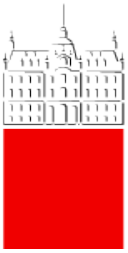
Univerza v Ljubljani
Fakulteta za matematiko in fiziko



6. naloga
Luščenje modelskih parametrov: linearni modeli

Miha Srdinšek

Povzetek

Gre za prileganje krivulj meritvam. Najprej si pogledamo enostaven primer, nato si pogledamo primer grdih podatkov, ki jim želimo zgolj najti čim boljši model, na koncu pa imamo dve datoteki podatkov, ki jih želimo tako sestaviti, da dobimo tretjo datoteko. Poznao dva absorbcijska spektra in ju želimo sestaviti tako, da dobimo naš absorbcijski spekter, ki je sestavljen iz teh dveh.

I. Farmakološke meritve

V prvi nalogi imamo zelo malo danih podatkov, katerim moramo prilagajati model

$$y = \frac{y_0 x}{x + a}, \tag{1}$$

pri čemer seveda iščemo konstanti y_0 in a. Pri tem ima vsaka meritev enako napako vredno 3 enote. Najprej enačbo linearizirамo, da bomo lažje naši prilagoditveno funkcijo. Zapišemo

$$\frac{1}{y} = \frac{1}{y_0} + \frac{a}{y_0}\frac{1}{x} \tag{2}$$

in pri tem pazimo, da se transformira tudi napaka, ki sedaj znaša

$$\frac{1}{y_i} = \frac{1}{y_i} \pm \frac{\sigma}{y_i^2}, \tag{3}$$

pri čemer i predstavlja indeks meritve, σ pa je tako ali tako ves čas velika 3 enote. Sedaj se namesto že pripravljenih programov za prilagajanje modelov, poslužimo kar ročnega programiranja metode najmanjših kvadratov. Za tako metodo moramo rešiti matrično enačbo

$$A\vec{a} = \vec{b} \tag{4}$$

za $\vec{a}$, kjer sta

$$A_{kj} = \sum_{i=1}^{N} \frac{\phi_j(x_i)\phi_k(x_i)}{\sigma_i^2} \qquad in \qquad b_k = \sum_{i=1}^{N} \frac{y_i \phi_k(x_i)}{\sigma_i^2}. \tag{5}$$

Pri tem sem z ϕ_j mislil čene linearne funkcije 2, v tem primeru to pomeni:

$$\phi_1(x_i) = 1 \qquad in \qquad \phi_2(x_i) = \frac{1}{x_i}. \tag{6}$$

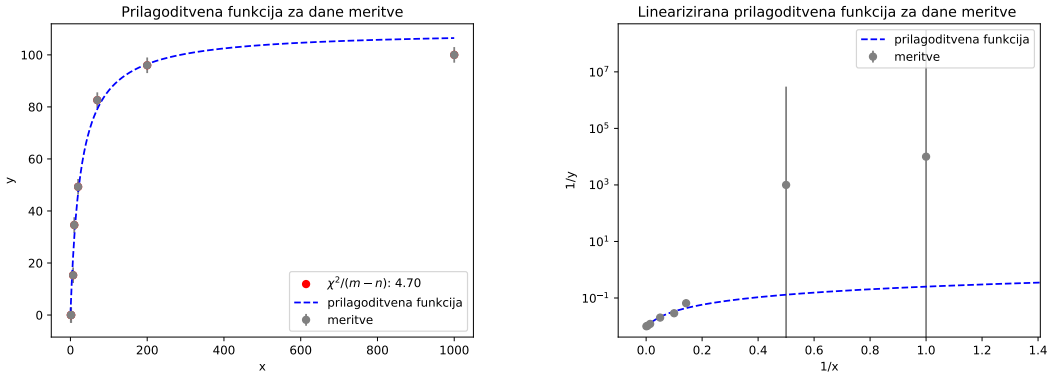
Ko to napišemo in rešimo sistem enačb 4 dobimo za dane podatke rešitvi

$$a = 26.6 \;\; in \;\; y0 = 109.3 \quad (\chi^2 = 23.5), \tag{7}$$

oziroma rešitve, če sistem reši7mo analitično

$$a = \frac{A_{xx}A_y - A_x A_{xy}}{A_{xx}A - A_x^2} = 26.6 \;\; in \;\; y0 = \frac{A A_{xy} - A_x A_y}{A_{xx}A - A_x^2} = 109.3 \quad (\chi^2 = 23.5), \tag{8}$$

ki data rešitve prikazani na grafih 1.



