6. Luščenje kemijskih parametrov

Domača naloga pri predmetu Modelska analiza I

Avtor: Matic Noč

15.11.2017

1 Uvod

Obravnavali bomo problem "fitanja"oziroma luščenja parametrov linearnega modela - to je model, ki je linearen v neznanih parametrih a_i, kar pomeni da lahko luščimo parametre modelu s vsemi funkcijami, dokler je ta model linearna kombinacija neznanih koeficentov. Ne bi pa mogli npr. fitati $y = a_1 sin(x - a_2)$. Tak problem ima analitično rešitev s minimizacijo residualov oz. *metodo najmanjših kvadratov - least squares*, ki pa nam ne da dober rezultat, če residuale ne obtežimo.

2 Odziv tkiva na reagente

2.1 Model

Pri prvi nalogi smo obravnavali odziv tkiv na različne reagente, kjer gre za reakcijo, pri kateri spremljamo vezavo molekul reagenta X na receptorje Y v tkivu:

$$\mathrm{Y} + \mathrm{X} \rightleftharpoons \mathrm{Y}^*. \tag{1}$$

V stacionarnem stanju tako dobimo zvezo

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

kjer pomeni y_0 nasičeni odziv tkiva in a koncentracijo, potrebno za odziv, ki je enak polovici nasičenega. Enačbo lahko predelamo na linearni model tako da obrnemo enačbo (2)

$$\begin{aligned} 1/y &= \frac{1}{y0} + \frac{a}{y0x} \\ \widetilde{y} &= a_1 \widetilde{x} + a_2 \end{aligned} \tag{3}$$

kjer je $\widetilde{y} = \frac{1}{y}$ in $\widetilde{x} = \frac{1}{x}$, $m = \frac{a}{y_0}$ in $n = \frac{1}{y_0}$. Pri tem se napake meritev σ_y transformirajo $\sigma_{\widetilde{y}}$, dobimo pa jih lahko tako da odvajamo enačbo $\widetilde{y} = 1/y$.

$$\begin{aligned} -1/y^2 dy &= d\widetilde{y} \\ \sigma_{\widetilde{y}} &= \sigma_y \frac{1}{y^2} \end{aligned} \tag{4}$$

2.2 Metoda najmanjših kvadratov

Oglejmo si sedaj problem. Na voljo imamo tabelo podatkov y_i, x_i, ki smo jih npr. izmerili v laboratoriju. Transformirajmo podatke v $\widetilde{y_i}, \widetilde{x_i}$ in jih ustavimo v enačbo (3). Dobimo predoločen sistem enačb za koeficienta m,n

$$\widetilde{y_i} = m\widetilde{x_i} + n \tag{5}$$

Ta sistem lahko zapišemo v matrični obliki

$$\mathbf{y} = X\mathbf{a} \tag{6}$$

,kjer je

$$\mathbf{y} = \begin{bmatrix} \widetilde{y_1} \\ \vdots \\ \widetilde{y_n} \end{bmatrix}.$$

in

$$X = \begin{bmatrix} 1 & \widetilde{x_1} \\ \vdots & \vdots \\ 1 & \widetilde{x_n} \end{bmatrix}$$

a pa je vektor neznanih koeficienotv

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}.$$

Rešitev poiščemo tako, da poiščemo take koeficiente a_i, ki minimizirajo kvadrat razlike napak med našim fitom in pravimi vrednostmi.

$$S = \sum (\widetilde{y_i} - f(x_i, a_1, a_2, ...))^2 \tag{7}$$

Torej je naša najboljša rešitev za a_i enaka rešitvi sistema enačb

$$\boldsymbol{\frac{\partial S}{\partial a_i} = 0}, \forall i = 0, ..., n_{meritev} \tag{8}$$

