

# PA081: Programování numerických výpočtů

## 8. Modelování experimentálních dat, metoda nejmenších čtverců

Aleš Křenek

jaro 2016

# Modelování experimentálních dat

- ▶ v experimentu naměříme v bodech  $x_i$  hodnoty  $y_i$ 
  - ▶  $x$  může být libovolná veličina: čas, napětí, poloha, ...
- ▶ chování systému popisujeme modelem  $y = \mathbf{M}(x)$
- ▶ model závisí na sadě parametrů  $a_i$ , tj.

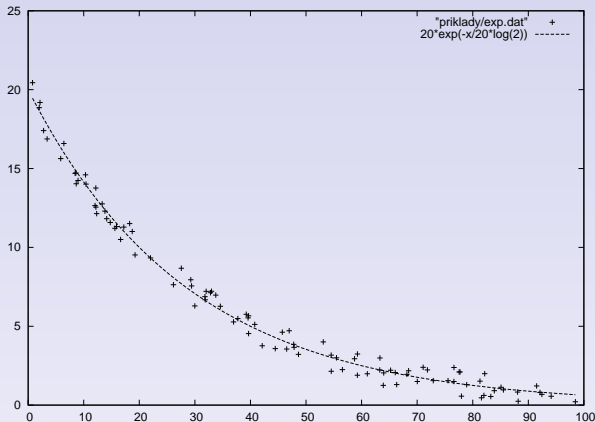
$$y = \mathbf{M}(x, a_1, \dots, a_M)$$

- ▶ hledáme takové hodnoty  $a_i$ , pro něž model odpovídá nejlépe experimentu

# Modelování experimentálních dat

## Příklad

- ▶ radioaktivní rozpad  $N = N_0 e^{-\frac{t \ln 2}{T}}$
- ▶  $N$  je počet atomů ve vzorku ( $N_0$  v čase  $t = 0$ ),  $T$  poločas rozpadu



# Metoda nejmenších čtverců

## Odvození

- ▶ „Jaká je pravděpodobnost, že konkrétní sada parametrů  $a_i$  je správná?“
  - ▶ špatně položená otázka
  - ▶ neexistuje „náhodná veličina modelů“
  - ▶ naopak, náhodnou veličinou jsou měřená data zatížena chybou
- ▶ tedy „Při daných parametrech  $a_i$ , jaká je pravděpodobnost měření  $(x_i, y_i)$ ?“

# Metoda nejmenších čtverců

## Odvození

- ▶ „Jaká je pravděpodobnost, že konkrétní sada parametrů  $a_i$  je správná?“
  - ▶ špatně položená otázka
  - ▶ neexistuje „náhodná veličina modelů“
  - ▶ naopak, náhodnou veličinou jsou měřená data zatížena chybou
- ▶ tedy „Při daných parametrech  $a_i$ , jaká je pravděpodobnost měření  $(x_i, y_i)$ ?“
  - ▶ nulová, je-li  $y$  spojitá veličina
  - ▶ musíme přidat „plus/minus odchylka měření  $\Delta y$ “
- ▶ model považujeme za správný, maximalizuje-li tuto pravděpodobnost
  - ▶ i tak je to velmi intuitivní konstrukce

# Metoda nejmenších čtverců

## Odvození

- ▶ předpokládáme normální rozložení chyby měření
- ▶ pravděpodobnost výskytu dané sady měření

$$\prod e^{-\frac{1}{2} \left( \frac{y_i - \mathbf{M}(x_i)}{\sigma} \right)^2} \Delta y$$

- ▶ maximalizace odpovídá minimalizaci logaritmu, tj.

$$\left( \sum \frac{(y_i - \mathbf{M}(x_i))^2}{2\sigma^2} \right) - N \ln \Delta y$$

