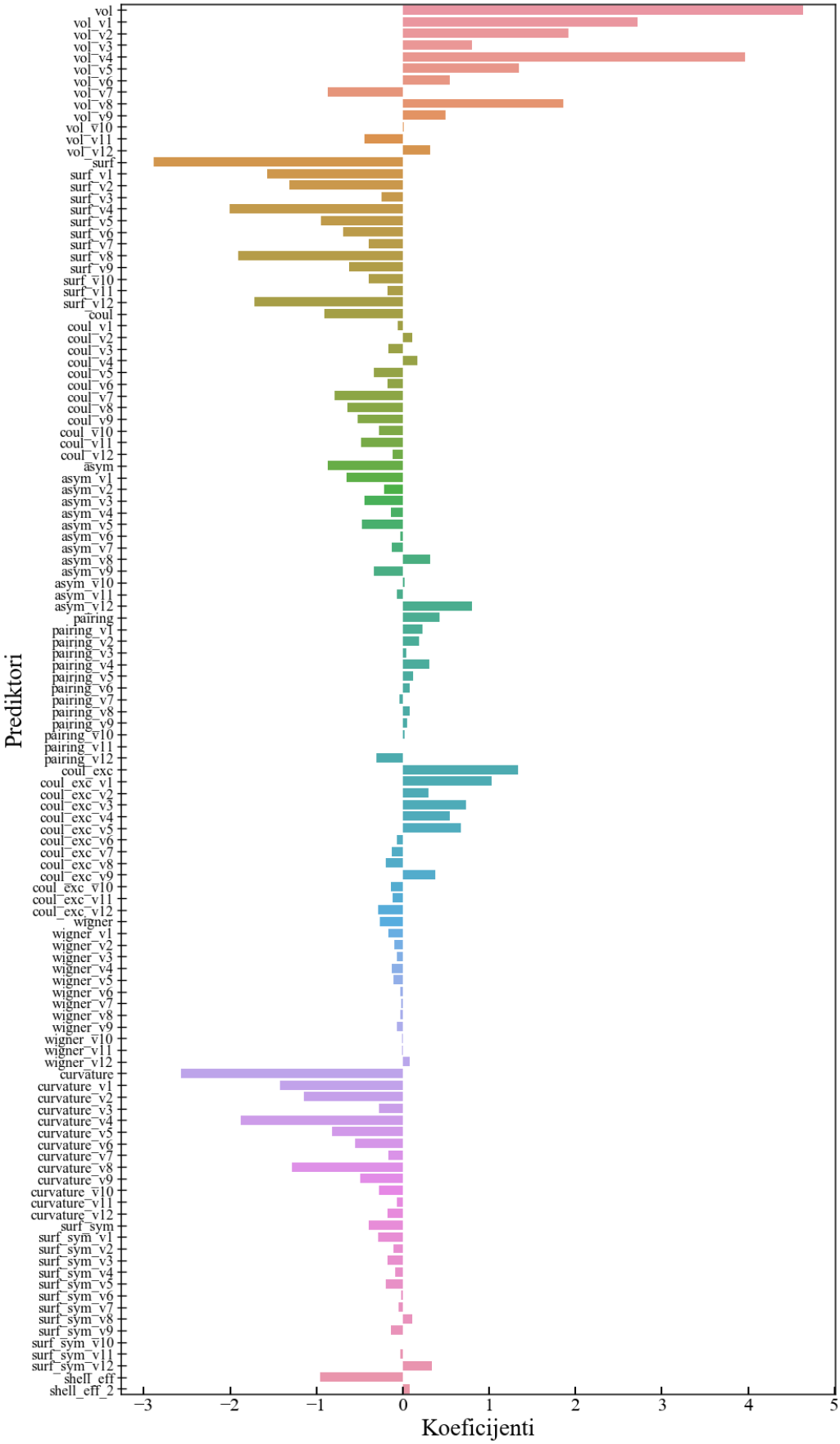


Slika 4.23: Koeficijenti u ovisnosti o parametru regularizacije alpha za hrbat-regresiju.

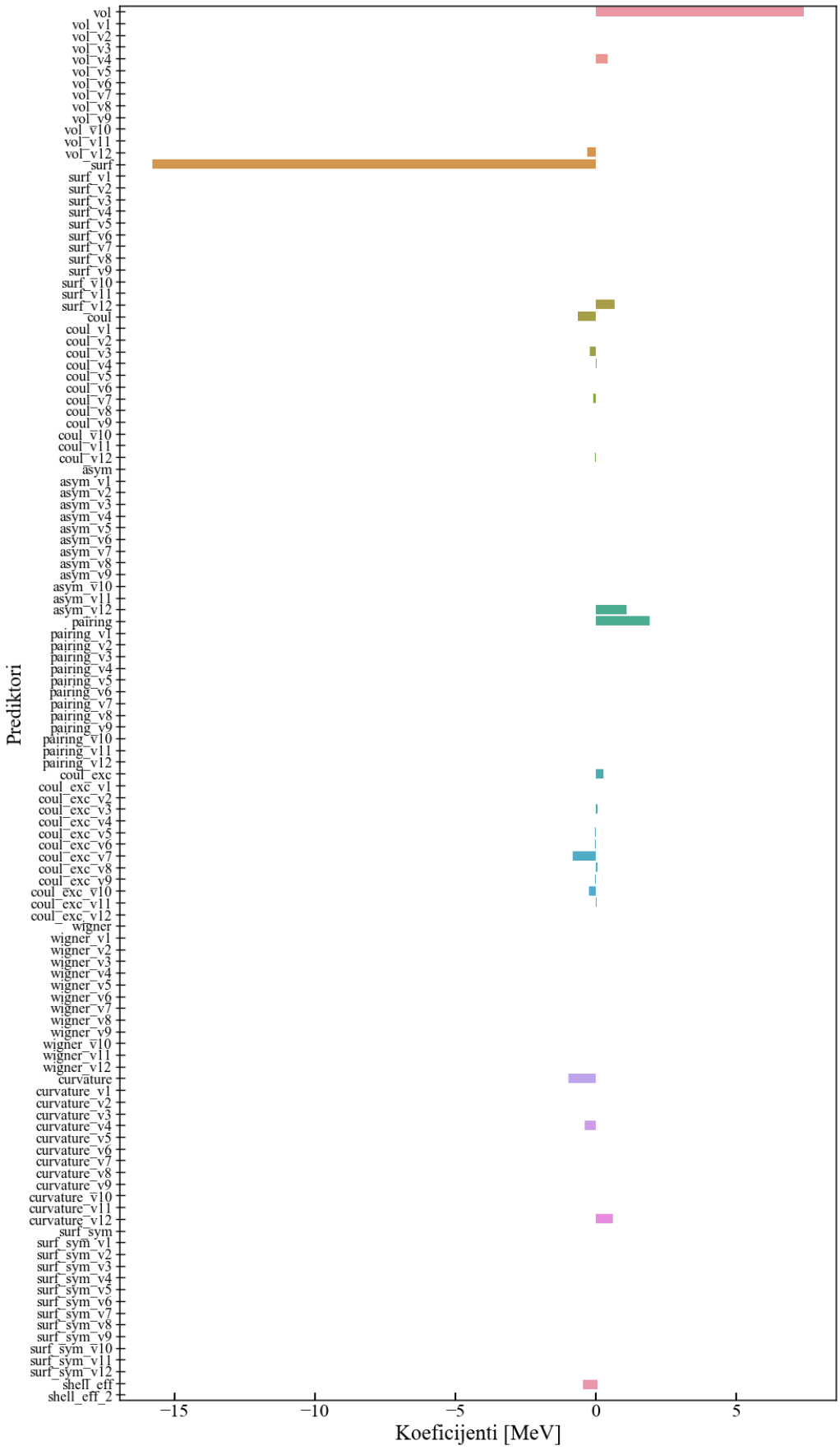


Slika 4.24: Koeficijenti u ovisnosti o parametru regularizacije alpha za LASSO.

