Optimal Robust Estimates using the Kullback-Leibler Divergence and the Hellinger Distance

Alfio Marazzi and Víctor J. Yohai
Universidad de Buenos Aires and CONICET

1 The Hampel optimality problem

In this talk we deal with the problem of finding optimal robust estimates for a multidimensional parameter. We give a new criterion based on a distance vetween distributions to define optimal estimates maximizing the efficiency under the model subject to a bound in the gross error sensitivity

Let $f(\mathbf{x}, \boldsymbol{\theta})$ be a family of densities, where $\mathbf{x} \in R^p$ and $\boldsymbol{\theta} \in \Theta \subset R^q$ and let $F_{\boldsymbol{\theta}}$ be the corresponding distribution functions.

Let $\mathcal{F}_p$ be the space of distribution functions on R^p. An estimating functional of $\boldsymbol{\theta}$ is a function $\mathbf{T} : \mathcal{F}_p \rightarrow \Theta$.

Then for each $F \in \mathcal{F}_p$, $\mathbf{T}(F) \in \Theta$

Let $\mathbf{T}$ be an estimating functional and let $\mathbf{x_1}, ..., \mathbf{x}_n$ be a random sample of a distribution F Then the estimate associated to the functional $\mathbf{T}$ is

$$\widehat{\boldsymbol{\theta}}_{\mathbf{T},n} = \mathbf{T}(F_n),$$

where

$$F_n(\mathbf{x}) = \frac{\#\{i : \mathbf{1} \leq i \leq n,\ \mathbf{x}_i \leq \mathbf{x}\}}{n}$$

is the empirical distribution. An example of an estimating functional with $p = q = 1$ is

$$T = E_F(x)$$

In this case

$$\widehat{\theta}_{T,n} = T(F_n) = E_{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Suppose that the functional $\mathbf{T}$ is weakly continuous. i.e.

$$F_n \to^D F, \text{ then } T(F_n) \to T(F)$$

Then, since

$$P(F_n \to^D F) = 1$$

we have

$$\widehat{\theta}_{\mathbf{T},n} = T(F_n) \to \mathbf{T}(F) \text{ a.s.}$$

$\mathbf{T}$ is Fisher consistent if

$$\mathbf{T}(F_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$$

Suppose now that $\mathbf{x}_1, ..., \mathbf{x}_n$ is a random sample of F_θ.

Then if the functional $\mathbf{T}$ is continuous and Fisher consistent we have

$$\widehat{\boldsymbol{\theta}}_{T,n} \rightarrow T(F_\theta) = \boldsymbol{\theta} \text{ a.s.}$$

Consider the contaminated distribution

$$F_{\boldsymbol{\theta},\mathbf{x},\varepsilon} = (1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon\delta_{\mathbf{x}}$$

where $\delta_{\mathbf{x}}$ is the point mass distribution at $\mathbf{x}$ that assigns probability one to $\mathbf{x}$

Then, **Hampel's influence function** is defined by

$$\mathbf{IF}(\mathbf{T},\mathbf{x},\boldsymbol{\theta}) = \left.\frac{\partial \mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon})}{\partial\varepsilon}\right|_{\varepsilon=0},$$

and then

$$\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}) - \mathbf{T}(F_{\boldsymbol{\theta}}) \simeq \mathbf{IF}(T,\mathbf{x},\boldsymbol{\theta})\varepsilon$$

Given a function $\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta})$: $R^p \times R^q \rightarrow R^q$, **an M-estimating functional** $\mathbf{T}_{\boldsymbol{\psi}}$ is defined by

$$E_F(\boldsymbol{\psi}(\mathbf{x}, \mathbf{T}_{\boldsymbol{\psi}}(F)) = 0$$

This estimating functional is Fisher consistent if

$$E_{\boldsymbol{\theta}}(\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta})) = 0$$

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be a random sample of $F_{\boldsymbol{\theta}}$ and let

$$\widehat{\boldsymbol{\theta}}_{\mathbf{T},n} = \mathbf{T}_{\boldsymbol{\psi}}(F_n),$$

where F_n is the empirical distribution. Then $\widehat{\boldsymbol{\theta}}_{\mathbf{T}_{\boldsymbol{\psi}},n}$ is the solution in $\boldsymbol{\theta}$ of

$$E_{F_n}(\boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta})) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{x}, \boldsymbol{\theta}) = 0$$

Hampel proved that under very general regularity conditions we have

$$n^{1/2}(\widehat{\boldsymbol{\theta}}_{T,n} - \boldsymbol{\theta}) \rightarrow_{\mathcal{D}} N(0, V(\mathbf{T}, \boldsymbol{\theta})),$$

where $\rightarrow_{\mathcal{D}}$ denotes convergence in distribution and

$$V(T, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\mathbf{IF}(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta}) IF(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta})'),$$

In the case that θ is a one-dimensional parameter ($q = 1$) we can define the gross error sensitivity (GES) by

$$\gamma(T, \theta) = \sup_{\mathbf{x}} |IF(T, \mathbf{x}, \theta)|.$$

$\gamma(T, \theta)$ is a measure of the robustness of T under infinitesimal contamination.

