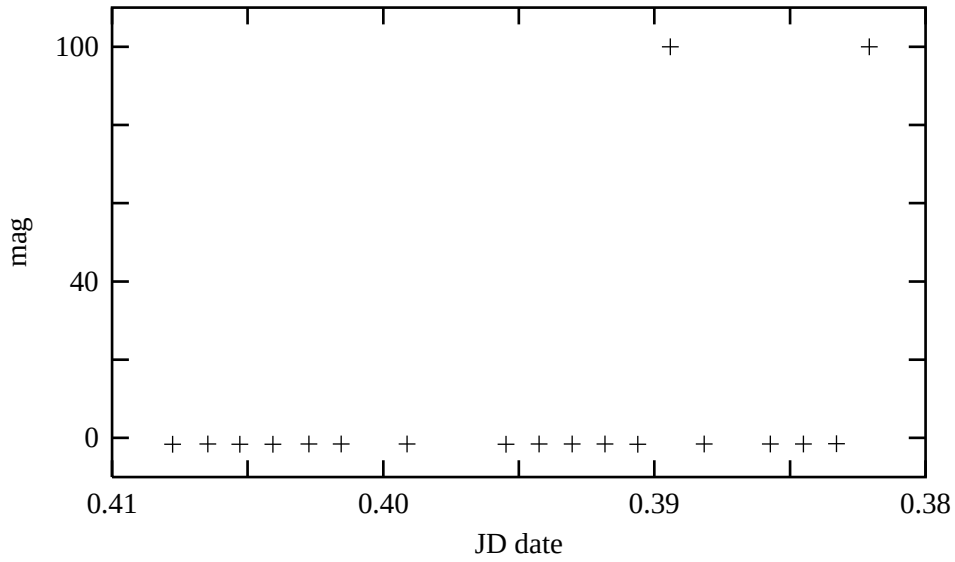
Robustní odhady statistických parametrů

“Někdy pracují dobře, jinde ne.”

Typická data - pozorování BL Lac




```
0.38223 -1.586 0.017      0.40550 -1.530 0.019
0.39453 -1.610 0.024      0.40671 -1.511 0.017
0.39575 -1.563 0.019      0.38353 -1.562 0.019
0.39697 -1.552 0.020      0.40792 99.999 9.999
0.39818 -1.556 0.019      0.38471 -1.618 0.023
0.39939 -1.590 0.014      0.38593 -1.612 0.026
0.40059 99.999 9.999      0.38726 -1.548 0.022
0.40184 -1.558 0.015      0.38845 -1.561 0.026
0.40428 -1.572 0.018      0.39088 -1.527 0.025
```

σ-clipping algoritmy

snaží se eliminovat příliš odlehlé hodnoty
Příklad použití na data z BL Lac (nevážené, $\sigma = 3.3$):

i	$\overline{x_i}$	σ_{x_i}
0	9.71900	7.741449
1	9.71900	7.741449

. . .

Příklad použití na data z BL Lac (nevážené, $\sigma = 2.0$):

i	$\overline{x_i}$	σ_{x_i}
0	9.719000	7.74144
1	-1.566000	2.913785
2	-1.566000	0.0077308

. . .

Typy na zlepšení: váhování dat, použití mediánu

Odhad aritmetického průměru

Vychází z metody největší věrohodnosti (Brandt 1970):

$$L = \prod_{i=1}^{N} f(\mathbf{x};\mathbf{t})$$

$$l = \ln L = \sum_{i=1}^{N} \ln f(x_i;\mathbf{t})$$

kde **t** jsou hledané parametry a x_i pak naměřená data. V případě minimalizace nejmenších čtverců použijeme

$$f(x_i;\overline{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-\overline{x})^2}{2}\right)$$

což vede na

$$\sum_{i=1}^{N} (x_i-\overline{x})^2 = 0$$

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Robustní odhad

Vychází opět z metody největší věrohodnosti. Základní rozdíl je v tom, že hledáme minimum určité funkce (Launer, Wilkinson 1979):

$$l = \ln L = \sum_{i=1}^{N} \ln f(x_i; \mathbf{t}) = -\sum_{i=1}^{N} \rho(x_i; \mathbf{t})$$