Uočava se kaotičan režim vrijednosti koeficijenata prilagodbe za male vrijednosti α. Povećanjem parametra alpha dolazi do suženja vrijednosti koeficijenata.
Pronalazak ravnoteže pristranosti i varijabilnosti, u svrhu smanjenja testne pogreške, radimo pomoću unakrsne provjere. Podaci su podijeljeni na deset jednakih dijelova. Odabran je interval α od 10^{-7} do 10^{5} za hrbat-regresiju te interval od 10^{-10} do 10^{-1} za LASSO. Dobivene su sljedeće vrijednosti $\alpha_{Hrbat} = 0.58$ te $\alpha_{LASSO} = 1.64 \times 10^{-4}$. Vrijednosti α uvrštene su u prilagodbu te su izračunate vrijednosti koeficijenata. Vrijednosti koeficijenta prikazane su na slikama 4.25 i 4.26.



Slika 4.25: Iznos koeficijenata prilagodbe za parametar regularizacije $\alpha_{hrbat} = 0.58$. Pogreška energije vezanja po nukleonu je $RMS_{hrbat-nukl.}(N = 119) = 0.022$ MeV-a.

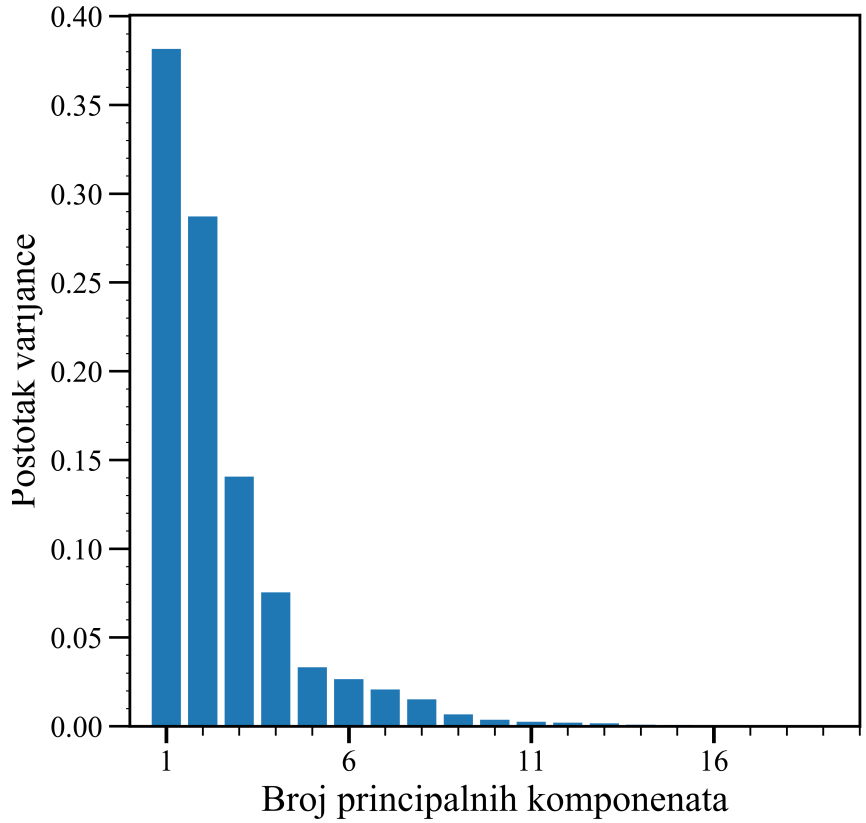


Slika 4.26: Iznos koeficijenata prilagodbe za parametar $\alpha_{LASSO} = 1.64 \times 10^{-4}$. Pogreška energije vezanja po nukleonu je $RMS_{LASSO-nukl.}(N = 119) = 0.027$ MeV-a.

Sa slika je vidljiva jasna razlika između dvije primijenjene metod regularizacije. Hrbat-regresija će u modelu ostaviti sve parametre, parametri se optimiziraju tako da se smanji vrijednost CV pogreške. Vidljiva je dominacija u vrijednosti koeficijenata povezanih sa osnovnim članovima. Proširenja su izraženija za volumni i površinski doprinos energiji vezanja. Imajući u vidu velik broj deformiranih jezgara u podacima, sami rezultati nisu iznenađujući. Unutar deformacija prevladavaju članovi koji su proporcionalni s brojem protona i neutrona. Najmanja promjena u vrijednosti koeficijenata je za efekt ljusaka, oni su ostali gotovo nepromjenjeni, dok su koeficijenti Wignerovog člana najviše sažeti. Da početnoj BW formuli nedostaje fizika koja bi opisivala preciznije deformacije vidimo pomoću LASSO regularizacije koja u modelu sa samo nekoliko članova ostavlja efekt ljusaka. Wignerov doprinos energiji vezanja je u potpunosti izbačen, što se i očekivalo kako je to najviše sažet član iz hrbat-regresijskog modela. Član površine ostao je najmanje izmijenjen u odnosu na vrijednost iz tablice A.3. Vodeći član asimetrije nije uključen u LASSO modelu, što nije čudno jer su članovi asimetrije i sparivanja u modelu $N = 5$ bili prvi isključeni.

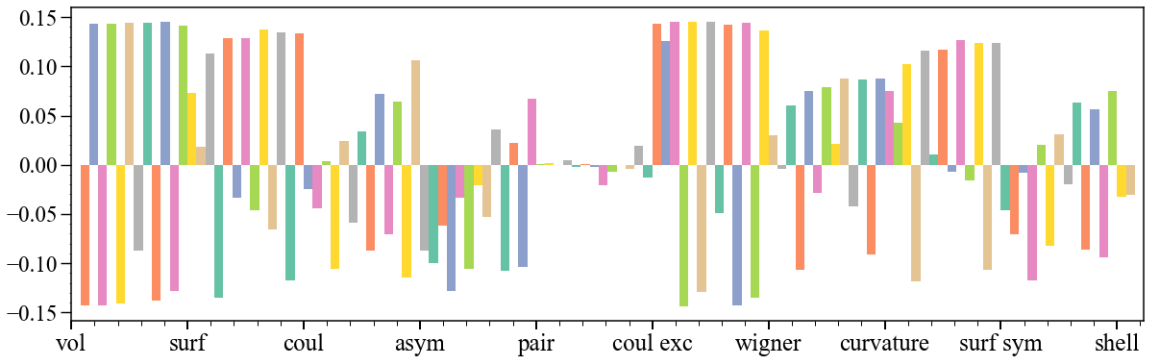
Analiza principalnih komponenata

Analiza principalnih komponenata predstavlja snažan statistički alat koji nam omogućava jednostavno rješavanje klasifikacijskih problema te služi za redukciju dimenzionalnosti. U ovom radu fokusirat ćemo se na redukciju dimenzionalnosti.
Prije samog početka podaci se skaliraju. Nakon toga slijedi odabir broja principalnih komponenata tj. odabiremo dimenziju koja je odgovarajuća za naš problem. Odabir dimenzije ćemo vršiti tako da početno pretpostavimo da nam je za opis potreban cjelokupan prostor prediktora, što je u našem slučaju $N_{dim} = 119$. Provjeravamo objašnjenu varijancu svih 119 novo odabranih komponenata. Na slici 4.27. je prikazan postotak objašnjene varijance pomoću principalnih komponenata.



