# Empirical likelihood-based inference for the difference of smoothed Huber estimators

J. Valeinis[1], M. Velina[1] and G. Luta[2]

[1]University of Latvia, Riga

[2]Georgetown University, Washington, D.C.

28th of June, 2011

# Recent results

1. Owen (1988) - empirical likelihood (EL) for Huber M-estimator
2. Valeinis (2007, 2010, 2011) - general two-sample EL method.
   - ROC curves, P-P and Q-Q plots;
   - difference of the two-sample means, distribution functions and quantiles, location-scale models;
   - R-code (collaboration with E. Cers) and package based on *smoothed estimating equations* for EL method.
3. Hampel *et al.* (2011) - smoothed Huber estimators.

**Idea** - to establish EL for the difference of two smoothed Huber estimators

# Empirical likelihood method

Empirical likelihood method was introduced by Art B. Owen in 1988. The idea - model data using distributions placing point masses on the data.

$$L(F) = \prod_{i=1}^{n} P(X = X_i) = \prod_{i=1}^{n} p_i, \quad \sum_{i=1}^{n} p_i = 1.$$

# Nonparametric (empirical) likelihood

**Definition**

Let $X_1, \ldots, X_n$ i.i.d. with unknown $F_0$. For distribution $F$ the nonparametric likelihood function is

$$L(F) = \prod_{i=1}^{n} (F(X_i) - F(X_i-)) = \prod_{i=1}^{n} p_i,$$

where $p_i = P(X = X_i)$ and $\sum_{i=1}^{n} p_i = 1$.

- $L(F)$ is maximized by ECDF $F_n$ with $p_i = 1/n$.
- IDEA: to express the parameter of interest $\theta$ as a functional from $F$, i.e., $\theta = \theta(F)$.

# Empirical likelihood (EL) for mean $\mu$

- $X_1, \ldots, X_n$ i.i.d. with $EX_i = \mu_0 \in \mathbb{R}$.
- Estimator $\hat{\mu} = \int x dF_n(x) = \bar{X}$.
- Profile empirical likelihood ratio

$$R(\mu) = \sup\left\{ \prod_{i=1}^{n} np_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i = \mu \right\}.$$

**Theorem (Owen, 1988)**

*$X_1, \ldots, X_n$ i.i.d. with $\mu_0 < \infty$. Then*

$$-2 \log R(\mu_0) \to_d \chi^2_1.$$

# EL in one-sample case: generalization

- Estimating function $g(X, \theta)$, $\theta \in \Theta \subset \mathbb{R}^p$ with

$$E_{F_0}\{g(X, \theta)\} = 0,$$

$$R(\theta) = \sup\Big\{ \prod_{i=1}^{n} np_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g(X_i, \theta) = 0 \}$$

- If $\theta_0 = \mu_0$ take $g(X_i, \theta) = X_i - \mu$.
- For Huber location M-estimator take $g(X_i, \theta) = \psi\left(\frac{X_i - \mu}{\hat{\sigma}}\right)$.

**Theorem (Qin and Lawless, 1994)**

*Under some (smoothness) conditions on $g(X, \theta)$,*

$$-2 \log R(\theta_0) \to_d \chi^2_p.$$

# M-estimators for location

**Definition**

Given a function $\rho$, an M-estimate of location is a solution of

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(X_i - \mu).$$

or equivalently if $\rho$ is differentiable, then

$$\sum_{i=1}^{n} \psi(X_i - \hat{\mu}) = 0.$$

**Definition**

M-estimate of location using scale estimators

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right) = 0.$$

# Huber's M-estimator

**Definition**

Huber's (1964) $\rho$ and $\psi$ functions

$$\rho_k(x) = \begin{cases} x^2 & \text{if } |x| \le k \\ 2k|x| - k^2 & \text{if } |x| > k, \end{cases}$$

$$\psi_k(x) = \begin{cases} x & \text{if } |x| \le k \\ sgn(x)k & \text{if } |x| > k. \end{cases}$$

- If $k \to \infty$ we obtain the sample mean;
- If $k \to 0$ we obtain the sample median;
- Standard choices for $k = \{0.86, 1.35, 1.5\}$.

