Robust tests for the comparison of regression curves

Juan Carlos Pardo-Fernández

Universidade de Vigo

Joint work with **Graciela Boente** (Universidad de Buenos Aires)

Symposium in honor to Ingrid Van Keilegom
A Coruña – June 15th, 2022

Universidade Vigo

Departamento de Estatística e Investigación Operativa

Fac. CC. Económicas e Empresariais
Campus de Vigo
E-36310 Vigo

Tel. 986 130 313
depc05.webs.uvigo.gal
udeconomicas@uvigo.gal

Vista a solicitude do Departamento de Matemáticas da Universidade da Coruña, o Departamento de Estatística e Investigación Operativa da Universidade de Vigo, reunido en Consello de Departamento o día 7 de outubro de 2021, acordou manifestar o seu apoio institucional á candidatura da profesora Ingrid Van Keilegom como Doutora Honoris Causa pola Universidade da Coruña.

Ao longo dos últimos case vinte anos, a profesora Van Keilegom mantivo e segue mantendo unha intensa colaboración científica con varios membros do noso Departamento. Visitounos en numerosas ocasións, impartiu seminarios e cursos de doutoramento e tamén acolleu a profesores e estudantes de doutoramento na Université catholique de Louvain e na Katholieke Universiteit Leuven (Bélxica).

Tal e como se pode comprobar no seu currículum vítae, os seus traballos foron publicados nas máis prestixiosas revistas da nosa área. A súa valía científica é amplamente recoñecida no campo da estatística non paramétrica, semiparamétrica e na análise de supervivencia.

En Vigo, a 7 de outubro de 2021

Juan Carlos Pardo Fernández
Director do Departamento de Estatística e Investigación Operativa
Universidade de Vigo


Bouw van een noodgemaal op de Molenbeek
verlast afkomstig v/d Molenbeek en Grote Beek
Stad Aalst - dienst openbare werken
Polder v/d Beneden - Dender met steun
nciebestuur, de streek Dendermonde en Aalst
het Vlaams gewest en Aquafin


Bouw van een noodgemaal op de Molenbeek
verlast afkomstig v/d Molenbeek en Grote Beek
Stad Aalst - dienst openbare werken
Polder v d Beneden - Dender met steun
nciebestuur, de streek Dendermonde en Aalst
het Vlaams gewest en Aquafin


Brugse Straffe Hendrik
Enfin
Pour les amateurs
Enfin
Pour les amateurs une sélection de Vins


dreft


On (semi)parametric modelling, estimation and inference for survival data subject to dependent censoring
Negera Wakgari Deresa
Ingrid Van Keilegom
KU Leuven
October 25, 2019
XIV Congreso Galego
de Estatística
e Investigación
de Operacións
SGAPEIO2019