Minimalizací pak dostáváme:

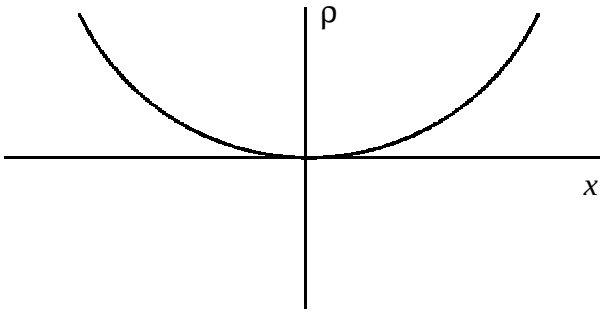
$$\sum_{i=1}^{N} \rho'(x_i; \mathbf{t}) \equiv \sum_{i=1}^{N} \psi(x_i; \mathbf{t}) = 0$$

Tato poslední rovnice je implicitní rovnicí vzhledem k parametrům. Její řešení je v praktických případech nutné provádět numericky.

Běžně používané volby ψ

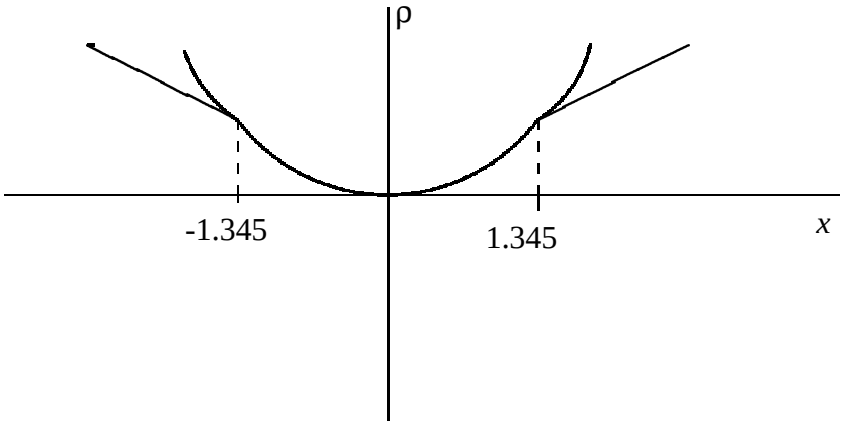
Gaussova funkce

$$\rho(x) = \frac{x^2}{2} \qquad \psi(x) = x$$



Huberova funkce (a = 1.645 na 95% hladině)

$$\rho = \begin{cases} \frac{x^2}{2}, & |x| \le a \\ a|x| - \frac{a^2}{2}, & |x| > a \end{cases} \qquad \rho = \begin{cases} -a, & x < -a \\ x, & |x| \le a \\ a, & x > -a \end{cases}$$



Robustní aritmerický průměr

Jejím základem je řešení rovnice

$$\sum_{i=1}^{N} \psi\left(\frac{x_i - \overline{x}}{s}\right) = 0$$

kde parameter s je normalizační parametr volený tak aby se blížil k sigma parametru gausova rozdělení u normálně rozdělených dat. Tuto rovnici lze nejlépe řešit Newtonovou metodou. Předtím ovšem musíme znát dobrý odhad kořene, v našem případě prumeru. Obyčejný průměr je k tomu nevhodný, nejlépe je použít median nebo nejčetnější hodnotu.

$$\overline{x}^{(0)} = \mathrm{med}(x_i), \qquad s = \mathrm{med}(x_i - \overline{x}^{(0)})/0.6745$$

$$\overline{x}^{(i+1)} = \overline{x}^{(i)} + \frac{s\sum \psi\left[(x_i - \overline{x}^{(i)})\right]}{\sum \psi'\left[(x_i - \overline{x}^{(i)})\right]}$$

