STAT 593
Robust statistics: Location and Scale

Joseph Salmon
http://josephsalmon.eu

Télécom Paristech, Institut Mines-Télécom
&
University of Washington, Department of Statistics
(Visiting Assistant Professor)

Outline

Location

Scale/dispersion

Simultaneous location and scale estimation

Asymptotic / minimax result

Table of Contents

Simplest location model[1]

$$x_i = \mu^* + \varepsilon_i, \text{ for } i = 1, \dots, n \tag{1}$$

- $\mu^* \in \mathbb{R}^p$ is the true parameter
- $x_1, \dots, x_n$ are n observations in $\mathbb{R}^p$; and $X = [x_1, \dots, x_n]$
- $\varepsilon_1, \dots, \varepsilon_n$ model the noise variables (also in $\mathbb{R}^p$) and are *i.i.d.* random variable having the same c.d.f. F

Consequence: $x_1, \dots, x_n$ are *i.i.d.* with c.d.f. $G(\cdot) = F(\cdot - \mu^*)$

Rem: when F has a density, we write it f

[1] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics: Theory and methods*. Chichester: John Wiley & Sons, 2006.

Contamination/mixture model

Imagine a proportion of $1-\alpha$ of the observations generated by a "normal" model, and a proportion α generated from an unknown c.d.f:

$$(1-\alpha)F + \alpha G$$

- F : models the "normal" observation, *e.g.*, F is the c.d.f. of a Gaussian distribution $\mathcal{N}(\mu^*, \sigma_*^2 \operatorname{Id}_p)$
- G : an arbitrary distribution (for instance a Gaussian with a way larger variance)
- α: contamination ratio/parameter

Rem: similarly, when the distributions F and G have densities f and g, the mixture density is $(1-\alpha)f + \alpha g$

Table of Contents

Maximum Likelihood Estimation (MLE)

Assuming model (1) such that f is the density (or p.d.f.) of F, the likelihood function is:

$$\mathcal{L}(x_1, \dots, x_n; \mu) = \prod_{i=1}^{n} f(x_i - \mu)$$

The Maximum Likelihood Estimation (MLE) of μ is defined by:

$$\hat{\mu}_n^{\mathrm{MLE}} \in \underset{\mu \in \mathbb{R}^p}{\arg\max}\, \mathcal{L}(x_1, \dots, x_n; \mu)$$

Rem: if F is known exactly, the MLE is "optimal" in the sense of asymptotic normality, see Section (10.8), Maronna *et al.* (2006).

Double objective: find estimators almost optimal when the model is not contaminated, but also almost optimal when it is.

More on MLE

Instead of maximizing a product of function, (convex) optimization would reformulated this equivalently as minimizing the (negative) log-likelihood:

$$\hat{\mu}_n^{\mathrm{MLE}} \in \operatorname*{arg\,min}_{\mu \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(x_i - \mu), \text{ where } \rho = -\log(f)$$

<u>Rem</u>: possibly additive / multiplicative constants can be removed

<u>Differentiable case</u>: If ρ is differentiable, then first order conditions (or Fermat's rule) ensure that:

$$0 = \frac{1}{n} \sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n), \text{ where } \psi = \rho'.$$

Examples ($p = 1$)

Distribution	$f(x)$	$\rho(x)$	$\psi(x)$	$\hat{\mu}_n$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$	$\frac{x^2}{2}$	x	$\bar{x}_n$
Laplace	$\frac{1}{2} \exp(-\lvert x \rvert)$	$\lvert x \rvert$	$\times$	$\mathrm{Med}_n(X)$

M-estimators for location parameter

Definition

We call M-estimator associated to a function ρ any estimator obtained as follows:

$$\hat{\mu}_n(\rho) \in \arg\min_{\mu \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(x_i - \mu), \text{ where } \rho \text{ is not necessarily } -\log(f)$$

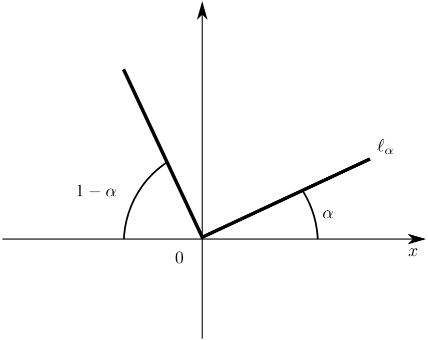
<u>Differentiable case</u>: If ρ is differentiable, then first order conditions (or Fermat's rule)

$$0 = \sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n), \text{ where } \psi = \rho'.$$

"Pinball loss" / quantile regression

$$\rho = \ell_\alpha \text{ where } \ell_\alpha(x) = \begin{cases} -(1-\alpha)x & \text{if } x \leq 0 \\ \alpha x & \text{if } x \geq 0 \end{cases}$$

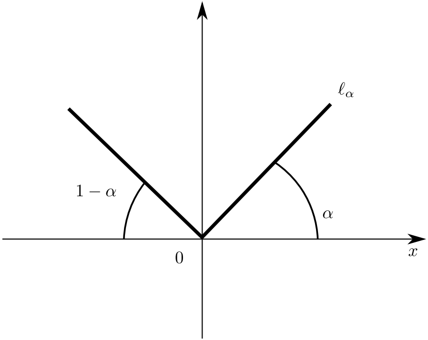
$$= \alpha |x| \mathbb{1}_{\{x \geq 0\}} + (1-\alpha)|x| \mathbb{1}_{\{x \leq 0\}}$$



Rem: we will discuss some more the case of non-differentiable but convex functions, sub-differentials, etc.

"Pinball loss" / quantile regression

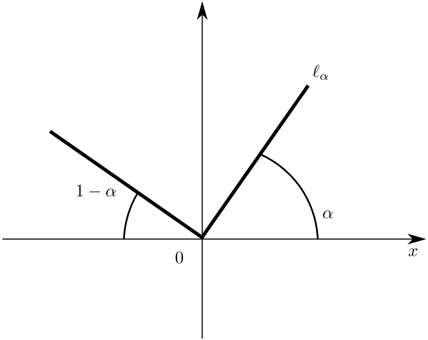
$$\rho = \ell_\alpha \text{ where } \ell_\alpha(x) = \begin{cases} -(1-\alpha)x & \text{if } x \leq 0 \\ \alpha x & \text{if } x \geq 0 \end{cases}$$
$$= \alpha|x|\mathbb{1}_{\{x\geq 0\}} + (1-\alpha)|x|\mathbb{1}_{\{x\leq 0\}}$$



Rem: we will discuss some more the case of non-differentiable but convex functions, sub-differentials, etc.

"Pinball loss" / quantile regression

$$\rho = \ell_\alpha \text{ where } \ell_\alpha(x) = \begin{cases} -(1-\alpha)x & \text{if } x \le 0 \\ \alpha x & \text{if } x \ge 0 \end{cases}$$
$$= \alpha|x|\mathbb{1}_{\{x\ge 0\}} + (1-\alpha)|x|\mathbb{1}_{\{x\le 0\}}$$



Rem: we will discuss some more the case of non-differentiable but convex functions, sub-differentials, etc.

Distribution of M-estimators

Define $\breve{\mu} \in \mathbb{R}^p$ as the theoretical counter part of the M-estimators: for $X \sim F$, it is defined by

$$\breve{\mu} := \breve{\mu}(F, \rho) \in \arg\min_{\mu \in \mathbb{R}^p} \mathbb{E}_F\big(\rho(X - \mu)\big)$$

$$\text{whereas} \qquad \hat{\mu}_n := \hat{\mu}_n(\rho) \in \arg\min_{\mu \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(x_i - \mu)$$

<u>Rem</u>:

- $\hat{\mu}_n(\rho) = \breve{\mu}(\hat{F}_n, \rho)$ where $\hat{F}_n$ is the empirical distribution based on the sample $x_1, \ldots, x_n$.
- In the MLE case : $\hat{\mu}_n^{\mathrm{MLE}} = \breve{\mu}(\hat{F}_n, -\log(f))$ where f is the p.d.f.
- $\breve{\mu}$ and $\hat{\mu}_n$ are translation equivariant

Example in 1D

- when $\rho(x) = \frac{x^2}{2}$ then $\breve{\mu}(F, \rho) = \mathbb{E}_F(X)$
- when $\rho(x) = |x|$ then $\breve{\mu}(F, \rho) = \mathrm{Med}_F(X)$ where $M = \mathrm{Med}_F(X)$ satisfies $F(M) = \dfrac{1}{2}$
- when $\rho(x) = \ell_\alpha(x) := \alpha|x|\mathbb{1}_{\{x \geq 0\}} + (1-\alpha)|x|\mathbb{1}_{\{x \leq 0\}}$ then $\breve{\mu}(F, \rho) = F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$ is[2] the α-quantile of the distribution F.

[2] R. Koenker and G. Bassett. "Regression quantiles". In: *Econometrica* 46.1 (1978), pp. 33–50.

Proof: assuming F has a density f

Goal: find $\quad \breve{\mu} \in \arg\min_{\mu \in \mathbb{R}} \mathbb{E}_F\big(\ell_\alpha(X-\mu)\big)$

$$\begin{aligned}
\mathbb{E}_F\big(\ell_\alpha(X-\mu)\big) &= \mathbb{E}_F(\alpha|X-\mu|\mathbb{1}_{\{X\geq\mu\}} + (1-\alpha)|X-\mu|\mathbb{1}_{\{X\leq\mu\}}) \\
&= \alpha \int_\mu^{+\infty} (x-\mu)f(x)dx - (1-\alpha)\int_{-\infty}^{\mu}(x-\mu)f(x)dx
\end{aligned}$$

Proof: assuming F has a density f

Goal: find $\quad \breve{\mu} \in \arg\min_{\mu \in \mathbb{R}} \mathbb{E}_F\big(\ell_\alpha(X-\mu)\big)$

$$\begin{aligned}
\mathbb{E}_F\big(\ell_\alpha(X-\mu)\big) &= \mathbb{E}_F(\alpha|X-\mu|\mathbb{1}_{\{X\geq\mu\}} + (1-\alpha)|X-\mu|\mathbb{1}_{\{X\leq\mu\}}) \\
&= \alpha \int_{\mu}^{+\infty} (x-\mu)f(x)dx - (1-\alpha)\int_{-\infty}^{\mu}(x-\mu)f(x)dx
\end{aligned}$$

Note that $\int_\mu^{+\infty}(x-\mu)f(x)dx = \int_\mu^{+\infty} xf(x)dx - \mu(1-F(\mu))$, so

Proof: assuming F has a density f

$$\text{Goal: find} \quad \breve{\mu} \in \arg\min_{\mu\in\mathbb{R}} \mathbb{E}_F\big(\ell_\alpha(X-\mu)\big)$$

$$\begin{aligned}\mathbb{E}_F\big(\ell_\alpha(X-\mu)\big) &= \mathbb{E}_F(\alpha|X-\mu|\mathbb{1}_{\{X\geq\mu\}} + (1-\alpha)|X-\mu|\mathbb{1}_{\{X\leq\mu\}}) \\ &= \alpha\int_\mu^{+\infty}(x-\mu)f(x)dx - (1-\alpha)\int_{-\infty}^{\mu}(x-\mu)f(x)dx\end{aligned}$$