- ▶  $N, \sigma, \Delta y$  jsou konstanty

# Metoda nejmenších čtverců

## Poznámky

- ▶ rozložení chyby všech měření nemusí být stejné
  - ▶ používá se modifikovaná funkce

$$\chi^2 = \sum \frac{(y_i - \mathbf{M}(x_i))^2}{2\sigma_i^2}$$

# Metoda nejmenších čtverců

## Poznámky

- ▶ rozložení chyby všech měření nemusí být stejné
  - ▶ používá se modifikovaná funkce

$$\chi^2 = \sum \frac{(y_i - \mathbf{M}(x_i))^2}{2\sigma_i^2}$$

- ▶ rozložení chyby nemusí být normální
  - ▶ počet měření v jednom bodě bývá příliš malý
  - ▶ zatížení chybou typu „někdo kopl do váhy“
  - ▶ metoda je na takové chyby nepřiměřeně citlivá
  - ▶ tzv. robustní statistiky



# Metoda nejmenších čtverců

## Poznámky

- ▶ rozložení chyby všech měření nemusí být stejné
  - ▶ používá se modifikovaná funkce

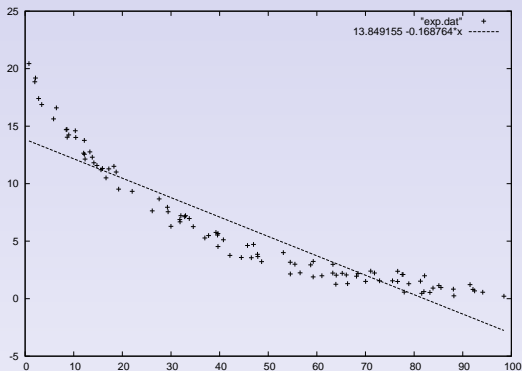
$$\chi^2 = \sum \frac{(y_i - \mathbf{M}(x_i))^2}{2\sigma_i^2}$$

- ▶ rozložení chyby nemusí být normální
  - ▶ počet měření v jednom bodě bývá příliš malý
  - ▶ zatížení chybou typu „někdo kopl do váhy“
  - ▶ metoda je na takové chyby nepřiměřeně citlivá
  - ▶ tzv. robustní statistiky
- ▶ systematická chyba
  - ▶ např. špatně kalibrovaný přístroj, závislost na související veličině

# Metoda nejmenších čtverců

## Zhodnocení vypočtených parametrů

- ▶ minimalizací  $\chi^2$  se téměř vždy hodnot  $a_i$  dopočítáme
- ▶ nevypovídá to ještě nic o kvalitě modelu
- ▶ např. lineární model radioaktivního rozpadu



# Metoda nejmenších čtverců

## Zhodnocení vypočtených parametrů

- ▶ „chi by eye“ nebo seriózní statistické zhodnocení
- ▶ hodnotíme pomocí *regularizované gamma funkce*
  - ▶ funkce konkrétního  $\chi^2$  a počtu stupňů volnosti ( $N - M$ )

$$Q\left(\frac{N - M}{2}, \frac{\chi^2}{2}\right)$$

- ▶ viz např. [http://en.wikipedia.org/wiki/Incomplete\\_gamma\\_function](http://en.wikipedia.org/wiki/Incomplete_gamma_function)
  - ▶ pravděpodobnost, že zcela náhodně vybraný vzorek  $(x_i, y_i)$  dá větší hodnotu  $\chi^2$
- ▶ čím větší tím lepší
  - ▶  $Q > 0.1$  je v pořádku
  - ▶  $Q \in [0.001, 0.1]$  je podezřelé, ale stále přijatelné, není-li distribuce chyb měření zcela normální, resp. je mírně podceněná
  - ▶  $Q < 0.001$  znamená špatný model nebo zcela nesmyslné měření

# Lineární regrese

- ▶ data prokládáme přímkou  $a + bx = 0$
- ▶ obecnější než se zdá na první pohled
  - ▶ data  $(x, y)$  lze předem libovolně (nelineárně) transformovat na  $(x', y')$
- ▶ minimalizovaná funkce

$$\chi^2(a, b) = \sum \frac{(y_i - a - bx_i)^2}{2\sigma_i^2}$$

- ▶ minimum v bodě nulových prvních derivací

$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum \frac{y_i - a - bx_i}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial b} = -2 \sum \frac{x_i(y_i - a - bx_i)}{\sigma_i^2}$$