# Smoothed M-estimator (Hampel *et al.*, 2011)

For a general $\psi$-function of an M-estimator define

$$\tilde{\psi}(x) = \int \psi(x+u) dQ_n(u),$$

where

- $Q_n$ may be chosen as a the distribution of the initial M-estimator
- $Q_n$ can be approximated by $N(0, V/n)$, where $V$ is asymptotic variance of the M-estimator.
- The smoothing principle can be applied to $\psi$ functions already smooth.

# Smoothed Huber M-estimator (Hampel *et al.*, 2011)

$$\begin{aligned}
\tilde{\psi}_k(x) &= k\Phi\left(\frac{x-k}{\sigma_n}\right) - k\left(1 - \Phi\left(\frac{x+k}{\sigma_n}\right)\right) \\
&\quad + x\left(\Phi\left(\frac{x+k}{\sigma_n}\right) - \Phi\left(\frac{x-k}{\sigma_n}\right)\right) \\
&\quad + \sigma_n\left(\phi\left(\frac{x+k}{\sigma_n}\right) - \phi\left(\frac{x-k}{\sigma_n}\right)\right),
\end{aligned}$$

where $\sigma_n = \sqrt{V/n}$, $\Phi$ and $\phi$ denote the cdf and pdf of $N(0,1)$.

- $V$ is the asymptotic variance of the nonsmooth Huber estimator
- For simulations Hampel *et al.* (2011) take $V = 2.046$ (an upper bound for contaminated normal models with $\epsilon = 0.2$).
- **Problem**: for practical applications $V$ has to be estimated!

# EL in two-sample setting: general framework

- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ i.i.d. with unknown $F_1$, $F_2$.
- $\Delta$ is a univariate parameter of interest, where $\theta_0$ is some nuisance parameter associated with $F_1$.

$$\mathbb{E}_{F_1} w_1(X, \theta_0, \Delta_0) = 0, \quad \mathbb{E}_{F_2} w_2(Y, \theta_0, \Delta_0) = 0.$$

- Profile EL ratio

$$R(\Delta) = \sup_{\theta, p, q} \prod_{i=1}^{n} (np_i) \prod_{j=1}^{m} (mq_j),$$

where $p_i, q_j \geq 0$, $\sum_{i=1}^{n} p_i = 1$, $\sum_{j=1}^{m} q_j = 1$ and

$$\sum_{i=1}^{n} p_i w_1(X_i, \theta, \Delta) = 0, \quad \sum_{j=1}^{m} q_j w_2(Y_j, \theta, \Delta) = 0.$$

# EL in two-sample setting: smooth Huber estimators

**Theorem (Valeinis, 2010, 2011)**

*Under some conditions (in particular, differentiability) on $w_1(X, \theta, \Delta)$ and $w_2(Y, \theta, \Delta)$,*

$$-2 \log R(\Delta_0, \hat{\theta}) \to_d \chi^2_1.$$

- Theorem holds for ROC, PP, QQ, difference of means etc.
- For the difference of two Huber smooth estimators, take $\Delta = \theta_1 - \theta_0$ and

$$w_1(X, \theta_0, \Delta) = \tilde{\psi}_k\left(\frac{X - \theta_0}{\hat{\sigma}_1}\right), \quad w_2(Y, \theta_0, \Delta) = \tilde{\psi}_k\left(\frac{Y - \Delta + \theta_0}{\hat{\sigma}_2}\right),$$

where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are scale estimators.

# Two-sample EL: computational issue

- To find $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\theta}$ we need to solve the equation system:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{w_1(X_i,\theta,\Delta)}{1+\lambda_1 w_1(X_i,\theta,\Delta)}=0,$$

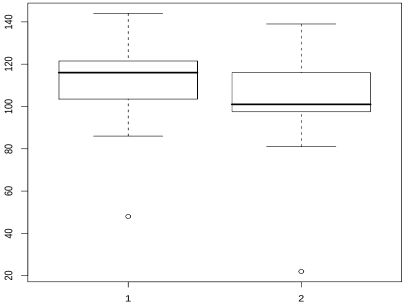
$$\frac{1}{m}\sum_{j=1}^{m}\frac{w_2(Y_j,\theta,\Delta)}{1+\lambda_2 w_2(Y_j,\theta,\Delta)}=0,$$

$$\sum_{i=1}^{n}\frac{\lambda_1\alpha_1(X_i,\theta,\Delta)}{1+\lambda_1 w_1(X_i,\theta,\Delta)}+\sum_{j=1}^{m}\frac{\lambda_2\alpha_2(Y_j,\theta,\Delta)}{1+\lambda_2 w_2(Y_j,\theta,\Delta)}=0.$$

