Robustní statistické metody

Populární úvod

Filip Hroch

Ústav teoretické fyziky a astrofyziky, MU Brno

28. říjen 2006, Vlašim

O co jde?

- Robustní znamená: “necitlivý k malým odchylkám od ideálních předpokladů na který je metoda odhadu optimalizována”.
- Praktický rozdíl pro PCVista vs. DAOPHOT.
- Tento způsob nazírání zavedl do statistiky G.E.P. Box v roce 1953.
- Životně důležitá metoda pro hromadné zpracování dat bez jakéhokoli zásahu obsluhy.
- Slovem “robustní” se označují často též metody numericky stabilní (řešení normálních rovnic singulárním nebo QR rozkladem místo LL eliminace).

Klasické metody

- "Někdy fungují dobře, jindy ne."
- Aritmerický průměr $\bar{x}$, střední odchylka, σ.
- Předpoklad: data mají normální rozdělení.
- Problém odlehlých hodnot se řeší filtrováním, např. σ-clipping.

$\sigma = 3.3$

i	$\bar{x}^{(i)}$	$\sigma^{(i)}$
0	9.71900	7.74145
1	9.71900	7.74145
2	9.71900	7.74145
...		

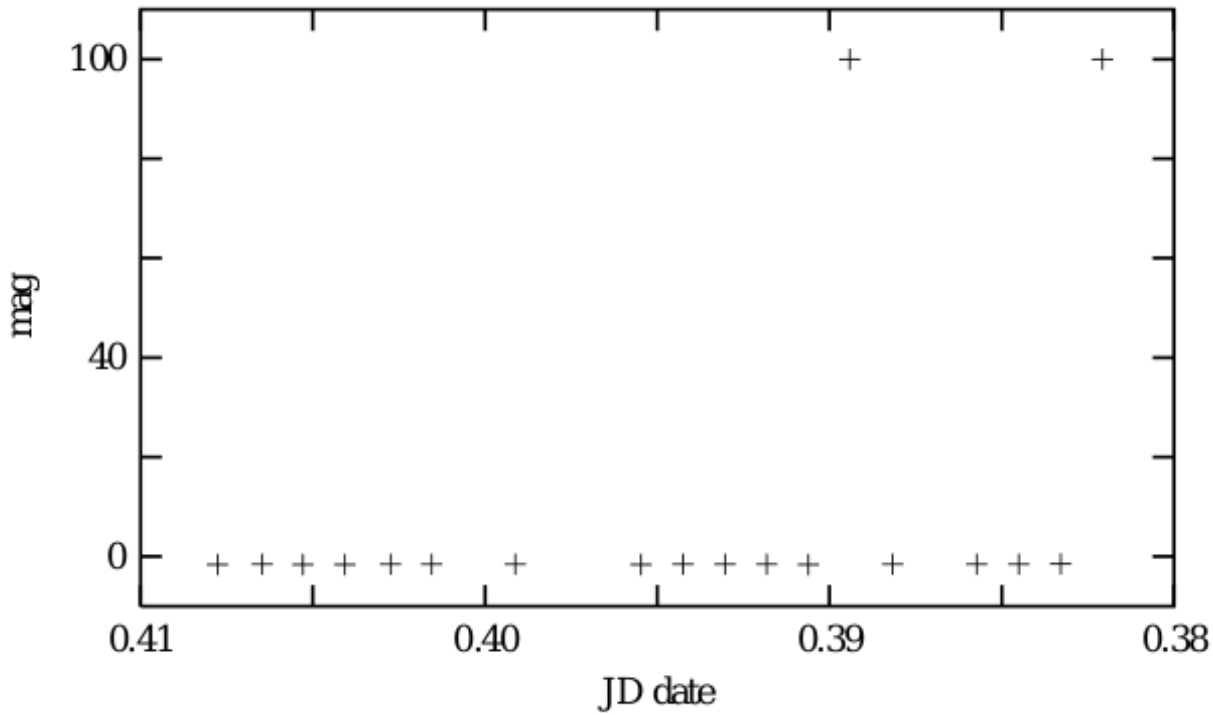
$\sigma = 2.0$

i	$\bar{x}^{(i)}$	$\sigma^{(i)}$
0	9.71900	7.74145
1	-1.56600	2.91379
2	-1.56600	0.00773
...		