- ▶ vhodným vyjádřením faktorů  $S, S_x, S_y, S_{xx}, S_{xy}$ 
  - ▶ součty zlomků konstruovaných z  $x_i, y_i, \sigma_i$
  - ▶ vše jsou to tady konstanty
- ▶ řešíme systém lineárních rovnic

$$\begin{aligned}Sa + S_x b &= S_y \\ S_x a + S_{xx} b &= S_{xy}\end{aligned}$$

- ▶ získáme řešení, ale nevíme nic o jeho kvalitě
  - ▶ odhad podle grafu nebo výpočet  $Q$
  - ▶ uvedená lineární regrese na radioaktivní rozpad začíná být přijatelná až když připustíme  $\sigma_i > 1.6$

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

Vícerozměrná  
data

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace

# Obecný model

- ▶  $\mathbf{M}(x, a_1, \dots, a_M)$  je nelineární funkce v  $a_i$ 
  - ▶ nelinearita v  $x$  by nevadila, viz dále
- ▶ výpočet minimálního  $\chi$  standardními optimalizačními metodami
- ▶ existují speciální varianty právě pro tvar funkce

$$\chi^2(a_1, \dots, a_M) = \sum \frac{(y_i - \mathbf{M}(x_i, a_1, \dots, a_M))^2}{2\sigma_i^2}$$

- ▶ včetně verzí s dostupnými prvními i druhými derivacemi
- ▶ díky speciálnímu použití další triky
- ▶ konkrétní metody
  - ▶ Levenberg-Marquardt
  - ▶ Moré
- ▶ např. NAG library

- ▶ model  $\mathbf{M}(x, a_1, \dots, a_M)$  je lineární kombinace

$$\mathbf{M}(x, a_1, \dots, a_M) = \sum a_j X_j(x)$$

- ▶ linearita ve smyslu parametrů modelu  $a_j$
- ▶ základní funkce  $X_j$  mohou být jakékoli
  - ▶ pro vyhodnocení modelu se použijí jen jejich konkrétní hodnoty v bodech  $x_i$
- ▶ opět minimalizujeme  $\chi^2$
- ▶ definujeme matici  $\mathbf{A}$  a vektor  $\mathbf{b}$

$$A_{ij} = \frac{X_j(x_i)}{\sigma_i} \quad b_i = \frac{y_i}{\sigma_i}$$

Modelování experimentálních dat

Metoda nejmenších čtverců

Lineární regrese

Obecný model

Lineární modely

SAXS

Vícerozměrná data

Metoda nejmenších absolutních odchylek

Regularizace

- ▶ derivováním dostáváme  $M$  rovnic pro  $k = 1, \dots, M$

$$0 = \sum_i \frac{1}{\sigma_i^2} \left( y_i - \sum_j a_j X_j(x_i) \right) X_k(x_i)$$

- ▶ po úpravách v maticovém vyjádření

$$(\mathbf{A}^T \mathbf{A}) \mathbf{a} = \mathbf{A}^T \mathbf{b}$$

- ▶ řešení bývá citlivé na zaokrouhlovací chyby
- ▶ preferovaná technika je QR rozklad



# Lineární modely

## Rozklad na singulární hodnoty

- ▶ model nemusí být dokonalý, mohou se objevit téměř lineární závislosti
  - ▶ některé základní funkce nebo jejich kombinace jsou pro danou datovou sadu irelevantní
- ▶ vede na téměř singulární matici
  - ▶ standardní metody inklinují k velkým hodnotám irelevantních parametrů
- ▶ SVD dokáže tyto problémy identifikovat
- ▶ algoritmus přímo hledá nejbližší řešení, tj. minimalizuje

