University *of Ljubljana*
Faculty *of Mathematics and Physics*

Modelska Analiza 1

6. naloga - Luščenje modelskih parametrov: Linearni modeli

Avtor: Matic Lubej
Asistent: dr. Simon Čopar
Predavatelj: prof. dr. Alojz Kodre

Ljubljana, 19.11.2013

Naloga:

Pri tej nalogi smo se spoznali z močno uporabnim orodjem prilagajanja funkcij podatkom. Delali smo z linearnimi metodami in se naučili, kako delujejo. V vseh primerih sem se prilagajanja lotil z uporabo SVD metode, ki mi je omogočala enostavno določitev vrednosti parametrov in njihovih napak. V prvem delu naloge smo imeli podatke merjenja odziva tkiva, kjer smo morali s prilagajanjem določiti vrednost koncentracije snovi v tkivu in vrednost nasičenega odziva tkiva. V drugem delu naloge smo iskali model, ki v odvisnosti od temperature in moči grelca opiše toplotno prevodnost jekla, v zadnjem delu naloge pa smo ocenjevali razmerje vezi Cd-O in Cd-S v danih vzorcih lista na podlagi priloženih standardnih spektrov.

Del I
Odziv tkiva

1 Naloga in metoda

V farmakologiji merijo odziv tkiva na različne reagente. Za večino teh pojavov lahko privzamemo, da gre za reakcijo, kjer spremljamo vezavo molekul reagenta X na receptorje Y v tkivu, kar lhko opišemo s kemijsko reakcijo:

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

kjer interpretiramo stanje Y^*, kot vzbujeno stanje tkiva. V stacionarnem stanju dobimo zvezo:

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

kjer y_0 pomeni nasičeni odziv tkiva in a koncentracijo, potrebno za odziv, ki je enak polovici nasičenega. Na voljo smo imeli merske podatke z meritveno napako 3 enot, iz katerih smo izluščili ta dva parametra.

Ker obravnavamo linearne modele, moramo najprej enačbo linearizirati, da bo uporaba metode sploh mogoča. Opazimo, da enačbo lineariziramo, če uporabimo transformacijo:

$$\begin{aligned} u &= \frac{1}{y}, \qquad t = \frac{1}{x}, \qquad \tilde{\sigma} = \sigma \left| \frac{\partial u}{\partial y} \right| = \sigma \left| \frac{1}{y^2} \right|, \\ & u = \frac{a}{y_0} t + \frac{1}{y_0} = kt + n, \end{aligned} \tag{3}$$

kjer smo definirali nova parametra k in n. Transformirane podatke sem nato pospravil v pravo obliko za nadaljno uporabo. Definiral sem:

$$A_{i,j} = \left[\frac{\varphi_i(x_j)}{\tilde{\sigma}_i} \right], \qquad b_i = \frac{u_i}{\tilde{\sigma}_i}, \tag{4}$$

kjer je $\varphi_i(x_j)$ i-ta testna funkcija pri j-ti meritvi. V našem primeru premice je imela matrika A sledečo obliko:

$$A = \begin{bmatrix} 1 & t \\ \vdots & \vdots \\ 1 & t \end{bmatrix}. \tag{5}$$

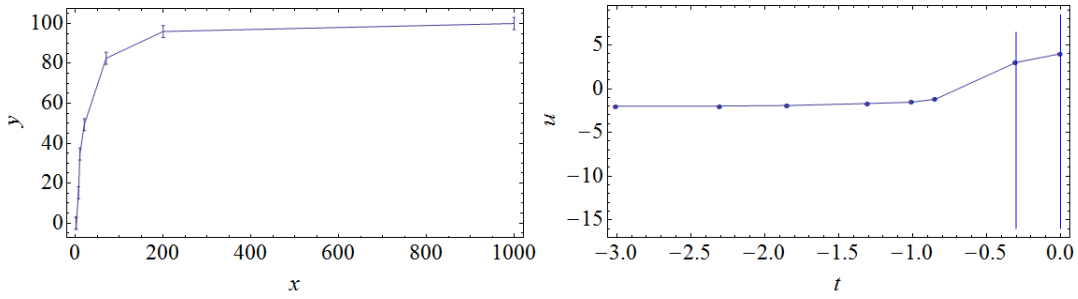
Z uporabo ukaza SingularValueDecomposition[] v matematičnem orodju Mathematica, sem definiral tri nove matrike:$UDV^\dagger = \mathrm{svd}(A)$, katere sem nato uporabil pri izračunavanju optimalnih parametrov in napak parametrov preko formul:

$$\begin{aligned} a &= \sum_{i=1}^{M} \left(\frac{U_i \cdot b}{d_i} \right) V_i, \\ \sigma^2(a_j) &= \sum_{i=1}^{M} \left(\frac{V_{ji}}{d_i} \right)^2, \end{aligned} \tag{6}$$