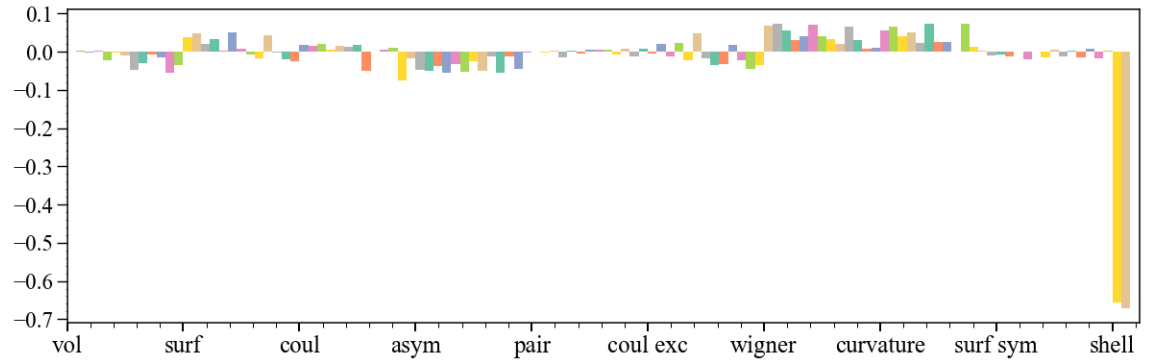
Slika 4.27: Postotak objašnjene varijance u ovisnosti o broju komponenata.

Za dobar opis modela potrebna je vrijednost varijabilnosti od 0.95, tj. s odabranim komponentama moramo moći opisati više od 95% promjena unutar podataka. Za naš model taj se uvjet postiže već nakon sedme komponente. Za objašnjavanje energije vezanja po nukleonu pomoću principalnih komponenata potreban je samo sedam dimenzionalni PC prostor. Rezultat predstavlja znatnu redukciju dimenzionalnosti.
Za razliku od regularizacijskih metoda, analiza principalnih komponenata ne vrši direktnu selekciju prediktora što možemo vidjeti prikažemo li odabrane principalne komponente u prostoru prediktora.
Na slici 4.28 je prikazana prva principalna komponenta koja sadrži najveću informaciju o promjeni podataka.



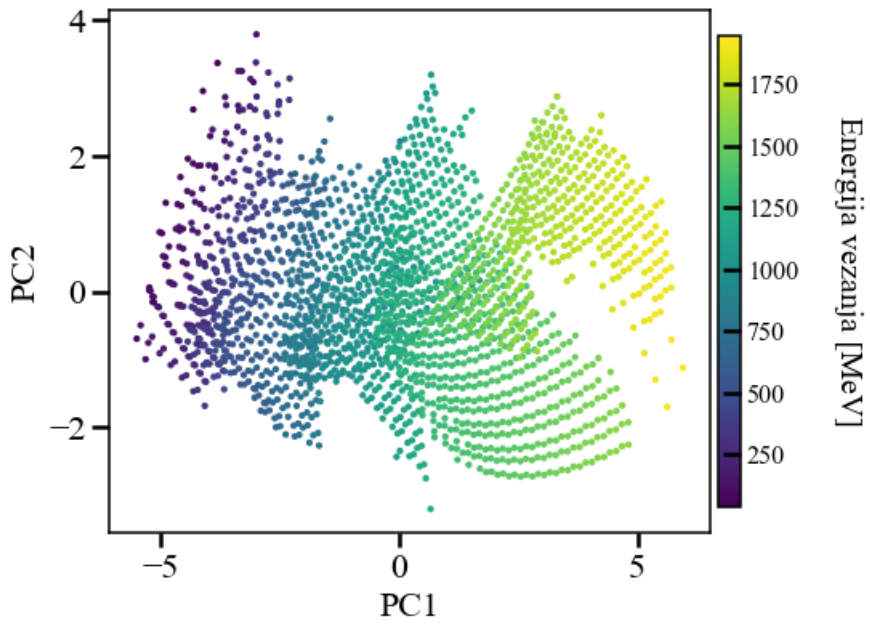
Slika 4.28: Prva komponenta prikazana u prostoru prediktora.

Principalna komponenta sadrži većinu početnih prediktora i relativno je teško odrediti koji prediktor je potrebno isključiti iz modela, no kao što je u početku naglašeno, to nije ni cilj ove metode. Ako se u modelu javi prediktor koji doprinosi varijabilnosti podataka, a teško ga je opisati pomoću neke od prijašnjih osi, metoda će odabrati novu os koja sadrži informaciju o promjeni podataka vezano uz taj prediktor. Na slici 4.29 prikazana je sedma i posljednja komponenta nužna za opis. Odabrana os opisuje isključivo promjenu podataka uslijed efekta ljusaka. Ova pojava je indikacija važnosti doprinosa efekta ljusaka energiji vezanja.



Slika 4.29: Sedma komponenta prikazana u prostoru prediktora.

Napomenuto je da se PCA često koristi za klasifikaciju podataka. Provjerimo kako metoda, bez navođenja informacije o odgovoru y odjeljuje naše podatke. Sedmero dimenzionalni prostor nemoguće je vizualizirati, ali klasifikaciju je dovoljno, u našem slučaju, prikazati pomoću dviju komponenti s najvećom varijabilnosti. Na slici 4.30 prikazana je klasifikacija podataka. Jezgre s većom energijom vezanja se nalaze na većim vrijednostima prve principalne komponente. Podskupovi jezgara sličnih energija vezanja bi bili jasnije odvojeni, ako bi problem bilo linearnije prirode.



Slika 4.30: Druga principalna komponenta u ovisnosti o prvoj komponenti. Prikazana je skala energije vezanja. Vidljivo je odvajanje podataka na podskupove ovisno o energiji vezanja.

Polinomna svojstva

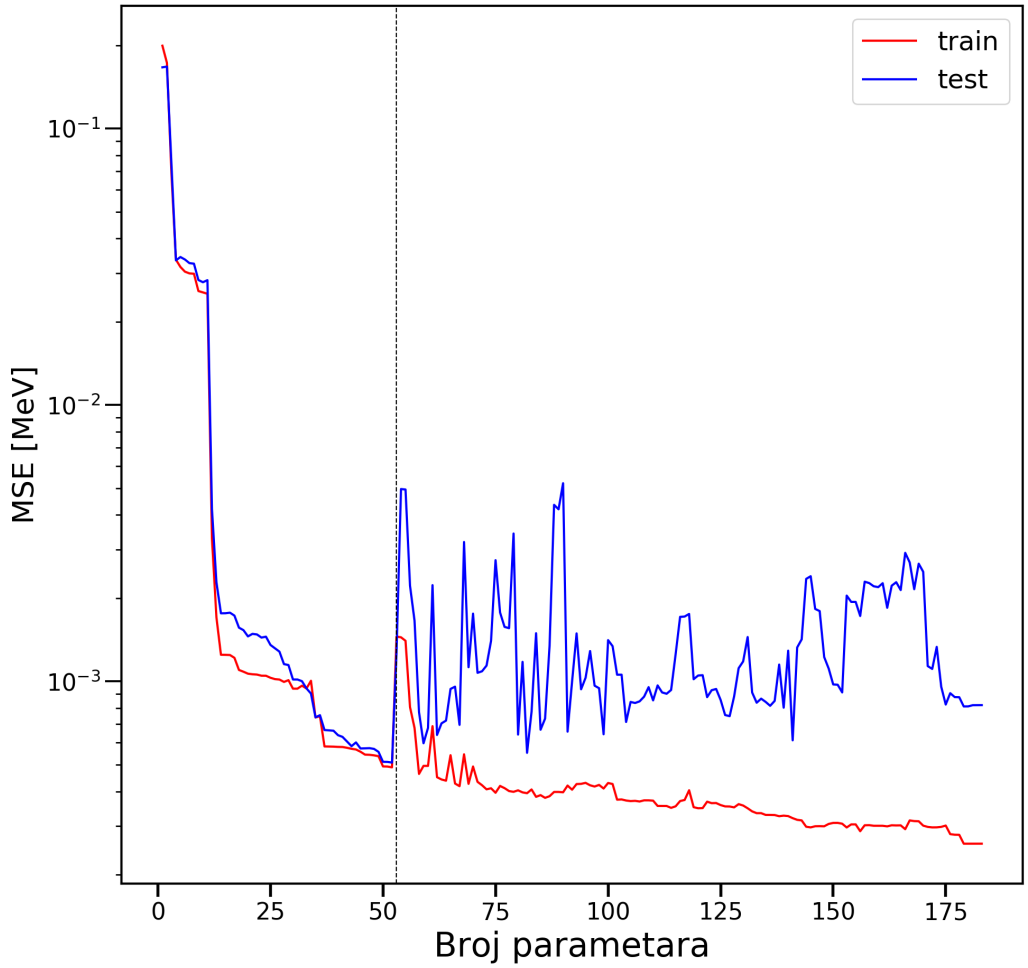
U podatkovnoj znanosti kreiranje novih varijabli iz postojećih uobičajena je metoda koja se koristi ako sumnjamo da u modelu postoje skriveni prediktori. Željeni broj prediktora kreiramo iz postojećih najjednostavnije pomoću funkcije *polinomna svojstva* (eng. *polynomial features*). "Polynomial features" ugrađena je funkcija iz paketa *"Scikit learn"* koja generira polinomne kombinacije ulaznih parametara. Na primjer, "polynomial features" za $n = 2$ tro-dimenzionalni ulazni parametar $[a, b, c]$ pretvara u $[1, a^2, b^2, c^2, ab, ac, bc]$ matricu.