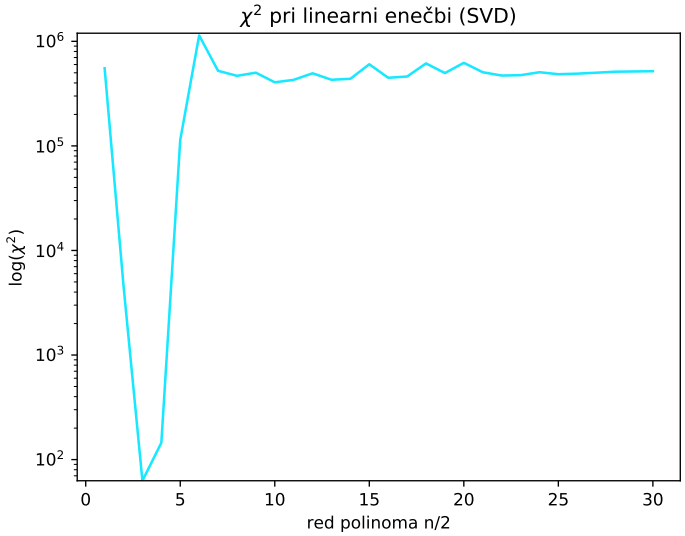
Slika 1: *Sliki prikazujeta prilagoditveno funkcijo in meritve katerim se prilega. Vidimo, da kljub velikim odstopanj, v linearizirani obliki (pri tem ostane funkcija znotraj intervala napak!) v dani obliki dobimo na oko čisto zadostno prileganje.*

II. Visokoločljivostni magnetni spektrometer

Pri drugem delu naloge se soočimo z zelo dolgim seznamom podatkov. Za vsako od treh količin, ki jih merimo imamo 11665 meritev. Gre za to, da v magnetnem polju uklanjamo snop nabitih delcev. Tem zmerimo položaj preden priletijo v magnetno polje in nato položaj kjer zadanejo tarčo. Tem žeimo prilagajati neko polinomsko vrsto potenc ali nekih drugih funkcij. Potenčna vrsta bi morala recimo biti oblike

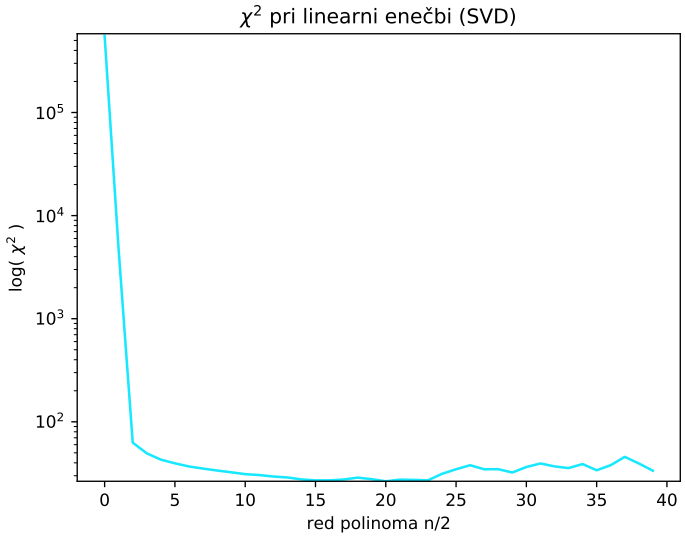
$$\theta_{tarca} = \sum_{i,j}^{n} a_{i,j} x^i_{detektor} \theta^j_{detektor}, \tag{9}$$

pri čemer je $q_{i,j}$ koeficient pred vsakim takim členom, katerega vrednost iščemo. Ker je podatkov tako zelo veliko se raje poslužimo SVD algoritma, ki je že vsebovan v pythonovi funkciji $numpy.linalg.lstsq()$. Ko poženemo algoritem za dane podatke dobimo prilegajočo funkcijo katere χ^2 je odvisen on stopnje polinoma (torej maksimalnega števila i in j v zgornji enačbi) tako kot je pirkazano na sliki 2. To je izredno grda prilegajoča funkcija, χ^2 pa se nižjim