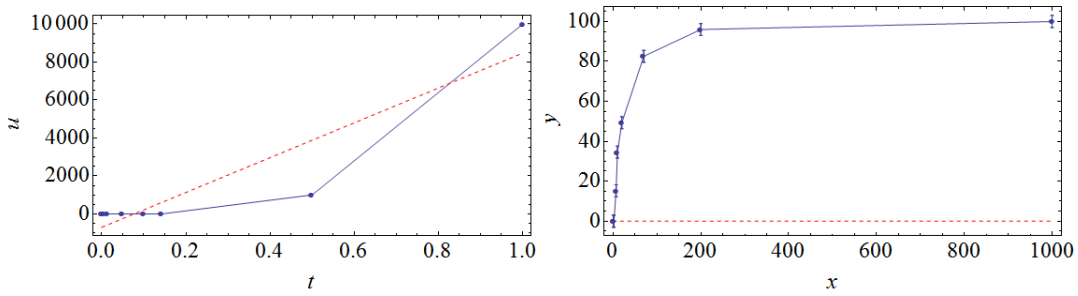
kjer je M manjša izmed dimenzij matrike D, U_i in V_i sta stolpca matrik U in V, d_i pa i-ti diagonalni element matrike D.

2 Rezultati

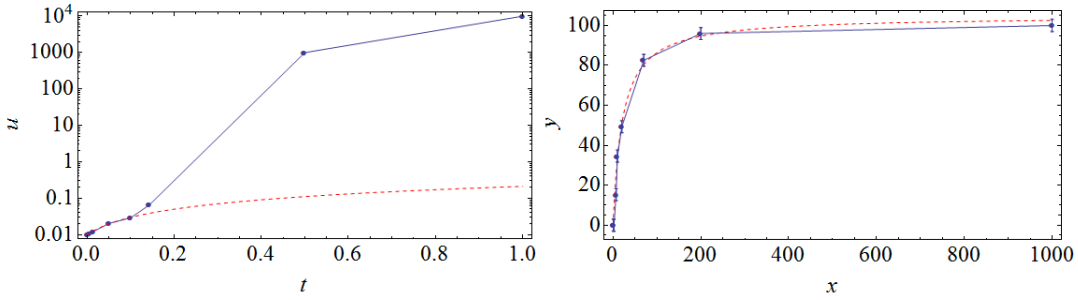
Najprej si oglejmo obliko funkcije z originalnimi in transformiranimi podatki:



Vidimo, da odziv na začetku hitro raste, nato pa saturira pri vrednosti y_0. Prikazane so tudi napake, ki pa nam na prvi pogled ne povedo veliko. Na desni sliki v logaritemski skali vidimo, da zadnji dve točki, ki sta v originalu prvi dve, odstopata iz skupine, poleg tega pa imata tudi ogromno napako. Oglejmo si rešitve v primeru, ko napak **ne** upoštevamo in v enačbah zgoraj delimo z 1:



Vidimo, da podatki v prvem primeru sploh ne ležijo na premici, vsaj ne vsi. Ker imajo vsi podatki enako utež, točke, ki so zelo oddaljene, možno potegnejo premico k sebi. Če parametre transformiramo nazaj v prvotno obliko vidimo, da krivulja sploh nima smisla, zato takšno reševanje opustimo. Popolnoma drugačno zgodbo dobimo, ko upoštevamo napake, sam boto imele točke z večjo natančnostjo dosti večjo utež od tistih točk, ki ležijo daleč stran in imajo veliko napako:



