

106: Luščenje modelskih parametrov

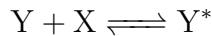
Peter Rupnik

21. november 2018

1 Prva naloga

1.1 Naloga

V farmakologiji merijo odziv tkiv na različne reagente. Za večino teh pojavov lahko privzamemo, da gre za reakcijo, kjer spremljamo vezavo molekul reagenta X na receptorje Y v tkivu.



V stacionarnem stanju dobimo zvezo

$$y(x; a, y_0) = \frac{y_0 x}{x + a}, \quad (1)$$

kjer pomeni y_0 nasičeni odziv tkiva in a koncentracijo, potrebno za odziv, ki je enak polovici nasičenega.

Iz merskih podatkov v datoteki `farmakoloski.dat` določi parametra y_0 in a . Napaka v meritvi odziva je v vsem področju enaka trem enotam. Zvezo lahko lineariziramo. Pazi, kako se pri tem transformirajo napake.

1.2 Linearizacija

Kot demonstrirano na predavanju sem zvezo (1) lineariziral z inverzijo:

$$y \longmapsto \frac{1}{y} \quad (2)$$

$$\frac{1}{y} = \tilde{y} = \frac{x + a}{y_0 x} = \frac{1}{y_0} \Phi_1(x) + \frac{a}{y_0} \Phi_2(x) \quad (3)$$

$$\Phi_1(x) = 1 \quad (4)$$

$$\Phi_2(x) = \frac{1}{x}, \quad (5)$$

pri transformaciji pa se spremenijo tudi napake:

$$y_i \pm \sigma_i \longmapsto \frac{1}{y_i \pm \sigma_i} = \frac{1}{y_0(1 \pm \frac{\sigma_i}{y_i})} = \frac{1}{y_i}(1 \mp \frac{\sigma_i}{y_i}) \quad (6)$$

$$= \frac{1}{y_i} \mp \frac{\sigma_i}{y_i^2} = \tilde{y}_i \mp \tilde{\sigma}_i \quad (7)$$

Iščem torej linearen fit za enačbo

$$\tilde{y} = a_1 \Phi_1(x) + a_2 \Phi_2(x), \quad (8)$$

pri čemer je

$$\tilde{y} = \frac{1}{y} \quad (9)$$

$$a_1 = \frac{1}{y_0} \quad (10)$$

$$a_2 = \frac{a}{y_0} \quad (11)$$

$$\Phi_1(x) = 1 \quad (12)$$

$$\Phi_2(x) = \frac{1}{x}. \quad (13)$$

1.3 Luščenje

Nadaljeval sem s formalizmom, ki je bil predstavljen na predavanju: tvoril sem matriko A in vektor \vec{b} :

$$A_{jk} = \sum_{i=1}^{N=2} \frac{\Phi_j(x_i)\Phi_k(x_i)}{\sigma_i^2} \quad (14)$$

$$b_j = \sum_{i=1}^{N=2} \frac{y_i \Phi_j(x_i)}{\sigma_i}. \quad (15)$$

A in \vec{b} izgledata tako:

$$A = \begin{bmatrix} 2.654 \cdot 10^7 & 1.818 \cdot 10^5 \\ 1.818 \cdot 10^5 & 4.660 \cdot 10^3 \end{bmatrix} \quad (16)$$

$$\vec{b} = \begin{bmatrix} 290346 \\ 2679 \end{bmatrix}, \quad (17)$$

preko njiju pa dobimo \vec{a} :

$$\vec{a} = A^{-1} \cdot \vec{b} \quad (18)$$

$$\vec{a} = \begin{bmatrix} \frac{1}{y_0} \\ \frac{a}{y_0} \end{bmatrix} = \begin{bmatrix} 0.00955264 \\ 0.20241427 \end{bmatrix} \quad (19)$$

Potrebujem tudi napake parametrov \vec{a} , le-te dobim preko kovariančne matrike $A^{-1} = C$:

$$A^{-1} = C = \begin{bmatrix} \text{var}(a_1) & \text{cov}(a_1, a_2) \\ \text{cov}(a_2, a_1) & \text{var}(a_2) \end{bmatrix} = \begin{bmatrix} 5.14135815 \cdot 10^{-08} & -2.00568648 \cdot 10^{-06} \\ -2.00568648 \cdot 10^{-06} & 2.92826399 \cdot 10^{-04} \end{bmatrix}. \quad (20)$$

Če imamo kovariančno matriko, lahko izračunamo tudi korelacijo med posameznimi parametri, saj je le-ta omejena na interval $[-1, 1]$, zaradi česar si jo ljudje lažje predstavljamo. Po nasvetu dobrih ljudi na <https://math.stackexchange.com/questions/186959/correlation-matrix> sem tvoril matriko ϱ :

$$\varrho = \begin{bmatrix} 1 & -0.51691484 \\ -0.51691484 & 1 \end{bmatrix}.$$

Korelacija med a_1 in a_2 je precejšnja, a to lahko pojasnimo z dejstvom, da smo v a_2 ‘zapakirali’ oba iskana parametra (glej enačbo (11), (22)).