Slika 2

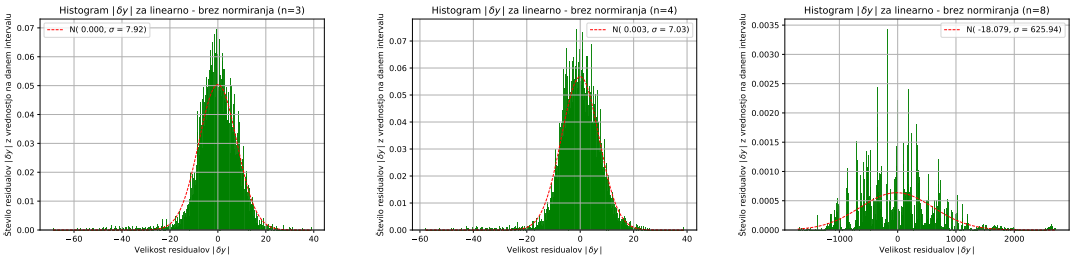
vrednostim približa le za načetku pri stopnji $n = 3$. Reducirani χ^2 ima tam vrednsot okoli 50. Nastalo situacijo se da rašiti z zelo preproto opazko. Najvišja številka s katero imamo opravka je vrednosti okoli 160 in seveda potence takšne vrednosti res močno skalirajo. Člen kjer se znajdeta dve številki velikostnega reda 10^2 oba pod potenco 5 je torej že reda 10^{20}. Pri tako velikih številih seveda pride takoj do napak, saj kar naenkrat operiramo z vrednostmi števil na očitno preširokem intervalu. Zato sem se odločil da vse meritve normiram s tem številom, kar mi bo prišlo prav tudi kasneje. Dobim precej lepšo odvisnost prikazano na sliki 3.



Slika 3

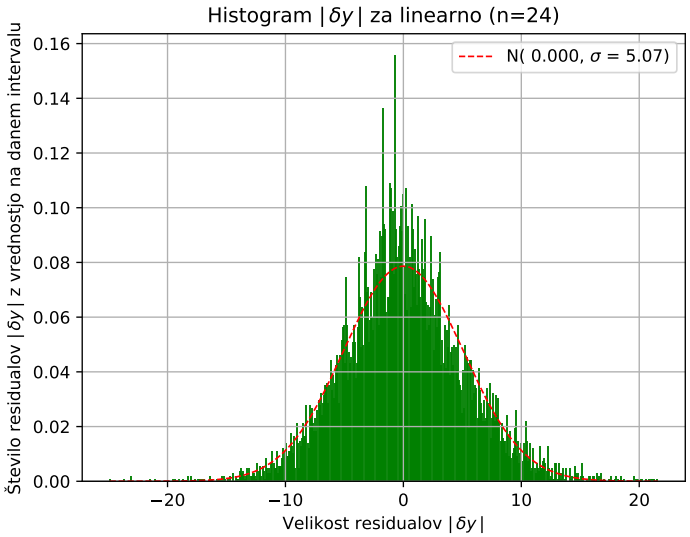
Tu se χ^2 lepo znižuje do stopnje 24, kjer vsebuje polinom kar 576 členov in prav toliko parametrov. Tam doseže reducirani χ^2 vrednost 31, nato pa se zopet prične dvigovati in plesati. Našli smo torej najbolšo stopnjo prilagoditvenega polinoma. Višji redovi polinoma, nam očitno že odstopajo od dejanskega zakona, ki se mu pokoravajo meritve.

Še en pogled na isto dogajanje je pogled skozi oči histogramov residualov. Pogledamo si koliko residualov se nahaja na nekem intervalu (koliko residualov je določene velikosti). Pričakovali bi, da za dobro prilagoditveno funkcijo dobimo lepo normalno porazdelitev residualov. Pri prem načinu je histogram zelo lep v točki z minimalnim χ^2, a že čim se premaknemo za eno stopnjo višje je porazdelitev manj normalna, pri čemer je pri stopnji 8 že močno razmetana. Napake so pri stopnji 8 tudi izredno velikoe reda 1000 kot lahko vidimo na sliki 4.



