A Comparison of Robust Location Estimators

by
William C. Wojciechowski

October 11, 1999

Motivation for Robust Statistics

- Never know the true distribution

- Assumptions are different than truth

- Performance of MLE diminishes

- Robust methods performance do not diminish

- Gain robustness, lose efficiency

Defining Robustness

Qualitative Robustness

- Hampel derives formal definition

- Prohorov distance

$$d(F_1, F_2) = \inf\{\varepsilon : F_1(S) \leq F_2(S^\varepsilon) + \varepsilon, \forall S \in \mathcal{S}\}$$

- Qualitative Robustness

$$\forall \varepsilon > 0, \exists \delta > 0, \forall F_2, \forall n \geq n_0 : d(F_1, F_2) < \delta \Rightarrow d(F_{\mathcal{I},1}(T_n), F_{\mathcal{I},2}(T_n)) < \varepsilon$$

- Qualitative Robustness

$$\forall \varepsilon > 0, \exists \delta > 0, \forall F_2, \forall n \geq n_0 : d(F_1, F_2) < \delta \Rightarrow |T_n(F_1) - T_n(F_2)| < \varepsilon$$

Defining Robustness

Quantitative Robustness

- Breakdown Points
- Maximum Asymptotic Bias
- Maximum Asymptotic Variance
- Influence Function

$$IF(x) = lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F+\varepsilon\delta(x))-T(F)}{\varepsilon}$$

Types of Robust Estimators

- L-Estimators

- R-Estimators

- M-Estimators

  $g(x;\theta)$ function of data and parameter

  $\min \Sigma_{i=1}^{n} g(x_i;T_n)$ or $\Sigma_{i=1}^{n} \psi(x_i;T_n) = 0$

  $g(x;\theta) = -log f(x;\theta)$ is MLE

Types of Robust Estimators

L_2 Distance

- $\int (f_\theta(z) - g(z))^2 dz$

- Empirical Version

$$\int f_\theta(z)^2 dz - 2 \int f_\theta(z) g(z) + \int g(z)^2$$

$$\int f_\theta(z)^2 dz - \frac{2}{n} \sum_{i=1}^{n} f_\theta(z_i)$$

Estimators considered

Normal L_2E

Mean

Median

Edgeworth

Outmean

10% Trim

15% Trim

25% Trim

Huber P15

Andrews AMT

Tukey Biweight

Hogg T1

Contamination considered

- $(1-\rho)N(0,1)+\rho N(\mu\sigma)$

- Mean shift

  $\mu=(2,4),\sigma=(1)$

- Standard deviation shift

  $\mu=(0),\sigma=(3,10)$

- $\rho=(0.05,0.10,0.15,0.25)$

- $n=(5,10,20,40,100,1000)$

Measures of performance

- Index of relative error ($\bar{\varepsilon}$)

$$\omega_j = \frac{1}{12} \sum_{i=1}^{12} |\hat{\mu}_{ij} - \mu_j|$$

$$\varepsilon_{ij} = \frac{|\hat{\mu}_{ij} - \mu_j|}{\omega_j}$$

$$\bar{\varepsilon}_i = \frac{1}{n} \sum_{j=1}^{n} \varepsilon_{ij}$$

- Index of relative rank ($\bar{\rho}$)

  Rank $|\hat{\mu}_{ij} - \mu_j|$ within each dataset

  ρ_{ij} = rank of estimator i for dataset j

$$\bar{\rho}_i = \frac{1}{n} \sum_{j=1}^{n} \rho_{ij}$$

Results

	Relative Error			*Relative Rank*		
Estimator	*All*	σ	μ	*All*	σ	μ
mean(MLE)	1.45	1.54	1.35	9.80	9.04	10.55
median	0.71	0.69	0.74	3.90	5.38	2.42
trim.10	0.99	0.94	1.03	7.63	7.22	8.04
trim.15	0.91	0.86	0.96	6.70	6.70	6.71
trim.25	0.76	0.71	0.82	4.31	4.88	3.75
huber	0.85	0.76	0.94	6.41	5.94	6.88
andrews	0.89	0.73	1.05	7.12	5.56	8.69
tukey	0.77	0.72	0.82	4.52	4.85	4.19
edgeworth	0.83	0.79	0.87	5.40	6.12	4.67
outmean	2.26	2.63	1.88	11.42	10.83	12.00
hogg	1.00	0.91	1.08	7.86	6.65	9.07
L_2E	0.58	0.71	0.45	2.94	4.83	1.04