Kvaliteto fita ocenim preko statistike χ^2 :

$$\chi^2 = \sum_{i=1}^N \frac{\left(y_i - \sum_{j=1}^M a_j \Phi_j(x_i) \right)^2}{\sigma_i^2} = 6.237, \quad (21)$$

to vrednost pa lahko primerjam z idealno, ki znaša $N - M$, kjer je N število merilnih točk (v našem primeru 8), M pa število parametrov (v našem primeru 2). Če numerično izvrednotim ($N - M$), ugotovim, da znaša vrednost točno 6. Spričo tega trdim, da je fit dober, a z možnostjo, da je naš model preskromen in bi ga morali dopolniti z dodatno kompleksnostjo.

1.4 Rezultati

Do sedaj vemo:

$$\vec{a} = \begin{bmatrix} \frac{1}{y_0} \\ \frac{a}{y_0} \end{bmatrix} = \begin{bmatrix} 0.00955264 \\ 0.20241427 \end{bmatrix} \quad (22)$$

$$\sigma_{\vec{a}}^2 = \begin{bmatrix} 5.14135815 \cdot 10^{-08} \\ 2.92826399 \cdot 10^{-04} \end{bmatrix}. \quad (23)$$

Če transformiram \vec{a} nazaj v prvotne spremenljivke s pripadajočimi napakami, dobim:

$$y_0 = 104.68 (1 \pm 2.34\%) \quad (24)$$

$$a = 21.189 (1 \pm 6.08\%) \quad (25)$$

1.5 Poskus fita brez linearizacije

Običajno ob zahtevah po fitanju posegamo po vgrajenih metodah programskih jezikov, brez predhodne linearizacije. Za primerjavo in vrednotenje dosedanje metodologije bom poskusil fitati parametra a in y_0 direktno na funkcijo (1). Moja hipoteza je, da zaradi enakomernosti negotovosti v direktnem prostoru izmerkov rezultati ne bodo bistveno odstopali od (24) in (25).

Z nelinearizirano optimizacijo (z metodo `scipy.optimize.curve_fit`) dobim vrednosti parametrov in kovariančno matriko:

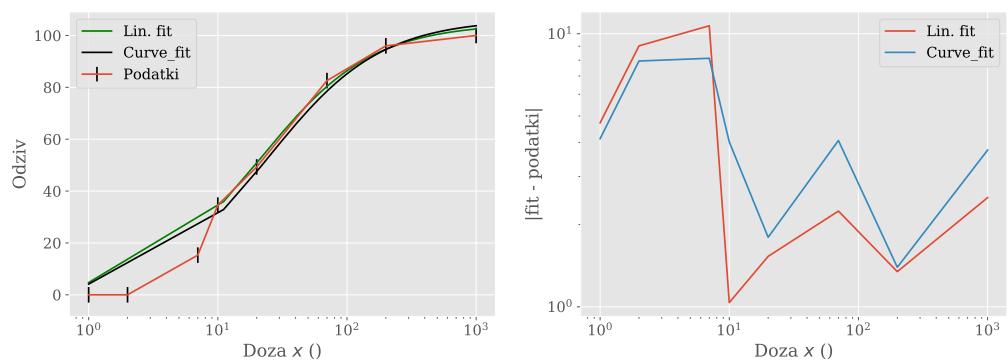
$$y_0 = 106.32 (1 \pm 4.475\%) \quad (26)$$

$$a = 24.763 (1 \pm 16.815\%) \quad (27)$$

$$C = \begin{bmatrix} 22.6473 & 13.2202 \\ 13.2202 & 17.3368 \end{bmatrix} \quad (28)$$

$$\varrho = \begin{bmatrix} 1 & 0.6671 \\ 0.6671 & 1 \end{bmatrix}. \quad (29)$$

Linearizacija se torej splača; nelinearna fitarija mi porodi večje negotovosti v parametrih in večjo korelacijo med parametri, pri vseh merilnih točkah so residualne vrednosti nad vrednostmi lineariziranega fita, kot je razvidno iz slike 1.



Slika 1: [LEVO] Originalni podatki z napakami in oba fita, lineariziran in nelineariziran.
[DESNO] Absolutne rezidualne vrednosti obej fitov.