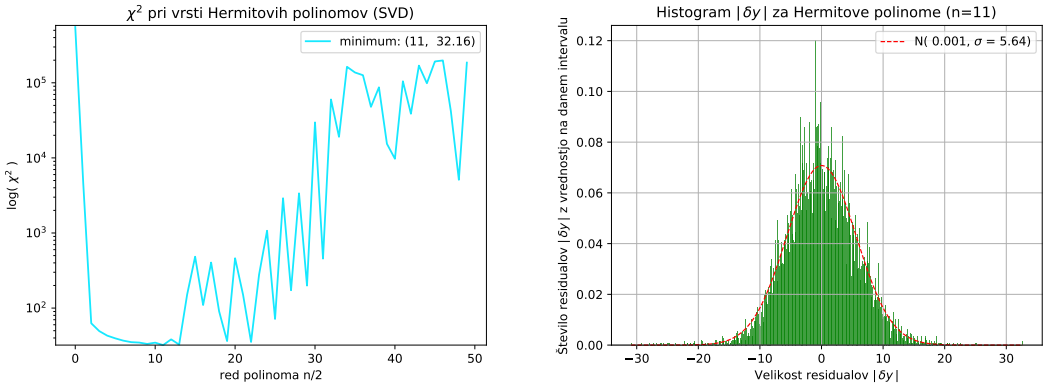
Slika 4: *Slike prikazujejo histogram residualov pri različnih radovih prilagoditvenega polinoma, če ne normiramo podatkov. Gre torej za histogram za točke na sliki 2. Pri vsaki je še izrisana najbližja normalna funkcija in v legendi parametri le te.*

Histogram za normirane podatke je pač zelo lep, kot lahko vidimo na sliki 5.



Slika 5: *Zraven je izrisana še najbližja normalna funkcija in v legendi parametri le te.*

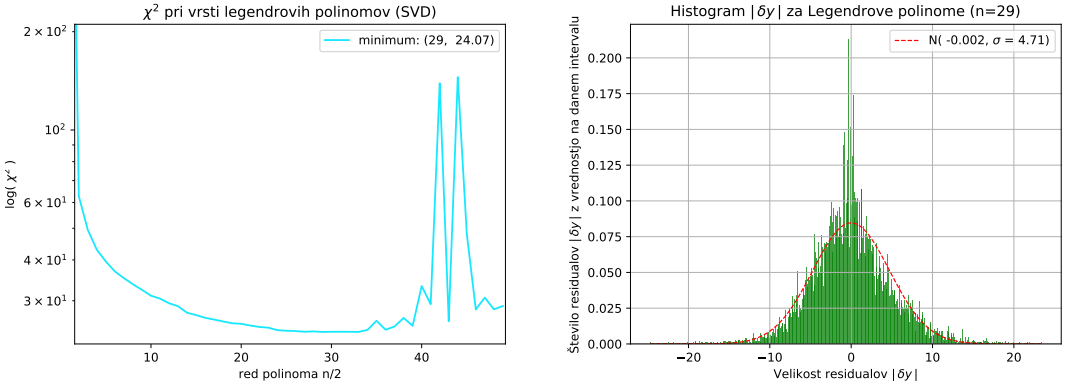
Slika 5 nas torej še dodatno prepriča, da je to primerna prilagoditvena funkcija. Vseeno si želimo pogledati še kakšne druga možnosti. Omenili smo, da so primerne tudi katerekoli druge funkcije, poleg tega pa sedaj zahtevamo še da so normirane, saj so s tem dosti lepše. Poglejmo si recimo za začetek Hermitove polinome. Pri tem zgolj zamenjamo potenco za argument oziroma stopnjo hermitovega polinoma in zopet vse meritve normiramo, saj funkcija srejme le vrednosti na intervalu $[-1, 1]$, in že lahko izvedemo metodo najmanjših kvadratov. Dobimo sliki 6, na katerih vidimo, da reduciranega χ^2 nismo uspeli pretirano zmanjšati in da prilegajoča vrsta zelo hitro podivja. To bi sicer lahko pričakovali, saj hermitovi polinomi res ne ikazujejo karakteristik, ki jih iščemo. Iščemo namreč polinome s precej visokimi frekvencami period, ker tudi naši podakti izgledajo raho šinusno". Zgornji zgled nas sicer ni pretirano opogumil glede iskanja



Slika 6: *Leva slika prikazuje odvisnost χ^2 od stopnje polinomov (v legendi je označena najnižja vrednost), desna pa histogram residualov, pri čemer je zraven izrisana še najbližja normalna funkcija in v legendi parametri le te.*