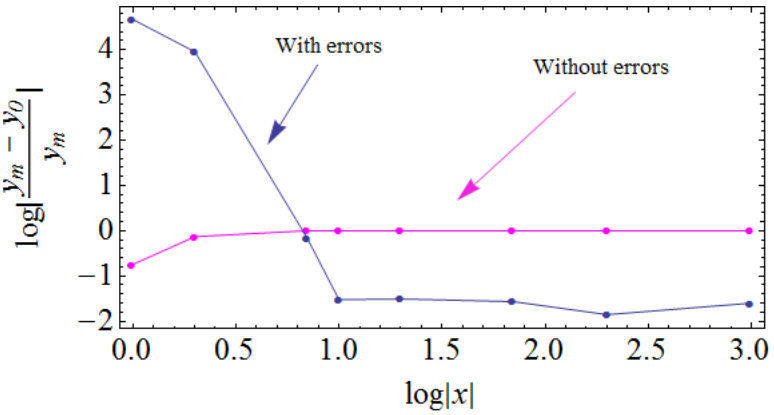
V logaritemski skali je sedaj očitno, da se krivulja bolje prilega natnačno izmerjenim podatkom, osamelce pa v veliki meri ignorira. Natančno ujemanje opazimo tudi v prvotni obliki podatkov. Zaključimo, da smo s takšnim rezultatom zadovoljni. O pravilnosti nam lahko veliko pove tudi vrednost testa χ^2, ki je definiran kot:

$$\chi^2 = \sum_i \left(\frac{y_i - y_{i,m}}{\sigma_i} \right)^2 . \tag{7}$$

Vrednosti tega testa za posamezen primer sta:

χ^2	3322.38	brez napak
χ^2	26.07	z napakami

Vidimo, da imamo razliko za več kot faktor 100. Oglejmo si še natančnost nove funkcije:



Čeprav ima pravilnejša metoda večjo napako v prvem delu, se natančnost hitro izboljša in spusti pod mejo tiste, kjer napak ne upoštevamo, hkrati pa ne smemo pozabiti, da videz v logaritemski skali vara, ker po natančnem pregledu vidimo, da natančnost metode, kjer napak nismo upoštevali, res ni tako dobra, medtem ko pri drugi metodi dosežemo natančnost na 2 decimalni mesti. Oglejmo si še vrednosti parametrov in njihove napake:

Parameter	Vrednost	Napaka
n	9.55×10^{-3}	$\pm 0.23 \times 10^{-3}$
k	0.20	± 0.02
y_0	104.7	± 2.5
a	21.2	± 1.3

Del II

Toplotna prevodnost jekla

1 Naloga

Dobili smo podane meritve in napake meritev toplotne prevodnosti jekla *Armco* v odvisnosti od temperature in moči grelca. Iz začetnih potenc spremenljivk T in P smo sestavili varčni model za meritve toplotne prevodnosti λ. Ker pravega modela ne poznamo, to naredimo približno v smislu Taylorjevega razvoja funkcije. Odločil sem se izčrpno preveriti vse konfiguracije potenc, kjer sem v različnih razdelih dodajal zraven različne funkcije, vedno pa sem obdržal konstantni člen in oba linearna člena. Testni model je imel obliko:

$$\lambda\left(T,P\right) = a_0 + a_1 T + a_2 T^2 + \cdots + b_1 P + b_2 P^2 + \cdots . \tag{8}$$

Potrebno je opomniti, da ima v tem primeru testna funkcija χ^2 pričakovano vrednost $N - p$, kjer je N število meritev, p pa število parametrov. V primeru, da se nam torej število parametrov pri različnih konfiguracijah spreminja, moramo testno funkcijo pravilno normirati, da jih lahko med seboj pravilno primerjamo.

Definiramo normirano testno funkcijo χ^2_{norm} kot:

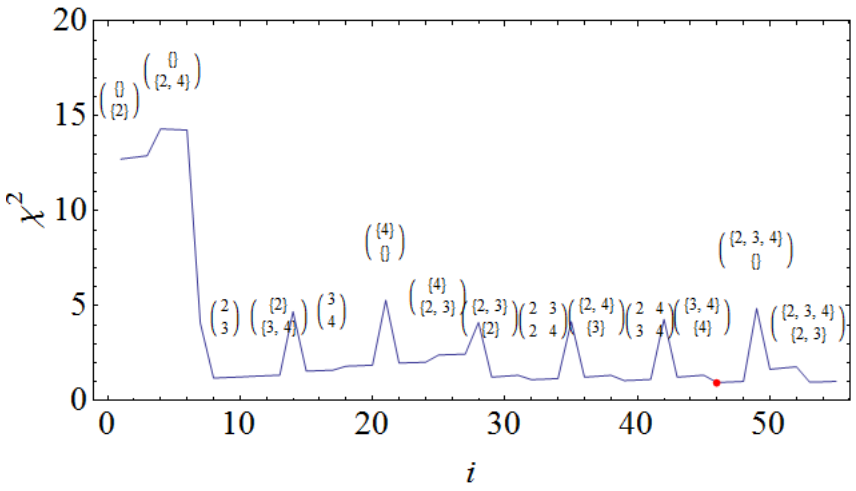
$$\chi^2_{norm} = \frac{\chi^2}{N-p}. \tag{9}$$