It is known that under general conditions the maximum likelihood estimate (MLE) is the one with smallest asymptotic variance. However in most the cases when T_{ML} is the functional corresponding to the ML estimate

$$\gamma(T_{ML}, \theta) = \infty$$

Note that the ML estimate of an M-estimate corresponding to the function

$$\psi_0(\mathbf{x}, \theta) = \frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta}$$

Hampel gave a criterion to find M-estimates which give an optimal trade off between efficiency under the model and infinitesimal robustness. He proposed to find the function $\psi^*(x, \theta)$ such that

$$V(T_{\psi^*}, \theta) = \text{minimum}$$

among all the M-functionals satisfying the Fisher consistency condition

$$E_{F_\theta}(\psi^*(\mathbf{x}, \theta)) = 0$$

and

$$\gamma(T_{\psi^*}, \theta) \leq C(\theta)$$

where $C(\theta)$ is a known function.

Define $h_c(t) : R^q \rightarrow R^q$ as the family of Huber functions given by

$$h_c(t) = \begin{cases} -c & \text{if} \quad t < c \\ t & \text{if} \quad |t| \leq c \\ c & \text{if} \quad t > c \end{cases} .$$

Hampel obtained the optimal ψ for this problem which has the following form

$$\psi^*(\mathbf{x}, \theta) = h_{m(\theta)}(\psi_0(\mathbf{x}, \theta) - d(\theta))$$

where $d(\theta)$ and $m(\theta)$ are chosen so that ψ^* satisfies the Fisher consistency condition

$$E_\theta(\psi^*(\mathbf{x}, \theta)) = 0$$

and

$$\gamma(T_{\psi^*}, \theta) = C(\theta)$$

Consider now the case where the dimension of $\boldsymbol{\theta}$ is $q > 1$ Suppose that we want to define optimal M-estimates similar to those proposed by Hampel for $q = 1$

Note that in this case since $\mathbf{T}$ is a vector

$$\mathbf{IF}(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta}) = \left. \frac{\partial \mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon})}{\partial \varepsilon} \right|_{\varepsilon=0},$$

is a vector too

Consider now the case where the dimension of $\boldsymbol{\theta}$ is $q > 1$ Suppose that we want to define optimal M-estimates similar to those proposed by Hampel for $q = 1$

Note that in this case since $\mathbf{T}$ is a vector

$$\mathbf{IF}(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta}) = \left. \frac{\partial \mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon})}{\partial \varepsilon} \right|_{\varepsilon=0},$$

The simplest way to generalize the concept of GES is to define the unstandardized GES as

$$\gamma_u(\mathbf{T}, \boldsymbol{\theta}) = \sup_{\mathbf{x}} ||IF(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta})||.$$

Stahel (1981) obtained optimal M-estimates by finding $\boldsymbol{\psi}^*(\mathbf{x}, \boldsymbol{\theta})$ such that

$$\text{trace}(V(\mathbf{T}_{\boldsymbol{\psi}^*}, \boldsymbol{\theta})) = \text{minimum}$$

subject to

$$E_F(\boldsymbol{\psi}^*(\mathbf{x}, \boldsymbol{\theta})) = 0$$

and

$$\gamma_u(\mathbf{T}_{\boldsymbol{\psi}^*}, \boldsymbol{\theta}) \leq C(\boldsymbol{\theta})$$

Since the definition of γ_u in not invariant under model reparametrizations, the corresponding optimal estimate is not equivariant either.

This means that if we reparametrize the family of distributions using the parameter

$$\boldsymbol{\lambda} = \mathbf{g}(\boldsymbol{\theta})$$

and $\mathbf{T}_1^*$ and $\mathbf{T}_2^*$ are the optimal estimates corresponding to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ respectively, it does not hold that

$$\mathbf{T}_2^*(F) = \mathbf{g}(\mathbf{T}_1^*(F)).$$

Stahel (1981) proposed two invariant definitions of GES. The first is the self-standardized GES, which is standardized using its own asymptotic covariance matrix $V(\mathbf{T}, \boldsymbol{\theta})$ and is given by

$$\gamma_s(\mathbf{T}, \boldsymbol{\theta}) = \sup_{\mathbf{x}} (\mathbf{IF}(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta})' V^{-1}(\mathbf{T}, \boldsymbol{\theta}) \mathbf{IF}(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta}))^{1/2}.$$

The second invariant definition of GES uses the information matrix $J(\boldsymbol{\theta})$ and its given by

$$\gamma_i(\mathbf{T}, \boldsymbol{\theta}) = \sup_{\mathbf{x}} (\mathbf{IF}(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta})' J(\boldsymbol{\theta}) \mathbf{IF}(\mathbf{T}, \mathbf{x}, \boldsymbol{\theta}))^{1/2}. \tag{1}$$

Stahel (1981) derived the optimal M-estimates using the three definitions of GES: γ_u, γ_s and γ_i

However, these standardizations are defined ad hoc and it is not clear what they mean .

However, these standardizations are defined ad hoc and it is not clear what they mean

To overcome this problem, in this paper we propose optimal M-estimates which use measures of robustness and efficiency based on the Kullback-Leibler divergence and Hellinger distance and we will show that they coincide with the optimal ones using $\gamma_i(\mathbf{T}, \boldsymbol{\theta})$.

2 Gross Error Sensitivity Based on a distance

Consider two distributions F^* and F on R^P with densities f^* and f respectively. Then the Kullback Leibler divergence from F^* to F

$$\begin{aligned} d_{kl}(F^*, F) &= \int_{R^p} \log\left(\frac{f(\mathbf{z})}{f^*(\mathbf{z})}\right) f(\mathbf{z}) dz. \\ &= E_F\left(\log\left(\frac{f(\mathbf{z})}{f^*(\mathbf{z})}\right)\right) \end{aligned}$$

We know that.

$$d_{kl}(F^*, F) \geq 0$$

and

$$d_{kl}(F^*, F) = 0 \iff F = F^*$$

The Hellinger distance is defined by

$$d_H(F^*, F) = 2\int_{R^p}(f^{*1/2}(\mathbf{z}) - f^{1/2}(\mathbf{z}))^2\mathbf{dz}.$$

In what follows d will denote indistinctly d_{kl} or d_H

Assume a parametric model with density $f(\mathbf{x}, \boldsymbol{\theta})$ as in Section 1 and let $F_{\boldsymbol{\theta}}(\mathbf{x})$ be the corresponding distribution function.