Če zapišemo S v matrični obliki dobimo

$$S = ||\mathbf{y} - X\mathbf{a}||^2 \tag{9}$$

Odvajamo po vektorju neznanih koeficientov in dobimo

$$\frac{dS}{d\mathbf{a}} = 2(\mathbf{y} - X\mathbf{a})X = 0 \tag{10}$$

oziroma matrični isti sistem enačb kot (8)

$$X^T X\mathbf{a} = X^T y \tag{11}$$

Torej rešitev **a** obstaja in je enaka

$$\mathbf{a} = (X^T X)^{-1} X^T y \tag{12}$$

Zaradi stabilnosti rešitev koeficiente ne dobimo s direktnim iskanjem inverza, ampak s reševanjem linearnega sistema enačb (11), ki ga lahko rešimo s LU, QR, SVD razcepom.

2.3 Utežitev

V osnovnem primeru, damo vsem točkam enako težo, kar seveda pomeni, da pri razpršenih podatkih ne bomo dobili dobrega fita, kot bomo kmalu videli. Utežimo zato residuale s $\sigma_{y_i} =: \sigma_i$. Če je napaka velika, se na tiste točke ne oziramo močno, saj ne pripomorejo k večanju S, če pa je napaka

majhna, pa je nujno da so točke fita in meritev blizu, saj bomo v nasprotnem primeru močno večali S.

$$S = \sum (\frac{\widetilde{y_i} - f(x_i, a_1, a_2, ...)}{\sigma_i})^2 \tag{13}$$

Vidimo torej, da lahko naše podatke $\mathbf{y}$, X le delimo s dano napako in uporabimo rešitev (12).

Tako tudi S postane pravi hi-kvadrat : $S = \sum (\frac{y_i - y_{expected}}{\sigma})^2 = \chi^2$, kjer je χ^2 porazdeljen po $N - M$ prostostnih stopenj.

Ko bo fit torej dober, mora biti naš χ^2 blizu najbolj verjetnega, to je $N - M$, zato lahko definiramo reduciran $hi - kvadrat$ $\widetilde{\chi}^2 = \chi^2/(N - M)$, ki mora biti pri dobrih parametrih enak 1 in je dobra mera za ugotavljanje dobrega fita.

2.4 Varianca koeficientov a

Recimo da y ne poznamo točno in zapišemo naše izmerke kot $\mathbf{y'} = \mathbf{y} + \epsilon$, kjer je y prava vrednost in ϵ_i nekoreliran beli šum, porazdeljen okoli 0, $< \epsilon >= 0$ z standardnim odklonom $< \epsilon_i \epsilon_j >= \sigma^2 \delta_{i,j}$. Torej je kovariančna matrika, v primeru enakih napak enaka $\epsilon^T \epsilon = \sigma^2 I$ Naša rešitev v tem primeru ne bo točna ampak ocena $\mathbf{a'} = (X^T X)^{-1} X^T (y + \epsilon)$, kjer je $\mathbf{a'}$ ocena in $\mathbf{a}$ prava vrednost koeficientov.

$$\mathbf{a'} = \mathbf{a} + (X^T)^{-1} X^T \epsilon \tag{14}$$

Pričakovana vrednost ocene je kar enaka pričakovani vrednosti rešitve (11) $< \mathbf{a'} >=< \mathbf{a} >$, napaka koeficienotv pa bo varianca (koren povprčene vrednosti kvadrata napak) razlike ocene in in prave vrednosti.

$$< (\mathbf{a'} - \mathbf{a})^2 >=< (\mathbf{a'} - \mathbf{a})(\mathbf{a'} - \mathbf{a})^T >=< (X^T X)^{-1} X^T < \epsilon \epsilon^T > X (X^T X)^{-1} \epsilon >$$

, kar je v primeru, ko je $< \epsilon \epsilon^T >= \sigma^2 I$ enako kar

$$var(\mathbf{a}) = M = \sigma^2 (X^T X)^{-1} \tag{15}$$

Če napak meritev ne poznamo, potem lahko ocenimo napako kota reduciran $\widetilde{\chi}^2 = \frac{\chi^2}{N-M}$, če pa uporabljamo uteži iz ocene napak pa

$$M = X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W M W^{\mathrm{T}} X (X^{\mathrm{T}} W^{\mathrm{T}} X)^{-1} \tag{16}$$