HSBC

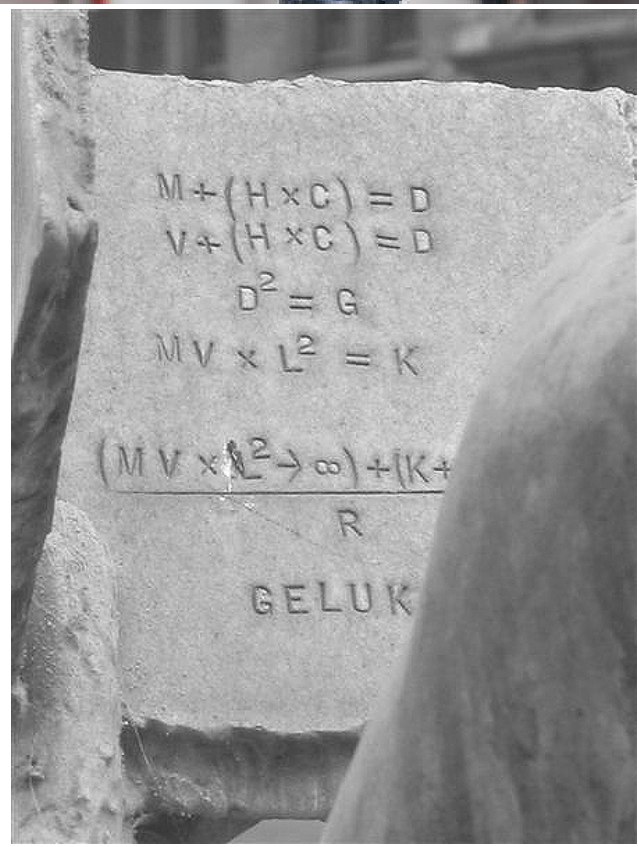

HSBC


SANI PASS TOURS


SANI PASS TOURS


SANI PASS TOURS


Highest Pub In Africa
TOP CHALET


M+(H×C)=D
V+(H×C)=D
D²=G
MV×L²=K
GELUK

– Outline of the talk

Robust estimation

Comparison of regression curves

A test based on characteristic functions

Robust version of the test

Some simulations

– Robust estimation [Maronna, Martin and Yohai (2006). Robust statistics]

Location model: $X = \mu + \varepsilon$

Assumption: ε and $-\varepsilon$ have the same distribution (no systematic errors), or, in other words, the distribution of ε is **symmetric around** 0.

Note that **no moments** are assumed.

Data: $\{X_i = \mu + \varepsilon_i, i = 1, \ldots, n\}$, where $\varepsilon_1, \ldots, \varepsilon_n$ is an i.i.d. sample of ε.

Objective: estimate μ, that is, find $\hat{\mu}$ "close to" μ.

– Robust estimation [Maronna, Martin and Yohai (2006). Robust statistics]

Location model: $X = \mu + \varepsilon$

Assumption: ε and $-\varepsilon$ have the same distribution (no systematic errors), or, in other words, the distribution of ε is **symmetric around** 0.

Note that **no moments** are assumed.

Data: $\{X_i = \mu + \varepsilon_i, i = 1, \ldots, n\}$, where $\varepsilon_1, \ldots, \varepsilon_n$ is an i.i.d. sample of ε.

Objective: estimate μ, that is, find $\hat{\mu}$ "close to" μ.

Well-behaved data (moments exist, no outliers): take the **least squares estimator**, that is, the solution to the minimization problem

$$\arg\min_a \sum_{i=1}^{n} (X_i - a)^2.$$

After taking derivatives, the solution satisfies the equation

$$\sum_{i=1}^{n} (X_i - a) = 0,$$

and is the sample mean $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$.

– Robust estimation

Other loss-functions can be considered.

In general, **M-estimators** are solutions of the problem

$$\arg\min_a \sum_{i=1}^{n} \rho(X_i - a),$$

where ρ is a **positive function**. At a population level, this means

$$\arg\min_a \mathbb{E}[\rho(X - a)].$$

– Robust estimation

Other loss-functions can be considered.

In general, **M-estimators** are solutions of the problem

$$\arg\min_a \sum_{i=1}^{n} \rho(X_i - a),$$

where ρ is a **positive function**. At a population level, this means

$$\arg\min_a \mathbb{E}[\rho(X - a)].$$

If ρ is differentiable with $\rho'(x) = \Psi(x)$ (usually called **score function**), then the solution, $\hat{\mu}_{\mathrm{R}}$ satisfies