Note that $\int_\mu^{+\infty}(x-\mu)f(x)dx = \int_\mu^{+\infty} xf(x)dx - \mu(1-F(\mu))$, so

$$\begin{aligned}\frac{d}{d\mu}\left(\int_\mu^{+\infty}(x-\mu)f(x)dx\right) &= \lim_{x\to+\infty} xf(x) - \mu f(\mu) - (1-F(\mu)) + \mu f(\mu) \\ &= -\,(1-F(\mu))\end{aligned}$$

Proof: assuming F has a density f

Goal: find $\quad \breve{\mu} \in \arg\min_{\mu\in\mathbb{R}} \mathbb{E}_F\big(\ell_\alpha(X-\mu)\big)$

$$
\begin{aligned}
\mathbb{E}_F\big(\ell_\alpha(X-\mu)\big) &= \mathbb{E}_F(\alpha|X-\mu|\mathbb{1}_{\{X\geq\mu\}} + (1-\alpha)|X-\mu|\mathbb{1}_{\{X\leq\mu\}}) \\
&= \alpha\int_\mu^{+\infty}(x-\mu)f(x)dx - (1-\alpha)\int_{-\infty}^{\mu}(x-\mu)f(x)dx
\end{aligned}
$$

Note that $\int_\mu^{+\infty}(x-\mu)f(x)dx = \int_\mu^{+\infty} xf(x)dx - \mu(1-F(\mu))$, so

$$
\begin{aligned}
\frac{d}{d\mu}\left(\int_\mu^{+\infty}(x-\mu)f(x)dx\right) &= \lim_{x\to+\infty} xf(x) - \mu f(\mu) - (1-F(\mu)) + \mu f(\mu) \\
&= -\,(1-F(\mu))
\end{aligned}
$$

Similarly, $\frac{d}{d\mu}\left(\int_{-\infty}^{\mu}(x-\mu)f(x)dx\right) = F(\mu)$ so the first order conditions of the associated minimization problem yield

Proof: assuming F has a density f

Goal: find $\quad \breve{\mu} \in \arg\min_{\mu \in \mathbb{R}} \mathbb{E}_F\big(\ell_\alpha(X-\mu)\big)$

$$\begin{aligned}
\mathbb{E}_F\big(\ell_\alpha(X-\mu)\big) &= \mathbb{E}_F(\alpha|X-\mu|\mathbb{1}_{\{X\geq\mu\}} + (1-\alpha)|X-\mu|\mathbb{1}_{\{X\leq\mu\}}) \\
&= \alpha\int_\mu^{+\infty}(x-\mu)f(x)dx - (1-\alpha)\int_{-\infty}^{\mu}(x-\mu)f(x)dx
\end{aligned}$$

Note that $\int_\mu^{+\infty}(x-\mu)f(x)dx = \int_\mu^{+\infty} xf(x)dx - \mu(1-F(\mu))$, so

$$\begin{aligned}
\frac{d}{d\mu}\left(\int_\mu^{+\infty}(x-\mu)f(x)dx\right) &= \lim_{x\to+\infty} xf(x) - \mu f(\mu) - (1-F(\mu)) + \mu f(\mu) \\
&= -\,(1-F(\mu))
\end{aligned}$$

Similarly, $\frac{d}{d\mu}\left(\int_{-\infty}^{\mu}(x-\mu)f(x)dx\right) = F(\mu)$ so the first order conditions of the associated minimization problem yield

$$0 = -\,\alpha(1-F(\breve{\mu})) + (1-\alpha)F(\breve{\mu})$$

Proof: assuming F has a density f

$$\text{Goal: find} \quad \breve{\mu} \in \arg\min_{\mu \in \mathbb{R}} \mathbb{E}_F\big(\ell_\alpha(X-\mu)\big)$$

$$\begin{aligned}
\mathbb{E}_F\big(\ell_\alpha(X-\mu)\big) &= \mathbb{E}_F(\alpha|X-\mu|\mathbb{1}_{\{X\geq\mu\}} + (1-\alpha)|X-\mu|\mathbb{1}_{\{X\leq\mu\}}) \\
&= \alpha\int_\mu^{+\infty}(x-\mu)f(x)dx - (1-\alpha)\int_{-\infty}^{\mu}(x-\mu)f(x)dx
\end{aligned}$$

Note that $\int_\mu^{+\infty}(x-\mu)f(x)dx = \int_\mu^{+\infty} xf(x)dx - \mu(1-F(\mu))$, so

$$\begin{aligned}
\frac{d}{d\mu}\left(\int_\mu^{+\infty}(x-\mu)f(x)dx\right) &= \lim_{x\to+\infty} xf(x) - \mu f(\mu) - (1-F(\mu)) + \mu f(\mu) \\
&= -\,(1-F(\mu))
\end{aligned}$$

Similarly, $\frac{d}{d\mu}\left(\int_{-\infty}^{\mu}(x-\mu)f(x)dx\right) = F(\mu)$ so the first order conditions of the associated minimization problem yield

$$0 = -\,\alpha(1-F(\breve{\mu})) + (1-\alpha)F(\breve{\mu}) \iff \alpha = F(\breve{\mu})$$

Proof: assuming F has a density f

Goal: find $\quad \breve{\mu} \in \arg\min_{\mu \in \mathbb{R}} \mathbb{E}_F\big(\ell_\alpha(X-\mu)\big)$

$$\begin{aligned}
\mathbb{E}_F\big(\ell_\alpha(X-\mu)\big) &= \mathbb{E}_F(\alpha|X-\mu|\mathbb{1}_{\{X\geq\mu\}} + (1-\alpha)|X-\mu|\mathbb{1}_{\{X\leq\mu\}}) \\
&= \alpha\int_\mu^{+\infty}(x-\mu)f(x)dx - (1-\alpha)\int_{-\infty}^{\mu}(x-\mu)f(x)dx
\end{aligned}$$

Note that $\int_\mu^{+\infty}(x-\mu)f(x)dx = \int_\mu^{+\infty} xf(x)dx - \mu(1-F(\mu))$, so

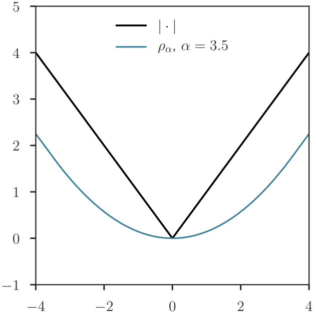
$$\begin{aligned}
\frac{d}{d\mu}\left(\int_\mu^{+\infty}(x-\mu)f(x)dx\right) &= \lim_{x\to+\infty} xf(x) - \mu f(\mu) - (1-F(\mu)) + \mu f(\mu) \\
&= -\,(1-F(\mu))
\end{aligned}$$

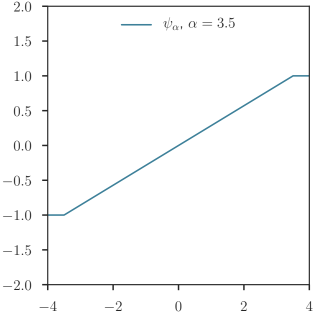
Similarly, $\frac{d}{d\mu}\left(\int_{-\infty}^{\mu}(x-\mu)f(x)dx\right) = F(\mu)$ so the first order conditions of the associated minimization problem yield

$$0 = -\,\alpha(1-F(\breve{\mu})) + (1-\alpha)F(\breve{\mu}) \iff \alpha = F(\breve{\mu}) \iff \breve{\mu} = F^{-1}(\alpha)$$

The Huber function

$$\rho_\alpha := \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \leq \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases} \qquad \psi_\alpha : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \leq \alpha \\ \text{sign}(x) & \text{if } |x| > \alpha \end{cases}$$

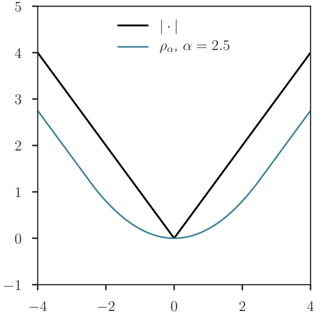


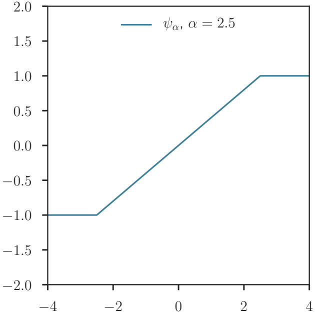


The Huber function

$$\rho_\alpha := \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \leq \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases}$$

$$\psi_\alpha : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \leq \alpha \\ \text{sign}(x) & \text{if } |x| > \alpha \end{cases}$$

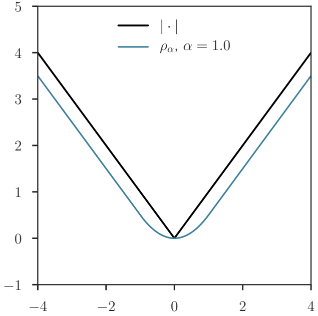


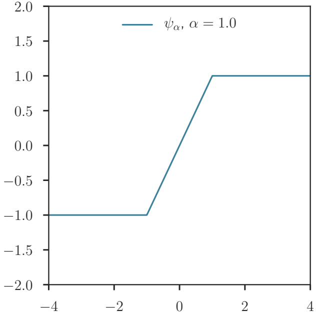


The Huber function

$$\rho_\alpha := \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \le \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases}$$

$$\psi_\alpha : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \le \alpha \\ \text{sign}(x) & \text{if } |x| > \alpha \end{cases}$$

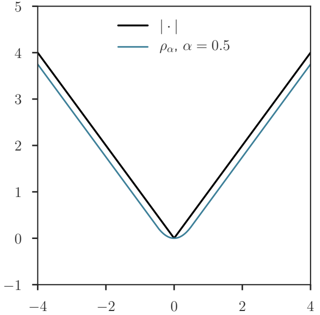


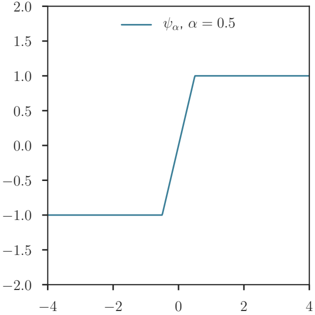


The Huber function

$$\rho_\alpha := \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \leq \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases}$$

$$\psi_\alpha : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \leq \alpha \\ \operatorname{sign}(x) & \text{if } |x| > \alpha \end{cases}$$