Zjednodužšená implementace odhadu z Munipacku:

```
subroutine prumer(n,x,t,dt)
    integer :: n        ! pocet dat
    real :: x(n)        ! hodnoty vyberu
    real :: t,dt        ! odhady parametru

    ! nulty odhad
    t = median(n,x)
    s = median(n,abs(x - t))
    s = s/0.6745

    do
       d = s*sum(psi((x - t)/s))/ &
           sum(dpsi((x - t)/s))
       t = t + d
       if( abs(d) < epsilon(d) exit
    enddo
    dt = sqrt(s**2*n/(n-1)* &
         sum(psi((x - t)/s)**2)/ &
         sum(dpsi((x - t)/s))**2)

end subroutine
```

Porovnání různých metod odhadu pro BL Lac

Aritmetický průměr: 9.719 ± 7.741

Vážený aritmetický průměr: -1.564530 ± 0.0180619

$3.3-\sigma$ clip bez vah: -1.566000 ± 0.0077308

Robustní průměr bez vah :

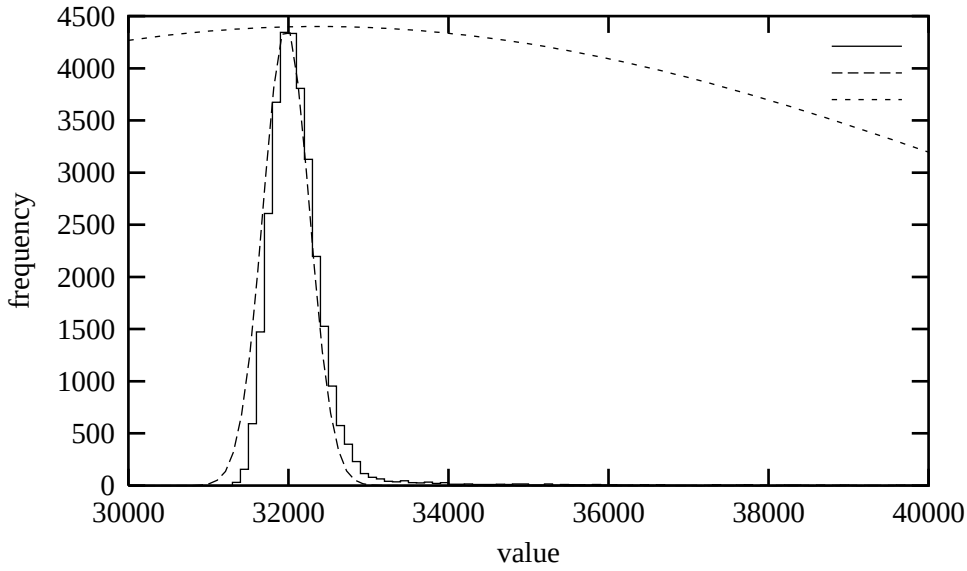
i	$\overline{x_i}$	σ_{x_i}
0	-1.561000	0.0250000
1	-1.565836	0.0079912
2	-1.565826	0.0079158

Odhad úrovně oblohy CCD snímku

Histogram intenzit náhodně vybraných ze snímku.

Obloha určena aritmetickým průměrem:
32361.26 ± 33.55

Obloha určena robustním průměrem:
31972.72 ± 1.634262



Přehled robustních metod

M-odhady představují odhady založené na metodě největší věrohodnosti, (*M*aximum likelihood)

L-odhady jsou *L*ineární kombinace několika statistik, například průměr, median, kvantily a pod

R-odhady jsou na základě fitování parametrů distribucí buď v histogramu nebo distribuční funkci, (*R*ank tests)

(Press, Flannery, Teukolsky, Vetterling 1986), (Launer, Wilkinson 1979)

Robustní metody a robustní metody

V praxi se lze setkat se dvojím typem tzv. robustních metod:

odhady robustní odhad o kterém uz byla řeč

metody to je pojem odvozený z různých sw balíků (Matlab, S, Octave,...) kde se tak označují fitovací metody pro vícerozměrnou minimalizaci

Vícerozměrná robustní minimalizace

Pricip metod vícerozměrné minimalizace je stejný jako jako u jednorozměrné.