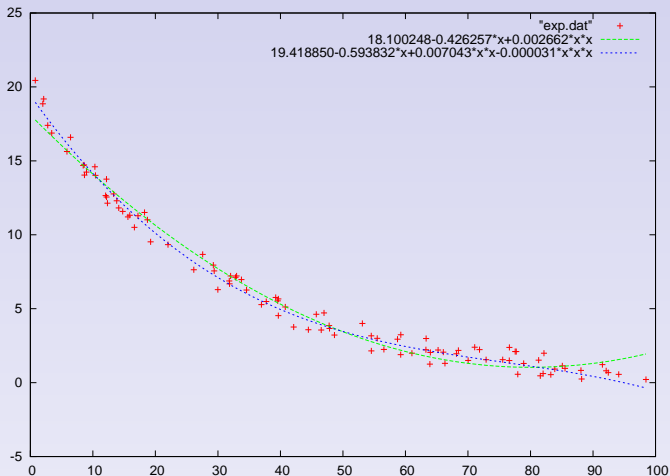
$$|\mathbf{Aa} - \mathbf{b}|^2$$

- ▶ zároveň detekuje problematické funkce

# Lineární modely

## Rozklad na singulární hodnoty

- ▶ stejná data, kvadratické a kubické funkce



- Modelování experimentálních dat
- Metoda nejmenších čtverců
- Lineární regrese
- Obecný model
- Lineární modely**
- SAXS
- Vícerozměrná data
- Metoda nejmenších absolutních odchylek
- Regularizace

- ▶ tabulka singulárních hodnot pro různé řády polynomu

řád	$\sigma_i$				
1	5.18	537.24			
2	39654.25	3.51	134.10		
3	31922200.00	66338.90	564.94	25.73	
4	2685350000.00	4133810.00	19913.70	272.78	99.50

- ▶ koeficienty u  $x^n$  pro  $n > 4$  jsou téměř nulové

# Komplexní příklad – SAXS

- ▶ *Small Angle X-ray Scattering*
- ▶ analytická metoda přinášející informace o struktuře molekuly
  - ▶ ve vodním roztoku – přirozené prostředí pro biochemii
  - ▶ relativně levný a nenáročný experiment
- ▶ RTG paprsek se v kontaktu s atomem rozptýlí
  - ▶ je-li v dráze odchyleného paprsku další atom, dochází k interferenci
  - ▶ detektor měří intenzitu záření v celém rozsahu rozptylu
- ▶ molekuly jsou v roztoku v náhodné orientaci
  - ▶ měřená křivka je průměrem všech možných orientací

Modelování experimentálních dat

Metoda nejmenších čtverců

Lineární regrese

Obecný model

Lineární modely

SAXS

Vícerozměrná data

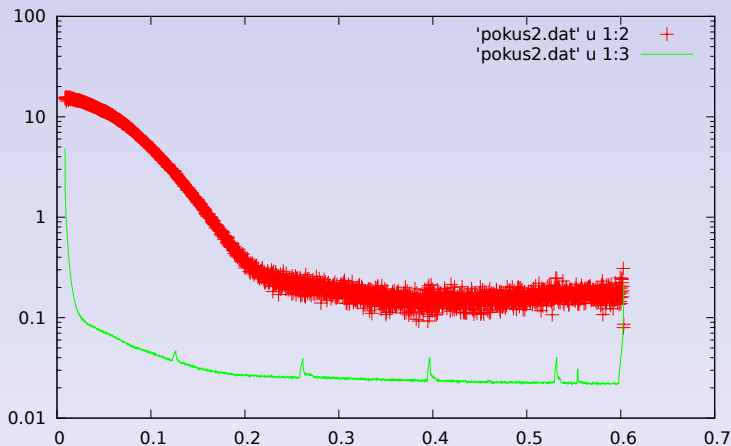
Metoda nejmenších absolutních odchylek

Regularizace

# Experimentální data

- ▶ sada obrázků
  - ▶ rozlišení  $\sim 20$  Mpix
  - ▶ monochromatické - jen intenzita
- ▶ křivka SAXS
  - ▶ závislost  $\log I$  na  $q = 4\pi \sin \theta / \lambda$   
( $\theta$  - úhel odchylení,  $\lambda$  - vlnová délka)
  - ▶ získána průměrem v kruzích
  - ▶ cca. 2000 bodů (hodnot  $q$ )
- ▶ informace o chybě
  - ▶ spolehlivá, průměrovaných pixelů je mnoho