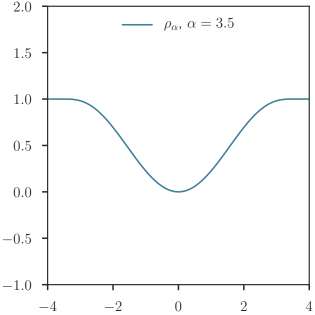


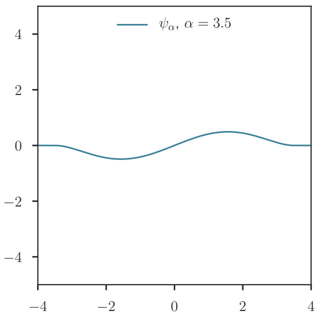


The Bisquare function

$$\rho_\alpha : \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \leq \alpha \\ 1 & \text{if } |x| > \alpha \end{cases}$$

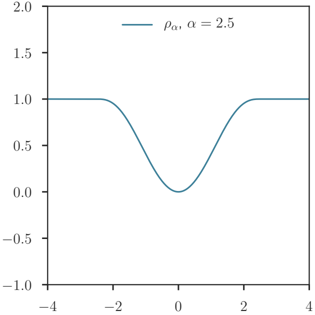
$$\psi_\alpha : \begin{cases} \frac{6x}{\alpha^2}[1 - (\frac{x}{\alpha})^2]^2 & \text{if } |x| \leq \alpha \\ 0 & \text{if } |x| > \alpha \end{cases}$$

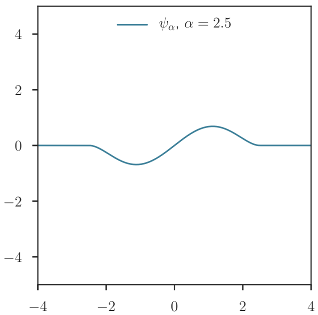




The Bisquare function

$$\rho_\alpha : \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \leq \alpha \\ 1 & \text{if } |x| > \alpha \end{cases} \qquad \psi_\alpha : \begin{cases} \frac{6x}{\alpha^2}[1 - (\frac{x}{\alpha})^2]^2 & \text{if } |x| \leq \alpha \\ 0 & \text{if } |x| > \alpha \end{cases}$$

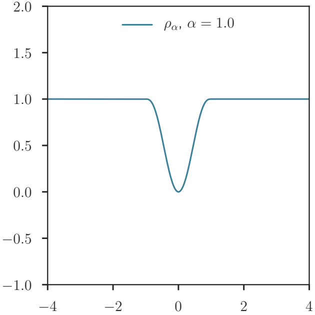


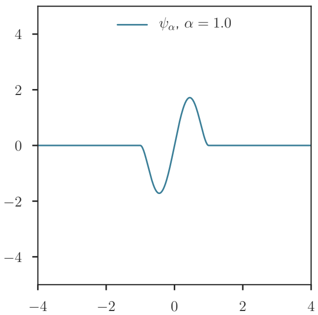


The Bisquare function

$$\rho_\alpha : \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \leq \alpha \\ 1 & \text{if } |x| > \alpha \end{cases}$$

$$\psi_\alpha : \begin{cases} \frac{6x}{\alpha^2}[1 - (\frac{x}{\alpha})^2]^2 & \text{if } |x| \leq \alpha \\ 0 & \text{if } |x| > \alpha \end{cases}$$

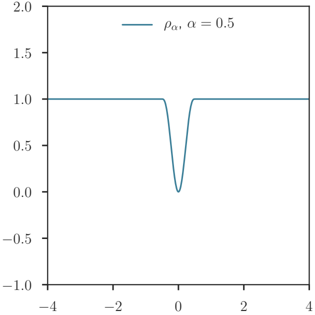




The Bisquare function

$$\rho_\alpha : \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \le \alpha \\ 1 & \text{if } |x| > \alpha \end{cases}$$

$$\psi_\alpha : \begin{cases} \frac{6x}{\alpha^2}[1 - (\frac{x}{\alpha})^2]^2 & \text{if } |x| \le \alpha \\ 0 & \text{if } |x| > \alpha \end{cases}$$



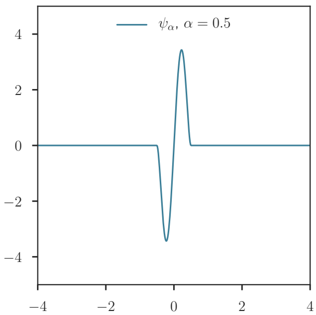


Table of Contents

Smoothness assumptions

For this section, we consider only the case where ρ is differentiable where $\psi = \rho'$, and ψ'' is bounded. Assume also $X \sim F$, and $x_1, \ldots, x_n$ are *i.i.d.* with the same distribution F.

Smoothness assumptions

For this section, we consider only the case where ρ is differentiable where $\psi = \rho'$, and ψ'' is bounded. Assume also $X \sim F$, and $x_1, \ldots, x_n$ are *i.i.d.* with the same distribution F.

Theorem

Under the previous smoothness assumption and provided ψ is non-decreasing

$$\sqrt{n}(\hat{\mu}_n - \breve{\mu}) \to_d \mathcal{N}(0, V^2), \text{where} \qquad V^2 = \frac{\mathbb{E}_F \left(\psi(X - \breve{\mu})^2\right)}{\left(\mathbb{E}_F \psi'(X - \breve{\mu})\right)^2}$$

is called the **asymptotic variance** of $\hat{\mu}_n$

Smoothness assumptions

For this section, we consider only the case where ρ is differentiable where $\psi = \rho'$, and ψ'' is bounded. Assume also $X \sim F$, and $x_1, \ldots, x_n$ are *i.i.d.* with the same distribution F.

Theorem

Under the previous smoothness assumption and provided ψ is non-decreasing

$$\sqrt{n}(\hat{\mu}_n - \breve{\mu}) \to_d \mathcal{N}(0, V^2), \text{where} \qquad V^2 = \frac{\mathbb{E}_F\left(\psi(X - \breve{\mu})^2\right)}{\left(\mathbb{E}_F \psi'(X - \breve{\mu})\right)^2}$$

is called the **asymptotic variance** of $\hat{\mu}_n$

Example : In the case $\rho(x) = x^2/2$, one recovers the CLT as $V^2 = \mathrm{Var}(X)$

Smoothness assumptions

For this section, we consider only the case where ρ is differentiable where $\psi = \rho'$, and ψ'' is bounded. Assume also $X \sim F$, and $x_1, \ldots, x_n$ are *i.i.d.* with the same distribution F.

Theorem

Under the previous smoothness assumption and provided ψ is non-decreasing

$$\sqrt{n}(\hat{\mu}_n - \breve{\mu}) \to_d \mathcal{N}(0, V^2), \text{where} \qquad V^2 = \frac{\mathbb{E}_F\left(\psi(X - \breve{\mu})^2\right)}{\left(\mathbb{E}_F \psi'(X - \breve{\mu})\right)^2}$$

is called the **asymptotic variance** of $\hat{\mu}_n$

Example : In the case $\rho(x) = x^2/2$, one recovers the CLT as $V^2 = \mathrm{Var}(X)$

Rem: since $\breve{\mu}$ is translation equivariant, in the translation model $x_i = \mu^* + \varepsilon_i$ then $V^2 =$ is independent of μ^*

Proof continued

By definition of $\breve{\mu}$ and $\hat{\mu}_n$

$$\mathbb{E}_F \psi(X - \breve{\mu}) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) = 0$$

Proof continued

By definition of $\breve{\mu}$ and $\hat{\mu}_n$

$$\mathbb{E}_F \psi(X - \breve{\mu}) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) = 0$$

Then the function $\breve{\lambda}$ and $\hat{\lambda}_n$ defined for any $s \in \mathbb{R}$

$$\breve{\lambda}(s) = \mathbb{E}_F \psi(X - s) \quad \text{and} \quad \hat{\lambda}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - s)$$

are non-increasing, $\breve{\lambda}(\breve{\mu}) = \hat{\lambda}_n(\hat{\mu}_n) = 0$ and $\lim_{n\to\infty} \hat{\lambda}_n(s) = \breve{\lambda}(s)$.

Proof continued

By definition of $\breve{\mu}$ and $\hat{\mu}_n$

$$\mathbb{E}_F \psi(X - \breve{\mu}) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) = 0$$

Then the function $\breve{\lambda}$ and $\hat{\lambda}_n$ defined for any $s \in \mathbb{R}$

$$\breve{\lambda}(s) = \mathbb{E}_F \psi(X - s) \quad \text{and} \quad \hat{\lambda}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - s)$$

are non-increasing, $\breve{\lambda}(\breve{\mu}) = \hat{\lambda}_n(\hat{\mu}_n) = 0$ and $\lim_{n\to\infty} \hat{\lambda}_n(s) = \breve{\lambda}(s)$.

Fact 1: $\boxed{\hat{\mu}_n \xrightarrow{p} \breve{\mu}}$

Proof continued

By definition of $\breve{\mu}$ and $\hat{\mu}_n$

$$\mathbb{E}_F \psi(X - \breve{\mu}) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) = 0$$

Then the function $\breve{\lambda}$ and $\hat{\lambda}_n$ defined for any $s \in \mathbb{R}$

$$\breve{\lambda}(s) = \mathbb{E}_F \psi(X - s) \quad \text{and} \quad \hat{\lambda}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - s)$$

are non-increasing, $\breve{\lambda}(\breve{\mu}) = \hat{\lambda}_n(\hat{\mu}_n) = 0$ and $\lim_{n\to\infty} \hat{\lambda}_n(s) = \breve{\lambda}(s)$.

Fact 1: $\boxed{\hat{\mu}_n \xrightarrow{p} \breve{\mu}}$

Proof: fix $\epsilon > 0$: since $\hat{\lambda}_n$ is non-increasing
$\mathbb{P}(\{\hat{\mu}_n < \breve{\mu} - \epsilon\}) \leq \mathbb{P}(\{\hat{\lambda}_n(\hat{\mu}_n) > \hat{\lambda}_n(\breve{\mu} - \epsilon)\}) = \mathbb{P}(\{0 > \hat{\lambda}_n(\breve{\mu} - \epsilon)\})$

Proof continued

By definition of $\breve{\mu}$ and $\hat{\mu}_n$

$$\mathbb{E}_F \psi(X - \breve{\mu}) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) = 0$$

Then the function $\breve{\lambda}$ and $\hat{\lambda}_n$ defined for any $s \in \mathbb{R}$

$$\breve{\lambda}(s) = \mathbb{E}_F \psi(X - s) \quad \text{and} \quad \hat{\lambda}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - s)$$

are non-increasing, $\breve{\lambda}(\breve{\mu}) = \hat{\lambda}_n(\hat{\mu}_n) = 0$ and $\lim_{n\to\infty} \hat{\lambda}_n(s) = \breve{\lambda}(s)$.