Klasické metody — BL Lac

Část měřených magnitud

t_{JD}	m_i	σ_i	t_{JD}	m_i	σ_i
0.38223	-1.586	0.017	0.40550	-1.530	0.019
0.40059	99.999	9.999	0.40671	-1.511	0.017
0.40184	-1.558	0.015	0.38353	-1.562	0.019
0.40428	-1.572	0.018	0.40792	99.999	9.999



Metoda největší věrohodnosti

Maximum likehood method

$$L = \prod_{i=1}^{N} f(\boldsymbol{x}; \boldsymbol{t}) \qquad (\max.)$$

Metoda nejmenších čtverců

$$f(x_i; \bar{x}) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-(x_i - \bar{x})^2/2}$$

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

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

Metoda největší věrohodnosti — rozdělovací funkce

Substituce funkcí

$$\ln L(\bar{x}) = \sum \ln f(x_i - \bar{x}) = -\sum \varrho(x_i - \bar{x})$$

$$\sum \psi(x_i - \bar{x}) = 0$$

Normální rozdělení — aritmerický průměr

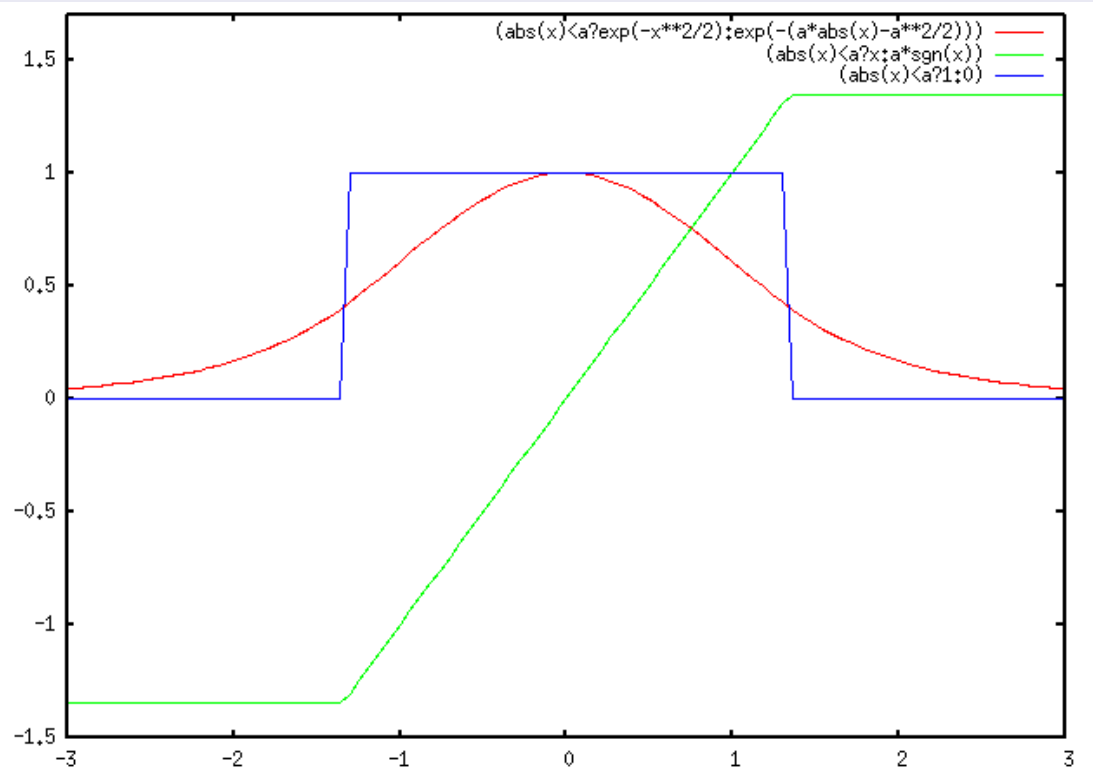
$$f(x) = exp(-x^2/2),\ \varrho = \frac{x^2}{2},\ \psi = x$$

Exponenciální rozdělení — medián

$$f(x) = exp(|x|),\ \varrho = |x|,\ \psi = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

Princip: minimalizujeme L pro speciální funkce.

$$\varrho(x) = \begin{cases} x^2/2, & |x| \leq a \quad a = 1.345 \\ a|x| - a^2/2, & |x| > a \end{cases}$$