Rozdíly jsou:

počátecní odhad nelze získávat medianem. Je nutné použít vícerozměrnou obdobu mediánu, kterou je minimalizace součtu absloutních odchylek

Je nutné použít vícerozměrnou verzi Newtonovy metody řešení rovnic. Všeobecně nejvíce uznávaná metoda na řešeni nelineárních rovnic jsou algoritmy MINPACKU založené na Marquard-Levendbergove metodě s výpočtem linearnich rovnic pomoci QR algoritmu.

Fitování profilu hvězdy

Byla fitována funkce

$$G(x,y|x_0,y_0,\sigma_x,\sigma_y,B,G_0) = G_0 e^{[(-(x-x_0)^2/2\sigma_x-(y-y_0)^2/2\sigma_y)]}+B$$

Výsledek pro LM metodu bez robustního odhadu s výpočtem matice LL faktorizací:

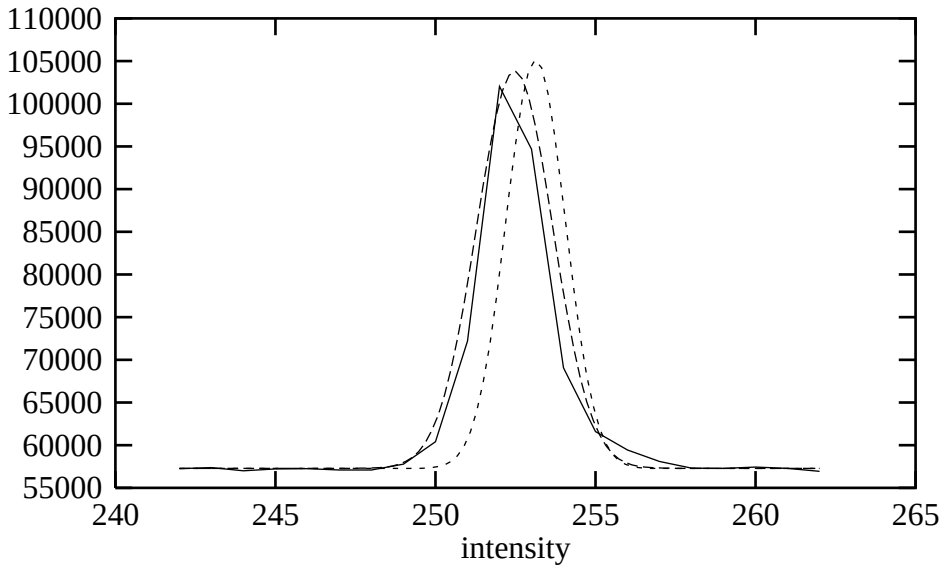
x_0	66.98005 ± 6.9058
y_0	253.1362 ± 7.326046
σ_x	0.8103039 ± 6.632534
σ_y	0.9281055 ± 7.338552
G_0	47895.81 ± 1169.781
B	57268.71 ± 389.8009

Residuální součet byl 6.915187E+08.

Robustní odhad MINPACKem s QR faktorizací:

x_0	66.96627 ± 2.690513
y_0	252.4740 ± 0.1612529
σ_x	0.7978429 ± 2.136275
σ_y	1.192521 ± 0.1654361
G_0	46561.96 ± 14156.46
B	57288.08 ± 18799.11

Residualní součet 1.641804E+07.


110000
105000
100000
95000
90000
85000
80000
75000
70000
65000
60000
55000
240
245
250
255
260
265
intensity

Literatura

S. Brandt: Statistical and Computational Methods in data analysis, Elsevier 1970

W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling : Numerical Recipes, Cambridge University Press 1986

R. L. Lauer, G. N. Wilkinson (eds.): Robustness in statistics, Academic Press 1979