Fact 1: $\boxed{\hat{\mu}_n \xrightarrow{p} \breve{\mu}}$

Proof: fix $\epsilon > 0$: since $\hat{\lambda}_n$ is non-increasing
$\mathbb{P}(\{\hat{\mu}_n < \breve{\mu} - \epsilon\}) \leq \mathbb{P}(\{\hat{\lambda}_n(\hat{\mu}_n) > \hat{\lambda}_n(\breve{\mu} - \epsilon)\}) = \mathbb{P}(\{0 > \hat{\lambda}_n(\breve{\mu} - \epsilon)\})$
Now remind that with the law of large number:
$\lim_{n\to\infty} \hat{\lambda}_n(\breve{\mu} - \epsilon) = \breve{\lambda}(\breve{\mu} - \epsilon) > 0$, so $\lim_{n\to\infty} \mathbb{P}(\{0 > \hat{\lambda}_n(\breve{\mu} - \epsilon)\}) = 0$.
Hence, $\lim_{n\to\infty} \mathbb{P}(\{\hat{\mu}_n < \breve{\mu} - \epsilon\}) = 0$ and similarly one can show that
$\lim_{n\to\infty} \mathbb{P}(\{\hat{\mu}_n > \breve{\mu} + \epsilon\}) = 0$ □

Proof (end)

By a Taylor expansion (Lagrange form) there exists $\tilde{\mu}$ such :

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\psi(x_i-\hat{\mu}_n) =&\frac{1}{n}\sum_{i=1}^{n}\psi(x_i-\breve{\mu}) + (\breve{\mu}-\hat{\mu}_n)\cdot\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\breve{\mu})\\ &+\frac{1}{2n}(\breve{\mu}-\hat{\mu}_n)^2\sum_{i=1}^{n}\psi''(x_i-\tilde{\mu})\end{aligned}$$

Proof (end)

By a Taylor expansion (Lagrange form) there exists $\tilde{\mu}$ such :

$$\frac{1}{n}\sum_{i=1}^{n}\psi(x_i-\hat{\mu}_n) = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i-\breve{\mu}) + (\breve{\mu}-\hat{\mu}_n)\cdot\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\breve{\mu})$$
$$+\frac{1}{2n}(\breve{\mu}-\hat{\mu}_n)^2\sum_{i=1}^{n}\psi''(x_i-\tilde{\mu})$$

Hence:

$$\sqrt{n}(\hat{\mu}_n-\breve{\mu}) = -\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi(x_i-\breve{\mu})\right)}{\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\breve{\mu}) + (\hat{\mu}_n-\breve{\mu})\frac{1}{2n}\sum_{i=1}^{n}\psi''(x_i-\tilde{\mu})}$$

Proof (end)

By a Taylor expansion (Lagrange form) there exists $\tilde{\mu}$ such :

$$\frac{1}{n}\sum_{i=1}^{n}\psi(x_i-\hat{\mu}_n) = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i-\breve{\mu}) + (\breve{\mu}-\hat{\mu}_n)\cdot\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\breve{\mu}) + \frac{1}{2n}(\breve{\mu}-\hat{\mu}_n)^2\sum_{i=1}^{n}\psi''(x_i-\tilde{\mu})$$

Hence:

$$\sqrt{n}(\hat{\mu}_n-\breve{\mu}) = -\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi(x_i-\breve{\mu})\right)}{\frac{1}{n}\sum_{i=1}^{n}\psi'(x_i-\breve{\mu}) + (\hat{\mu}_n-\breve{\mu})\frac{1}{2n}\sum_{i=1}^{n}\psi''(x_i-\tilde{\mu})}$$

Provided ψ'' is bounded, the numerator converges to $\mathbb{E}_F\psi'(X-\breve{\mu})$. Hence, with Slutsky's lemma and the CLT,

$$\sqrt{n}(\hat{\mu}_n-\breve{\mu}) \to_d \mathcal{N}(0,V^2), \text{where} \quad V^2 = \frac{\mathbb{E}_F\left(\psi(X-\breve{\mu})^2\right)}{\left(\mathbb{E}_F\psi'(X-\breve{\mu})\right)^2}$$

Intuitive view on M-estimators

Assume for simplicity that $\psi(0) = 0$ and that $\psi'(0)$ exists, then one can defined

$$W(x) = \begin{cases} \frac{\psi(x)}{x} & \text{if } x \neq 0 \\ \psi'(0) & \text{if } x = 0 \end{cases}$$

and then

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) \iff 0 = \sum_{i=1}^{n} W(x_i - \hat{\mu}_n)(x_i - \hat{\mu}_n)$$

$$\iff \boxed{\hat{\mu}_n = \frac{\sum_{i=1}^{n} W(x_i - \hat{\mu}_n) x_i}{\sum_{i'=1}^{n} W(x_{i'} - \hat{\mu}_n)}}$$

<u>Interpretation</u>: this is a weighted average with weights (often) decaying when $x_i - \hat{\mu}_n$ is large (*i.e.*, for outlying observations)

(Normalized) Weights examples

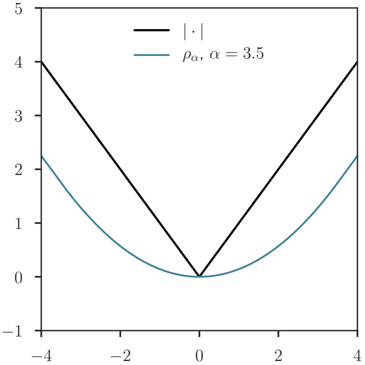
- Mean case: $W(x) = 1$ all data points are weighted equally

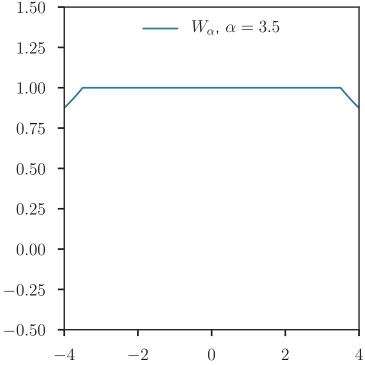
(Normalized) Weights examples

- Huber case:

$$\rho_\alpha = \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \le \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases}$$

$$W_\alpha(x) = \min\left(1, \frac{\alpha}{|x|}\right)$$

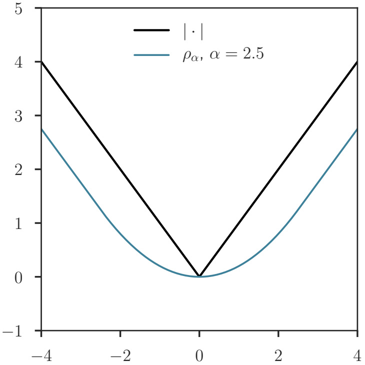


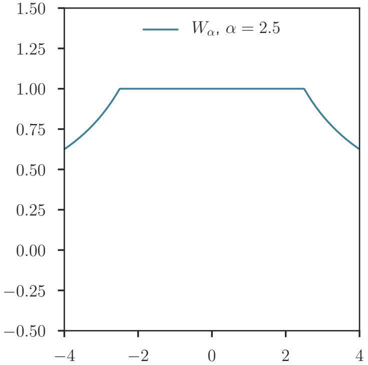


(Normalized) Weights examples

- Huber case:

$$\rho_\alpha = \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \leq \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases} \qquad W_\alpha(x) = \min\left(1, \frac{\alpha}{|x|}\right)$$



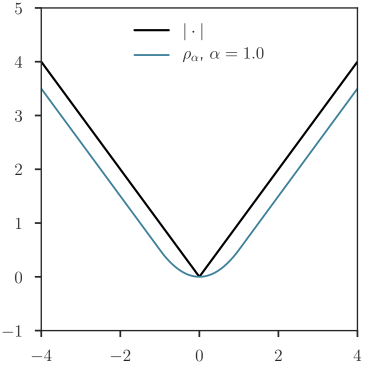


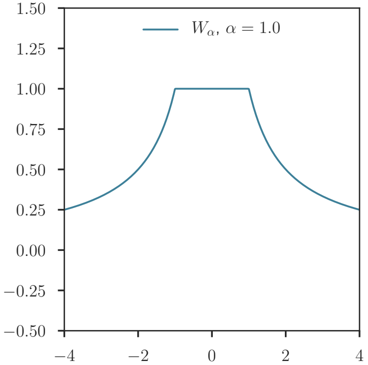
(Normalized) Weights examples

- Huber case:

$$\rho_\alpha = \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \leq \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases}$$

$$W_\alpha(x) = \min\left(1, \frac{\alpha}{|x|}\right)$$



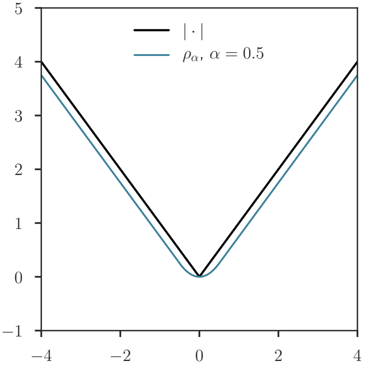


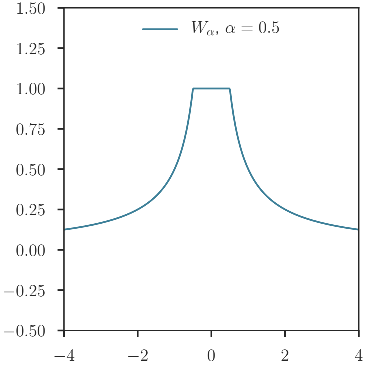
(Normalized) Weights examples

- Huber case:

$$\rho_\alpha = \begin{cases} \frac{x^2}{2\alpha} & \text{if } |x| \leq \alpha \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha \end{cases}$$

$$W_\alpha(x) = \min\left(1, \frac{\alpha}{|x|}\right)$$

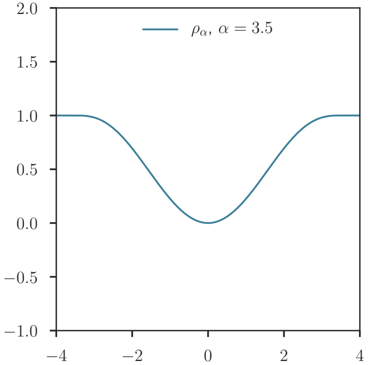


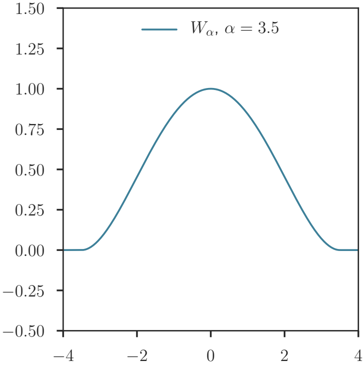


(Normalized) Weights examples

- Bi-square case:

$$\rho_\alpha = \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \leq \alpha \\ 1 & \text{if } |x| > \alpha \end{cases} \qquad W_\alpha(x) = \left[1 - \frac{x^2}{\alpha^2}\right]^2 \mathbb{1}_{[-\alpha,\alpha]}(x)$$