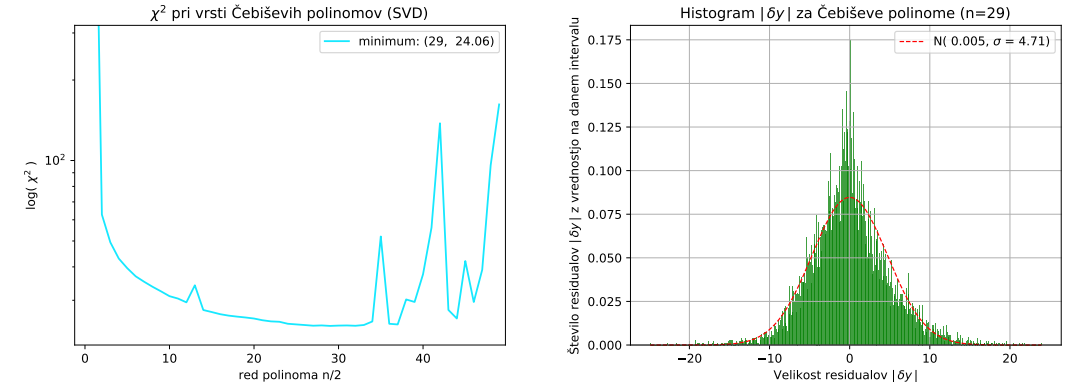
novih ortoganilnih funkcij za prilagodiveni polinom, a splača se vstrajati in poskusiti z bolj očitnim poskusom - Legendrovimi funkcijami. Če z njimi zapišemo funkcijo po dozdajšnjem zgledu dobimo sliki 7. Sedaj smo uspeli reducirani χ^2 opazno zmanšati. Te vrednosti so nam že čisto povšeči. POleg tega vidimo, da precej nizke vrednosti dosežemo že pri nižjih stopnjah

polinomov, tako da ne rabimo izvesti ravno najbolj natančne prilagoditvene funkcije pa smo še vedno bolj natančni kot pri potenčni vrsti (produktu dveh potenčnih vrst). Pogum, ki smo ga



Slika 7: *Leva slika prikazuje odvisnost χ^2 od stopnje polinomov (v legendi je označena najnižja vrednost), desna pa histogram residualov, pri čemer je zraven izrisana še najbližja normalna funkcija in v legendi parametri le te.*

dobili s tem poskusom nas vodi v iskanje novih ortogonalih fukcij s katerimi bi lahko zapisali takšen polinom. Očiten poskus bi bili polinomi Čebiševa, saj smo jih omenjali že tudi na vajah. Z njimi dobimo, kot vidimo na sliki 8, najboljši model do sedaj. Dobimo reducirani $\chi^2 = 24,06$.



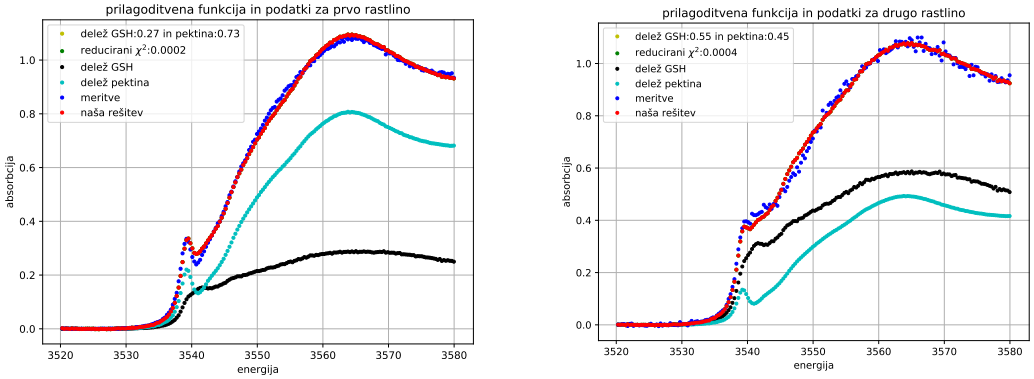
Slika 8: *Leva slika prikazuje odvisnost χ^2 od stopnje polinomov (v legendi je označena najnižja vrednost), desna pa histogram residualov, pri čemer je zraven izrisana še najbližja normalna funkcija in v legendi parametri le te.*

Takšen rezultat je že res dober. Poskusili bi lahko še s kakšnimi ortogonalnimi funkcijami, ampak prav verjetno nebi bii uspešni. Sam sem recimo poskusil še s funkcijami *mqthieuce* in *se*, pri isti vrednosti parametra q, a sem dobil porazne rezultate, zato jih tu sploh ne bom prikazal. Zaključil bi, da so verjetno polinomi Čebiševa najboljše feunkcije za te meritve.