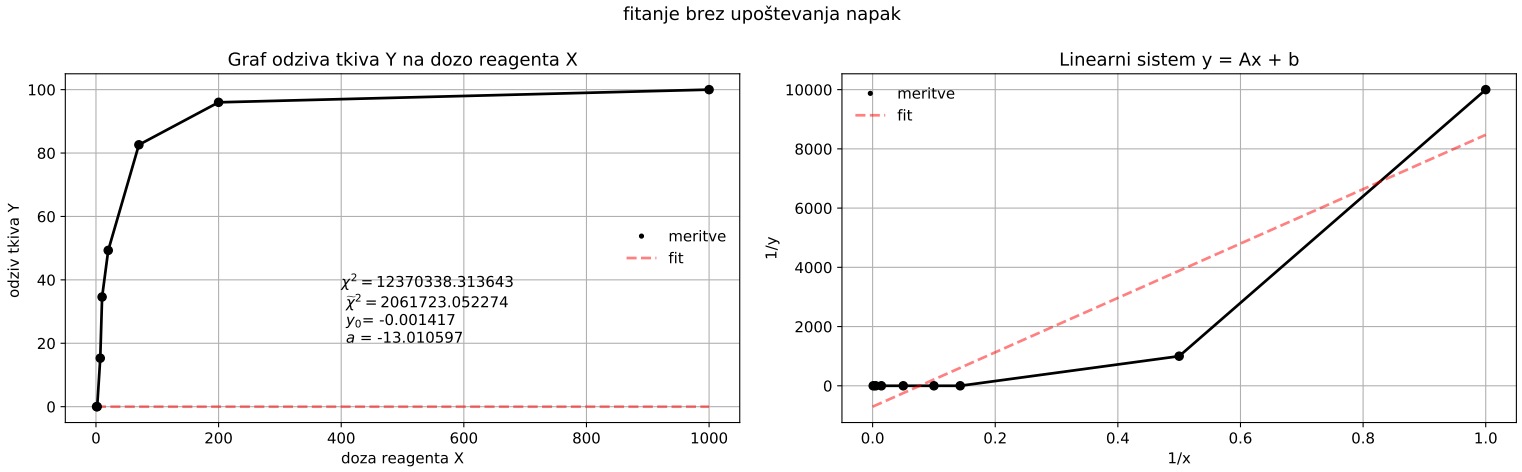
Če so naše uteži kar napake $W = M^{-1}$:

$$M = (X^{\mathrm{T}} W X)^{-1} \tag{17}$$

Torej v primeru ko so naše uteži napake lahko le izračunamo inverz matrike $(X^{\mathrm{T}} W X)^{-1}$

2.5 Reševanje brez upoštevanja napak

Podatke sem obrnil in na matriki X uporabil metodo *np.linalg.lstsq*, ki uporabi SVD razcep za najhitrejši izračun rešitve sistema (11)



Že na grafu vidimo, da so ti parametri napačni, kar nam pove tudi vrednost $\tilde{\chi}^2$. Ker ni utežitve morajo parametri vsem točkam zagotoviti enako dobro bližino, kar pomeni, da se ne skoncentrira na točke s majhno napako.