Given $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$ we define

$$D(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = d(F_{\boldsymbol{\theta}^*}, F_{\boldsymbol{\theta}})$$

Assume a parametric model with density $f(\mathbf{x}, \boldsymbol{\theta})$ and let $F_{\boldsymbol{\theta}}(\mathbf{x})$ be the corresponding distribution function.

Given $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$ we define

$$D(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = d(F_{\boldsymbol{\theta}^*}, F_{\boldsymbol{\theta}})$$

This measure is invariant with respect to parameter transformations. If we define $\boldsymbol{\lambda} = g(\boldsymbol{\theta})$ and $\boldsymbol{\lambda}^* = g(\boldsymbol{\theta}^*)$, and we put $\overline{F}_{\lambda} = F_{g^{-1}(\theta)}$ we have

$$d(\overline{F}_{\boldsymbol{\lambda}^*}, \overline{F}_{\boldsymbol{\lambda}}) = d(F_{\boldsymbol{\theta}^*}, F_{\boldsymbol{\theta}})$$

Let $\mathbf{T}$ be a Fisher-consistent, estimating functional of $\boldsymbol{\theta}$

$$\mathbf{T}(F_{\boldsymbol{\theta}}) = \boldsymbol{\theta},$$

then

$$D(\mathbf{T}(F_{\boldsymbol{\theta}}), \boldsymbol{\theta}) = d(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}}) = 0$$

Consider now

$$F_{\boldsymbol{\theta},\mathbf{x},\varepsilon} = (1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon\delta_{\mathbf{x}}$$

Since in general $\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}) \neq \boldsymbol{\theta}$

$$D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}), \boldsymbol{\theta}) > 0$$

and we can define the bias function based on a distance d by

$$B_d(\mathbf{T}, \boldsymbol{\theta}, \varepsilon) = \sup_{\mathbf{x}} D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}), \boldsymbol{\theta})$$

as a measure of the robustness of the estimating functional $\mathbf{T}$.

Since $D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}), \boldsymbol{\theta})$ is in general complicated we will try to obtain an approximation using a Taylor expansion

The we will calculate the first and second derivatives of $D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}), \boldsymbol{\theta})$ at $\varepsilon = 0$.

Lemma. Let d be indistinctly d_{kl} or d_H and

$$F_{\boldsymbol{\theta},\mathbf{x},\varepsilon} = (1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon\delta_{\mathbf{x}}$$

Then for any $\mathbf{x} \in R^p$ we have

$$\left.\frac{\partial D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta})}{\partial\varepsilon}\right|_{\varepsilon=0} = 0.$$

Since the first order derivative of $D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta})$ at ε is 0, in order to approximate $D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta})$ we need to compute the second order derivative

Lemma. Let d be indistinctly d_{kl} or d_H. Then

$$\left.\frac{\partial^2 D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta})}{\partial \varepsilon^2}\right|_{\varepsilon=0} = \mathbf{IF}(\mathbf{T},\mathbf{x},\boldsymbol{\theta})' J(\boldsymbol{\theta}) \mathbf{IF}(\mathbf{T},\mathbf{x},\boldsymbol{\theta}), \tag{2}$$

where $J(\boldsymbol{\theta})$ is the information matrix

$$J(\boldsymbol{\theta}) = E\left(\left(\frac{\partial \log(f(\mathbf{x},\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}\right)^2\right)$$

Then, according to the above Lemmas, for small ε we have

$$\begin{aligned} D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta}) &\cong \frac{1}{2}\left.\frac{\partial^2 D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta})}{\partial\varepsilon^2}\right|_{\varepsilon=0}\varepsilon^2 \\ &= \frac{1}{2}IF(\mathbf{T},\mathbf{x},\boldsymbol{\theta})'J(\boldsymbol{\theta})IF(\mathbf{T},\mathbf{x},\boldsymbol{\theta})\varepsilon^2. \end{aligned}$$

Then, according to the above Lemmas, for small ε we have

$$\begin{aligned} D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta}) &\cong \frac{1}{2}\frac{\partial^2 D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta})}{\partial\varepsilon^2}\bigg|_{\varepsilon=0}\varepsilon^2 \\ &= \frac{1}{2}IF(\mathbf{T},\mathbf{x},\boldsymbol{\theta})'J(\boldsymbol{\theta})IF(\mathbf{T},\mathbf{x},\boldsymbol{\theta})\varepsilon^2. \end{aligned}$$

Then we define the GES based on the distance d of the functional $\mathbf{T}$ as

$$\begin{aligned} \gamma_d(\mathbf{T},\boldsymbol{\theta}) &= \sup_{\mathbf{x}}\left(\frac{\partial^2 D(\mathbf{T}(F_{\boldsymbol{\theta},\mathbf{x},\varepsilon}),\boldsymbol{\theta}))}{\partial\varepsilon^2}\bigg|_{\varepsilon=0}\right)^{1/2} && (3) \\ &= \sup_{\mathbf{x}}\Big(IF(\mathbf{T},\mathbf{x},\boldsymbol{\theta})'J(\boldsymbol{\theta})IF(\mathbf{T},\mathbf{x},\boldsymbol{\theta})\Big)^{1/2} && (4) \\ &= \gamma_i(\mathbf{T},\boldsymbol{\theta}). && (5) \end{aligned}$$

and we have

$$B_d(\mathbf{T}, \boldsymbol{\theta}, \varepsilon) \approx \frac{1}{2}\gamma_i^2(\mathbf{T}, \boldsymbol{\theta})\varepsilon^2$$

3 Asymptotic efficiency Based on d

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be a random sample of $F_{\boldsymbol{\theta}}$ and let $\mathbf{T}$ be an estimating functional of $\boldsymbol{\theta} \in R^q$.