Ulazni parametri odabirani su tako da izlazna matrica sadrži glavne doprinose iz semi-empirijske formule. Doprinosi sparivanja (eng. Pairing, Pair.) te efekt ljusaka (eng. Shel effect, Shell.) dodani su kao članovi bez funkcijske ovisnosti o A, N ili Z. Ulazni članovi glase

$$\begin{aligned}[A^{-4/3}, A^{-1/2}, A^{-1/3}, A^{4/3}, A^{1/2}, A^{1/3},& \\ Z, Z^2, Z^{4/3}, Z^{1/2}, N^{1/2}, (N-Z)^2, N^2 - Z^2,& \\ Pair., Shell.]&\end{aligned} \tag{4.5}$$

Pomoću funkcije kreirano je 264 parametara, neki od tih parametra se ponavljaju te su stoga izbačeni iz konačne matrice. Dodatno smo izbacili sve paramente koji ne zadovoljavaju $VIF \leqslant 15$ kriterij. Konačno, ostaje 183 parametra.

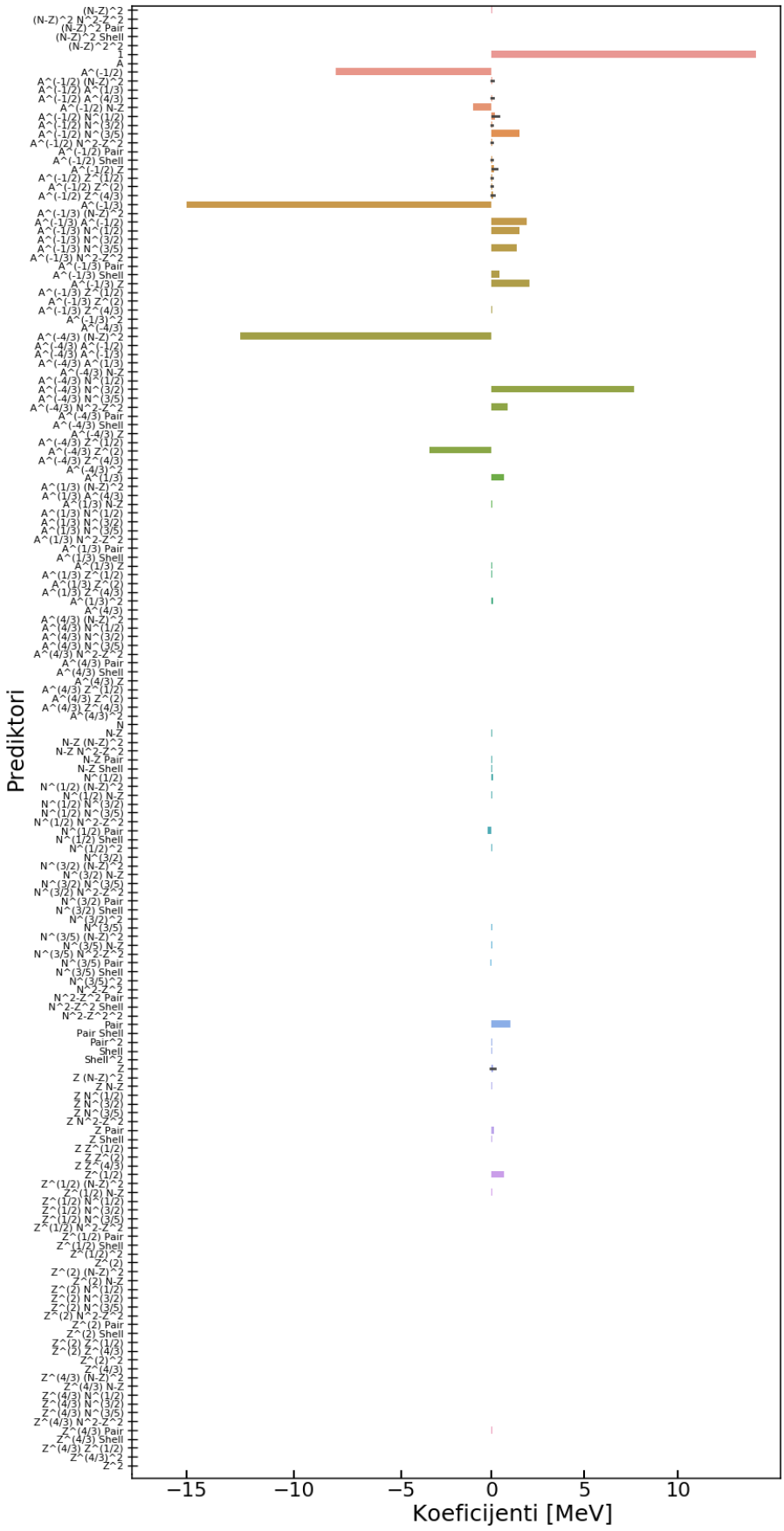
Na izdvojenih 183 parametara primijenjena je linearna regresija te je dobivena pogreška energije vezanja po nukleonu od $RMS_{poly-nuk.}(N = 183) = 0.0162$ MeV-a. Ukupna pogreška energije vezanja je $RMS_{poly}(N = 183) = 1.31$ MeV-a. Kako u modelu imamo veći broj prediktora moramo provjeriti dolazi li do preprilagođenosti, što provjeravamo pomoću trening-test podjele podataka.



Slika 4.31: Vrijednost MSE pogreške u ovisnosti o broju parametara. Plavom bojom prikazana je pogreška test seta, crvenom je pogreška treniranog seta. Test set sadrži 490 nasumično odabranih jezgra.

Sa slike 4.31 iščitavamo da do preprilagođenosti dolazi nakon N=53 uključenog prediktora.

Kako bi dobili bolji pregled važnijih doprinosa koristit ćemo LASSO regularizaciju. Unakrsnom provjerom određen je $\alpha_{LASSO} = 7.38 \times 10^{-3}$, dobiveni iznosi koeficijenata prikazani su na slici 4.32.



Slika 4.32: Koeficijenti prilagodbe za parametar regularizacije $\alpha_{LASSO} = 7.38 \times 10^{-3}$. Pogreška energije vezanja za odabran α_{LASSO} iznosi $RMS_{LASSO}(N = 183) = 1.974$ MeV-a.

LASSO regularizacija zadržala je 51 prediktor. Uz odabran parametar regularizacije $\alpha_{LASSO} = 7.38 \times 10^{-3}$ dobiva se pogreška energije vezanja $RMS_{poly-LASSO}(N = 183) = 1.97$ MeV-a. Prema iznosu koeficijenta izdvajamo konstantan član, $A^{-1/3}$. Primjećujemo da se radi o volumnom te površinskom doprinosu. U modelu je prisutan Coulombov doprinos, $Z^2A^{-4/3}$, te nepoznati član $A^{-4/3}(N-Z)^2$. Nepoznati član opisuje razliku naboja jezgre kroz njen radijus te bi mogao služiti kao dopuna članu asimetrije. Značajan član koji se pojavljuje je $N^{3/5}A^{-4/3}$, koji uz konstantan volumni član ima najznačajniji doprinos energiji vezanja po nukleonu. Član sparivanja se u modelu pojavljuje kao samostalan član te uz potencije $A^{-1/3}$, Z, $Z^{4/3}$.
Hrbat-regresija zadržava sve članove u modelu stoga nije pogodna za analizu pojedinačnih doprinosa.