2 Potence T in P

Pri tem modelu sem privzel funkcijo

$$\lambda\left(T,P\right) = a_0 + a_1 T + b_1 P + a_n T^n + b_m P^m, \qquad \forall n \in [2,5]\,,\ \forall m \in [2,5]\,. \tag{10}$$

Za vsako konfiguracijo sem izračunal vrednost χ^2_{norm}. Oglejmo si obliko te funkcije v odvisnosti od vrstnega reda konfiguracij za tak model na sliki spodaj:



Opomniti je treba, da odvisnost vrednosti χ^2_{norm} ni odvisna od mesta konfiguracija i, saj je to popolnoma odvisno od tega, na kakšen način izvajamo zanko po konfiguracijah. Na sliki je prikazanih še nekaj konfiguracij, prav tako pa točka minimalne vrednosti χ^2_{norm}, ki je označena z rdečo piko. Vidimo, da se minimalni vrednosti približa veliko konfiguracij, ki pa so lahko po obliki funkcije med seboj zelo različne. Izkazalo se je, da je najbolj natančen model funkcija oblike:

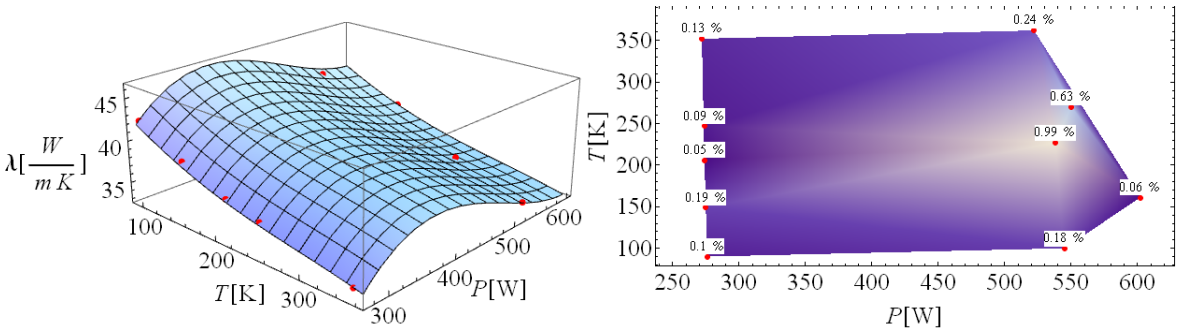
$$\lambda\left(T,P\right) = a_0 + a_1 T + b_1 P + a_3 T^3 + a_4 T^4 + b_2 P^2 + b_3 P^3. \tag{11}$$

Vrednost izbranih parametrov, njihove napake ter vrednost χ^2_{norm} so prikazani v naslednji tabeli:

Parameter	Vrednost	Napaka
a_0	-42	± 55
a_1	-0.0643	± 0.0085
a_3	5.18×10^{-7}	$\pm 1.99 \times 10^{-7}$
a_4	-8.79×10^{-10}	$\pm 4.18 \times 10^{-10}$
b_1	0.63	± 0.39
b_2	-0.0013	± 0.0009
b_3	9.14×10^{-7}	$\pm 6.32 \times 10^{-7}$
χ^2_{norm}	0.965151	/

Vidimo, da so nekatere napake relativno majhne, veliko pa je tudi takšnih ki so relativno ogromne. V nobenem primeru nam ti podatki, razen mogoče a_0 in χ^2_{norm} ne povejo nič pametnega, zato jih ne bom več prikazoval.

