Numerické metody a programování

Lekce 11

Modelování dat

fitování dat modelem

model

- množina vhodných funkcí
- teoretický model

nalezení koeficientů modelu

- volba míry vhodnosti (kvality) modelu
- optimalizace
- stanovení neurčitostí parametrů modelu
- ověření vhodnosti optimálního modelu

Test dobré shody (goodness of fit)

model s M parametry

$$y(x) = y(x; a_1, a_2, \ldots, a_M)$$

naměřené hodnoty y_i v N bodech x_i

χ^2 statistika

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i; a_1, \ldots, a_M)}{\sigma_i} \right)^2$$

pravděpodobnost, že součet ν kvadrátů normálních náhodných veličin ($\sigma = 1$) překročí χ^2

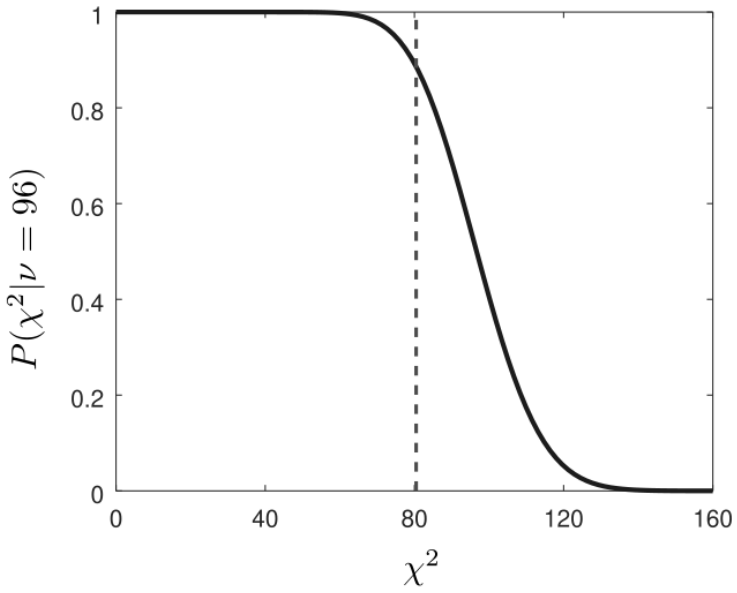
$$P(\chi^2 \mid \nu) = Q\left(\frac{\nu}{2}, \frac{\chi^2}{2} \right)$$

počet stupňů volnosti

$$\nu = N - M$$

regularizovaná neúplná gama funkce

$$Q(a, x) = \frac{1}{\Gamma(a)} \int_x^\infty e^{-t}\, t^{a-1}\, dt$$



model je zavržen, pokud je hodnota $P(\chi^2|\nu)$ příliš malá (např. $P < 0.005$)

χ^2 *fit přímkou*

model

$$y(x) = a\ x + b$$

kvalita fitu

$$\chi^2 = \sum_i \frac{1}{\sigma_i^2}(y_i - a x_i - b)^2$$

optimalizace

$$\frac{\partial \chi^2}{\partial a} = 0\ , \quad \frac{\partial \chi^2}{\partial b} = 0$$

statistické sumy

$$S = \sum_i \frac{1}{\sigma_i^2}, \quad S_x = \sum_i \frac{x_i}{\sigma_i^2}, \quad S_y = \sum_i \frac{y_i}{\sigma_i^2}, \quad S_{xx} = \sum_i \frac{x_i^2}{\sigma_i^2}, \quad S_{xy} = \sum_i \frac{x_i y_i}{\sigma_i^2}$$

extremální rovnice

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

řešení

$$a = \frac{S_{xy} S - S_x S_y}{\Delta}, \quad b = \frac{S_{xx} S_y - S_x S_{xy}}{\Delta}$$

$$\Delta = S_{xx} S - S_x^2$$

propagace chyb

$$\sigma_f^2 = \sum_i \sigma_i^2 \left(\frac{\partial f}{\partial y_i} \right)^2$$

neurčitosti parametrů modelu

$$\sigma_a^2 = \frac{S}{\Delta}, \quad \sigma_b^2 = \frac{S_{xx}}{\Delta}$$

χ^2 *obecný lineární fit*

model vytvořen lineární kombinací bázových funkcí

model

$$y(x) = \sum_{k=1}^{M} a_k X_k(x)$$

např. $\{X_k\} = \{1, x, x^2, \ldots\}$

kvalita fitu

$$\chi^2 = \sum_i^N \left[\frac{y_i - \sum_k^M a_k X_k(x_i)}{\sigma_i} \right]^2$$