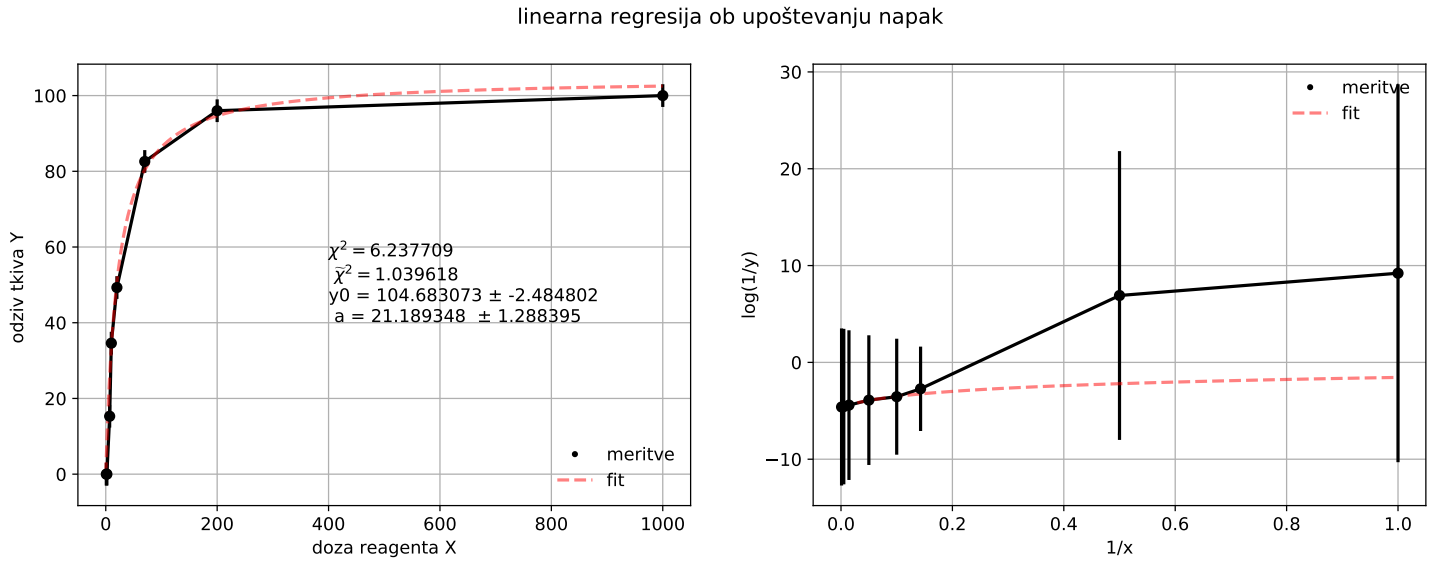
2.6 Reševanje s napakami

$$X = \begin{bmatrix} \frac{1}{\sigma} & \frac{\tilde{x}}{\sigma} \\ \vdots & \vdots \\ \frac{1}{\sigma} & \frac{\tilde{x}}{\sigma} \end{bmatrix}$$

in

$$\mathbf{y} = \begin{bmatrix} \frac{\tilde{y}}{\sigma} \\ \vdots \\ \frac{\tilde{y}}{\sigma} \end{bmatrix}.$$

Za varianco napak sem še vedno sem dobiti inverz matrike $(X^T W X)$, ki sem ga dobil s metodo *np.linalg.inv*. Tako sem dobil tudi korelacijske koeficiente in napake parametrov.



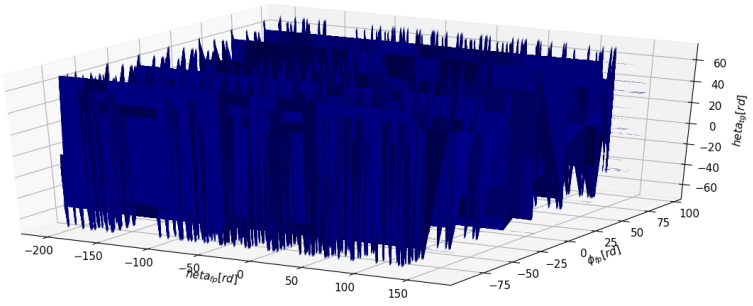
Vidimo, da tokrat je rešitev veliko bolj pravilna, saj je točke, ki imajo veliko napako zaradi utežitve lahko izpustil (desni graf je prikazan v logaritemski skali!), $\widetilde{\chi}^2$ pa je zelo blizu 1, kar pomeni, da je zelo malo verjetno, da naš fit ni pravi in je posledica statističnega naključja.

