PA081: Programování numerických výpočtů

10. Modelování experimentálních dat, metoda nejmenších čtverců

Aleš Křenek

jaro 2011

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.

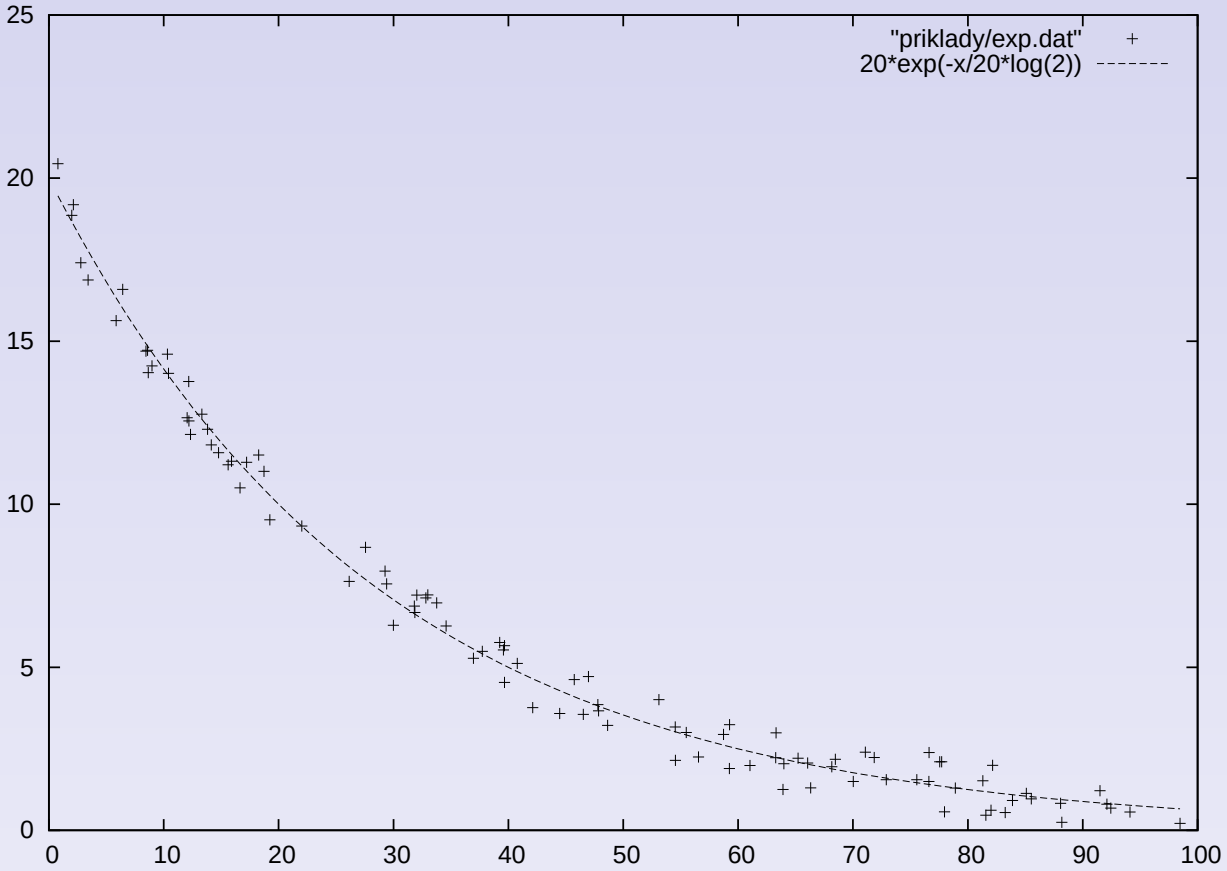
$$y = \mathbf{M}(x, a_1, \ldots, a_M)$$

- 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 zatížená 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 zatížená 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 Δy“
- 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