2 Druga naloga

2.1 Naloga

Za uporabo visokoločljivostnega magnetnega spektrometra potrebujemo preslikavo, ki iz izmerjenih količin rekonstruira parametre trajektorije delcev, potrebne za izračun energije in drugih kinematičnih količin. V datoteki `thtg-xfp-thfp.dat` najdete kalibracijske podatke s stolpci ϑ_{tg} (disperzijski kot na tarči v stopinjah) ter f_p (položaj v goriščni ravnini v milimetrih) in ϑ_{fp} (kot v goriščni ravnini v stopinjah). Natančnosti meritev kotov so velikostnega reda miliradianov, položajev pa okrog milimetra. Sestavi varčni model za preslikavo

$$(x_{fp}, \vartheta_{fp}) \mapsto \vartheta_{tg}$$

Uporabiš lahko na primer najnižje potence x_{fp} in ϑ_{fp} ali pa kakšne druge funkcije teh dveh spremenljivk.

2.2 Tipanje problema

Nalogo bom pričel reševati z uvozom podatkov in ogledom izmerkov (slika 2).

2.3 Pристop k reševanju

Kot namiguje naslov sklopa (“Linearni modeli”), se splača k luščenju pristopiti z linearnim modelom. V svojem redosledu sem prekrstil spremenljivke:

$$x_{fp} \longmapsto x_1 \tag{30}$$

$$\vartheta_{fp} \longmapsto x_2 \tag{31}$$

$$\vartheta_{tg} \longmapsto y, \tag{32}$$

nato pa sem tvoril funkcijo

$$f(x) = \sum_{i=1}^N a_{p,r} \Phi_{p,r}(x_1, x_2), \tag{33}$$

kjer so

$$\Phi_{p,r}(x_1, x_2) = x_1^p \cdot x_2^r. \tag{34}$$

Zaradi narave problema se zdi smiselno predstaviti parametre \vec{a} iz prejšnje naloge z matriko $\underline{\underline{a}}$, kjer elementi $a_{i,j}$ stojijo pred členi $\Phi(x_1, x_2)$ s potencami x_1^i in x_2^j .

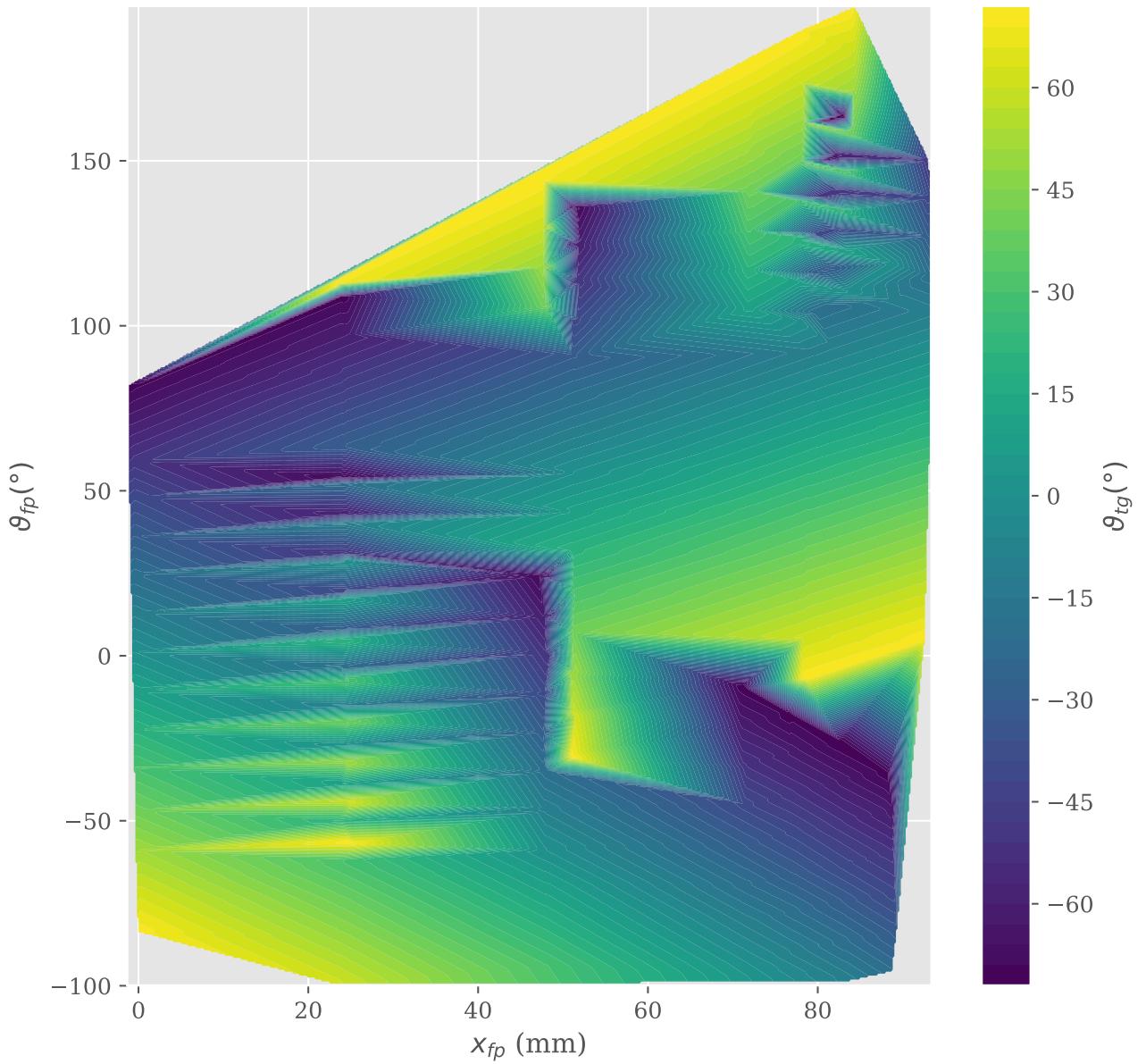
Bazne funkcije Φ sem si pripravil z računalnikom, in sicer v obliki zanke, ki mi je izpisovala nene funkcije

```
lambda x1,x2: (x1**0)*(x2**0),      #fi 0,0
```