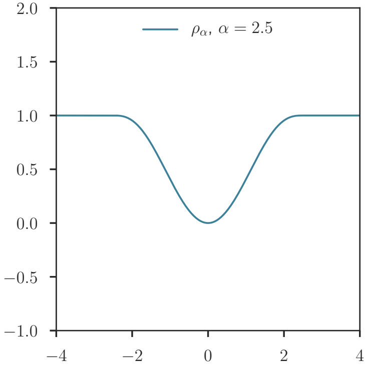


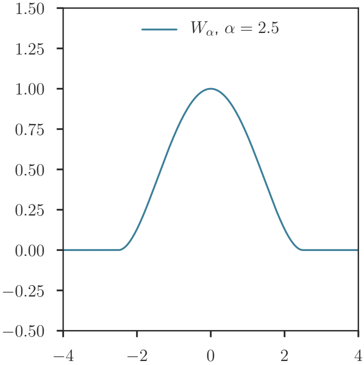


(Normalized) Weights examples

- Bi-square case:

$$\rho_\alpha = \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \leq \alpha \\ 1 & \text{if } |x| > \alpha \end{cases} \qquad W_\alpha(x) = \left[1 - \frac{x^2}{\alpha^2}\right]^2 \mathbb{1}_{[-\alpha,\alpha]}(x)$$

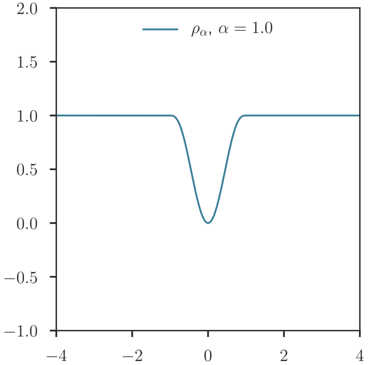


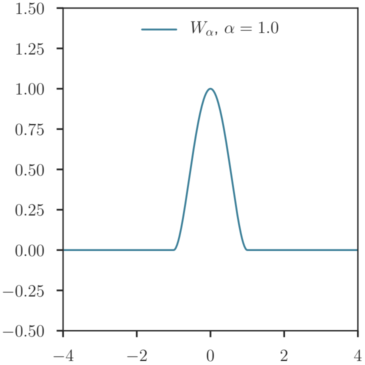


(Normalized) Weights examples

- Bi-square case:

$$\rho_\alpha = \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \leq \alpha \\ 1 & \text{if } |x| > \alpha \end{cases} \qquad W_\alpha(x) = \left[1 - \frac{x^2}{\alpha^2}\right]^2 \mathbb{1}_{[-\alpha,\alpha]}(x)$$

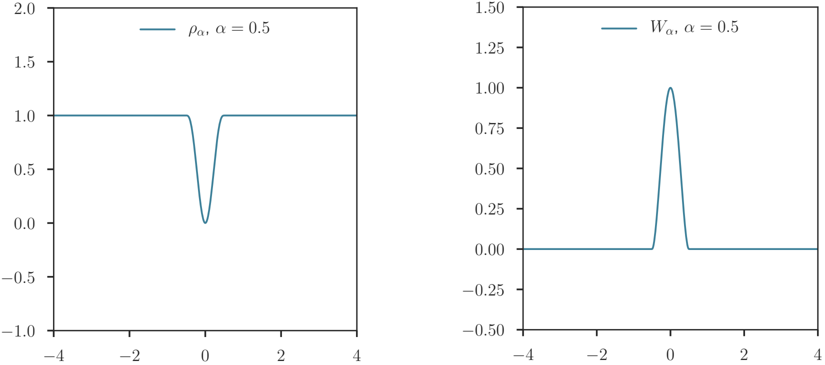




(Normalized) Weights examples

- Bi-square case:

$$\rho_\alpha = \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \leq \alpha \\ 1 & \text{if } |x| > \alpha \end{cases} \qquad W_\alpha(x) = \left[1 - \frac{x^2}{\alpha^2}\right]^2 \mathbb{1}_{[-\alpha,\alpha]}(x)$$



Another interpretation

$$\hat{\mu}_n = \hat{\mu}_n + \frac{1}{n}\sum_{i=1}^{n}\psi(x_i - \hat{\mu}_n) = \frac{1}{n}\sum_{i=1}^{n}\zeta(x_i, \hat{\mu}_n)$$

where $\zeta(x, \mu) = \mu + \psi(x - \mu)$

Example : for the Huber case $\zeta(x, \mu) = \mu + \alpha\psi_\alpha(x - \mu)$ where

$$\psi_\alpha(x) : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \leq \alpha \\ \operatorname{sign}(x) & \text{if } |x| > \alpha \end{cases}$$

so

$$\zeta(x, \mu) := \Pi_{[\mu-\alpha, \mu+\alpha]}(x) = \begin{cases} \mu - \alpha & \text{if } x < \mu - \alpha \\ x & \text{if } \mu - \alpha \leq x \leq \mu + \alpha \\ \mu + \alpha & \text{if } x > \mu + \alpha \end{cases}$$

Rem: this is connected to "Windsorizing"

Huber and "Windsorizing" p.79, Huber (1964)

Let $\hat{\mu}_n(X)$ the M-estimator for the Huber function:

$\sum_{i=1}^{n} \psi_\alpha(x_i - \hat{\mu}_n(X)) = 0$ where $\psi_\alpha(x) : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \leq \alpha \\ \mathrm{sign}(x) & \text{if } |x| > \alpha \end{cases}$

Huber and "Windsorizing" p.79, Huber (1964)

Let $\hat{\mu}_n(X)$ the M-estimator for the Huber function:

$\sum_{i=1}^n \psi_\alpha(x_i - \hat{\mu}_n(X)) = 0$ where $\psi_\alpha(x) : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \leq \alpha \\ \operatorname{sign}(x) & \text{if } |x| > \alpha \end{cases}$

Then:

$$0 = \sum_{i:|x_i - \hat{\mu}_n(X)| \leq \alpha} \alpha(x_i - \hat{\mu}_n(X)) + \sum_{i: x_i > \hat{\mu}_n(X) + \alpha} \alpha + \sum_{i: x_i < \hat{\mu}_n(X) - \alpha} -\alpha$$

Huber and "Windsorizing" p.79, Huber (1964)

Let $\hat{\mu}_n(X)$ the M-estimator for the Huber function:

$\sum_{i=1}^n \psi_\alpha(x_i - \hat{\mu}_n(X)) = 0$ where $\psi_\alpha(x) : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \leq \alpha \\ \text{sign}(x) & \text{if } |x| > \alpha \end{cases}$

Then:

$$0 = \sum_{i:|x_i - \hat{\mu}_n(X)| \leq \alpha} \alpha(x_i - \hat{\mu}_n(X)) + \sum_{i:x_i > \hat{\mu}_n(X)+\alpha} \alpha + \sum_{i:x_i < \hat{\mu}_n(X)-\alpha} -\alpha$$

$$0 = \sum_{i:|x_i - \hat{\mu}_n(X)| \leq \alpha} \alpha(x_i - \hat{\mu}_n(X)) + \sum_{i:x_i > \hat{\mu}_n(X)+\alpha} \alpha + \hat{\mu}_n(X) - \hat{\mu}_n(X)$$
$$+ \sum_{i:x_i < \hat{\mu}_n(X)-\alpha} -\alpha - \hat{\mu}_n(X) + \hat{\mu}_n(X)$$

Huber and "Windsorizing" p.79, Huber (1964)

Let $\hat{\mu}_n(X)$ the M-estimator for the Huber function:

$\sum_{i=1}^n \psi_\alpha(x_i - \hat{\mu}_n(X)) = 0$ where $\psi_\alpha(x) : \begin{cases} \frac{x}{\alpha} & \text{if } |x| \leq \alpha \\ \mathrm{sign}(x) & \text{if } |x| > \alpha \end{cases}$

Then:

$$0 = \sum_{i:|x_i - \hat{\mu}_n(X)| \leq \alpha} \alpha(x_i - \hat{\mu}_n(X)) + \sum_{i:x_i > \hat{\mu}_n(X)+\alpha} \alpha + \sum_{i:x_i < \hat{\mu}_n(X)-\alpha} -\alpha$$

$$0 = \sum_{i:|x_i - \hat{\mu}_n(X)| \leq \alpha} \alpha(x_i - \hat{\mu}_n(X)) + \sum_{i:x_i > \hat{\mu}_n(X)+\alpha} \alpha + \hat{\mu}_n(X) - \hat{\mu}_n(X)$$

$$+ \sum_{i:x_i < \hat{\mu}_n(X)-\alpha} -\alpha - \hat{\mu}_n(X) + \hat{\mu}_n(X)$$

Interpretation: $\hat{\mu}_n(X)$ is the empirical mean of the modified

dataset $\tilde{X}$ where $\tilde{x}_i = \begin{cases} \hat{\mu}_n(X) - \alpha & \text{if } \ x_i < \hat{\mu}_n(X) - \alpha \\ x_i & \text{if } |x_i - \hat{\mu}_n(X)| \leq \alpha \\ \hat{\mu}_n(X) + \alpha & \text{if } \ x_i < \hat{\mu}_n(X) + \alpha \end{cases}$

Computational difficulties

- some methods are convex and smooth ($\rho(x) = x^2/2$, Huber)
- some methods are convex but non-smooth (pinball, $\rho(x) = |x|$, etc.)
- some methods are non-convex but smooth (bi-square)

Numerical “recipes” will be investigated later on.

Table of Contents

Simplest dispersion model

$$x_i = \sigma_* \varepsilon_i, \text{ for } i = 1, \ldots, n \tag{2}$$

- $\sigma_* \in \mathbb{R}_{++}$ is the (true) scale parameter
- $x_1, \ldots, x_n$ are n observations in $\mathbb{R}^p$; and $X = [x_1, \ldots, x_n]$
- $\varepsilon_1, \ldots, \varepsilon_n$ and are *i.i.d.* random variable having the same c.d.f. F and density f

Consequence: $x_1, \ldots, x_n$ are *i.i.d.* with density $\frac{1}{\sigma} f(\frac{\cdot}{\sigma})$

MLE for scale estimation

Assuming model (2) such that f is the density (or p.d.f.) of F, the likelihood function is:

$$\mathcal{L}(x_1, \ldots, x_n; \sigma) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f\left(\frac{x_i}{\sigma}\right)$$

The Maximum Likelihood Estimation (MLE) of σ is defined by:

$$\hat{\sigma}_n^{\mathrm{MLE}} \in \underset{\sigma \in \mathbb{R}_{++}}{\arg\max}\, \mathcal{L}(x_1, \ldots, x_n; \sigma)$$

Transforming using $-\log$ then

$$\hat{\sigma}_n^{\mathrm{MLE}} \in \underset{\sigma \in \mathbb{R}_{++}}{\arg\min}\, \frac{1}{n} \log(\sigma) - \sum_{i=1}^{n} \log\left(f\left(\frac{x_i}{\sigma}\right)\right)$$

For smooth function f (*i.e.*, when f' exist) $\hat{\sigma}_n^{\mathrm{MLE}}$ satisfies:

$$\frac{1}{n} \sum_{i=1}^{n} \nu\left(\frac{x_i}{\hat{\sigma}_n^{\mathrm{MLE}}}\right) = 1, \quad \text{where} \quad \nu(x) = -\frac{x \cdot f'(x)}{f(x)}$$

Example

Distribution	$f(x)$	$\nu(x)$	$\hat{\sigma}_n^{\mathrm{MLE}}$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$	x^2	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$
Laplace	$\frac{1}{2}\exp(-\lvert x\rvert)$	$\lvert x\rvert$	$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i\rvert$

Table of Contents

M-estimators of scale

Definition

We call M-estimator of scale associated to a function ν any estimator $\hat{\sigma}_n$ obtained solving the following equation w.r.t. σ:

$$\frac{1}{n}\sum_{i=1}^{n}\nu\left(\frac{x_i}{\sigma}\right) = 1$$

- when $\forall i \in [n], x_i = 0$, it is natural to set $\hat{\sigma}(0,\ldots,0) = 0$
- $\hat{\sigma}$ is then scale equivariant $\hat{\sigma}_n(\alpha x_1,\ldots,\alpha x_n) = \alpha\hat{\sigma}_n(x_1,\ldots,x_n)$
- when n is even and $\nu = 2\mathbb{1}_{[-1,1]^c}$, then $\hat{\sigma}_n = \mathrm{Med}_n(|x_1|,\ldots,|x_n|)$.