# Experimentální data



Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

Vícerozměrná  
data

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace

- ▶ předpokládaná struktura molekuly
  - ▶ 3D model
  - ▶ typy atomů
  - ▶ jejich polohy  $\Rightarrow$  vzájemné vzdálenosti
- ▶ dodatečné parametry
  - ▶ chování solventu (vody)
  - ▶ anomálie na povrchu molekuly
    - ▶ voda se zahušťuje nebo ředí podle povrchových vlastností
  - ▶ měřítko křivky
    - ▶ jde o tvar, absolutní hodnota nemá přímý smysl
- ▶ funkce  $cI(q, c_1, c_2)$

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

Vícerozměrná  
data

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace

- ▶ lineární model pouze v  $c$ , nelineární v  $c_1, c_2$
- ▶ minimalizujeme funkci

$$\chi^2 = \frac{1}{2(N-1)} \sum_{i=1}^N \left( \frac{I_{\text{exp}}(q_i) - cI(q_i, c_1, c_2)}{\sigma_i} \right)^2$$

- ▶ v rozumném rozsahu  $c, c_1, c_2$  pouze 1 lokální minimum
  - ▶ bylo by možné přímo optimalizovat
  - ▶ neděláme kvůli složitějšímu případu (viz dále)

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

Vícerozměrná  
data

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace



- ▶ lineární model pouze v  $c$ , nelineární v  $c_1, c_2$
- ▶ minimalizujeme funkci

$$\chi^2 = \frac{1}{2(N-1)} \sum_{i=1}^N \left( \frac{I_{\text{exp}}(q_i) - cI(q_i, c_1, c_2)}{\sigma_i} \right)^2$$

- ▶ v rozumném rozsahu  $c, c_1, c_2$  pouze 1 lokální minimum
  - ▶ bylo by možné přímo optimalizovat
  - ▶ neděláme kvůli složitějšímu případu (viz dále)
- ▶ zafixujeme  $c_1, c_2$  a derivujeme

$$\frac{d\chi^2}{dc} = \frac{1}{2(N-1)} \sum_{i=1}^N \frac{2(I_{\text{exp}}(q_i) - cI(q_i, c_1, c_2))}{\sigma_i} \frac{(-I(q_i, c_1, c_2))}{\sigma_i}$$

Modelování ex-  
perimentálních  
datMetoda  
nejmenších  
čtvercůLineární  
regrese

Obecný model

Lineární  
modely

SAXS

 Vícerozměrná  
dataMetoda  
nejmenších  
absolutních  
odchylek

Regularizace

- ▶ optimální hodnota

$$c = \frac{\sum_{i=1}^N \frac{I_{\text{exp}}(q_i) I(q_i, c_1, c_2)}{\sigma_i^2}}{\sum_{i=1}^N \frac{I(q_i, c_1, c_2)^2}{\sigma_i^2}}$$

- ▶ počítáme opakovaně pro různé hodnoty  $c_1, c_2$ 
  - ▶ držíme se smysluplných intervalů
  - ▶ stačí  $20 \times 80$  výpočtů
- ▶ ze všech vybereme tu s nejmenším  $\chi^2$
- ▶ celkem  $1600N$  vyhodnocení  $I(q, c_1, c_2)$ 
  - ▶ složitost  $I(q, c_1, c_2)$  je  $O(M^2)$  v počtu atomů