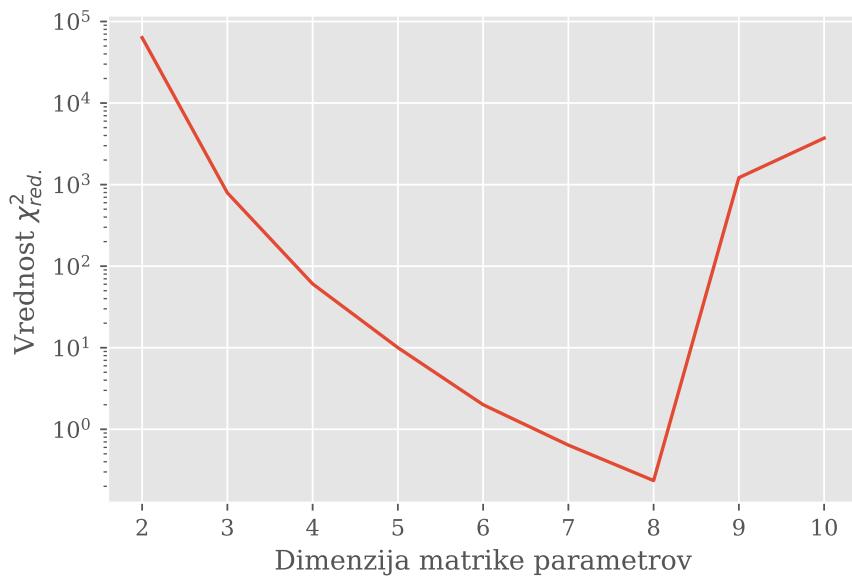
s spremenljajočimi potencami, te funkcije sem nato prekopiral v seznam baznih funkcij in delal s slednjim.

Z istim formalizmom kot v prvi nalogi sem izračunal matriko $A_{j,k}$ (enačba (14)), s to razliko, da sem v imenovalcu popravil $\sigma_1^2 \mapsto \sigma_{x_1} \sigma_{x_2}$, da upoštevam obe dimenziji.

S povečevanjem števila koeficientov sem do neke točke zniževal $\chi^2_{\text{red.}}$, od neke točke naprej (točneje pri številu koeficientov nad 8²) pa prične strmo naraščati, fit izgubi vsakršno podobnost z originalom in uide v $\pm\infty$, kar pripisujem nestabilnosti potenciranja števil na visoke potence. Za ilustracijo prilagam sliko 3.



Slika 2: Takoimenovan ‘contour plot’ našega prostora $x_{fp}, \vartheta_{fp}, \vartheta_{tg}$: na obeh oseh sta prikazani neodvisni spremenljivki x_{fp} in ϑ_{fp} , barva pa prikazuje vrednost naše fitane spremenljivke ϑ_{tg} . Vidimo, da naloga nikakor ne bo enostavna zaradi velike iregularnosti v podatkih.



Slika 3: Spreminjanje vrednosti reducirane statistike χ^2_{red} , izračunane z napačnimi podatki.