$$\sum_{i=1}^{n} \Psi(X_i - \hat{\mu}_{\mathrm{R}}) = 0.$$

Or, at a population level

$$\mathbb{E}[\Psi(X - \hat{\mu}_{\mathrm{R}})] = 0.$$

– Robust estimation

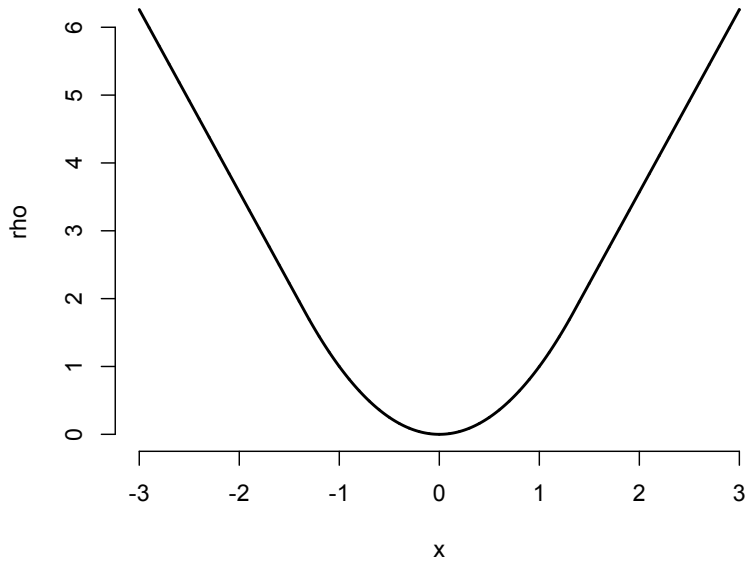
Examples of ρ-functions:

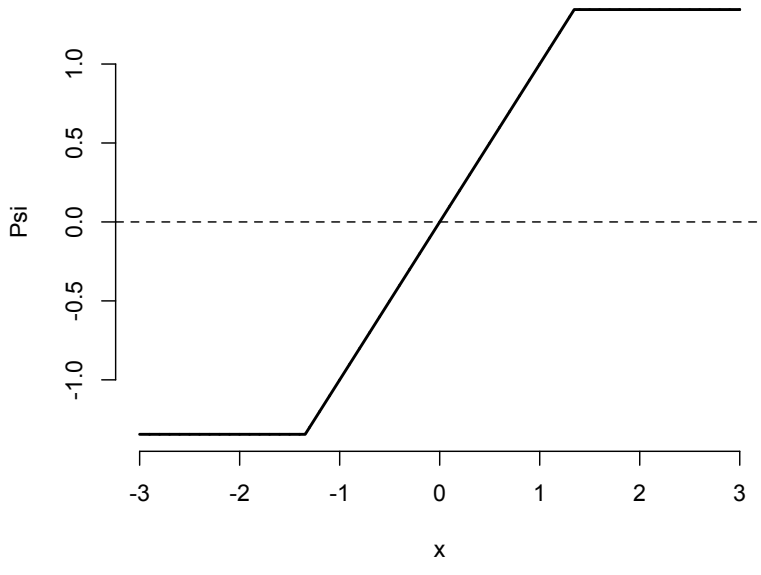
- $\rho(x) = x^2$, then the M-estimator is the **sample mean**
- $\rho(x) = |x|$, then the M-estimator is the **sample median**
- Huber's function
- Tukey's bisquare function
- etc.

– Robust estimation

Example: **Huber's ρ- and Ψ-functions**

$$\rho(x) = \begin{cases} x^2 & \text{if } |x| \leq k \\ 2k|x| - k^2 & \text{if } |x| > k \end{cases} \qquad \Psi(x) = \begin{cases} x & \text{if } |x| \leq k \\ \operatorname{sgn}(x)k & \text{if } |x| > k \end{cases}$$