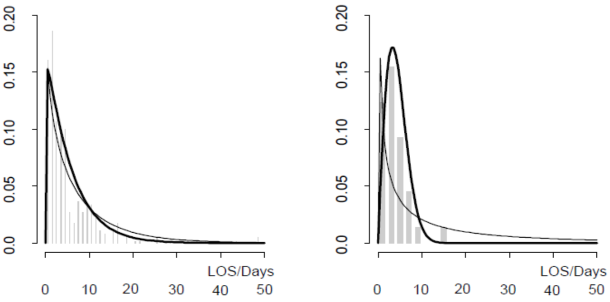
- $\alpha_1$ and $\alpha_2$ are derivatives of $w_1$ and $w_2$ with respect to $\theta$.
- Difficult to find "right solution" for this system.

# Data example: IQ dataset (Heritier *et al.*, 2009)

- Group 1: 15 children have mothers suffering from postpatal depression;
- Group 2: 79 children have healthy mothers.
- $H_0$: no difference between the mean IQs across the two groups.



- Group 1: 315 length of stay in days (LOS) of patients hospitalized in Belgium during 1988 for certain “disorders of the nervous system”;
- Group 2: 32 LOS of patients hospitalized in Switzerland.
- $H_0$: no difference between LOS in Belgium and Switzerland



# Data example: Stigler (1977)

- Newcomb's Third Series of measurements ($n_1 = 20, n_2 = 20, n_3 = 26$) of the passage time of light, made July 24, 1882 to Sept. 5, 1882 with the corresponding true value 33.02;
- $H_0$: no difference between three measurements.

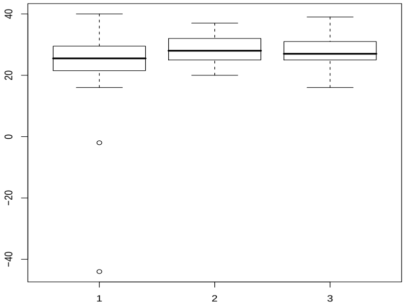


**Table 1**. Bootstraped variances $V$ with different scale estimators $\hat{\sigma} = 1, sd$ and $MAD = Med\{|\mathbf{X} - Med(\mathbf{X})|\}$ for nonsmooth Huber estimator, $k = 1.35$ based on $10,000$ bootstrap samples.

| | s=1 | | s=MAD | | s=sd | |
|---|---|---|---|---|---|---|
| Data | sample 1 | sample 2 | sample 1 | sample 2 | sample 1 | sample 2 |
| IQ dataset | 167.42 | 318.89 | 189.19 | 357.86 | 173.97 | 465.76 |
| data9-10 | 59.06 | 37.86 | 58.04 | 32.23 | 160.48 | 33.08 |
| data10-11 | 37.86 | 13.14 | 32.23 | 17.51 | 33.08 | 17.04 |
| data9-11 | 59.06 | 13.14 | 58.04 | 17.51 | 160.48 | 17.04 |
| Marazzi1 | 15.46 | 9.85 | 42.41 | 14.73 | 74.25 | 2690.92 |
| Marazzi1* | 16.67 | 6..95 | 42.41 | 8.28 | 74.06 | 106.48 |
| Marazzi2 | 5.51 | 11.22 | 7.42 | 13.03 | 47.97 | 14.73 |
| Marazzi3 | 6.69 | 38.05 | 7.87 | 59.37 | 36.89 | 50.34 |

# $p$-values for different datasets

**Table 2**. $p$-values for testing $H_0$ with different scale estimators $\hat{\sigma} = 1, sd$ and $MAD$ for the smooth Huber estimator, $k = 1.35$ based on 10,000 simulations, $V$ estimated by nonparametric bootstrap using $MAD$.