Robustní aritmetický průměr

Postup řešení

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

0

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

i

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

∞

$$\sigma^2 = s^2 \frac{n}{n-1} \frac{\sum \psi^2}{(\sum \psi')^2}$$

Robustní metody — přehled

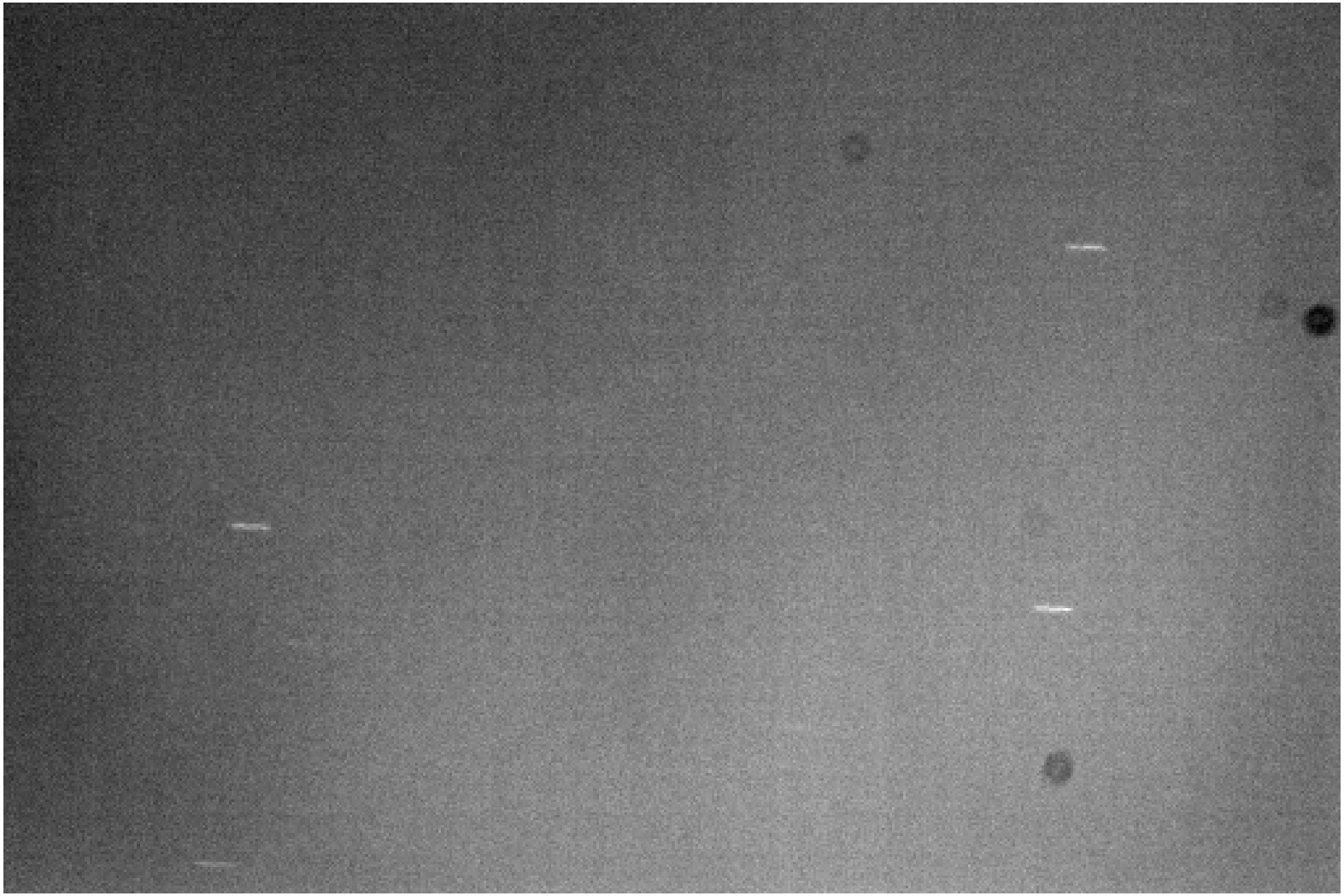
Huberova klasifikace

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

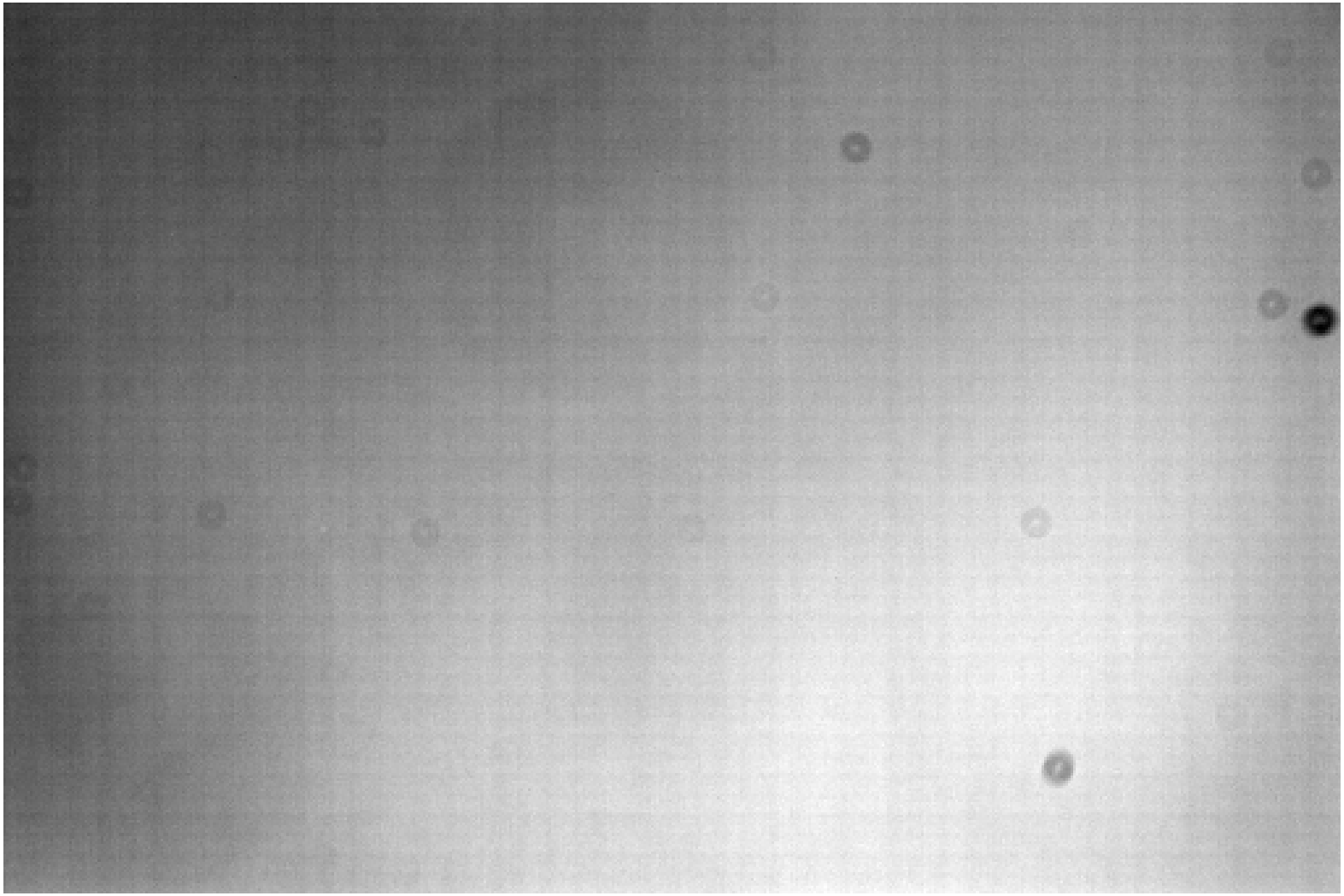
L-odhady jsou *L*ineární kombinace několika statistik, například průměr, median, kvantily a pod.

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

Použití robustních metod — flat field — originál



Použití robustních metod — flat field — průměr



Použití robustních metod — BL Lac

Průměrná magnituda

- Aritmetický průměr: 9.719 ± 7.741
- Vážený aritmetický průměr: -1.565 ± 0.018
- $2.0 - \sigma$ clip bez vah: -1.566 ± 0.008
- Robustní průměr bez vah :