# Regrese modelu

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

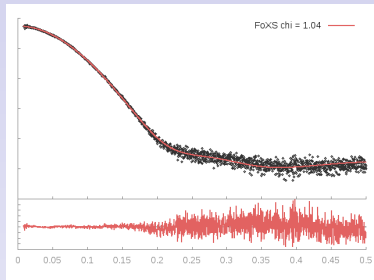
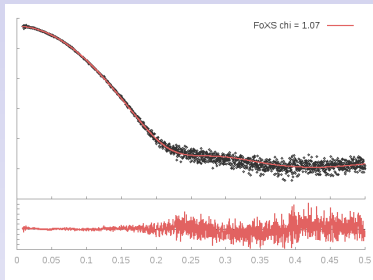
Lineární  
modely

SAXS

Vícerozměrná  
data

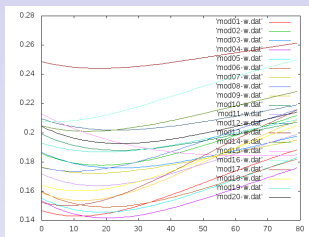
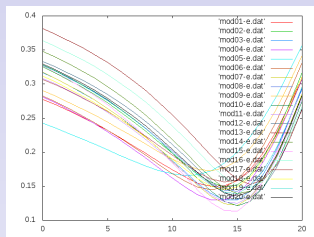
Metoda  
nejmenších  
absolutních  
odchylek

Regularizace



- ▶ molekula může existovat ve více tvarech
  - ▶ spontánně přechází mezi nimi
  - ▶ v roztoku existují všechny současně
  - ▶ SAXS křivka je jejich průměrem
- ▶ počítáme s více modely  $I_j(q, c_1, c_2), j = 1, \dots, K$
- ▶ každému přiřadíme váhu  $w_j$  a kombinujeme je lineárně
- ▶ omezení  $\sum w_j = 1$

- ▶ různá citlivost na změny  $c_1, c_2$



- ▶ vzniká mnoho lokálních minim

- ▶ opět iterujeme různé hodnoty  $c_1, c_2$
- ▶ zavedeme substituci  $w'_j = cw_j$ 
  - ▶ parametr  $c$  následně zrekonstruujeme  $c = \sum w'_j$
  - ▶ můžeme si dovolit - regrese netáhne k falešnému řešení
- ▶ řešíme lineární regresi pro parametry  $w_1, \dots, w_K$

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

Vícerozměrná  
data

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace

- ▶ opět iterujeme různé hodnoty  $c_1, c_2$
- ▶ zavedeme substituci  $w'_j = cw_j$ 
  - ▶ parametr  $c$  následně zrekonstruujeme  $c = \sum w'_j$
  - ▶ můžeme si dovolit - regrese netáhne k falešnému řešení
- ▶ řešíme lineární regresi pro parametry  $w_1, \dots, w_K$
- ▶ lze čekat, že některá  $w_j$  budou irelevantní
  - ▶ preferovaná metoda SVD
- ▶ doleštění vícerozměrnou optimalizací
  - ▶ známá citlivost na  $c_1, c_2$  - jsme blízko minima
  - ▶ výpočet  $I(q, c_1, c_2)$  není příliš složitý
    - ▶ parciální derivace, sdružené gradienty
    - ▶ zapouzdřeno v hotovém programu, zkusíme Powella

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

Vícerozměrná  
data

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace

- ▶ místo dvojic  $(x_i, y_i)$  máme  $(\mathbf{x}_i, y_i)$ 
  - ▶  $\mathbf{x}$  je  $k$ -rozměrný vektor
- ▶ model  $\mathbf{M}$  je funkce  $\mathbb{R}^k \rightarrow \mathbb{R}$
- ▶ jinak se nic nemění
  - ▶ minimalizujeme vůči parametrům
  - ▶ základní funkce se pouze vyhodnocují v  $\mathbf{x}_i$
  - ▶ není třeba derivovat podle složek  $\mathbf{x}_i$