Suppose that $\widehat{\boldsymbol{\theta}}_{\mathbf{T},n} = \mathbf{T}(F_n)$ satisfies

$$n^{1/2}(\widehat{\boldsymbol{\theta}}_{\mathbf{T},n} - \boldsymbol{\theta}) \rightarrow_{\mathcal{D}} N(0, V(\mathbf{T}, \boldsymbol{\theta})), \tag{6}$$

One invariant way to measure the performance of $\widehat{\boldsymbol{\theta}}_n^{\mathbf{T}}$ when the true parameter is $\boldsymbol{\theta}$ is to use

$$D(\widehat{\boldsymbol{\theta}}_{\mathbf{T},n}, \boldsymbol{\theta}) = d(F_{\widehat{\boldsymbol{\theta}}_{\mathbf{T},n}}, F_{\boldsymbol{\theta}}).$$

The following result gives the asymptotic distribution of this measure .

Lemma. Let d be indistinctly d_{kl} or d_H. Assume that $n^{1/2}(\widehat{\boldsymbol{\theta}}_{\mathbf{T},n} - \boldsymbol{\theta}) \to_{\mathcal{D}} N(0, V_{\mathbf{T}}(\boldsymbol{\theta}))$. Then

$$nD(\widehat{\boldsymbol{\theta}}_{\mathbf{T},n}, \boldsymbol{\theta}) \to_{\mathcal{D}} \frac{1}{2} \sum_{i=1}^{q} \lambda_i(\boldsymbol{\theta}) v_i^2, \tag{7}$$

where

$$v_1, ...v_q \text{ are i.i.d. r.v. with distribution } \mathsf{N}(0,1)$$

and

$$\lambda_i(\boldsymbol{\theta}), 1 \leq i \leq q,$$

are the eigenvalues of

$$J(\boldsymbol{\theta}) V(\mathbf{T}, \boldsymbol{\theta}).$$

Note that if $\mathbf{T}_0$ is the maximum likelihood estimator, $V(\mathbf{T}_0, \boldsymbol{\theta}) = J^{-1}(\boldsymbol{\theta})$, and then $\lambda_1(\boldsymbol{\theta}) = ... = \lambda_q(\boldsymbol{\theta}) = 1$

Then we define the following measure of asymptotic efficiency of $\widehat{\boldsymbol{\theta}}_n^{\mathbf{T}}$ based on the distance d

$$
\begin{aligned}
\mathsf{AE}_d(\mathbf{T}, \boldsymbol{\theta}) &= \frac{E\left(1/2 \sum_{i=1}^{q} v_i^2\right)}{E\left(1/2 \sum_{i=1}^{q} \lambda_i(\boldsymbol{\theta}) v_i^2\right)} \\
&= \frac{q}{\sum_{i=1}^{q} \lambda_i(\boldsymbol{\theta})} \\
&= \frac{q}{\mathsf{trace}(J(\boldsymbol{\theta}) V_{\mathbf{T}}(\boldsymbol{\theta}))}. \qquad (8)
\end{aligned}
$$

Observe that since by Rao-Cramer

$$V_{\mathbf{T}}(\boldsymbol{\theta}) - J^{-1}(\boldsymbol{\theta}) \text{ is positive semidefinite}$$

then

$$\lambda_i(\boldsymbol{\theta}) \geq 1 \text{ for } 1 \leq i \leq q,$$

and therefore

$$\mathrm{AE}_d(\mathbf{T}, \boldsymbol{\theta}) \leq 1.$$

Observe that since by Rao-Cramer

$$V_{\mathbf{T}}(\boldsymbol{\theta}) - J^{-1}(\boldsymbol{\theta}) \text{ is positive semidefinite}$$

then

$$\lambda_i(\boldsymbol{\theta}) \geq 1 \text{ for } 1 \leq i \leq q,$$

and therefore

$$\mathrm{AE}_d(\mathbf{T}, \boldsymbol{\theta}) \leq 1.$$

Besides we have that

$$\lambda_i(\boldsymbol{\theta}) = 1 \text{ for } 1 \leq i \leq q \text{ iff } V(\mathbf{T}, \boldsymbol{\theta}) = J^{-1}(\boldsymbol{\theta}),$$

and in this case $\mathrm{AE}_d(\mathbf{T}, \boldsymbol{\theta}) = 1$. This happens when $\mathbf{T}$ is the functional associated with the MLE.

4 Optimal M-estimates in the case of a multidimensional parameter

A natural way to define equivariant optimal robust estimates using the Hampel approach is as follows.