| | t-test | | EL Mean | EL Smoothed Huber | | |
|---|---|---|---|---|---|---|
| | var.eq=F | var.eq=T | | s=1 | s=MAD | s=sd |
| IQ dataset | 0.12 | 0.02 | 0.05 | 0.05 | **0.05** | 0.04 |
| data9-10 | 0.11 | 0.11 | 0.02 | 0.22 | **0.09** | 0.04 |
| data10-11 | 0.63 | 0.63 | 0.62 | 0.58 | **0.56** | 0.62 |
| data9-11 | 0.15 | 0.10 | 0.03 | 0.32 | **0.14** | 0.06 |
| Marazzi1 | 0.19 | 0.00 | 0.02 | 0.68 | **0.97** | 0.03 |
| Marazzi1* | 0.95 | 0.93 | 0.95 | 1.00 | **0.52** | 0.51 |
| Marazzi2 | 0.08 | 0.08 | 0.02 | 0.05 | **0.09** | 0.02 |
| Marazzi3 | 0.89 | 0.89 | 0.89 | 0.81 | **0.09** | 0.32 |

**Table 3**. 95% empirical coverage accuracy of $X \sim N(0,1)$ and contaminated model $Y \sim (1-\epsilon)N(0,1) + \epsilon N(\mu,1)$ with $\epsilon = 0.06$, $k = 1.35$ and $\mu = 1, 5, 10$ based on $10,000$ simulations.

| $n = m = 50$ | t.int | el.int | el.hub1 | el.hub2 | sm.boot | V1 | V2 |
|---|---|---|---|---|---|---|---|
| $\mu = 1$ | 0.95 | 0.94 | 0.94 | 0.94 | 0.91 | 0.94 | 1.07 |
| $\mu = 5$ | 0.65 | 0.64 | 0.66 | 0.65 | 0.64 | 0.97 | 1.11 |
| $\mu = 10$ | 0.14 | 0.13 | 0.15 | 0.15 | 0.20 | 0.94 | 1.07 |
| $n = m = 100$ | | | | | | | |
| $\mu = 1$ | 0.93 | 0.93 | 0.93 | 0.93 | 0.88 | 0.94 | 1.05 |
| $\mu = 5$ | 0.42 | 0.42 | 0.45 | 0.45 | 0.43 | 0.94 | 1.07 |
| $\mu = 10$ | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.87 | 1.07 |

# Contaminated gamma models (Marazzi, 2002): empirical coverage accuracy

**Table 4**. 95% empirical coverage accuracy, $X \sim Gamma(\alpha = 1, \sigma)$ and $Y \sim (1-\epsilon)Gamma(\alpha = 5, \sigma = 1) + \epsilon Uniform[0, 50]$, $\epsilon = 0.06$, $k = 1.35$ and $\sigma = 5, 6, 7, 8, 9$ based on $10,000$ simulations.

| | t.int | | EL.int | | sm.boot | | EL.hub1 | | EL.hub2 | | V |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma$ | acc | len | acc | len | acc | len | acc | len | acc | len | |
| 5 | 0.62 | 3.05 | 0.68 | 3.05 | 0.36 | 2.98 | 0.42 | 2.66 | 0.31 | 2.55 | 23.51 |
| 6 | 0.69 | 3.56 | 0.75 | 3.55 | 0.38 | 3.47 | 0.48 | 3.10 | 0.35 | 2.95 | 33.64 |
| 7 | 0.74 | 4.09 | 0.78 | 4.08 | 0.45 | 3.99 | 0.55 | 3.56 | 0.41 | 3.39 | 45.53 |
| 8 | 0.78 | 4.62 | 0.81 | 4.60 | 0.48 | 4.50 | 0.61 | 4.03 | 0.49 | 3.83 | 61.10 |
| 9 | 0.81 | 5.19 | 0.84 | 5.16 | 0.50 | 5.02 | 0.67 | 4.55 | 0.55 | 4.32 | 74.27 |

# References


1. J. Valeinis, E. Cers, J. Cielens (2010). Two-sample problems in statistical data modelling. *Mathematical modelling and analysis*, **15**(1), pages 137-151.
2. J. Valeinis and E. Cers (2011). Extending the two-sample empirical likelihood (preprint, available on homepage)
3. F. Hampel, C. Henning and E. Ronchetti (2011). A smoothing principle for the Huber and other location $M$-estimators. *Computational Statistics and Data Analysis*, **55**(1), pages 324-337.
4. A. Marazzi (2002). Bootstrap tests for robust means of asymmetric distributions with unequal shapes. *Computational Statistics and Data Analysis*, **39**(4), pages 503-528.
5. S. M. Stigler (1977). Do robust estimators work with real data? *The Annals of Statistics*, **5**(6), pages 1055 - 1098.