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

Vícerozměrná  
data

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace



- ▶ místo dvojic  $(x_i, y_i)$  máme  $(\mathbf{x}_i, y_i)$ 
  - ▶  $\mathbf{x}$  je  $k$ -rozměrný vektor
- ▶ model  $\mathbf{M}$  je funkce  $\mathbb{R}^k \rightarrow \mathbb{R}$
- ▶ jinak se nic nemění
  - ▶ minimalizujeme vůči parametrům
  - ▶ základní funkce se pouze vyhodnocují v  $\mathbf{x}_i$
  - ▶ není třeba derivovat podle složek  $\mathbf{x}_i$
- ▶ obecněji  $\mathbf{M}: \mathbb{R}^k \rightarrow \mathbb{R}^l$ 
  - ▶ potřebujeme vhodnou metriku na  $\mathbb{R}^l$
  - ▶ počítáme se vzdáleností  $|\mathbf{y}_i, \mathbf{M}(\mathbf{x}_i)|$  podle ní

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

Vícerozměrná  
data

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace

# Komplexní příklad – ekvalizace elektronegativity

Obhajoba DP Tomáše Račka

PA081:  
Programování  
numerických  
výpočtů

A. Křenek

Modelování ex-  
perimentálních  
dat

Metoda  
nejmenších  
čtverců

Lineární  
regrese

Obecný model

Lineární  
modely

SAXS

**Vícerozměrná  
data**

Metoda  
nejmenších  
absolutních  
odchylek

Regularizace

# Metoda nejmenších absolutních odchylek

- ▶ metoda nejmenších čtverců minimalizovala

$$\sum (y_i - \mathbf{M}(x_i, a_1, \dots, a_M))^2$$

tj. normu řádu 2

- ▶ metoda **nejmenších absolutních odchylek** minimalizuje normu řádu 1

$$\sum |y_i - \mathbf{M}(x_i, a_1, \dots, a_M)|$$

# Metoda nejmenších absolutních odchylek

## Výhody a nevýhody

- ▶ větší robustnost
  - ▶ méně citlivá na odlehlé případy
  - ▶ nemají kvadratickou váhu, snáz se „přebijí“
- ▶ menší stabilita
  - ▶ malý posun v  $x$  může mít velký vliv na výsledné řešení
- ▶ nejednoznačné řešení
  - ▶ lineární členy se vzájemně kompenzují proti posunu v ose  $y$
- ▶ interaktivní srovnání na lineární regresi:
  - ▶ [http://www.math.wpi.edu/Course\\_Materials/SAS/tables/7.3/73\\_choices.html](http://www.math.wpi.edu/Course_Materials/SAS/tables/7.3/73_choices.html)

# Metoda nejmenších absolutních odchylek

## Výpočet

- ▶ nejmenší čtverce
  - ▶ derivace kvadratické funkce
  - ▶ vede na systém lineárních rovnic
- ▶ absolutní odchylky
  - ▶ úloha lineárního programování
  - ▶ např. Barrodale-Robertsův algoritmus



- ▶ místo minimalizace výrazu  $\|A\mathbf{a} - \mathbf{b}\|^2$  minimalizujeme

$$\|A\mathbf{a} - \mathbf{b}\|^2 + \|\Gamma\mathbf{a}\|^2$$

s vhodně volenou maticí  $\Gamma$  (Tikhonovova regularizace)

- ▶ podobně jako u nejmenších čtverců je řešením

$$(A^T A - \Gamma^T \Gamma)^{-1} A^T \mathbf{b}$$

- ▶ volbou  $\Gamma$  zvýhodňujeme nějaké řešení
  - ▶ např.  $\Gamma = \alpha I$  preferuje menší normu

# Regularizace

## Vztah k SVD

- ▶ rozklad  $A$  na signgulární hodnoty  $A = U\Sigma V^T$
- ▶ potom řešení regularizovaného problému s  $\Gamma = \alpha I$  je

$$V\Sigma'U^T b \quad \text{kde prvky } \Sigma \text{ jsou } \frac{\sigma_i}{\sigma_i^2 + \alpha^2}$$

tj.  $\alpha$  významněji ovlivní právě „skoro nulové“ singulární hodnoty