3 Visokoločljivostni magnetni spektrometer

Ko delce spustimo skoz magnetno polje, se njihov tir ukrivi. Če izmerimo kje se nahajajo na detektorju, lahko napovemo kje so se nahajali delci na tarči, in iz katere smeri so (pri)leteli, iz tega pa njihove energije in kinematske količine.
Torej imamo dva seta koordinat $(x_{tg}, y_{tg}, \theta_{tg}, \phi_{tg})$ in v goriščni ravnini detektorja: $(x_{fp}, y_{fp}, theta_{fp}, \phi_{fp})$, kjer so x,y položaj trka na detektorju in ϕ, θ azimutalni in zenitni kot trajektorje. V splošnem je vsaka koordinata na tarči kombinacija vseh ostalih, zato se bomo posvetili le eni od tarčevskih koordinat θ_{tg}. Izmerjeni kalibracijski podatki so x_{fp}, ϕ_{fp} torej iščemo preslikavo

$$(x_{fp}, \phi_{fp}) => \theta_{tg} \tag{18}$$

Vzeli bomo nekaj različnih preslikav in skušali najti tako, ki nam bo dala $\widetilde{\chi}^2 \sim 1$. Najprej se prepričajmo da nam podatki ne prikažejo ničesar, saj jih lahko plotamo na 3D grafu



Slika 1: iz podatkov ne vidimo nič

3.1 Linearni model

Vzemimo najprej naiven linearni model

$$\theta_{tg} = a_0 + a_1\theta_{fp} + a_2 x_{fp} \tag{19}$$

Pričakujemo, da ni linearne ravnine, ki nam dobro opisala zgornjo sliko.

Tabela 1: Linearni testni model

$\widetilde{\chi}^2$	354.488
a_0	-6.642701150896979101e+00 ± 0.0026895426110164398
a_1	-8.402619006911082877e-01 ± 5.2782693856238785e-05
a_2	9.736583733416463993e-01 ±7.22595806240393e-05 ,

Tako kot smo pričakovali tak fit nima smisla

3.2 Izboljšava modela

Vzemimo, da je naša ravnina lahko ukrivljena in tako začnimo s dodatnimi kvadratnimi potencami, napak fita zaradi preglednost in ne bistvenosti ne bom več pisal

$$\theta_{tg} = a_0 + a_1\theta_{fp} + a_2 x_{fp} + a_3 x_{fp}^2 + a_4\theta_{fp}^2 \tag{20}$$

Tabela 2: Kvadratni model

$\widetilde{\chi}^2$	8.11274154
a_0	1.05023609151
a_1	-0.849920201119
a_2	-2.93380021016e-05
a_3	0.985348592305
a_4	-0.00180972046672,

Vidimo, da se je fit bistveno izboljšal vendar je še vedno daleč od 1.

Vzemimo še višje potence, mešane člene.

$$\theta_{tg} = a_0 + a_1\theta_{fp} + a_2 x_{fp} + a_3 x_{fp}^2 + a_4\theta_{fp}^2 + a_5 x_f p\theta_{fp} + a_6 x_{fp}^2\theta_{fp} + a_7 x_{fp}\theta_{fp}^2 \tag{21}$$

Tabela 3: Višje potence

$\widetilde{\chi}^2$	2.50906764
a_0	1.68357878e+00
a_1	-8.50988358e-01
a_2	-2.66485613e-04
a_3	9.79401322e-01
a_4	-2.14775172e-03
a_5	5.47218368e-04
a_6	4.01308547e-06
a_7	-2.02687574e-06 ,

$\widetilde{\chi}^2$ se še zmanjša vendar pa višje potence ne bodo nikoli dobro fitale problema, saj potenčne funkcije med seboj niso ortogonalne in ne bodo pokrile vseh možnosti. Smisleno bi bilo poskusiti s ortogonalnimi funkcijami npr