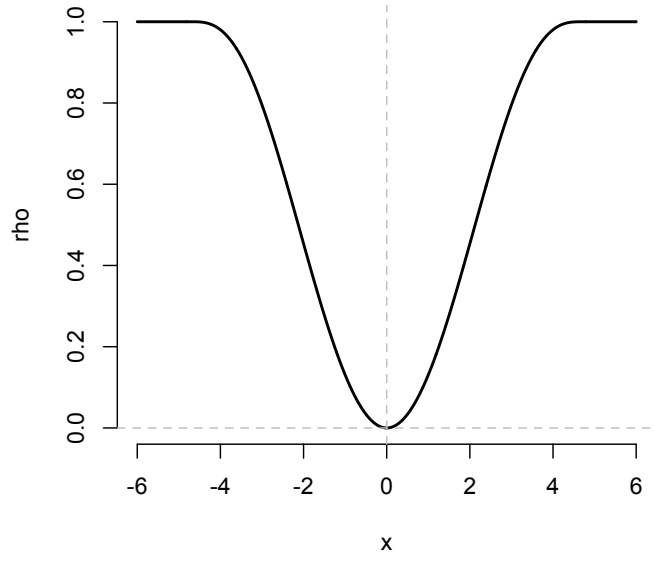


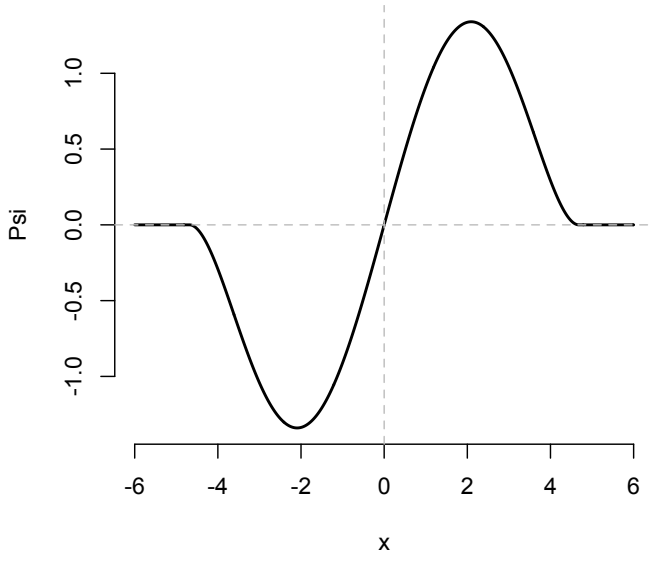


– Robust estimation

Example: **Tukey's biweight ρ- and Ψ-functions**

$$\rho(x) = \begin{cases} 1-(1-(x/k)^2)^3 & \text{if } |x| \le k \\ 1 & \text{if } |x| > k \end{cases} \qquad \Psi(x) = \begin{cases} x(1-(x/k)^2)^2 & \text{if } |x| \le k \\ 0 & \text{if } |x| > k \end{cases}$$

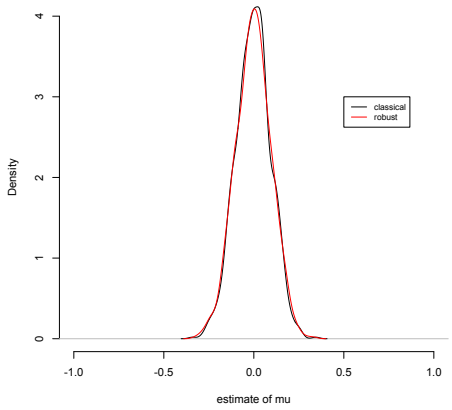




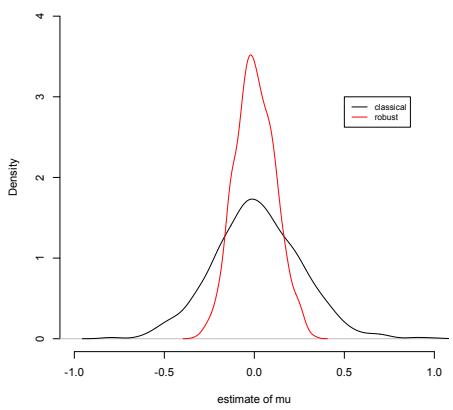
– Robust estimation

Estimated densities of $\hat{\mu}_{\mathrm{CL}} = \bar{X}$ (in black) and the robust M-estimator based on Huber's function with $k = 1.345$, $\hat{\mu}_{\mathrm{R}}$ (in red), obtained from 1000 simulated data sets of size $n = 100$.