Find a function $\boldsymbol{\psi}^*(\mathbf{x}, \boldsymbol{\theta})$ such that

$$\mathrm{AE}_d(\mathbf{T}_{\boldsymbol{\psi}}, \boldsymbol{\theta}) = \text{maximum}$$

subject to that $\mathbf{T}_{\boldsymbol{\psi}^*}$ is Fisher-consistent i.e., it satisfies

$$E_{\boldsymbol{\theta}}(\boldsymbol{\psi}^*(\mathbf{x}, \boldsymbol{\theta})) = 0$$

and to

$$\gamma_d(\mathbf{T}_{\boldsymbol{\psi}}, \boldsymbol{\theta}) \leq C(\boldsymbol{\theta}).$$

where $C(\boldsymbol{\theta})$ is a fixed function.

According to what have seen, this problem is the same as finding the Fisher-consistent M-estimate such that

$$\text{trace}(J(\boldsymbol{\theta})V_{\mathbf{T}_{\boldsymbol{\psi}}}(\boldsymbol{\theta})) = \text{ minimum}$$

subject to

$$\gamma_i(\boldsymbol{\theta}) \leq C(\boldsymbol{\theta}).$$

This is precisely the problem of finding the optimal M-estimate using the standardized GES $\gamma_i(\boldsymbol{\theta})$ studied by Stahel.

Define $\mathbf{H}_c(\mathbf{t}) : R^q \rightarrow R^q$ as the multivariate Huber function given by

$$\mathbf{H}_c(\mathbf{t}) = \begin{cases} \mathbf{t} & \text{if } ||\mathbf{t}|| \leq c \\ \frac{c}{||\mathbf{f}||}\mathbf{t} & \text{if } ||\mathbf{t}|| > c \end{cases} .$$

The optimal $\boldsymbol{\psi}$ for this problem which has the following form

$$\boldsymbol{\psi}^*(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{H}_{C(\boldsymbol{\theta})}(A(\boldsymbol{\theta})(\boldsymbol{\psi}_0(\mathbf{x}, \boldsymbol{\theta}) - d(\boldsymbol{\theta}))),$$

where $d(\boldsymbol{\theta}) : \Theta \rightarrow R^q$, $A(\boldsymbol{\theta}) : \Theta \rightarrow R^{q \times q}$ is a non-singular matrix,

Define $\mathbf{H}_c(t) : R^q \longrightarrow R^q$ as the multivariate Huber function given by

$$\mathbf{H}_c(\mathbf{t}) = \begin{cases} \mathbf{t} & \text{if } ||\mathbf{t}|| \leq c \\ \frac{c}{||\mathbf{f}||}\mathbf{t} & \text{if } ||\mathbf{t}|| > c \end{cases} .$$

The optimal $\boldsymbol{\psi}$ for this problem which has the following form

$$\boldsymbol{\psi}^*(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{H}_{C(\boldsymbol{\theta})}(A(\boldsymbol{\theta})(\boldsymbol{\psi}_0(\mathbf{x}, \boldsymbol{\theta}) - d(\boldsymbol{\theta}))),$$

where $d(\boldsymbol{\theta}) : \Theta \longrightarrow R^q$, $A(\boldsymbol{\theta}) : \Theta \longrightarrow R^{q \times q}$ is a non-singular matrix, The functions $d(\boldsymbol{\theta})$ and $A(\boldsymbol{\theta})$ must be chosen so that $\boldsymbol{\psi}^*$ satisfies

$$E_{\boldsymbol{\theta}}(\boldsymbol{\psi}^*(\mathbf{x}, \boldsymbol{\theta})) = 0$$

and

$$\gamma_d(\mathbf{T}_{\boldsymbol{\psi}^*}, \boldsymbol{\theta}) = \gamma_i(\mathbf{T}_{\boldsymbol{\psi}^*}, \boldsymbol{\theta}) = C(\boldsymbol{\theta})$$

One practical problem is the choice of the bound $C(\boldsymbol{\theta})$.

One possible solution to this problem, is to choose for each $\boldsymbol{\theta}$ the constant $C(\boldsymbol{\theta})$ so that

$$\mathsf{AE}_d(\mathbf{T}_{\boldsymbol{\psi}^*}, \boldsymbol{\theta}) = 1 - \alpha$$

For example we can take $\alpha = 0.95$

Open Problem To find the optimal estimates using other distances. For example

Total variation

Kolmogorov

Prohorov

5 Example

We have dataset with the number of days at the hospital of 32 patients classified as "disorders of the nervous system". We fit a negative binomial model using the ML estimate and the optimal estimates with 95% and 80% of efficiency.