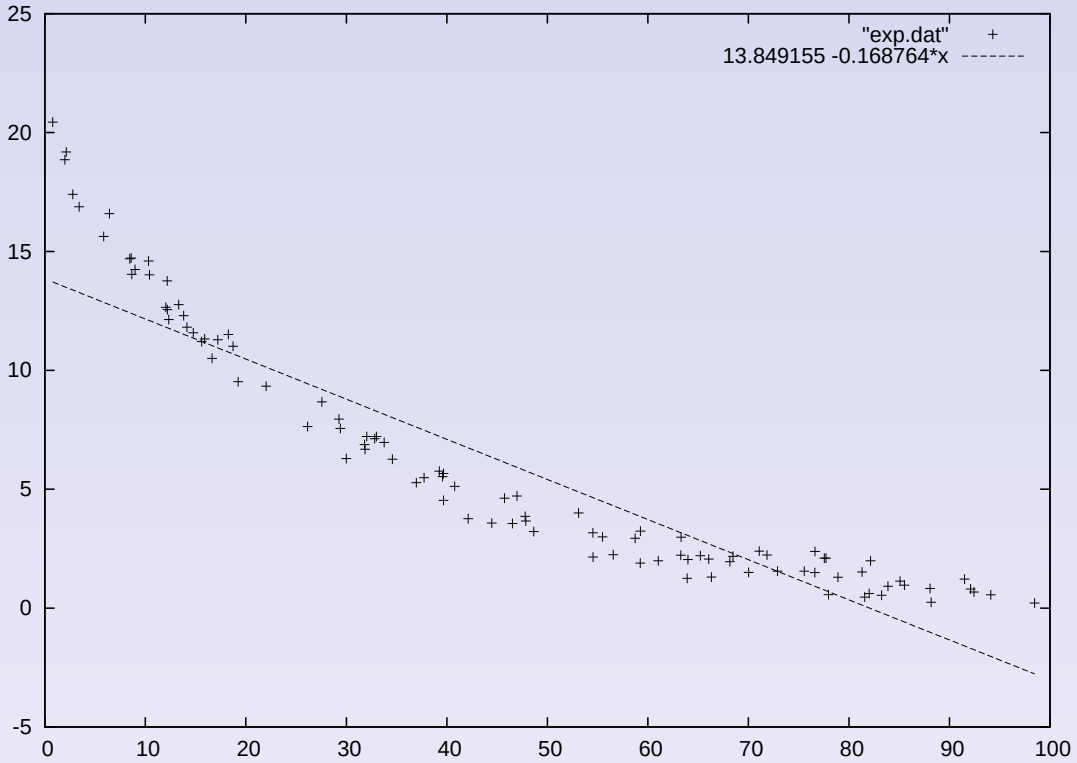
$$\chi^2 = \sum \frac{(y_i - \mathbf{M}(x_i))^2}{2\sigma_i^2}$$

- 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 χ^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 seriozní statistické zhodnocení
- hodnotíme pomocí *regularizované gamma funkce*
  - funkce konkrétního χ^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`
  - pravděpodobnost, že zcela náhodně vybraný vzorek (x_i, y_i) dá větší hodnotu χ^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

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

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

Lineární regrese

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

$$Sa + S_x b = S_y$$
$$S_x a + S_{xx} b = S_{xy}$$

- 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 $\sigma_i > 1.6$

Obecný model

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

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

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

Lineární modely

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

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

- 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 χ^2
- definujeme matici $\mathbf{A}$ a vektor $\mathbf{b}$