5 Zaključak

Statističke metode učenja pokazale su se kao izrazito dobar alat prilikom analize semi-empirijske formule masa.

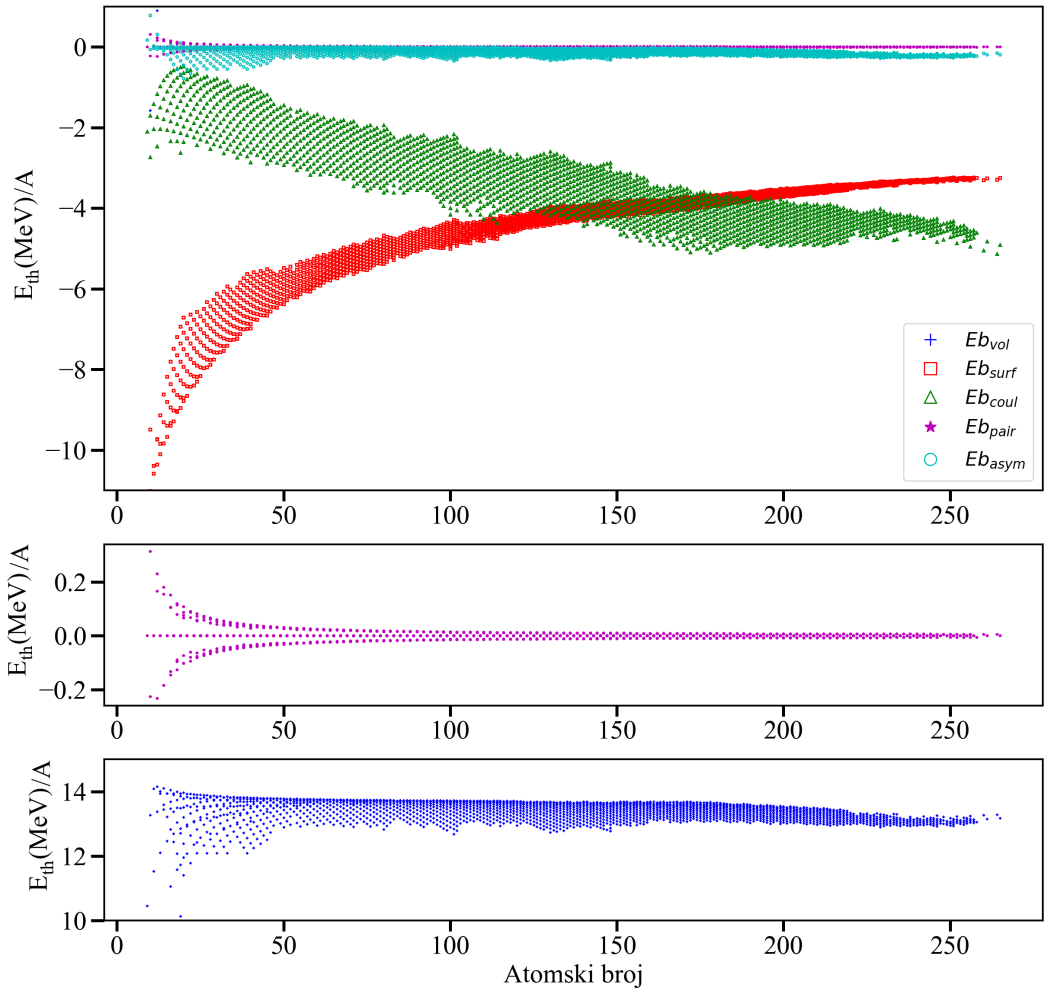
Kod sustava gdje je ulazni broj prediktora malen, kao što je bio slučaj za BW formulu samo s pet osnovnih članova, metoda najmanjih kvadrata pronašla je iznose koeficijenata (tablica A.2.) te je ukupna pogreška energije vezanja takvog modela $RMS(N = 5) = 3.004$ MeV-a. Ovaj model služio je kao polazna točka daljnjim istraživanjima kako smo povećavali broj prediktora. Modelu od $N = 5$ prediktora pridruženo je dodatnih šest članova koji su najzastupljeniji unutar znanstvene zajednice za opis makroskopskih efekata koji utječu na energiju vezanja jezgre. Pogreška energije vezanja za taj model iznosila je $RMS(N = 11) = 2.256$ MeV-a. Model od $N = 11$ članova linearno je proširen članovima po uzoru na deformacije uvedene u članku [5]. Dobiven model sadrži 119 članova tj. 9 osnovnih članova i njihove deformacije te dva neproširena člana: efekt ljusaka i njegov kvadrat.

Sumiramo proširenja za pojedini doprinos energiji vezanja po nukleonu

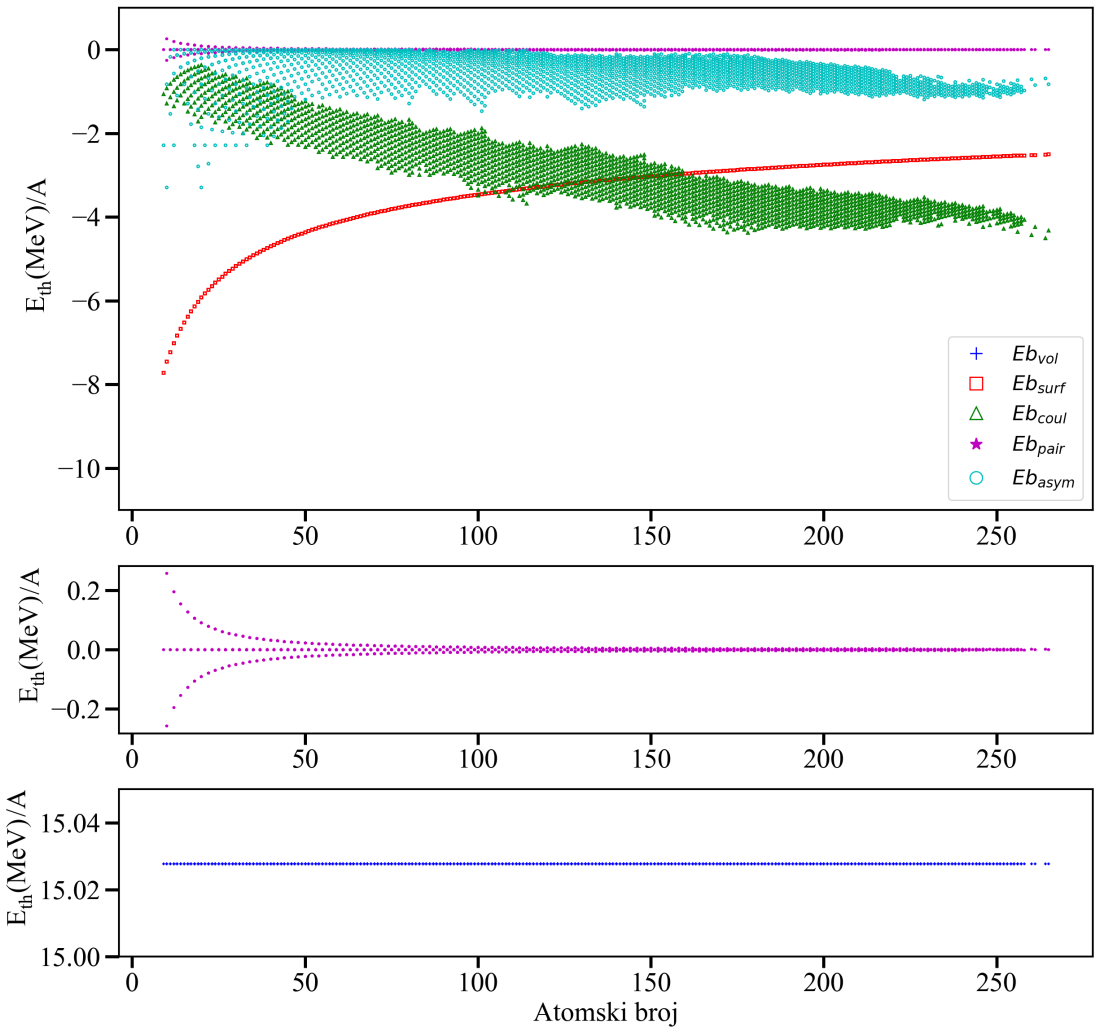
$$E_k/A = \sum_{j=0}^{12} a_j^k k v_j \tag{5.1}$$