Bolj zanimivo si je ogledati 3D model te funkcije in konturno sliko napake na slikah spodaj:



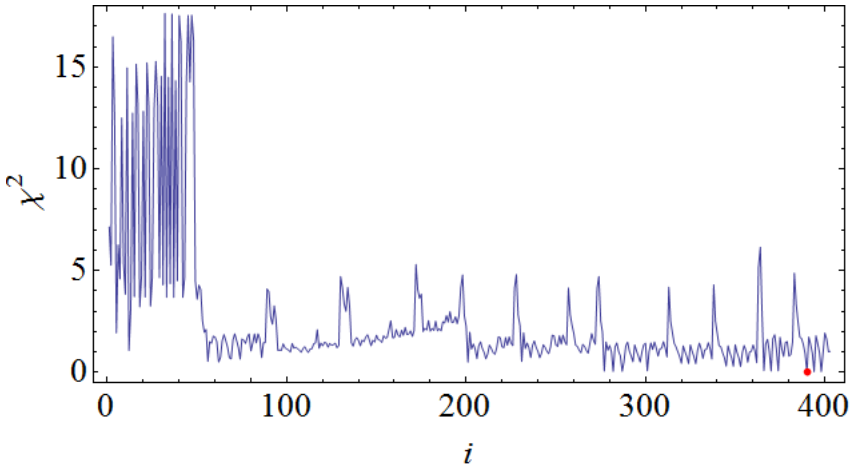
Iz zgornjega modela lahko sklepamo tako na odvisnost modela od T kot na odvisnost od P. Vidimo, da bi bilo pametno imeti še vmesno meritev, ker bi tako lahko ta model ovrgli ali pa potrdili. Na desni sliki vidimo še vrednosti napak v merjenih točkah. Največja relativna napaka pri tej konfiguraciji je 1 %.

3 Potence T, P in $T \cdot P$

Poleg tega da dodamo še mešani člen, lahko Taylorjevo vrsto posplošimo tudi na Laurentovo, kjer upoštevamo še negativne potence. Pri tem modelu sem privzel funkcijo:

$$\begin{aligned} \lambda(T,P) &= a_0 + a_1 T + b_1 P + a_n T^n + b_m P^m + c_k (T \cdot P)^k , \\ & \forall n \in [-1,4] , \ \forall m \in [-1,4] , \ \forall k \in [-2,2] . \end{aligned} \tag{12}$$

Ponovno sem za vsako konfiguracijo sem izračunal vrednost χ^2_{norm}. Oglejmo si obliko te funkcije v tem primeru:



Vidimo, da je v tem primeru porazdelitev bolj gosta, saj je tudi več možnih konfiguracij. Ponovno je prikazana točka minimalne vrednosti χ^2_{norm}, ki je označena z rdečo piko. Izkazalo se je, da je najbolj natančen model funkcija oblike:

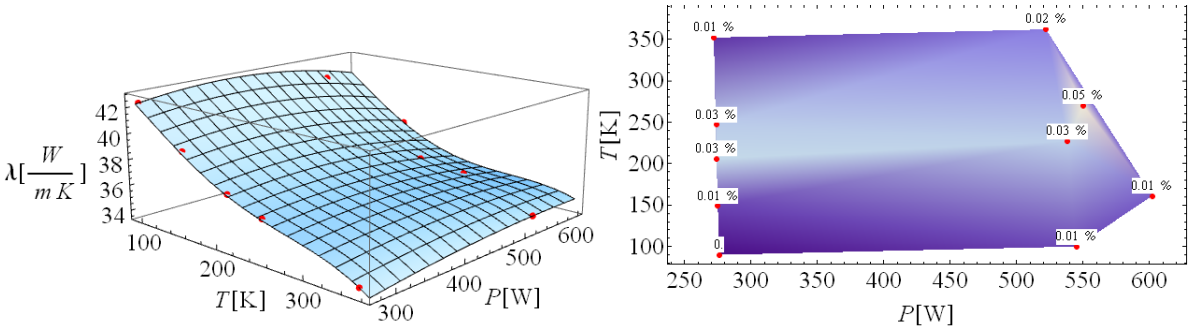
$$\lambda(T,P) = a_0 + a_1 T + b_1 P + a_2 T^2 + a_3 T^3 + a_4 T^4 + b_2 P^2 + c_1 (T \cdot P) + c_2 (T \cdot P)^2 . \tag{13}$$

Vidimo, da se v tem primeru negativne potence niso izkazale za optimalno izbiro.