Intuitive view on M-estimators

Assume for simplicity that $\nu'(0) = 0$ and that $\nu''(0) > 0$, then one can defined

$$W(x) = \begin{cases} \frac{\nu(x)}{x^2} & \text{if } x \neq 0 \\ \nu''(0) & \text{if } x = 0 \end{cases}$$

and then

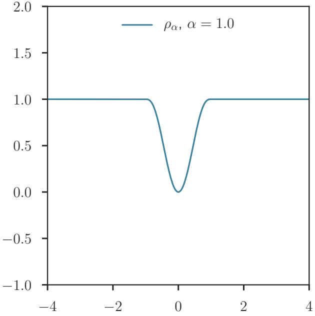
$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} W\left(\frac{x_i}{\hat{\sigma}_n}\right) x_i^2$$

Interpretation: this is a weighted average, with weights (often) decaying when $\frac{x_i}{\hat{\sigma}_n}$ is large (*i.e.*, for outlying observations)

Classical weights

- square case $\rho(x) = x^2$, $W(x) = 1$
- Bi-square case $(\alpha = 1)$

$$\rho(x) : \begin{cases} 1 - [1 - (\frac{x}{\alpha})^2]^3 & \text{if } |x| \leq \alpha \\ 1 & \text{if } |x| > \alpha \end{cases} \quad W(x) = \min\left(3 - 3x^2 + x^4, \frac{1}{x^2}\right)$$



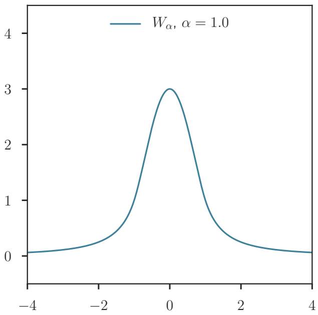


Table of Contents

Location/dispersion model

$$x_i = \mu^* + \sigma_* \varepsilon_i, \text{ for } i = 1, \ldots, n \tag{3}$$

- $\mu^* \in \mathbb{R}^p$ is the true parameter
- $x_1, \ldots, x_n$ are n observations in $\mathbb{R}^p$; and $X = [x_1, \ldots, x_n]$
- $\varepsilon_1, \ldots, \varepsilon_n$ model the noise variables (also in $\mathbb{R}^p$) and are *i.i.d.* random variable having the same density f

Consequence: $x_1, \ldots, x_n$ are *i.i.d.* with p.d.f. $\frac{1}{\sigma_*} f\left(\frac{\cdot - \mu^*}{\sigma_*}\right)$

Table of Contents

Simultaneous MLE

Assuming model (1) such that f is the density (or p.d.f.) of F, the simultaneous MLE estimators of location and scale are:

$$(\hat{\mu}_n^{\text{MLE}}, \hat{\sigma}_n^{\text{MLE}}) \in \underset{(\mu,\sigma)\in\mathbb{R}^p\times\mathbb{R}_{++}}{\arg\max} \left[\frac{1}{\sigma^n}\prod_{i=1}^{n} f\left(\frac{x_i-\mu}{\sigma}\right)\right]$$

or equivalently:

$$(\hat{\mu}_n^{\text{MLE}}, \hat{\sigma}_n^{\text{MLE}}) \in \underset{(\mu,\sigma)\in\mathbb{R}^p\times\mathbb{R}_{++}}{\arg\min} \left[\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{x_i-\mu}{\sigma}\right) + \log\sigma\right]$$

where $\rho = -\log(f)$.

Simultaneous M-estimators

Simultaneous M-estimators of location and estimation are $\hat{\mu}$ and $\hat{\sigma}$ satisfying for functions $\psi = \rho'$ and ν the following system of equation:

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \psi\left(\dfrac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right) = 0 \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \nu\left(\dfrac{x_i - \hat{\mu}_n}{\hat{\sigma}_n}\right) = 1 \end{cases}$$

Rem: in the MLE case $\psi(x) = -\frac{f'(x)}{f(x)}$ and $\nu(x) = -\frac{x \cdot f'(x)}{f(x)}$

Computational difficulties

Note that even if ρ is a convex function the function $\sigma \to \rho(z/\sigma) + \log(\sigma)$ is often non-convex

Several ways can be used to alleviate that:

- Change of variable: $\gamma = \frac{1}{\sigma}$
- Concomitant estimation, see Section 7.7, Huber (1981):

Substitute
$$\underset{(\mu,\sigma)\in\mathbb{R}^p\times\mathbb{R}_{++}}{\arg\min} \left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{x_i-\mu}{\sigma}\right)+\log\sigma\right]$$
by
$$\underset{(\mu,\sigma)\in\mathbb{R}^p\times\mathbb{R}_{++}}{\arg\min} \left[\frac{1}{n}\sum_{i=1}^{n}\sigma\cdot\rho\left(\frac{x_i-\mu}{\sigma}\right)\right]$$

Rem: $(\mu,\sigma) \to \sigma\cdot\rho\left(\frac{x-\mu}{\sigma}\right)$ is jointly convex as it is the **perspective** function of $\rho(x-\cdot)$, see for instance Section 3.2.6, Boyd and Vandenberghe (2004)

Table of Contents

Asymptotic for M-estimation

Assume from now on that $\breve{\mu} = 0$ and that $X \sim F$. Then,

$$\sqrt{n}\hat{\mu}_n \rightarrow_d \mathcal{N}(0, V^2(\psi, F)), \quad \text{where} \quad V^2(\psi, F) = \frac{\mathbb{E}_F\left(\psi(X)^2\right)}{\left(\mathbb{E}_F \psi'(X)\right)^2}$$

Theorem

$$V^2(\psi, F) \geq \frac{1}{\mathbb{E}_F\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]}, \text{ with equality when } \psi \propto -\frac{f'}{f}$$

Rem: equality holds when one choses ψ (or ρ) associated to the MLE, leading to the best asymptotic performance. Though, one needs to know the distribution of F to consider $\rho = -\log(f)$

Sketch of proof

First, note that by integration by part:

$$-\mathbb{E}_F\left(\psi'(X)\right) = -\int \psi'(t)f(t)dt = \int \psi(t)f'(t)dt$$

Sketch of proof

First, note that by integration by part:

$$
\begin{aligned}
-\mathbb{E}_F\left(\psi'(X)\right) &= -\int \psi'(t)f(t)dt = \int \psi(t)f'(t)dt \\
&= \int \psi(t)\frac{f'(t)}{f(t)}f(t)dt = \mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right)
\end{aligned}
$$

Sketch of proof

First, note that by integration by part:

$$\begin{aligned} -\mathbb{E}_F\left(\psi'(X)\right) &= -\int \psi'(t)f(t)dt = \int \psi(t)f'(t)dt \\ &= \int \psi(t)\frac{f'(t)}{f(t)}f(t)dt = \mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right) \end{aligned}$$

Using Cauchy-Schwartz inequality:

Sketch of proof

First, note that by integration by part:

$$
\begin{aligned}
-\mathbb{E}_F\left(\psi'(X)\right) &= -\int \psi'(t)f(t)dt = \int \psi(t)f'(t)dt \\
&= \int \psi(t)\frac{f'(t)}{f(t)}f(t)dt = \mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right)
\end{aligned}
$$

Using Cauchy-Schwartz inequality:

$$
\left[\mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right)\right]^2 = \left(\int \psi(t)\frac{f'(t)}{f(t)}f(t)dt\right)^2
$$

Sketch of proof

First, note that by integration by part:

$$\begin{aligned} -\mathbb{E}_F\left(\psi'(X)\right) &= -\int \psi'(t) f(t) dt = \int \psi(t) f'(t) dt \\ &= \int \psi(t) \frac{f'(t)}{f(t)} f(t) dt = \mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right) \end{aligned}$$

Using Cauchy-Schwartz inequality:

$$\begin{aligned} \left[\mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right)\right]^2 &= \left(\int \psi(t)\frac{f'(t)}{f(t)} f(t) dt\right)^2 \\ &\leq \left(\int \psi^2(t) f(t) dt\right)\left(\int \left(\frac{f'(t)}{f(t)}\right)^2 f(t) dt\right) \end{aligned}$$

Sketch of proof

First, note that by integration by part:

$$
\begin{aligned}
-\mathbb{E}_F\left(\psi'(X)\right) &= -\int \psi'(t)f(t)dt = \int \psi(t)f'(t)dt \\
&= \int \psi(t)\frac{f'(t)}{f(t)}f(t)dt = \mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right)
\end{aligned}
$$

Using Cauchy-Schwartz inequality:

$$
\begin{aligned}
\left[\mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right)\right]^2 &= \left(\int \psi(t)\frac{f'(t)}{f(t)}f(t)dt\right)^2 \\
&\leq \left(\int \psi^2(t)f(t)dt\right)\left(\int \left(\frac{f'(t)}{f(t)}\right)^2 f(t)dt\right) \\
&= \mathbb{E}_F\left(\psi^2(X)\right)\mathbb{E}_F\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]
\end{aligned}
$$

Sketch of proof

First, note that by integration by part:

$$-\mathbb{E}_F\left(\psi'(X)\right) = -\int \psi'(t) f(t) dt = \int \psi(t) f'(t) dt$$
$$= \int \psi(t) \frac{f'(t)}{f(t)} f(t) dt = \mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right)$$

Using Cauchy-Schwartz inequality:

$$\left[\mathbb{E}_F\left(\psi(X)\frac{f'(X)}{f(X)}\right)\right]^2 = \left(\int \psi(t)\frac{f'(t)}{f(t)} f(t) dt\right)^2$$
$$\leq \left(\int \psi^2(t) f(t) dt\right)\left(\int \left(\frac{f'(t)}{f(t)}\right)^2 f(t) dt\right)$$
$$= \mathbb{E}_F\left(\psi^2(X)\right)\mathbb{E}_F\left[\left(\frac{f'(X)}{f(X)}\right)^2\right]$$

Hence, $\mathbb{E}_F\left[\left(\frac{f'(X)}{f(X)}\right)^2\right] \geq \frac{(\mathbb{E}_F \psi'(X))^2}{\mathbb{E}_F(\psi(X)^2)}$ $\square$

Table of Contents

Minimax / Game theory point of view[3]

- Players: Practitioner vs. (adversarial) Nature
- Known parameters: G and $\epsilon \in [0, 1[$ are known (to both nature and practitioner) and samples are drawn according to $F_\epsilon := F = (1 - \epsilon)G + \epsilon H$ with a corruption level ϵ, *i.e.*, $f_\epsilon := f = (1 - \epsilon)g + \epsilon h$, where $-\log(g)$ is convex; *i.e.*, g is **log-concave**
- Objective: the player aims at minimizing the asymptotic variance $V^2(\psi, F)$
- Practitioner's action: picks ψ for "optimal" M-estimation
- Nature's action: picks the distribution H that harms the asymptotic variance the most

[3] P. J. Huber. "Robust estimation of a location parameter". In: *Ann. Math. Statist.* 35 (1964), pp. 73–101.