Na tej točki sem podvomil vase in kolege povprašal, ali prepoznajo monstrum na sliki 2. Prijazni kolegi so potrdili moje sume in mi postavili diagnozo idiota, ki ne zna uvažati podatkov iz datotek s fiksno širino stolpca.

2.4 Pravilen pristop k reševanju

Namesto preko uvoza podatkov s knjižnjico

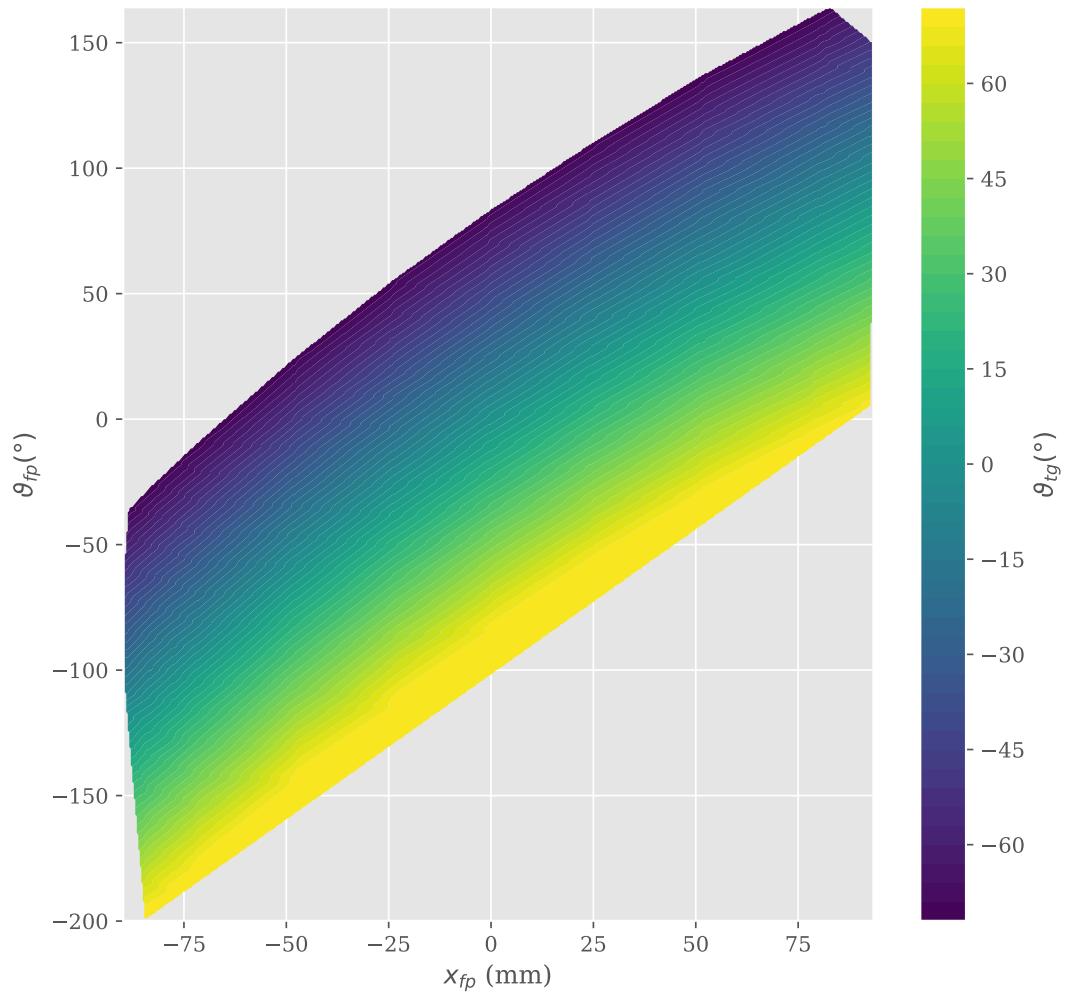
```
pandas.read_fwf()
```

sem uporabil

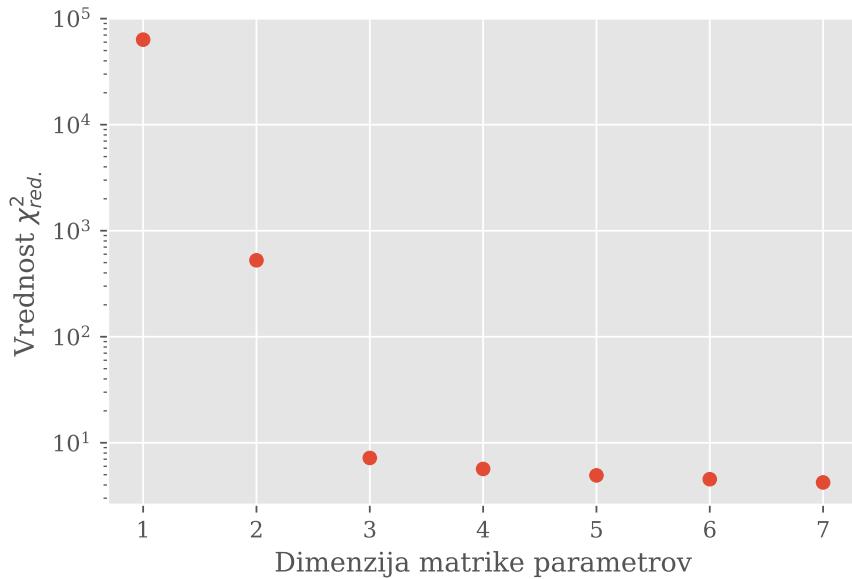
```
numpy.loadtxt()
```

Kot je očitno s slike 4, dá tak uvoz dosti bolj smiselne podatke. Verjamem, da bi tudi na prvi način šlo, a ne brez brskanja po dokumentaciji, eksperimentiranja in razhroščevanja.