III. Rendgenski absorbcijski robovi

Pri tej nalogi imamo podane meritve absorbcijskih spektrov kadmija (Cd), ki se skriva v restlinah v družbi dveh različnih elemetov. Ravno od tega, kateri element je v bližini Cd je odvisen absorbcijski spekter, zato lahko ocenimo deleže teh primesi. Absorbcijski spekter je odvisen od parov Cd-O in Cd-S, zato imamo podane tudi meritve absorbcijskih spektrov za ločene primesi Cd-O in Cd-S. Vemo, da je absorbcijski spekter sestavljen linearno in zato lahko s prilagajanjem teh ločenih meritev, absorbcijskemu spekru rastlin, določimo deleže žvepla in kisika. Opazujemo dve vrsti rastlin, od katerih je ena (Thlaspi) znana kot hiperkumulator težkih kovin, o drugi pa nimamo posebnih podatkov. Iz samih meritev bomo torej lahko morda celo zaključili katere meritve pripadajo kateri rastlini in s tem potrdili pravilno razporeditev podanih podatkov.

Če kar takoj izvedemo ta poskus se znajdemo pred problemom, da je reducirani χ^2 kar za tri velikostne rede manjši od 1, zato moramo najprej oceniti novo napako meritev, da bomo dobili boljše rešitve. A vseeno si poglejmo kako izgledajo rešitve sedaj, da jih bomo lahko primerjali s tem kar bomo dobili s popravkom (slika 9).



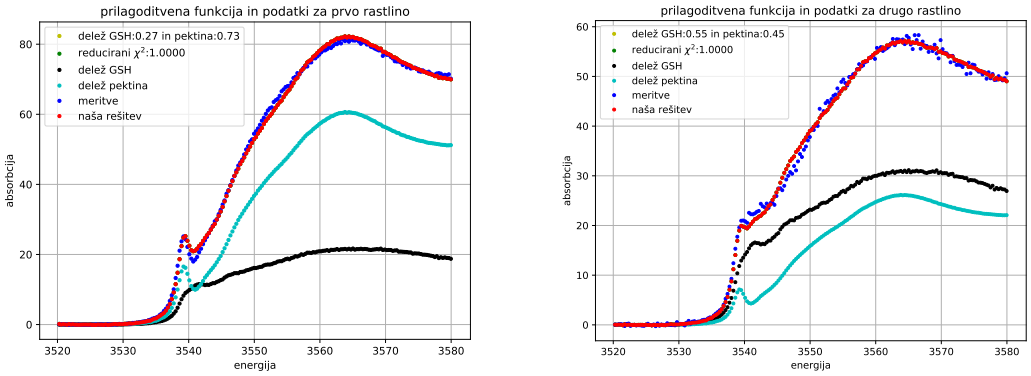
Slika 9: *Na slikah vidimo prispevke različnih sosednjih elementov k spektru.*

Sedaj pa iz našega χ^2 ugotovimo kakšna je optimalna izbira napake s tem, da zahtevamo da mora biti vrednost enaka 1.

$$\frac{\chi^2}{\sigma^2} = 1 \tag{10}$$

Dobili smo torej novo napako, ki bo znašala za prvo rastlino $\sigma^2 = 0,000177294784184$ in za drugo rastlino $\sigma^2 = 0,000355617427079$.

Vidimo na slikah 10, da je bilo torej to popolnoma nepotrebno, kar bi lahko sklepali že od začetka, saj smo imeli na začetku povsod enako napako in če povsod isto napako spremenimo v neko novo povsod enako napako, ne bomo nič spremenili. Uteži bodo še vedno v enakem razmerju med sabo, le absoutna veliksot se bo spremenila. Gejmo zatorej raje sliko 9, saj ima vse vrednosti lepo normirane. in si pač mislimo, da težko govorimo o tem kolikšna je dejanska vrednost χ^2. Sliki 9 nam lepo pokažeta, da ima prva rastlina precej nižjo koncentracijo mešanice GSH, ker pomeni Cd-O in zelo veliko koncentracijo Cd-S. Pri drugi rastlini se koncentracija O precej poveča, a se tudi koncentracija S zmanjša, tako da imata obe približno enako koncentracijo.Lahko bi torej zaključili, da druga rastlina poskra vase bolj enakomerno vse mešanice, pri čemer ima prva rastlina bolj sofisticirane metode črpanja mešanice s O kot mešanice z S.



Slika 10: *Na slikah vidimo prispevke različnih sosednjih elementov k spektru.*