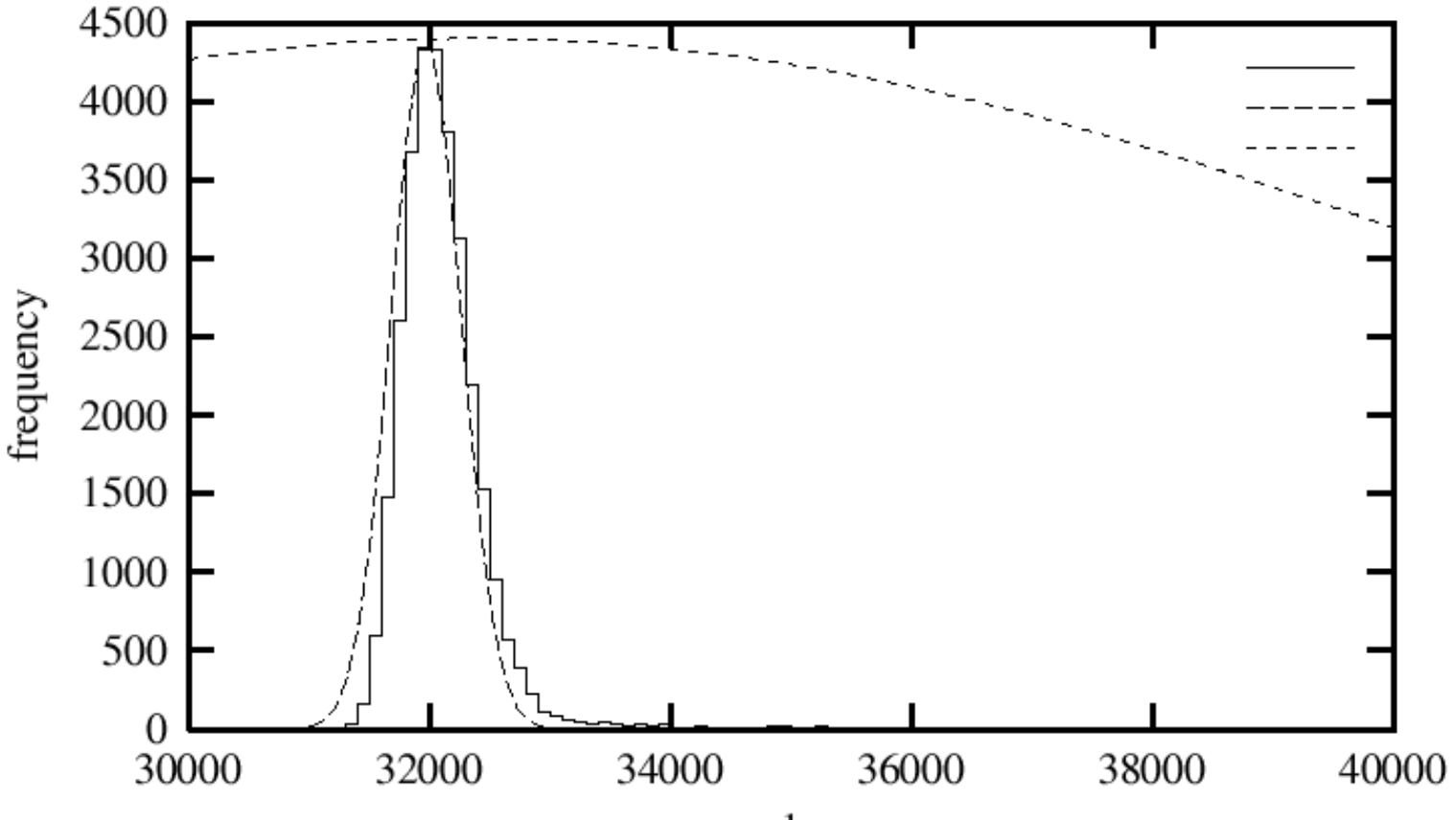
i	$\bar{x}^{(i)}$	$\sigma^{(i)}$
0	-1.56100	0.02500
1	-1.56583	0.00799
2	-1.56582	0.00791

Použití robustních metod — jas oblohy

Úroveň oblohy

- aritmetický průměr: 32361.3 ± 33.6
- robustní průměr: 31972.7 ± 1.6

Histogram intenzit náhodně vybrané části snímku.