Ponovim postopek luščenja parametrov, tokrat seveda bistveno bolj uspešno. Na sliki 5 vidimo sprva močno padanje statistike χ^2_{red} , nato pa koleno in bistveno počasnejše zmanjševanje. Na vodoravnji osi je prikazana dimenzija matrike, ki ustreza najvišji potenci posamezne količine, zato število vseh prostih parametrov, ki jih luščimo, narašča kot kvadrat dimenzije matrike. Odločil sem se, da bom uporabil dimenzijo največ 3, saj nato z večanjem prostostnih stopenj ne pridobimo več dosti.



Slika 4: Prikaz istih količin kot na sliki 2, le da tokrat s pravilno prebranimi podatki. Vidimo, da so podatki lepi, zvezni, brez diagonalnih robov in hipnotičnih cik–akov.



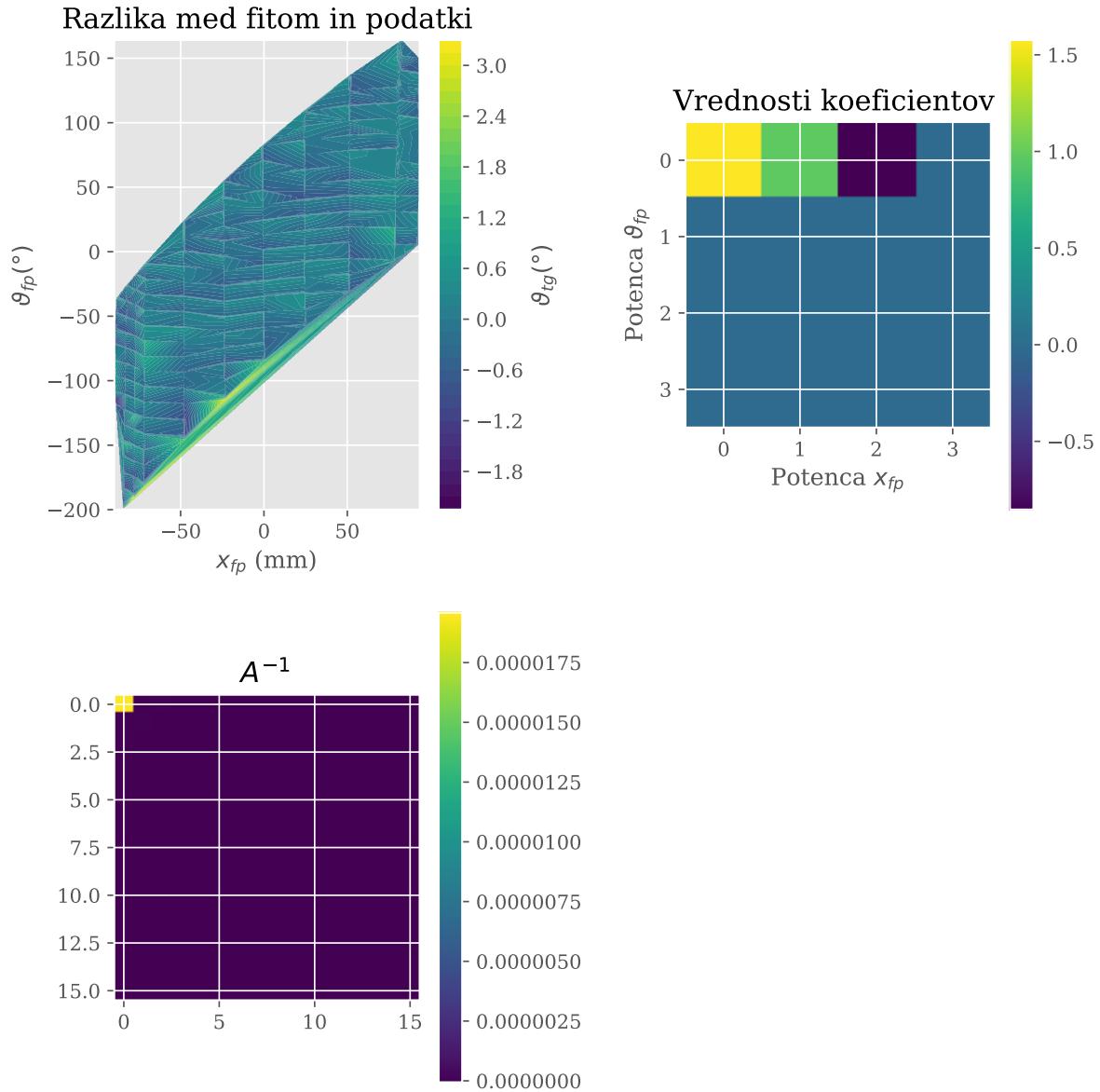
Slika 5: Spreminjanje vrednosti reducirane statistike χ_{red}^2 , izračunane s pravilno uvoženimi podatki.