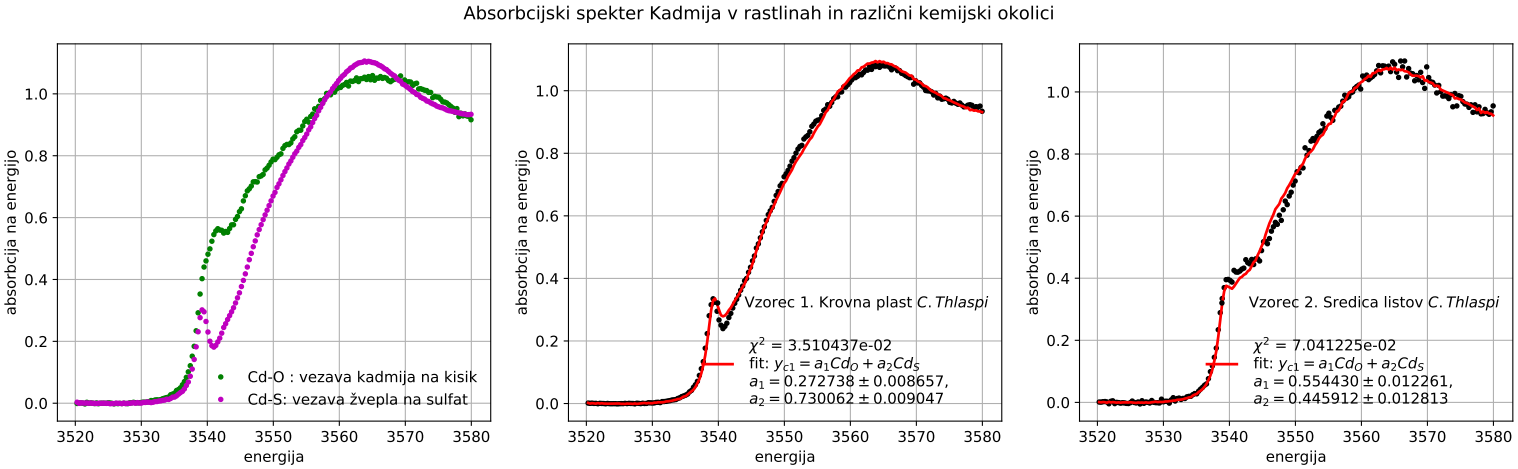
$$\theta_{tg} = \sum a_n sin(n x_{fp}) + \sum b_n cos(n\theta_{fp})) \tag{22}$$

kjer vzamemo visoko število n, toda ti ne dajo dobre rešitve, saj sem se s 300 sinisi približal šele $\widetilde{\chi}^2 = 1.277904083153612591e + 00$. Vseeno pa je bil to najboljši fit.