Equilibrium

Theorem

There exists $F_0 = (1-\epsilon)G + \epsilon H_0$ and ψ_0 s.t.

$$\forall F \text{ s.t. } \mathbb{E}_F(\psi_0) = 0, \quad V^2(\psi_0, F) \leq V^2(\psi_0, F_0) \leq V^2(\psi, F_0)$$

Let $[t_0, t_1]$ be the largest interval such that $|g'/g| \leq \alpha$ and let

$$(1-\epsilon)^{-1} = \int_{t_0}^{t_1} g(t)dt + \frac{g(t_0) + g(t_1)}{\alpha}$$

$$f_0(t) = \begin{cases} (1-\epsilon)g(t_0)e^{\alpha(t-t_0)} & \text{if } t \leq t_0 \\ (1-\epsilon)g(t) & \text{if } t_0 < t < t_1 \\ (1-\epsilon)g(t_1)e^{-\alpha(t-t_1)} & \text{if } t \geq t_1 \end{cases}$$

and $\psi_0 = -f_0'/f_0$ is monotone and bounded by α.

Rem: $V^2(\psi_0, F_0) \leq V^2(\psi, F_0)$ was proved earlier noticing that the best choice is $\psi = \psi_0 := -f_0'/f_0$

Huber loss as an equilibrium

Corollary

Assume that $g(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$. Then the equilibrium is reached for $F_0 = (1-\epsilon)G + \epsilon H_0$ and ψ_0 s.t.

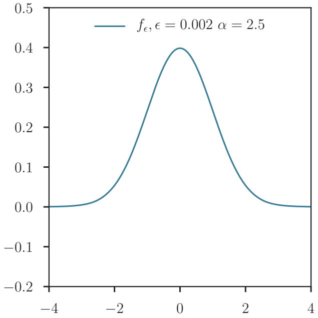
$$(1-\epsilon)^{-1} = \int_{-\alpha}^{\alpha} g(t)dt + \frac{g(-\alpha)+g(\alpha)}{\alpha}$$

$$f_0(t) = \begin{cases} (1-\epsilon)g(-\alpha)e^{\alpha(t+\alpha)} & \text{if } t \leq -\alpha \\ (1-\epsilon)g(t) & \text{if } -\alpha < t < \alpha \\ (1-\epsilon)g(\alpha)e^{-\alpha(t-\alpha)} & \text{if } t \geq \alpha \end{cases}$$

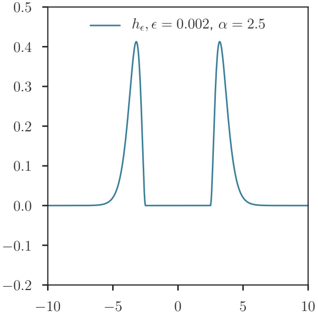
Rem: $\psi_0(x) := -f_0'(x)/f_0(x) = \min(\max(-\alpha, x), \alpha)$ is, up to re-scaling, the Huber ψ_α function introduced earlier

Worst adversarial distribution : Gaussian case

$$f_0(t) \propto \exp(-\rho_\alpha(t))$$

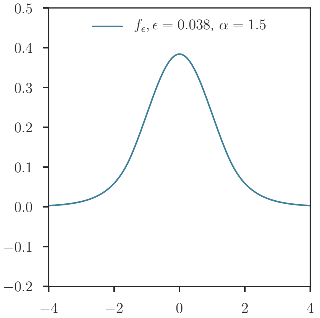


$$h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$$

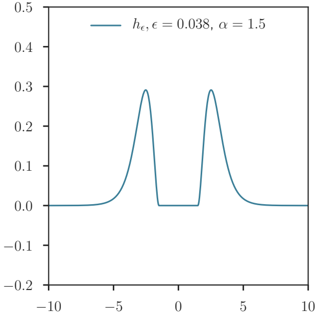


Worst adversarial distribution : Gaussian case

$$f_0(t) \propto \exp(-\rho_\alpha(t))$$

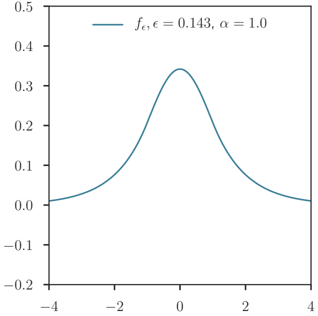


$$h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$$

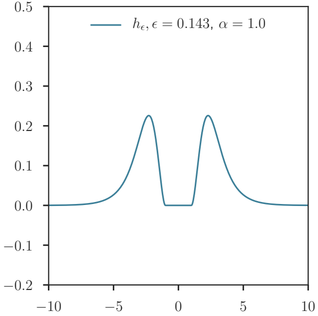


Worst adversarial distribution : Gaussian case

$$f_0(t) \propto \exp(-\rho_\alpha(t))$$

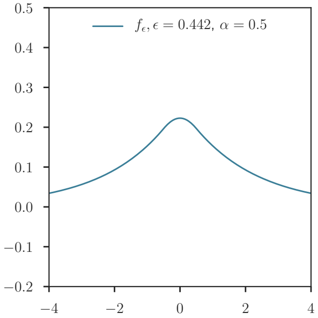


$$h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$$

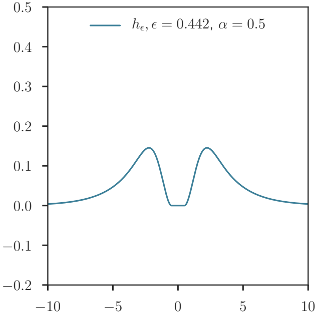


Worst adversarial distribution : Gaussian case

$$f_0(t) \propto \exp(-\rho_\alpha(t))$$

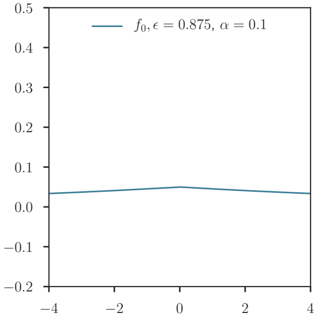


$$h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$$

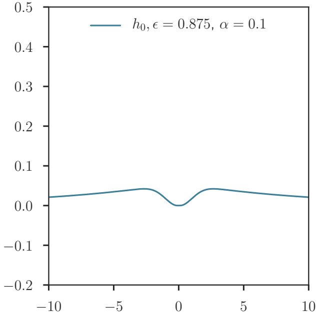


Worst adversarial distribution : Gaussian case

$$f_0(t) \propto \exp(-\rho_\alpha(t))$$



$$h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$$



Proof of theorem

$$\frac{1}{1-\epsilon} = \int_{t_0}^{t_1} g(t)dt + \frac{g(t_0)+g(t_1)}{\alpha}$$

$$f_0(t) = \begin{cases} (1-\epsilon)g(t_0)e^{\alpha(t-t_0)} & \text{if } t \leq t_0 \\ (1-\epsilon)g(t) & \text{if } t_0 < t < t_1 \\ (1-\epsilon)g(t_1)e^{-\alpha(t-t_1)} & \text{if } t \geq t_1 \end{cases}$$

Proof of theorem

$$\frac{1}{1-\epsilon} = \int_{t_0}^{t_1} g(t)dt + \frac{g(t_0)+g(t_1)}{\alpha}$$

$$f_0(t) = \begin{cases} (1-\epsilon)g(t_0)e^{\alpha(t-t_0)} & \text{if } t \le t_0 \\ (1-\epsilon)g(t) & \text{if } t_0 < t < t_1 \\ (1-\epsilon)g(t_1)e^{-\alpha(t-t_1)} & \text{if } t \ge t_1 \end{cases}$$

Fact 1: f_0 is a p.d.f.

Proof of theorem

$$\frac{1}{1-\epsilon} = \int_{t_0}^{t_1} g(t)dt + \frac{g(t_0)+g(t_1)}{\alpha}$$

$$f_0(t) = \begin{cases} (1-\epsilon)g(t_0)e^{\alpha(t-t_0)} & \text{if } t \leq t_0 \\ (1-\epsilon)g(t) & \text{if } t_0 < t < t_1 \\ (1-\epsilon)g(t_1)e^{-\alpha(t-t_1)} & \text{if } t \geq t_1 \end{cases}$$

Fact 1: f_0 is a p.d.f.

Proof:

- f is non-negative since g is non-negative
- $\int f_0(t)dt = 1$ since ϵ is constructed for that

Proof continued

Fact 2: $h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$ defined below is a p.d.f.

$$h_0(t) := \begin{cases} \frac{1-\epsilon}{\epsilon}[g(t_0)e^{\alpha(t-t_0)} - g(t)] & \text{if } t \leq t_0 \\ 0 & \text{if } t_0 < t < t_1 \\ \frac{1-\epsilon}{\epsilon}[g(t_1)e^{-\alpha(t-t_1)} - g(t)] & \text{if } t \geq t_1 \end{cases}$$

Proof continued

Fact 2: $h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$ defined below is a p.d.f.

$$h_0(t) := \begin{cases} \frac{1-\epsilon}{\epsilon}[g(t_0)e^{\alpha(t-t_0)} - g(t)] & \text{if } t \le t_0 \\ 0 & \text{if } t_0 < t < t_1 \\ \frac{1-\epsilon}{\epsilon}[g(t_1)e^{-\alpha(t-t_1)} - g(t)] & \text{if } t \ge t_1 \end{cases}$$

Proof: $\int h_0 = \frac{1}{\epsilon}\int [f_0 - (1-\epsilon)g] = 1$ since $\int f_0 = \int g = 1$.

Proof continued

Fact 2: $h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$ defined below is a p.d.f.

$$h_0(t) := \begin{cases} \frac{1-\epsilon}{\epsilon}[g(t_0)e^{\alpha(t-t_0)} - g(t)] & \text{if } t \le t_0 \\ 0 & \text{if } t_0 < t < t_1 \\ \frac{1-\epsilon}{\epsilon}[g(t_1)e^{-\alpha(t-t_1)} - g(t)] & \text{if } t \ge t_1 \end{cases}$$

Proof: $\int h_0 = \frac{1}{\epsilon}\int[f_0 - (1-\epsilon)g] = 1$ since $\int f_0 = \int g = 1$.