Vrednost χ^2_{norm} je prikazana v naslednji tabeli:

Parameter	Vrednost
a_0	38±6
χ^2_{norm}	0.016819

Opazimo, da je v tem primeru vrednost χ^2_{norm} dosti manjša, iz česar lahko sklepamo na večjo natančnost. Na isti način si sedaj spet oglejmo obliko 3D modela in napake v konturni sliki:



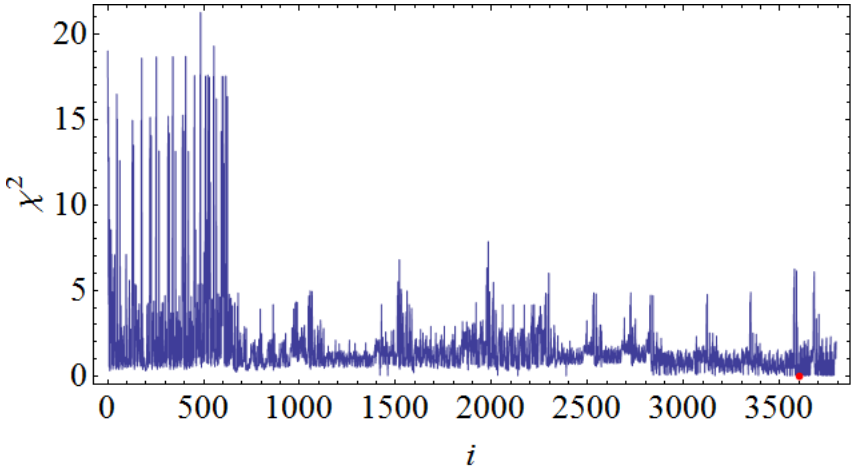
Spet lahko iz zgornjega modela sklepamo na odvisnost od T in od P. V tem primeru imamo lahko divergenco, zato dodatne meritve ne bi škodile. Na desni sliki vidimo še vrednosti napak v merjenih točkah, kjer je največja relativna napaka 0.05 %.

4 Potence T, P, $T \cdot P$ in P/T

Prejšnje modele sem obogatil še z dodatnim členom P/T. Pri tem modelu sem privzel model oblike:

$$\begin{aligned} \lambda(T,P) &= a_0 + a_1 T + b_1 P + a_n T^n + b_m P^m + c_k (T \cdot P)^k + d_l (P/T)^l , \\ & \forall n \in [-1,4] , \ \forall m \in [-1,4] , \ \forall k \in [-2,2] , \ \forall l \in [-2,2] . \end{aligned} \tag{14}$$

Tudi v tem primeru sem za vsako konfiguracijo sem izračunal vrednost χ^2_{norm}. Oglejmo si obliko te funkcije še za zadnji primer:



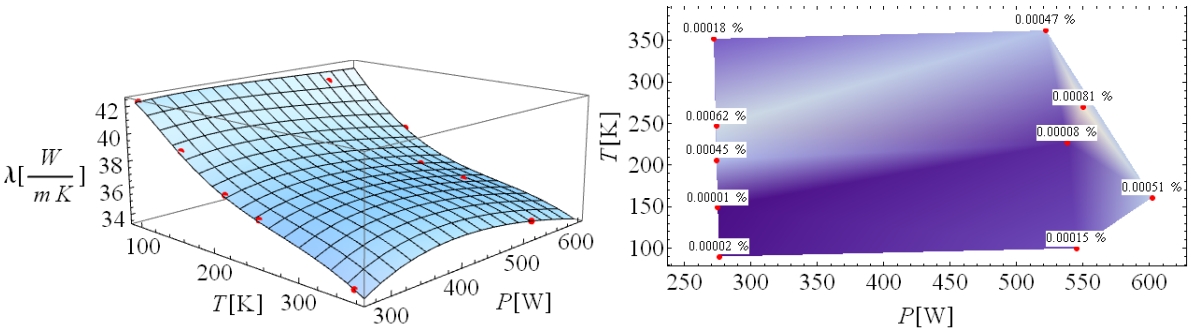
Ker je konfiguracij v tem primeru spet več, je porazdelitev toliko gostejša. Ponovno je prikazana točka minimalne vrednosti χ^2_{norm}, ki je označena z rdečo piko. Izkazalo se je, da je najbolj natančen model funkcija oblike:

$$\lambda\left(T,P\right)=a_0+a_{-1}T^{-1}+a_1T+a_2T^2+a_3T^3+c_1\left(T\cdot P\right)+d_{-1}\left(T\cdot P\right)^{-1}+d_2\left(T\cdot P\right)^2. \qquad (15)$$

V tem primeru sploh nimamo odvisnosti od samostojne spremenljivke P, se pa le-ta skriva v mešanih členih. Vidimo, da se je za optimalno izbiro tokrat izkazala tudi vključitev negativnih potenc. Vrednost χ^2_{norm} je prikazana v naslednji tabeli:

Parameter	Vrednost
a_0	76±22
χ^2_{norm}	5.15306×10^{-6}

Zopet lahko sklepamo na še večjo natančnost, saj je vrednost testa v tem primeru res najmanjša. Na isti način si sedaj spet oglejmo obliko 3D modela in napake v konturni sliki:



Model se glede na prejšnjega ne spremeni drastično, opazimo pa, da je največja napaka v tem primeru borih 8.1×10^{-4} %.