4 Absorbcijski spekter Kadmija v okolici različnih kemijskih elementov

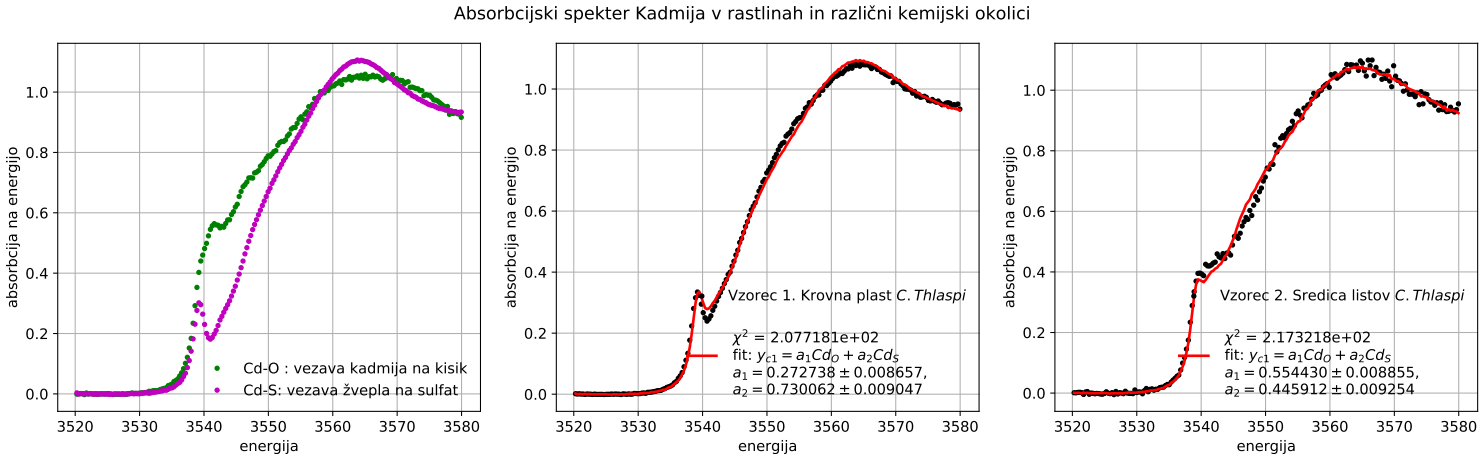
V tej nalogi smo želili preučiti abosrbcijski spekter Kadmija v rastlinah. Vemo, da je absorbcijski spekter Kadmija odvisen od elementov, ki so v okolici, vendar teoretično še ne znamo dobiti pravo

odvisnost, zato si pomagamo s luščenjem parametrov, saj vemo, da je spekter linearna kombinacija dveh spektrov, ko je Kadmij okolici sulfata in vezan na žveplo in ko je v okolici pektina in vezan na kisik. Ta predpostavka pomeni, da se v rastlinah pretežno veže le na te dva elementa.



Slika 2: Levo: absorbcijska spektra Kadmija ob različnih vezavah elementov, desno: izmerjen absorbcijski spekter krovne plasti in sredice listov. Opazimo več stvari: 1)fit je na oko kar dober, čeprav se zdi da smo naleteli na "$overfit$", torej popolno prileganje podatkom, pa je to le posledica, da nismo obtežili problema s napakam, ki so majhne, saj so majhni tudi podatki. 2) v krovni plasti je več kadmija vezanega na žveplo, torej je v okolici več sulfata z glutationom, medtem ko je v listni sredici malce več kadmija vezanega na kisik. Za variančno matriko sem potreboval napako, za katero sem vzel kar $\widetilde{\chi}^2$.

Realno vrednost χ^2 bi dobili, če bi podali neko oceno za napake. Tu je nujno da so vse napake enake, ker nam enakomerna utežitev da dober rezultat. Glede na to da želimo imeti $\widetilde{\chi}^2 \sim 1$, lahko obratno ocenimo napake tako da izračunamo $\sigma^2 = \frac{\chi^2}{N-M}$, saj je $\chi^2 \sim N-M$, ko je fit dober (vidimo iz slike). Torej je naša napaka $\sigma_1 = 0.013$ in $\sigma_2 = 0.018$. Ker je utežitev enakomerna vpliva le na vrednost χ^2 in ne na obliko grafa.



Slika 3: Vidimo, da napaka le zveča $\chi^2 \sim 200 = N - M$

5 Zaključek

Naučili smo se luščiti linearne parmetre in ocenjevati kako dober fit je, saj vidimo, da v več dimenzionalnem prostoru ne moremo kar na grafu videti ali je fit dober. Dobro bi bilo implementirati neke vrste strojno učenje, kjer bi avtomatizirali postopek; torej bi model sam dodajal različne parametre in na testnih podatkih iskal najboljši fit, ter se tako "*naučil*" najboljše preslikave. Zaradi velikega števila podatkov, nikjer nismo opazili "overfita", ko vzameš preveč parametrov in začne χ^2 zaradi tega močno padati. Videli smo da je ocena napak zelo pomembna, saj nam da veliko informacij o tem katere točke so najbolj pomembne.