za k= vol, surf, coul, pair, asym, gdje je $v_0 = 1$, a a_j^k su prediktori dobiveni hrbat-regresijom uz parametar sažimanja $\alpha_{hrbat} = 0.58$. Dobivena suma predstavlja doprinos pojedinog člana *k* i njegovih deformacija, što je prikazano na slici 5.1.

Rezultate možemo usporediti s doprinosima koji su dobiveni linearnom regresijom. Prvu bitnu razliku uočavamo kod volumnog doprinosa, on je puno širi i maksimalna vrijednost mu je niža nego što je to slučaj kod neproširenog člana. Deformacije površinskog člana povećavaju površinski doprinos za male vrijednosti masenog broja. Deformacije vezane u površinski i volumni doprinos značajnije su po iznosu od ostalih deformacija. Član asimetrije reguliran je smanjenjem vrijednosti njegovih deformacija.



Slika 5.1: Utjecaj proširenja na pojedinačan član iz BW formule. Korišteni koeficijenti su dobiveni hrbat-regresijom za parametar regularizacije $\alpha_{Hrbat} = 0.58$, pogreška energije vezanja po nukleonu ovog modela je $RMS_{nucl.-Hrbat}(N = 119) = 0.022$ MeV-a, pogreška energije vezanja je $RMS_{Hrbat}(N = 119) = 1.772$ MeV-a.



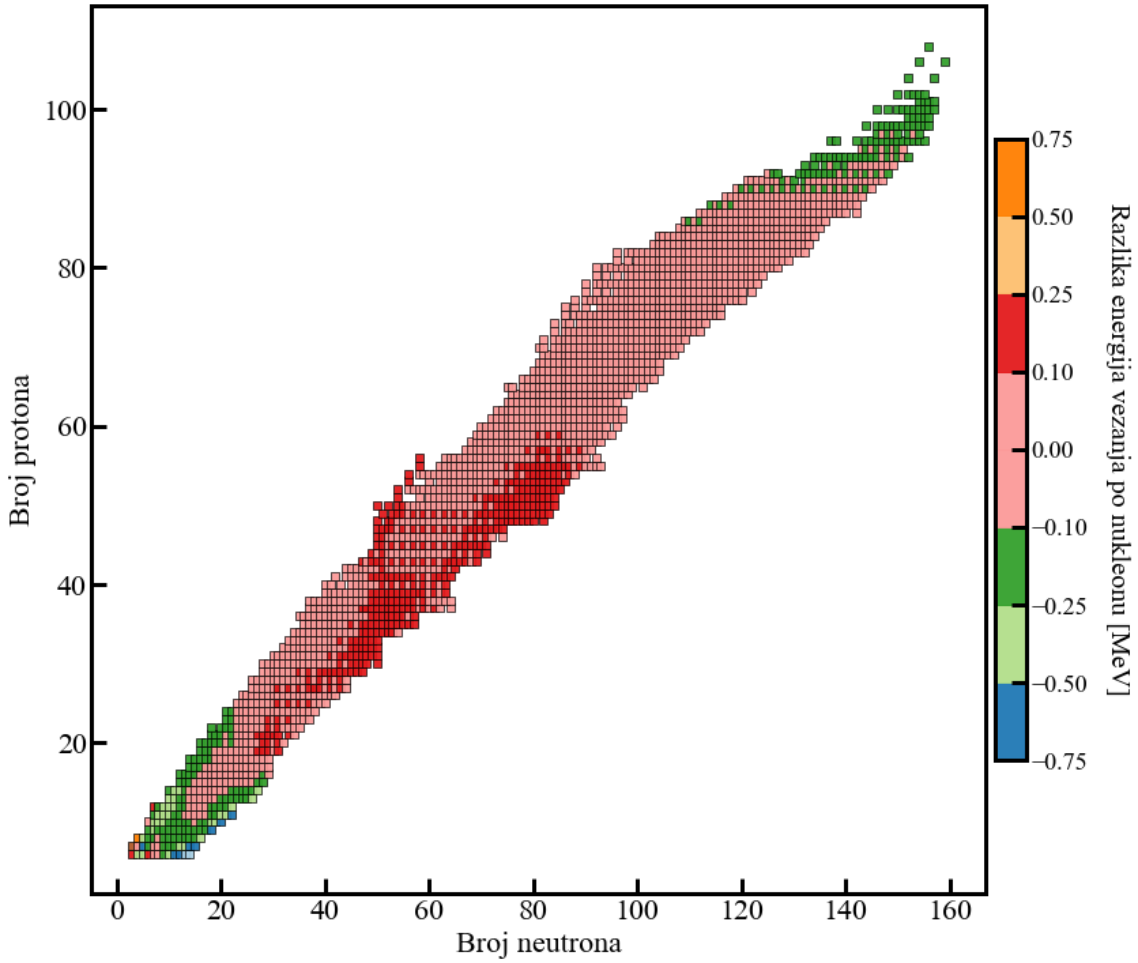
Slika 5.2: Doprinosi pojedinačnih članova za BW formulu. Korišteni su koeficijenti dobiveni linearnom regresijom prikazani u tablici A.2. Pogreška energije vezanja je $RMS(N = 5) = 3.004$ MeV-a.

Metoda najmanjih kvadrata je za model od $N = 119$ članova davala vrijednosti koeficijenata koji nisu u skladu s ranijim istraživanjima. Modeli dobiveni MNK su visoko fleksibilni te posjeduju jako malu moć generalizacije, stoga je u sustav nužno uvesti regularizacije.

Kvaliteta modela koji se dobije regularizacijom ovisan je o odabiru parametra sažimanja, α. Unakrsna provjera predstavljala je optimalno rješenje za taj problem, jer nije bilo potrebno izdvojiti određen broj jezgara za testiranja.