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

Lineární modely

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

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

- 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

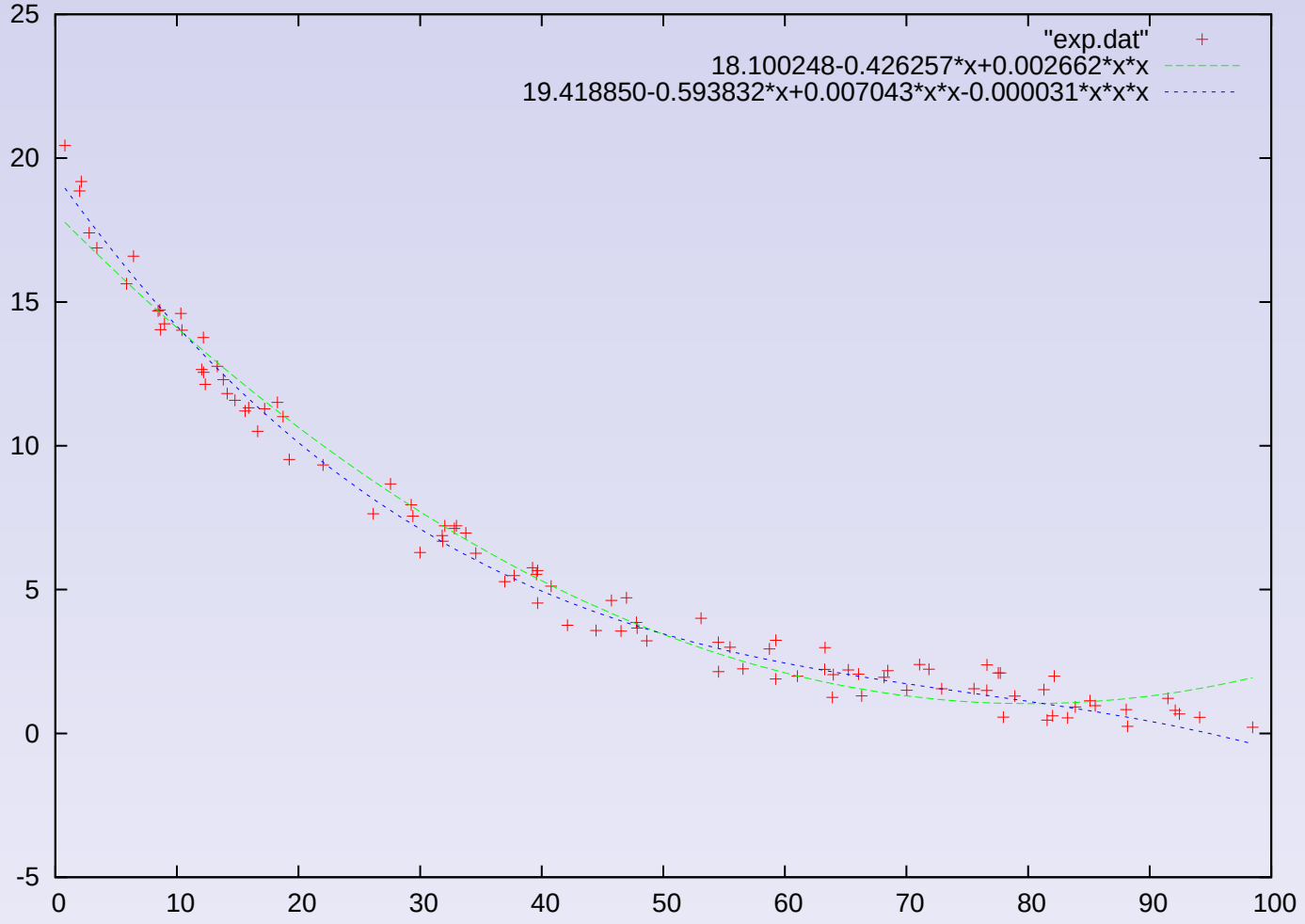
$$|\mathbf{Aa} - \mathbf{b}|^2$$

- zároveň detekuje problematické funkce

Lineární modely

Rozklad na singulární hodnoty

- stejná data, kvadratické a kubická funkce



Lineární modely

- tabulka singulárních hodnot pro různé řády polynomu

řád	σ_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

Vícerozměrná data

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