Now $-\log(g)$ is convex so this function is lower bounded by its tangent at t_0, and for any $t \le t_0$

Proof continued

Fact 2: $h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$ defined below is a p.d.f.

$$h_0(t) := \begin{cases} \frac{1-\epsilon}{\epsilon}[g(t_0)e^{\alpha(t-t_0)} - g(t)] & \text{if } t \leq t_0 \\ 0 & \text{if } t_0 < t < t_1 \\ \frac{1-\epsilon}{\epsilon}[g(t_1)e^{-\alpha(t-t_1)} - g(t)] & \text{if } t \geq t_1 \end{cases}$$

Proof: $\int h_0 = \frac{1}{\epsilon}\int[f_0 - (1-\epsilon)g] = 1$ since $\int f_0 = \int g = 1$.

Now $-\log(g)$ is convex so this function is lower bounded by its tangent at t_0, and for any $t \leq t_0$

$$-\log(g)(t) \geq -\log(g)(t_0) + \frac{\partial}{\partial t}[-\log g(t_0)](t - t_0)$$

Proof continued

<u>Fact 2</u>: $h_0 := \frac{1}{\epsilon}[f_0 - (1-\epsilon)g]$ defined below is a p.d.f.

$$h_0(t) := \begin{cases} \frac{1-\epsilon}{\epsilon}[g(t_0)e^{\alpha(t-t_0)} - g(t)] & \text{if } t \leq t_0 \\ 0 & \text{if } t_0 < t < t_1 \\ \frac{1-\epsilon}{\epsilon}[g(t_1)e^{-\alpha(t-t_1)} - g(t)] & \text{if } t \geq t_1 \end{cases}$$

<u>Proof:</u> $\int h_0 = \frac{1}{\epsilon}\int[f_0 - (1-\epsilon)g] = 1$ since $\int f_0 = \int g = 1$.

Now $-\log(g)$ is convex so this function is lower bounded by its tangent at t_0, and for any $t \leq t_0$

$$\begin{aligned} -\log(g)(t) &\geq -\log(g)(t_0) + \frac{\partial}{\partial t}[-\log g(t_0)](t - t_0) \\ &\geq -\log(g)(t_0) - \alpha(t - t_0) \end{aligned}$$

where we have used $\frac{\partial}{\partial t}[-\log g(t_0)] = -g'(t_0)/g(t_0) \geq -\alpha$. Hence $h_0(t) \geq 0$ when $t \leq t_0$; similarly $h_0(t) \geq 0$ when $t \geq t_1$.

Proof (continued II)

Fact 4:

$$V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof (continued II)

Fact 4:

$$V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof:

$$V^2(\psi_0, F_0) = \frac{\mathbb{E}_{F_0}\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_{F_0}\psi_0'(X)\right)^2}$$

Proof (continued II)

Fact 4:

$$V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof:

$$V^2(\psi_0, F_0) = \frac{\mathbb{E}_{F_0}\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_{F_0}\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_{H_0}\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_{H_0}\psi_0'(X)\right]^2}$$

Proof (continued II)

Fact 4:

$$V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof:

$$V^2(\psi_0, F_0) = \frac{\mathbb{E}_{F_0}\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_{F_0}\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_{H_0}\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_{H_0}\psi_0'(X)\right]^2}$$

Then, reminding $\psi_0(x) := -\frac{f_0'(x)}{f_0(x)} = \min\left(\max\left(-\alpha, -\frac{g'(x)}{g(x)}\right), \alpha\right)$

Proof (continued II)

Fact 4:

$$V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof:

$$V^2(\psi_0, F_0) = \frac{\mathbb{E}_{F_0}\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_{F_0}\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_{H_0}\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_{H_0}\psi_0'(X)\right]^2}$$

Then, reminding $\psi_0(x) := -\frac{f_0'(x)}{f_0(x)} = \min\left(\max\left(-\alpha, -\frac{g'(x)}{g(x)}\right), \alpha\right)$

- $|\psi_0(t)| = \alpha$ and $\psi_0'(t) = 0$ for $t \notin [t_0, t_1]$

Proof (continued II)

Fact 4:

$$V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof:

$$V^2(\psi_0, F_0) = \frac{\mathbb{E}_{F_0}\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_{F_0}\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_{H_0}\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_{H_0}\psi_0'(X)\right]^2}$$

Then, reminding $\psi_0(x) := -\frac{f_0'(x)}{f_0(x)} = \min\left(\max\left(-\alpha, -\frac{g'(x)}{g(x)}\right), \alpha\right)$

- $|\psi_0(t)| = \alpha$ and $\psi_0'(t) = 0$ for $t \notin [t_0, t_1]$
- $h_0(t) = 0$ for $t \in [t_0, t_1]$

Proof (continued II)

Fact 4:

$$V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof:

$$V^2(\psi_0, F_0) = \frac{\mathbb{E}_{F_0}\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_{F_0}\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_{H_0}\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_{H_0}\psi_0'(X)\right]^2}$$

Then, reminding $\psi_0(x) := -\frac{f_0'(x)}{f_0(x)} = \min\left(\max\left(-\alpha, -\frac{g'(x)}{g(x)}\right), \alpha\right)$

- $|\psi_0(t)| = \alpha$ and $\psi_0'(t) = 0$ for $t \notin [t_0, t_1]$
- $h_0(t) = 0$ for $t \in [t_0, t_1]$

Hence, $\mathbb{E}_{H_0}\left[\psi_0(X)^2\right] = \alpha^2$ and $\mathbb{E}_{H_0}\psi_0'(X) = 0$, and

$$V^2(\psi_0, F_0) \leq \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

$\square$

Proof (continued III)

Fact 5: for any F with $\mathbb{E}_F(\psi_0) = 0$

$$V^2(\psi_0, F) \le V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof (continued III)

Fact 5: for any F with $\mathbb{E}_F(\psi_0) = 0$

$$V^2(\psi_0, F) \leq V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof: fix H so

$$V^2(\psi_0, F) = \frac{\mathbb{E}_F\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_F\psi_0'(X)\right)^2}$$

Proof (continued III)

Fact 5: for any F with $\mathbb{E}_F(\psi_0) = 0$

$$V^2(\psi_0, F) \leq V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof: fix H so

$$V^2(\psi_0, F) = \frac{\mathbb{E}_F\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_F\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_H\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_H\psi_0'(X)\right]^2}$$

Proof (continued III)

Fact 5: for any F with $\mathbb{E}_F(\psi_0) = 0$

$$V^2(\psi_0, F) \leq V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof: fix H so

$$V^2(\psi_0, F) = \frac{\mathbb{E}_F\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_F\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_H\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_H\psi_0'(X)\right]^2}$$

Then, reminding $\psi_0(x) := -\frac{f_0'(x)}{f_0(x)} = \min\left(\max\left(-\alpha, -\frac{g'(x)}{g(x)}\right), \alpha\right)$

Proof (continued III)

Fact 5: for any F with $\mathbb{E}_F(\psi_0) = 0$

$$V^2(\psi_0, F) \leq V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G \psi_0'(X)\right]^2}$$

Proof: fix H so

$$V^2(\psi_0, F) = \frac{\mathbb{E}_F\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_F \psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_H\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G \psi_0'(X) + \epsilon\mathbb{E}_H \psi_0'(X)\right]^2}$$

Then, reminding $\psi_0(x) := -\frac{f_0'(x)}{f_0(x)} = \min\left(\max\left(-\alpha, -\frac{g'(x)}{g(x)}\right), \alpha\right)$

- $|\psi_0(x)| \leq \alpha$ and $\mathbb{E}_H\left[\psi_0(X)^2\right] \leq \alpha^2$

Proof (continued III)

Fact 5: for any F with $\mathbb{E}_F(\psi_0) = 0$

$$V^2(\psi_0, F) \le V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof: fix H so

$$V^2(\psi_0, F) = \frac{\mathbb{E}_F\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_F\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_H\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_H\psi_0'(X)\right]^2}$$

Then, reminding $\psi_0(x) := -\frac{f_0'(x)}{f_0(x)} = \min\left(\max\left(-\alpha, -\frac{g'(x)}{g(x)}\right), \alpha\right)$

- $|\psi_0(x)| \le \alpha$ and $\mathbb{E}_H\left[\psi_0(X)^2\right] \le \alpha^2$
- $\psi_0' \ge 0$ since $\frac{\partial^2}{\partial t^2}[-\log g(t)] \ge 0$ by convexity of $-\log g$

Proof (continued III)

Fact 5: for any F with $\mathbb{E}_F(\psi_0) = 0$

$$V^2(\psi_0, F) \leq V^2(\psi_0, F_0) = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

Proof: fix H so

$$V^2(\psi_0, F) = \frac{\mathbb{E}_F\left[\psi_0(X)^2\right]}{\left(\mathbb{E}_F\psi_0'(X)\right)^2} = \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\mathbb{E}_H\left[\psi_0(X)^2\right]}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X) + \epsilon\mathbb{E}_H\psi_0'(X)\right]^2}$$

Then, reminding $\psi_0(x) := -\frac{f_0'(x)}{f_0(x)} = \min\left(\max\left(-\alpha, -\frac{g'(x)}{g(x)}\right), \alpha\right)$

- $|\psi_0(x)| \leq \alpha$ and $\mathbb{E}_H\left[\psi_0(X)^2\right] \leq \alpha^2$
- $\psi_0' \geq 0$ since $\frac{\partial^2}{\partial t^2}[-\log g(t)] \geq 0$ by convexity of $-\log g$

Hence, $\mathbb{E}_H\psi_0'(X) \geq 0$ for any H, and

$$V^2(\psi_0, F) \leq \frac{(1-\epsilon)\mathbb{E}_G\left[\psi_0(X)^2\right] + \epsilon\alpha^2}{\left[(1-\epsilon)\mathbb{E}_G\psi_0'(X)\right]^2}$$

□

Complements on Huber function

- More on variational formulations : Section 2.4, Hampel *et al.* (1986) after we introduce influence functions
- Connections with convex analysis and smoothing for non-smooth function, as in Nesterov (2005) Beck and Teboulle (2012), will be made later

References I

- Beck, A. and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.
- Boyd, S. and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004, pp. xiv+716.
- Hampel, F. R. et al. *Robust statistics: The Approach Based on Influence Functions*. Wiley series in probability and statistics. Wiley, 1986.
- Huber, P. J. "Robust estimation of a location parameter". In: *Ann. Math. Statist.* 35 (1964), pp. 73–101.
- – .*Robust Statistics*. John Wiley & Sons Inc., 1981.
- Koenker, R. and G. Bassett. "Regression quantiles". In: *Econometrica* 46.1 (1978), pp. 33–50.
- Maronna, R. A., R. D. Martin, and V. J. Yohai. *Robust statistics: Theory and methods*. Chichester: John Wiley & Sons, 2006.

References II

- Nesterov, Y. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.