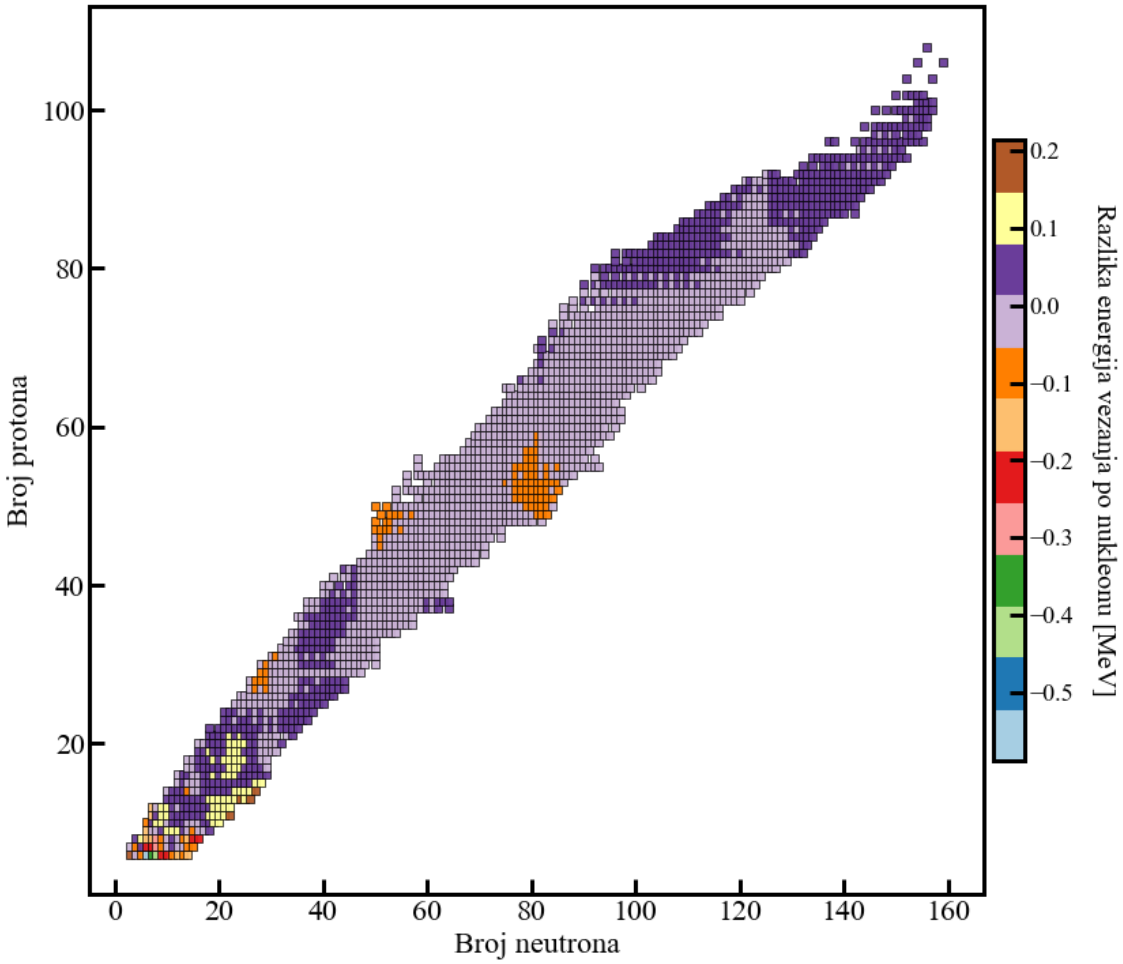
Hrbat-regresija, za pronađeni parametar α pomoću unakrsne provjere, je dala sažete koeficijente koje smo mogli analizirati. Dominantni članovi bili su neproširени članovi te je primijećen snažan utjecaj efekta ljusaka na ukupnu energiju vezanja.

Regularizacije su efikasan alat i kada imamo vrlo malo informacije o mogućim prediktorima. Kreiran je model od 264 parametara kod kojih većina članova gotovo i nema fizikalnu pozadinu, tj. kreirani članovi su služili kao šum članovima čija fizikalna pozadina je poznata. LASSO regresijom izdvojeni su isključivo prediktori poznate fizike, manji broj prediktora bio je nepoznat. Dominantniji, nepoznati članovi imaju ponašanja proporcionalna s razlikom protonskog i neutronskog broja te su obrnuto proporcionalna radijusu jezgre što može biti indikacija na kvantne učinke koji nisu poznati na makroskopskoj skali energije vezanja. Pregled kod kojih jezgara imamo najveća odstupanja modela od eksperimenta prikazan je na mapi nuklida, slike 5.3., 5.4., 5.5 i 5.6.



Slika 5.3: Mapa nuklida za model $N = 119$ prediktora prikazana u energijskoj skali razlike modela i eksperimenta. Iznosi koeficijenata dobiveni su hrbat-regresijom, parametar regularizacije $\alpha_{Hrbat} = 0.58$.

Sa slika 5.3. i 5.4. vidimo da do najvećeg odstupanja dolazi u području A od 70 do 150, gdje znamo da su prisutne deformacije jezgara. LASSO regularizacija nešto bolje opisuje to područje na karti te su odstupanja svedena na područje $A \approx 140$.



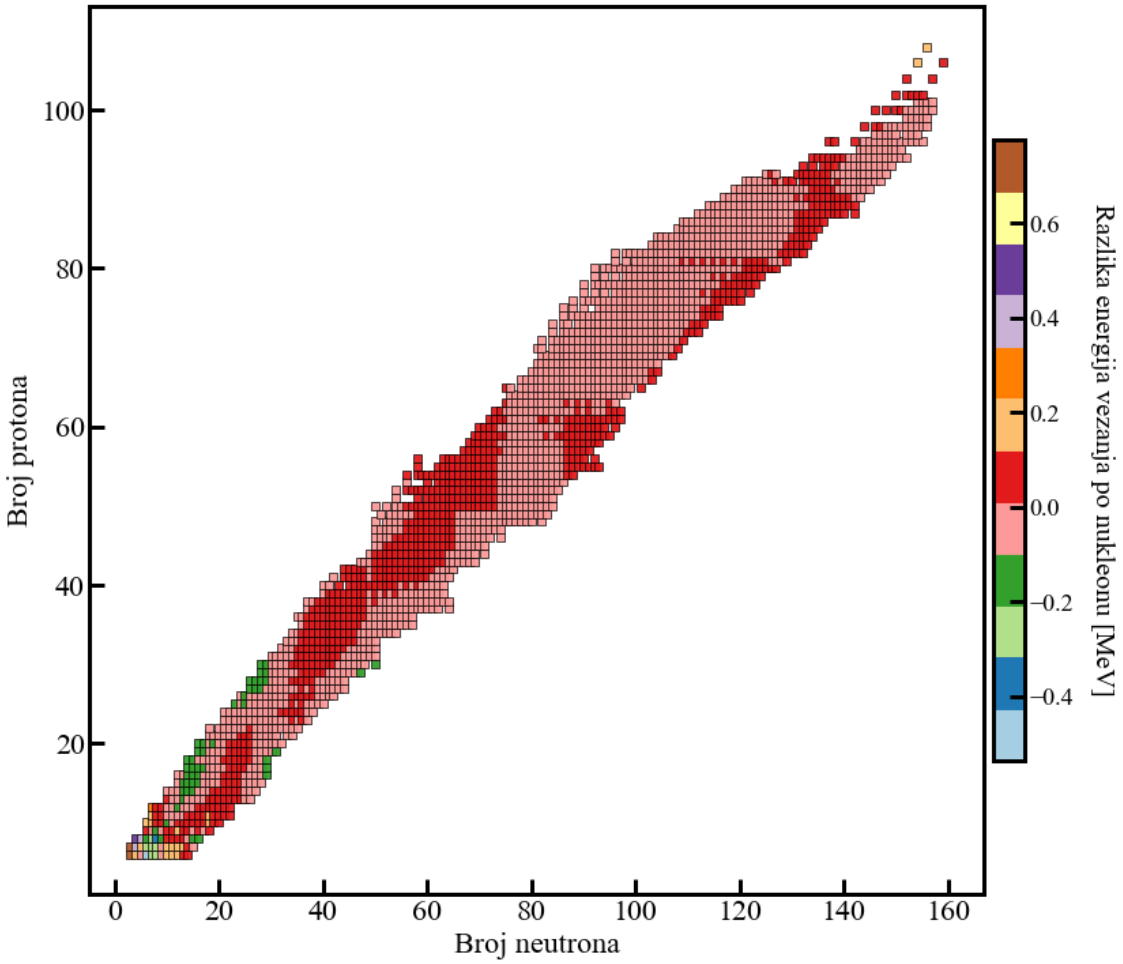
Slika 5.4: Mapa nuklina za model $N = 119$ prediktora prikazana u energijskoj skali razlike modela i eksperimenta. Iznosi koeficijenata dobiveni su LASSO regresijom, parametar regularizacije $\alpha_{LASSO} = 1.64 \times 10^{-4}$.

Kod modela dobivenog za $N = 183$ prediktora situacija je drugačija zato što nemamo jedinstvene prediktore koji bi unijeli točnu fiziku u model, stoga su razlike predviđanja i eksperimenta prisutne duž više jezgara.