zavedeme

$$b_i = \frac{y_i}{\sigma_i}, \quad A_{ik} = \frac{X_k(x_i)}{\sigma_i}$$

dostaneme

$$\chi^2 = \sum_i^N \left(b_i - \sum_k^M A_{ik} a_k \right)^2 = \left\| A \vec{a} - \vec{b} \right\|^2$$

minimalizace χ^2 → SVD rozklad → pseudoinverze

řešení

$$\vec{a} = \underbrace{(A^T A)^{-1}}_{C} A^T \vec{b} = A^- \vec{b}, \quad A^- \ldots \text{pseudoinverze}$$

Jakobián

$$J_{kj} = \frac{\partial a_k}{\partial b_j}$$

propagace kovarianční matice

$$\Gamma^a = J \Gamma^b J^T$$

pro model výše (nezávislá data y_i) platí

$$J = A^-, \quad \Gamma^b = \hat{1}$$

a tedy

$$\Gamma^a = J J^T = C A^T A C^T = C$$

speciální případ

$$\sigma^2(a_j) = C_{jj}$$

např. pro fitování přímkou

$$y(x) = a_1 + a_2 x \quad \rightarrow \quad X_1(x) = 1, \quad X_2(x) = x \quad \rightarrow \quad A_{i1} = \frac{1}{\sigma_i}, \quad A_{i2} = \frac{x_i}{\sigma_i}$$

statistické sumy

$$A^T A = \begin{pmatrix} \sum_i \frac{1}{\sigma_i^2} & \sum_i \frac{x_i}{\sigma_i^2} \\ \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{x_i^2}{\sigma_i^2} \end{pmatrix} = \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix}$$

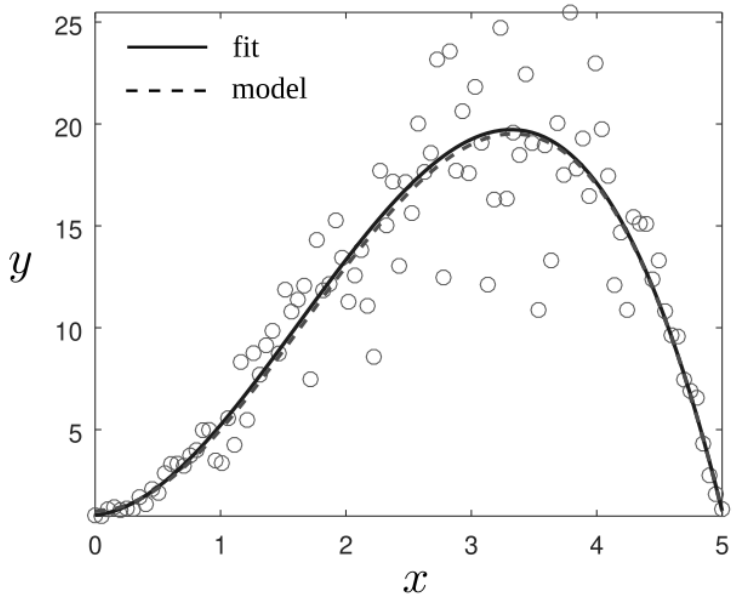
kovarianční matice

$$C = (A^T A)^{-1} = \frac{1}{\Delta} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}$$

variance ve shodě s výše odvozeným výsledkem

$$\sigma^2_{a_1} = C_{11} = \frac{S_{xx}}{\Delta}, \quad \sigma^2_{a_2} = C_{22} = \frac{S}{\Delta}$$

příklad:



skutečný model:

$y = -x^3 + 5x^2 + 1$

variance:

$\sigma = y/5$

počet měření:

$N = 100$

bázové funkce:

$1, x, x^2, x^3$

koeficienty:

$0{,}805\,;\, 0{,}617\,;\, 4.782\,;\, -0.979$

chyby parametrů:

$0{,}113\,;\, 0{,}450\,;\, 0{,}297\,;\, 0{,}045$

$\chi^2 = 80{,}46$

$P(\chi^2 \mid \nu) = 0{,}886$