2.5 Rezultati

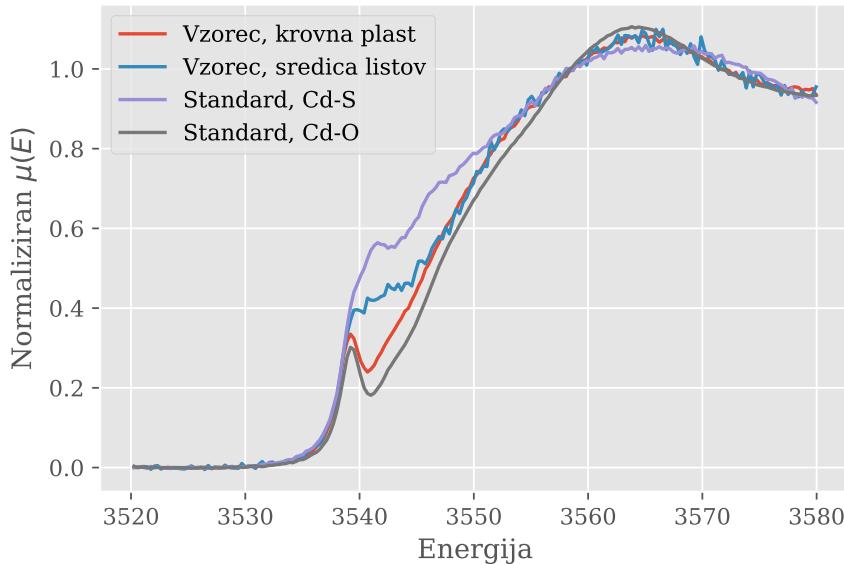
Grafično rezultate predstavljam na sliki 6, numerično pa jih podajam spodaj:

Parameter	Vrednost	Negotovost
$a_{0,0}$	1.577	0.0044
$a_{0,1}$	0.974	1.447e-4
$a_{0,2}$	-0.853	0.000105
$a_{0,3}$	0.001	2.74e-06
$a_{1,0}$	-0.001	1.73e-06

V tabelo sem vključil vse koeficiente, ki so bili po zaokroževanju na 3 decimalna mesta neničelni. Za njihov prispevek k statistiki χ^2 bi jih lahko posamezno odstranjeval iz množice $\{\Phi_{r,p}(x_1, x_2)\}$ in gledal, kakšno povečanje doprinese takša eliminacija, ali pa bi luščenje opravil s postopkom SVD.



Slika 6: Grafična ponazoritev rezultatov luščenja: [ZGORAJ LEVO:] razlika med podatki in našim fitom z linearno kombinacijo različnih potenc obeh spremenljivk, [ZGORAJ DESNO:] Numerične vrednosti koeficientov: opazimo, da največ prispevajo potence x_{fp} , [SPODAJ:] Kovariančna matrika $C = A^{-1}$, ki nam pove napako parametrov in njihovo koreliranost.



Slika 7: Spektri dveh standardov in dveh vzorcev. Za vsak spekter imamo 200 merilnih točk pri istoležnih energijah.