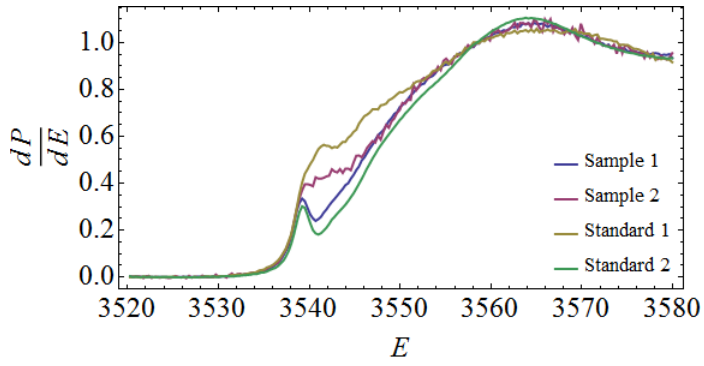
Del III
Absorpcijski spekter listnih vzorcev

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 danih podakih smo dobili 4 absorpcijske spektre kadmija na robu L_3 iz študije, 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 sestavljena. Preko danih podatkov smo določili odstotno razmerje vezi Cd—O in Cd—S v obeh listnih vzorcih.

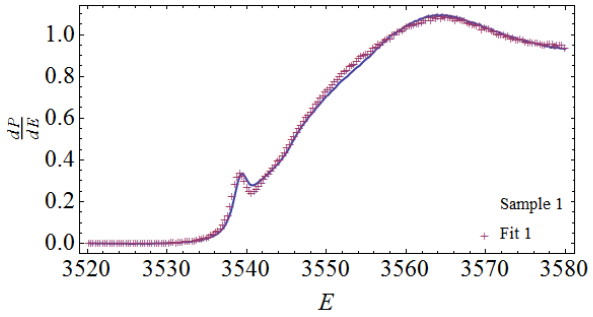
Najprej si oglejmo obliko teh absorbcijskih vzorcev:

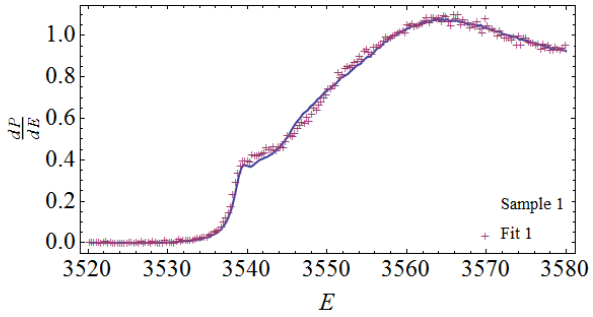


Vidimo, da so si v osnovi precej podobni. Naša naloga je torej najti koeficiente prisotnosti vezi Cd—O in Cd—S v vzorcu tako, da vzorec sestavimo iz linearne funkcije obeh standardov:

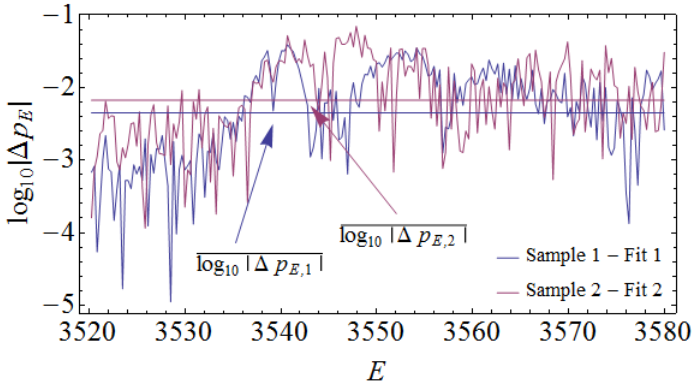
$$\begin{aligned} Sample_1(E) &= k_1 Standard_1(E) + k_2 Standard_2(E), \\ Sample_2(E) &= k_3 Standard_1(E) + k_4 Standard_2(E), \end{aligned}$$

kjer so k_{1-4} parametri, ki predstavljajo prisotnost posamezne vezi v vzorcih. Po isti metodi in na isti način kot že pri prejšnjih delih naloge, smo rešili tudi ta problem. Dobimo rešitve, ki so prikazane spodaj:





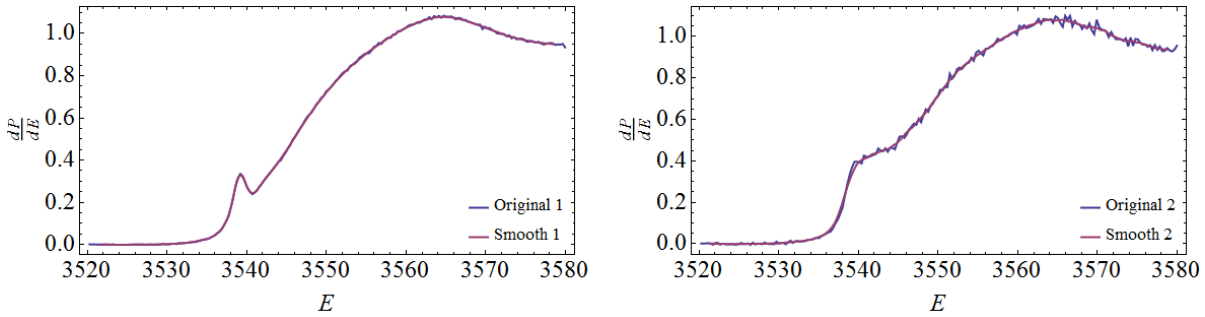
Vidimo, da se naša funkcija dobro prilagaja podatkom, majhne nevšečnosti so le pri večjih spremembah. Opazimo večjo razgibanost podatkov pri drugem vzorcu, kar nam predstavlja izvor napake in zato slabše rezultate. Oglejmo si še relativno napako med podatki in funkcijo, sestavljeno iz standardov:



Rešitev je precej natančna, v najslabšem mestu na 1 decimano mesto. Prav tako se iz grafa vidi, da ima drugi vzorec v splošnem večjo napako, saj ima tudi bolj „razgibane“ podatke.

2 Parametri in ocena njihove napake

Z rešitvijo dobimo tudi optimalne parametre, ki pa niso kaj prida vredni, saj nimamo podanih merskih napak. Pri nalogi sem se odločil, da bom kar se da relno poskusil oceniti napako, zato da bomo posledično imeli idejo o napaki parametrov. Teoretično jih lahko izračunamo, vendar bo izvor napake samo na račun korelacij med testnimi funkcijami. Ker ne želimo tvegati, bomo ocenjeno napako posplošili na vse meritve, saj enako obtežene meritve dajo vedno enak rezultat, ne glede na vrednost uteži. Enaka utež po vrednosti vpliva le na napako optimalnih parametrov. Najbolj primerno se mi je zdelo, da si ogledam „razgibanost“ meritev, in to nekako pretvorim v kvantitativen podatek. Med izborom funkcij, ki jih ima matematično orodje Mathematica, sem našel ukaz MovingAverage[], ki ne naredi nič drugega, kot izračuna povprečno vrednost sosednjih točk. Z vsakim izvršenjem tega ukaza tako izgubimo eno točko, vendar dobimo vse bolj gladko krivuljo. To krivuljo sem nato interpoliral z ukazom Interpolation, in izračunal razliko vrednosti v skupnih točkah. Dobil sem tabelo absolutnih napak, iz katere sem nato izračunal RMS vrednost, to pa vzel za povprečno „širino“ naših podatkov. Poglejmo si obliko zglajenih podatkov v primerjavi z originalnimi meritvami:



Vidimo, da funkcijo lepo zgladimo, pri tem pa izgubimo praktično zanemarljivo informacij. Končno lahko izluščimo vrednosti parametrov in grobe ocene njihovih napak:

Parameter	Vrednost	Napaka
k_1	0.2727	± 0.0021
k_2	0.7301	± 0.0022
k_3	0.5544	± 0.0075
k_4	0.4459	± 0.0079
$r_1 = k_2/k_1$	2.68	± 1.13
$r_2 = k_4/k_3$	0.80	± 0.21

Tako smo prišli do razmerij vezi Cd—O in Cd—S v obeh listnih vzorcih, hkrati pa smo tudi na grobo ocenili napako teh razmerij.