Table 1: Length of stay of 32 hospital patients

LOS	1	2	3	4	5	6	7	8	9	16	115	198	374
frequency	2	6	5	5	4	2	2	1	1	1	1	1	1

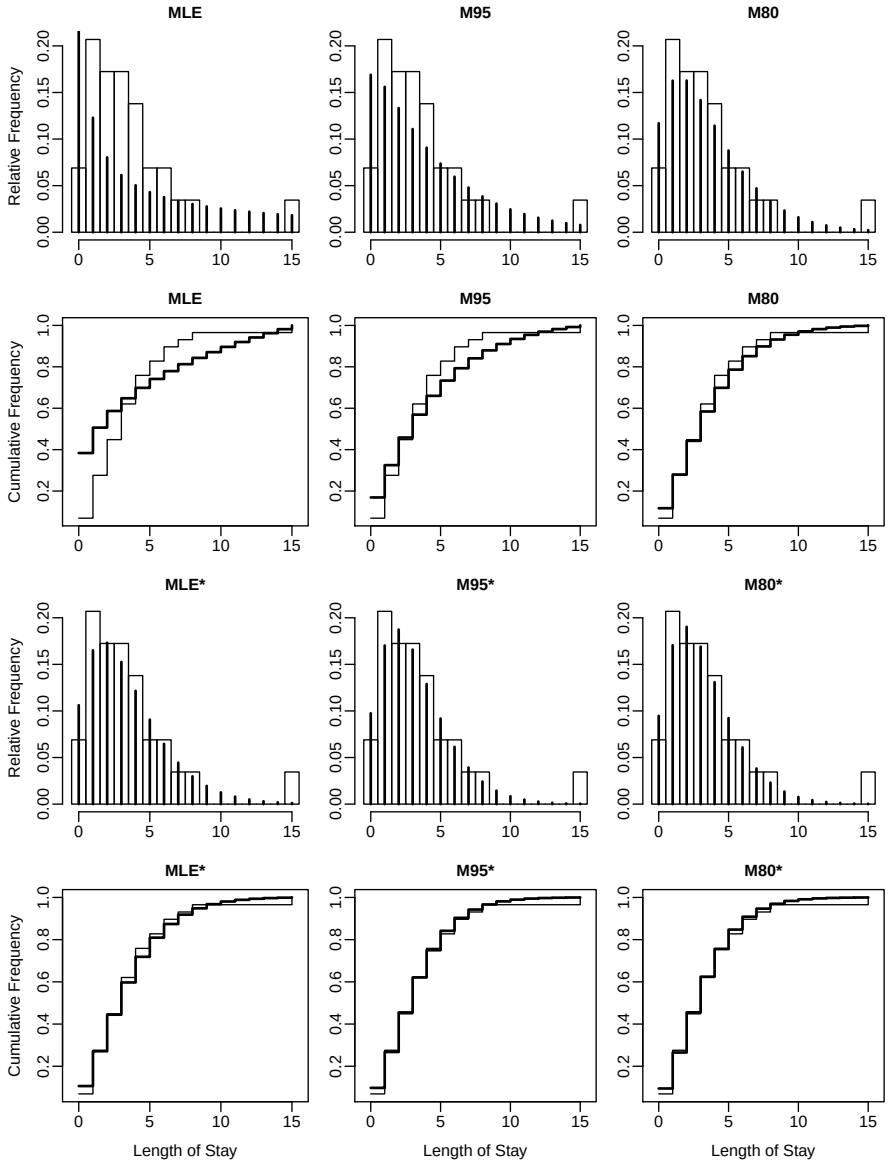


Fig. 1 Frequency and cdf plots; bold lines indicate estimated frequencies.

affected by the outliers. This feature of the optimal M-estimates is due to the monotonicity of Huber's function $h_c(t)$ which does not clearly reject the outliers. M80 provides the best fit and an expected LOS of 4.58 days.

An alternative model to describe the LOS distribution is the zero truncated negative binomial model (Hilbe, 2008)

$$g_{\alpha,\mu}(x) = \frac{1}{1 - p_0(\alpha, \mu)} f_{\alpha,\mu}(x), \quad x = 1, 2, ... \tag{14}$$

6 Estimates with minimum GES.

Now we come back to the case of a one-dimensional parameter

Consider a location model where we have a family of densities

$$f(x-\theta)$$

where f is symmetric and unimodal. Huber (1964) showed that the

median

is the estimate with

smallest maximum bias

among all the location equivariant estimates

This implies that the M-estimate based on the sign function is the one

$$\gamma(T_\psi, \theta) = \text{minimum}$$

Consider now an arbitrary family of densities

$$f(x, \theta),\ \theta \in \Theta \subset R$$

and suppose that we want to find the M-estimate with

$$\gamma(T_\psi, \theta) = \text{minimum}$$

. In Maronna, Martin and Yohai (2006) the following result was proved

Theorem. Consider the score function

$$\psi_0(\mathbf{x}, \theta) = \frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta}$$

and suppose that

$$C(\theta, \theta^*) = \mathrm{med}_\theta(\psi_0(\mathbf{x}, \theta^*))$$

is strictly monotone with respect to θ^* and continuous

Then the M-estimating functional $T(F)$ with minimum $\gamma(T_\psi, \theta)$ is obtained by solving

$$\mathsf{med}_F(\psi_0(\mathbf{x}, \theta)) = \mathsf{med}_\theta(\psi_0(\mathbf{x}, \theta))$$

Then the M-estimating functional $T(F)$ with minimum $\gamma(T_\psi, \theta)$ is obtained by solving

$$\mathsf{med}_F(\psi_0(\mathbf{x}, \theta)) = \mathsf{med}_\theta(\psi_0(x, \theta))$$

Recall that the maximum likelihood estimating functional $T_0(F)$ is given by

$$\mathsf{E}_F(\psi_0(x, \theta)) = \mathsf{E}_\theta(\psi_0(x, \theta))) = 0$$

Note the similarity between the two estimates: the one with the minimum variance and the one with the minimum GES

Corollary. Suppose that $C(\theta, \theta^*)$ is strictly monotone on θ^* and continuous and that $\psi_0(x, \theta)$ is strictly monotone on x. Then the M-estimating functional $T(F)$ with minimum $\gamma(T_\psi, \theta)$ is obtained by solving

$$\text{med}_F(x) = \text{med}_\theta(x),$$

This extends the result for the location model

Families of distributions with support on the nonnegative integers $Z_{\geq 0}$

The assumption that $C(\theta, \theta^*) = \text{med}_\theta(\psi_0(\mathbf{x}, \theta^*))$ is strictly monotone on θ^* is not satisfied for the case of a discrete family of distributions

Families of distributions with support on the nonnegative integers $Z_{\geq 0}$

The assumption that $C(\theta, \theta^*) = \text{med}_\theta(\psi_0(\mathbf{x}, \theta^*))$ is strictly monotone on θ^* is not satisfied for the case of a discrete family of distributions

In fact, the usual definition of medianis

$$\text{med}(F) = \text{min}\{k,\ F(k) \geq 0.5\}$$

Since $\text{med}(F_\theta)$ take integer values, changing θ a little in general the median does not change.