3 Tretja naloga

3.1 Naloga

Podrobnosti v profilu rentgenskih absorpcijskih robov so odvisne od kemijske okolice elementa. Teorijske napovedi profila še niso dovolj natančne in zanesljive, zato si pri analizah snovi pomagamo s standardi. V datoteki `CdL3_1infit.norm` so zbrani 4 absorpcijski spektre kadmija na robu L_3 iz studije, kako ta kovina učinkuje na rastline. V prvih dveh vzorcih so izolirane celične stene iz krovne plasti in iz sredice listov rastline *C. Thlaspi*, ki je znan hiperakumulator težkih kovin. Zadnja dva spektra sta dobljena na standardih, kompleksih Cd sulfata z glutationom (GSH) in pektinom: v prvem je Cd vezan izključno na žveplo, v drugem na kisik. V listnih vzorcih dopuščamo obe vrsti vezave, vemo pa, da sta prispevka obeh v spektru linearno sestavljenega. Določi odstotno razmerje vezi Cd—O in Cd—S v obeh listnih vzorcih.

3.2 Tipanje problema

Uvozil sem vse štiri spektre, dva vzorčna in dva standarda. Njihov izgled lahko vidimo na sliki 7.

3.3 Reševanje problema

Nalogo sem reševal z metodo SVD (*singular value decomposition*), zaradi česar sem si najprej pripravil matriko

$$A = \begin{bmatrix} \frac{\tilde{S}_1}{\sigma_{s1}} & \frac{\tilde{S}_2}{\sigma_{s2}} \end{bmatrix}, \quad (35)$$

kjer je \vec{S}_1 vektor izmerkov prvega standarda, \vec{S}_2 vektor izmerkov drugega standarda, σ_{s1} in σ_{s2} pa negotovosti teh standardov, zatem pa še matriko izmerkov:

$$\vec{b}_1 = \begin{bmatrix} \vec{v}_1 \\ \sigma_{v1} \end{bmatrix} \quad (36)$$

$$\vec{b}_2 = \begin{bmatrix} \vec{v}_2 \\ \sigma_{v2} \end{bmatrix}, \quad (37)$$

kjer sta vektorja \vec{b}_1 in \vec{b}_2 samo izmerjena spektra vzorcev, skalirana z njuno negotovostjo. Iskane parametre \vec{a}_1 in \vec{a}_2 nato dobim preko zvez

$$A = UwV^T \quad (38)$$

$$\vec{a} = \sum_{i=1}^M \frac{U_i \cdot \vec{b}}{w_i} V_i \quad (39)$$

$$\sigma^2(a_j) = \sum_{i=1}^M \left(\frac{V_{j,i}}{w_i} \right)^2 \quad (40)$$

3.3.1 Ocena negotovosti

Negotovosti nimam, a jih potrebujem. To sem premostil s sledečim trikom: σ^2 je za skalarno konstanto definirana kot vsota kvadratov odmika izmerkov od povprečja, skalirana s kvadratom števila izmerkov. Za moj primer spektra sem računal vsoto kvadratov odmika točk na spektru od lokalne povprečne vrednosti spektra, v praksi to pomeni, da sem se čez vektorje izmerkov standardov in vzorcev zapeljal z bločnim povprečenjem 5-ih točk in nato računal odmike točk od tega povprečja. To sem storil v eni vrstici s

```
s1_sigma = np.sum((data.s1.rolling(5).mean().shift(-2)-data.s1)**2)/len(data.s1-5).
```

Dodatno previdnost zahteva le dejstvo, da pri vsakem povprečenju izgubimo eno točko, zaradi tega moramo računati odmike med ustrezno zamaknjениm spektrom, podobno kot to storimo, kadar primerjamo originalne in filtrirane signale po FIR filtriranju. Rezultati tega postopka mi dajo ocene negotovosti:

σ_{s1}	1.819 e-5
σ_{s2}	1.242 e-5
σ_{v1}	1.823 e-5
σ_{v2}	1.054 e-4

3.4 Rezultati

Spekter vsakega vzorca bom popisal z linearno kombinacijo obeh standardov, torej

$$\vec{v} = k_1 \vec{S}_1 + k_2 \vec{S}_2, \quad (41)$$

kjer sta vektorja \vec{S} spektra standardov, koeficiente k pa nam povesta zastopanost posamezne komponente v spektru vzorca.

3.4.1 Vzorec 1

V prvem vzorcu dobim sledeče rezultate:

Parameter	Vrednost	Negotovost
k_1	0.27227	0.002773
k_2	0.60460	0.002404

Trdim torej, da razmerje med zastopanostjo vezi Cd-S in Cd-O znaša 0.450 ± 0.06 .

3.4.2 Vzorec 2

V drugem vzorcu dobim sledeče rezultate:

Parameter	Vrednost	Negotovost
k_1	0.23042	0.002773
k_2	0.153738	0.002404

Trdim torej, da razmerje med zastopanostjo vezi Cd-S in Cd-O znaša 1.499 ± 0.021 .