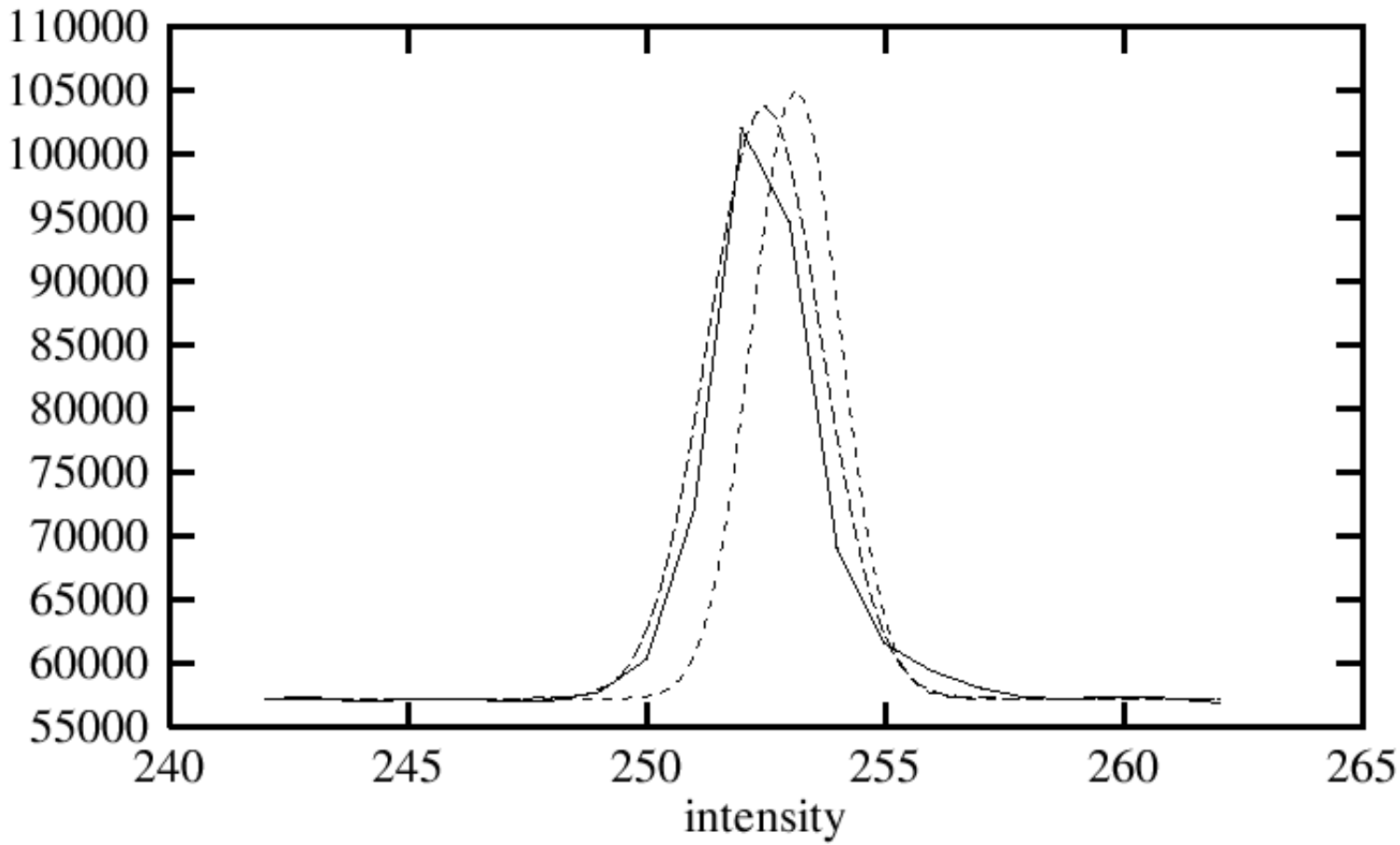


Změny

- Medián je transformován na součet absolutních hodnot odchylek (minimalizace negradientní metodou).
- Vícerozměrná metoda řešení rovnic — MINPACK (QR + Marquart Levendberg).

Použití robustních metod ve více rozměrech — profil (1)

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



Použití robustních metod ve více rozměrech — profil (2)

LM+LL*

x_0	67.0	$\pm$	6.9
y_0	253.1	$\pm$	7.3
d_x	0.810	$\pm$	6.6
d_y	0.928	$\pm$	7.3
G_0	47900	$\pm$	1170
B	57300	$\pm$	390
S_0	$6.91 \cdot 10^8$		

* Levendberg – Marquartova metoda + LL rozklad při výpočtu matic.

MINPACK+QR*

x_0	67.0	$\pm$	2.7
y_0	252.47	$\pm$	0.16
d_x	0.798	$\pm$	2.130
d_y	1.192	$\pm$	0.165
G_0	46600	$\pm$	14200
B	57300	$\pm$	19000
S_0	$1.64 \cdot 10^7$		

*Upravená Levendberg – Marquartova metoda + QR rozklad při výpočtu matic.

Literatura

- S. Brandt: Statistical and Computational Methods in data analysis, Elsevier 1970
- P. J. Huber: Robust Statistics, Wiley, New York 1981
- R. L. Lauer, G. N. Wilkinson (eds.): Robustness in statistics, Academic Press 1979
- W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling: Numerical Recipes, Cambridge University Press 1986