Therefore the median can not identify θ.

This suggest to change the definition of median so that the parameter θ can be identified.

This suggest to change the definition of median so that the parameter θ can be identified.

Definition. Given a distribution F with support on the nonnegative integers $Z_{\geq 0}$ and probability function p, the smooth median of F (will be denoted by smed(F)) is the median of the continuous variable with density

$$f(x) = p(k), \text{ if } k - 0.5 < x \leq k + 0.5$$

This suggest to change the definition of median so that the parameter θ can be identified.

Definition. Given a distribution F with support on the nonnegative integers $Z_{\geq 0}$ and probability function p, the smooth median of F (will be denoted by smed(F)) is the median of the continuous variable with density

$$f(x) = p(k), \text{ if } k - 0.5 < x \leq k + 0.5$$

This means that we compute the median of a distribution spreading uniformly the mass assigned to the value k within the interval $(k - 0.5, k + 0.5]$

This smooth median turns out to be

$$\text{smed}(F) = \text{med}(F) - 0.5 + \frac{0.5 - F(\text{med}(F) - 1)}{p(\text{med}(F))},$$

Then

$$\text{med}(F) - 0.5 < \text{smed}(F) \leq \text{med}(F) + 0.5$$

Given a family F_θ of distributions with values in $Z_{\geq 0}$, we can define an estimating functional of θ by

$$\text{smed}(F) = \text{smed}(F_\theta)$$

We can state the following Theorem.

Given a family F_θ of distributions with values in $Z_{\geq 0}$, we can define an estimating functional of θ by

$$\text{smed}(F) = \text{smed}(F_\theta)$$

We can state the following Theorem.

Theorem Assume that $\psi_0(x, \theta)$ is continuous on θ and strictly monotone on x and θ. Then the estimating functional of θ with smallest GES is defined by $T(F) = \theta$ satisfying

$$\text{smed}(F) = \text{smed}(F_\theta). \tag{9}$$

Influence function

Let $T(F)$ be the functional defined by the value of θ satisfying

$$\text{smed}(F) = \text{smed}(F_\theta). \tag{10}$$

The influence curve of this estimate is very similar to the one of the median . Under very general conditions

(a) If $F_\theta(\text{med}(F_\theta)) > 0.5$

$$\text{IF}(T, \theta, x) = \begin{cases} -m_0(\theta) & \text{if} \quad x < \text{med}(F) \\ m_1(\theta) & \text{if} \quad x = \text{med}(F) \\ m_0(\theta) & \text{if} \quad x > \text{med}(F) \end{cases}$$

where

$$|m_1(\theta)| \leq m_0(\theta)$$

(a) If $F_\theta(\text{med}(F_\theta)) > 0.5$

$$\text{IF}(T, \theta, x) = \begin{cases} -m_0(\theta) & \text{if} \quad x < \text{med}(F) \\ m_1(\theta) & \text{if} \quad x = \text{med}(F) \\ m_0(\theta) & \text{if} \quad x > \text{med}(F) \end{cases}$$

where

$$|m_1(\theta)| \leq m_0(\theta)$$

(b)If $F_\theta(\text{med}(F_\theta)) = 0.5$

$$\text{IF}(T, \theta, x) = \begin{cases} -m_0^*(\theta) & \text{if} \quad x \leq \text{med}(F) \\ m_1^*(\theta) & \text{if} \quad x > \text{med}(F) \end{cases}$$

Asymptotic distribution

Let $\widehat{\theta}_n$ be defined by

$$\text{smed}(F_n) = \text{smed}(F_\theta)$$

Then under very general conditions

(a) If $F_\theta(\text{med}(F_\theta)) > 0.5$

$$n^{1/2}(\widehat{\theta}_n - \theta) \to^{\mathcal{D}} N\left(0, \sigma^2(\theta)\right)$$

(b) If $F_\theta(\text{med}(F_\theta)) = 0.5$,

$$n^{1/2}(\widehat{\theta}_n - \theta) \to^{\mathcal{D}} G_\theta$$

where G_θ has a density g_θ of the form

$$g_\theta(x) = \begin{cases} \phi_0(x/\sigma_1(\theta)) & \text{if } x < 0 \\ \phi_0(x/\sigma_2(\theta)) & \text{if } x > 0 \end{cases}$$