Dodavanje novih članova sustavu je opravdano, ako novo pridodan član unosi novo fizikalno ponašanje u model. Povećavanjem kompleksnosti, povećavamo varijabilnost na štetu generalizacije, zato su recimo modeli, s većim brojem prediktora, koji su dobiveni metodom najmanjih kvarata, manje praktični za interpretaciju te je teže dobiti novi uvid u samu fiziku atomske jezgre. Preprilagođenost podataka efikasno se izbjegava regularizacijama, želimo li jednostavniji model koristiti ćemo LASSO regularizaciju, dok će hrbat-regresija davati nešto preciznije rezultate.

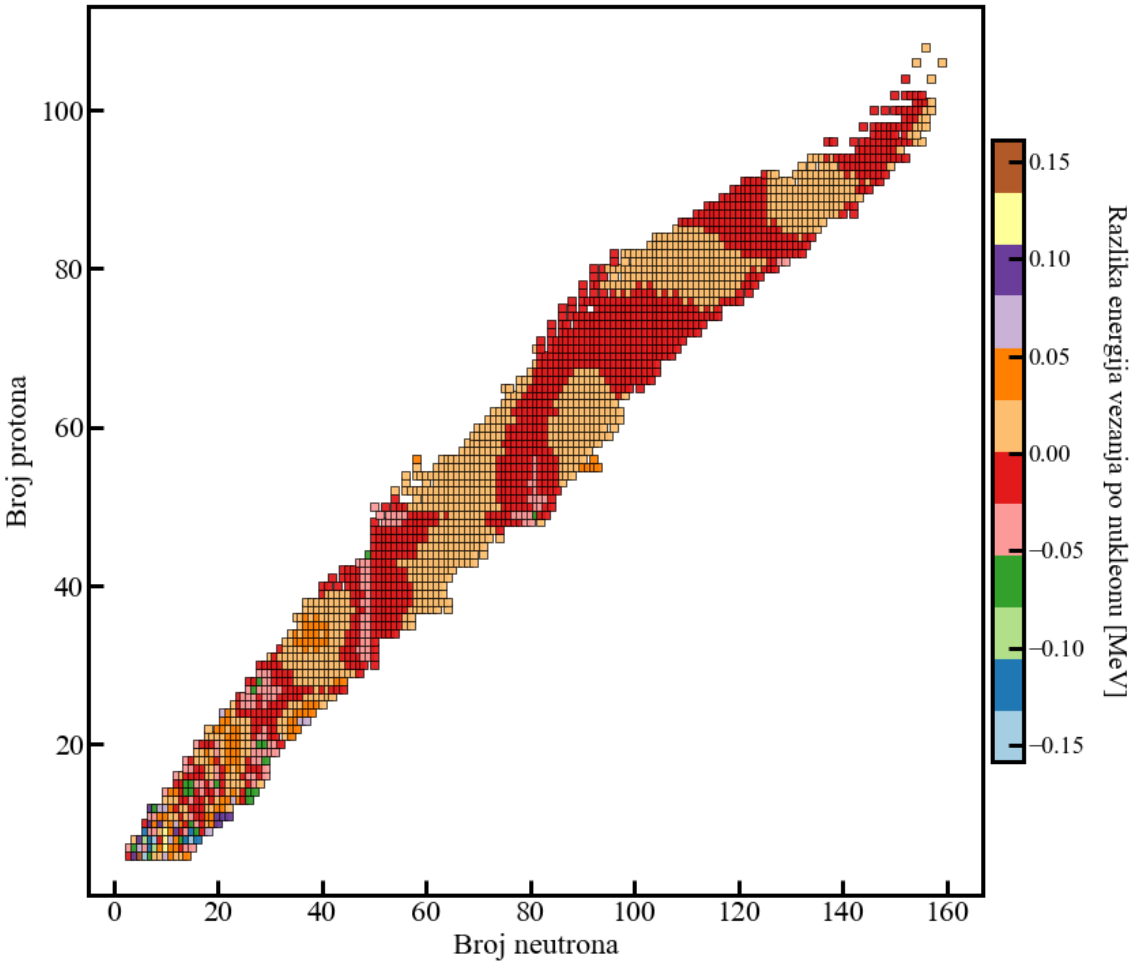
Rezultati ovog rada predstavljaju polaznu točku za daljnja istraživanja, ako je dovoljno uključiti samo linearna proširenja.

Linearna proširivanja, u usporedbi s nelinearnim, pokazala su se tehnički lakšima



Slika 5.5: Mapa nuklida dobivena za model $N = 183$ prediktora. Iznosi koeficijenata dobiveni su LASSO regresijom, parametar regularizacije $\alpha_{LASSO} = 7.38 \times 10^{-3}$.

za izvedbu na vrlo malu štetu pogreške. Članak [5] je imao tehnički vrlo zahtjevne nelinearne modele gdje je pogreška u iznosila 1.11-1.53 MeV-a, dok je ovim radom postignuta minimalna pogreška, uz regularizacije $RMS_{Hrbat}(N = 119) = 1.772$ MeV-a.



Slika 5.6: Mapa nuklida dobivena za model $N = 183$ prediktora. Iznosi koeficijenata dobiveni su hrbat-regresijom, parametar regularizacije $\alpha_{Hrbat} = 4.51$.

Literatura

[1] Yarkoni, T.; Westfall, J. : Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning // Perspectives on Psychological Sci., Vol 12., 6(2017), str. 1-23.

[2] Thi- Vu, M.A.; Tulay, A.; Demba, B.; Buzsaki, G.; Carlson, D. : A Shared Vision for Machine Learning in Neuroscience // JNeurosci. Vol. 38, 7(2018), str. 1601-1607

[3] Kaur, G.; Srivastava, M.; Kumar, A. : Speaker and Speech Recognition using Deep Neural Network // International Journal of Emerging Research in Management and Technology. Vol. 6, 8(2017), str. 118-123.

[4] Vossen, A. : Basics of Feature Selection and Statistical Learning for High Energy Physics // Part of the proceedings of the Track Computational Intelligence for HEP Data Analysis, `https://cds.cern.ch/record/1094673/files/p1.pdf`, 16.5.2008.

[5] Mavrodiev, S. Cht. : Improved Numerical Generalization of Bethe- Weizsacker Mass Formula // Nuclear Theory 35 (2016) 288. arXiv:1607.07217

[6] Kirson W. M. : Mutual influence of terms in a semi-empirical mass formula // Nuclear Physics A, 798 (2008), str. 29-60

[7] Jean-Louis Basdevant, J. L.; Rich, J.; Spiro, M. Fundamentals in Nuclear Physics: From Nuclear Structure to Cosmology., 1st ed. New York : Springer, 2005.

[8] Semi-empirical mass formula, `https://en.wikipedia.org/wiki/Semi-empirical_mass_formula`, 16.7.2012.

[9] Hastie, T.; Tibshirani, R.; Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed.: Springer, 2016.

[10] James G.; Witten D.; Hastie T.; Tibshirani R. An Introduction to Statistical Learning: with Applications in R. 1st ed. 7th printing : Springer, 2017.

[11] The 2016 Atomic Mass Evaluation, Atomic Mass Data Center, http://amdc.in2p3.fr/masstables/Ame2016/mass16.txt, 1.5.2017.

[12] K. Heyde Basic Ideas and Concepts in Nuclear Physics: An Introductionary Approach, 2nd ed. : Bristol and Philadelphia, 1999.

[13] Von Weizsacker, C. F. : Zur Theorie der Kernmassen // Zeitschrift für Physik, 96 (1935), str. 7-8.

[14] Bailey, D. : Semi-empirical Nuclear Mass Formula // String and Binding Energy, 3rd ed. Toronto: Springer, 2011.

[15] D. J. Galić : Statističko učenje, diplomski rad // voditelj rada: prof. dr. sc. M. Huzak, Sveučilište u Zagrebu, Matematički odsjek, 2018.