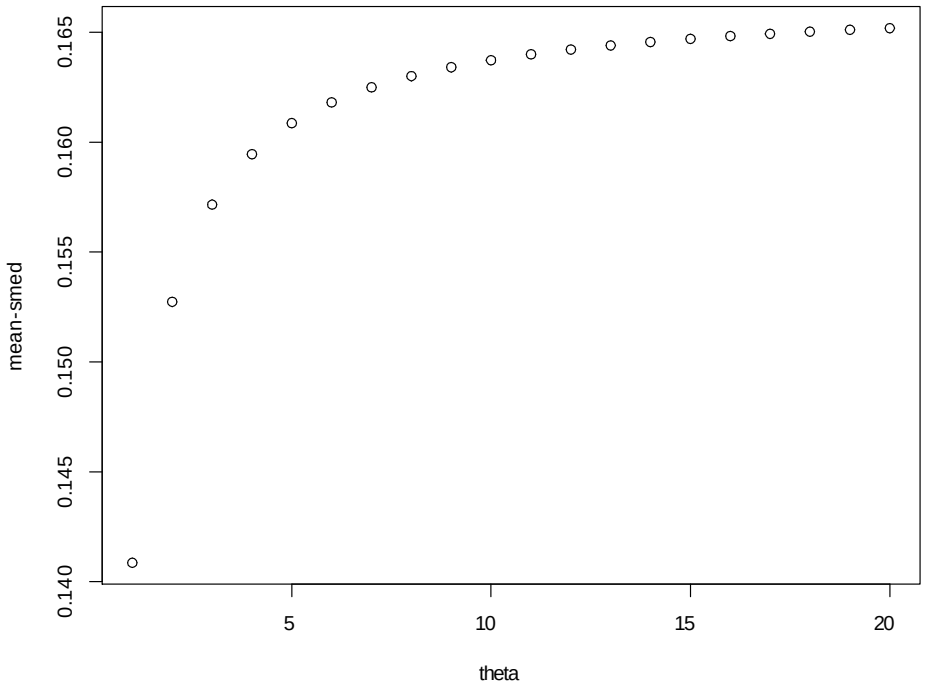
and where $\phi_0(x)$ is the density of a $N(0, 1)$ distribution.

7 Application to the Poisson distribution

Consider now the case of the Poisson family

$$p(x, \theta) = \frac{e^{-\theta}\theta^x}{x!}$$

In the next Figure we show $r(\theta) = \theta - \text{smed}(F_\theta)$.



It can be noted that for $\theta \geq 1$, $r(\theta)$ is quite constant.

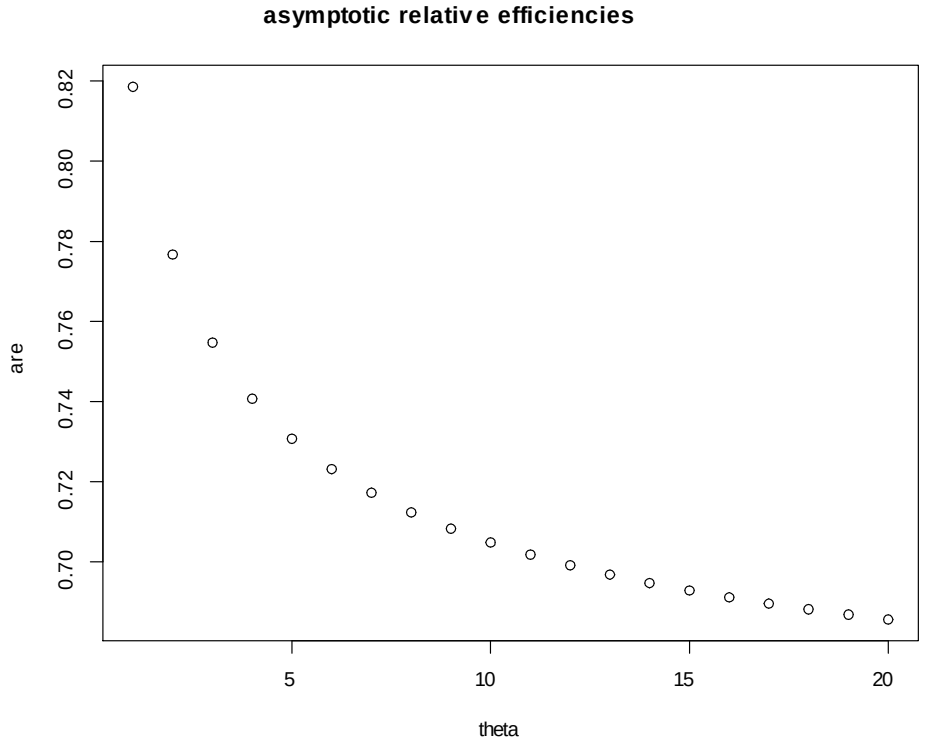
In fact, we have $0.14 \leq r(\theta) \leq 0.165$.

Let $\widehat{\theta}_n$ the estimate defined by

$$\text{smed}(F_n) = \text{smed}(F_\theta)$$

In the next Figure we display the asymptotic relative efficiency of $\widehat{\theta}_n$ with respect to the maximum likelihood estimate (the sample mean) $\overline{x}$. We denote this efficiency by

$$\text{are}(\theta) = \frac{\text{avar}(\overline{x})}{\text{avar}(\widehat{\theta})}$$



We note that $0.685 \leq \text{are}(\theta) \leq 0.82$.

For large values of λ it is very close to the efficiency of the median as estimate of the mean of a normal distribution 0.6366

8 References

Hampel, F. R. (1974) The influence curve and its role in robust estimation. *J.Amer. Stat. Assoc.*, **69**, 383-394.

Hampel, F.R., Ronchetti, E. M., Rousseeuw, P.J., Stahel, W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006) *Robust Statistics: Theory and Methods*, Wiley, Chichister.

Stahel, W.A. (1981) Robust estimation, infinitesimal optimality and covariance matrix estimators. Ph.D. Thesis, ETH, Zurich. In german

Yohai, V.J. (2008) Optimal robust estimates using the Kullback–Leibler divergence. *Statistical and Probability Letters*, **78**, 1811-1816..

Marazzi, A. and Yohai, V. J. "Optimal robust estimates using the Hellinger distance" Advances in Data Analysis and Classification, 4, 169-179, 2010