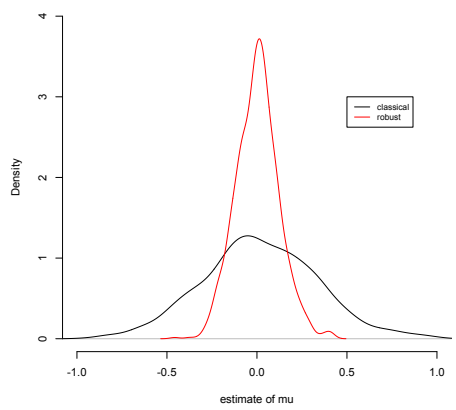
$$X \sim N(0,1) \qquad X \sim \begin{cases} N(0,1) \text{ w.p. } 0.95 \\ N(0,10) \text{ w.p. } 0.05 \end{cases} \qquad X \sim \begin{cases} N(0,1) \text{ w.p. } 0.90 \\ N(0,10) \text{ w.p. } 0.10 \end{cases}$$



$$nVar(\hat{\mu}_{\mathrm{CL}}) = 0.991$$
$$nVar(\hat{\mu}_{\mathrm{R}}) = 1.055$$



$$nVar(\hat{\mu}_{\mathrm{CL}}) = 5.897$$
$$nVar(\hat{\mu}_{\mathrm{R}}) = 1.225$$



$$nVar(\hat{\mu}_{\mathrm{CL}}) = 10.142$$
$$nVar(\hat{\mu}_{\mathrm{R}}) = 1.439$$

– Robust estimation: nonparametric regression

Let (X, Y) be a random vector that follows the **homoscedastic nonparametric regression model** given by

$$Y = m(X) + \sigma\, \varepsilon$$

where

- m is a nonparametric smooth function,
- the error ε is independent of the covariate X, has a **symmetric** distribution and has scale 1.
- σ is the scale.

– Robust estimation: nonparametric regression

Let (X, Y) be a random vector that follows the **homoscedastic nonparametric regression model** given by

$$Y = m(X) + \sigma\, \varepsilon$$

where

- m is a nonparametric smooth function,
- the error ε is independent of the covariate X, has a **symmetric** distribution and has scale 1.
- σ is the scale.

Robust estimation: **avoid moment conditions on** ε.

– Robust estimation: nonparametric regression

Let Ψ be a **bounded** and continuous function and define

$$\lambda(x, a, \sigma) = \mathbb{E}\left[\Psi\left(\frac{Y-a}{\sigma}\right) \mid X = x\right].$$

Note that under our regression model, if Ψ is an **odd** function –such as, for example, Tukey's bisquare function– and the error has a **symmetric** distribution, then

$$\lambda(x, m(x), \sigma) = \mathbb{E}\left[\Psi\left(\frac{Y-m(x)}{\sigma}\right) \mid X = x\right] = \mathbb{E}\left[\Psi\left(\varepsilon\right)\right] = 0.$$

for any $\sigma > 0$.

– Robust estimation: nonparametric regression

Let Ψ be a **bounded** and continuous function and define

$$\lambda(x, a, \sigma) = \mathbb{E}\left[\Psi\left(\frac{Y-a}{\sigma}\right) \mid X = x\right].$$

Note that under our regression model, if Ψ is an **odd** function –such as, for example, Tukey's bisquare function– and the error has a **symmetric** distribution, then

$$\lambda(x, m(x), \sigma) = \mathbb{E}\left[\Psi\left(\frac{Y-m(x)}{\sigma}\right) \mid X = x\right] = \mathbb{E}\left[\Psi\left(\varepsilon\right)\right] = 0.$$

for any $\sigma > 0$.

Data: i.i.d observations (X_i, Y_i), $i = 1, \ldots, n$, from (X, Y).

Boente and Fraiman (1989, *JMVA*) proposed the **robust nonparametric estimator of** $m(x)$ as

$$\text{the solution } \widehat{m}(x) \text{ of the equation } \widehat{\lambda}(x, \widehat{m}(x), \widehat{\sigma}) = 0,$$

where

$$\widehat{\lambda}(x, a, \sigma) = \sum_{i=1}^{n} K_h\left(x - X_i\right) \Psi\left(\frac{Y_i - a}{\sigma}\right).$$

– Robust estimation: nonparametric regression

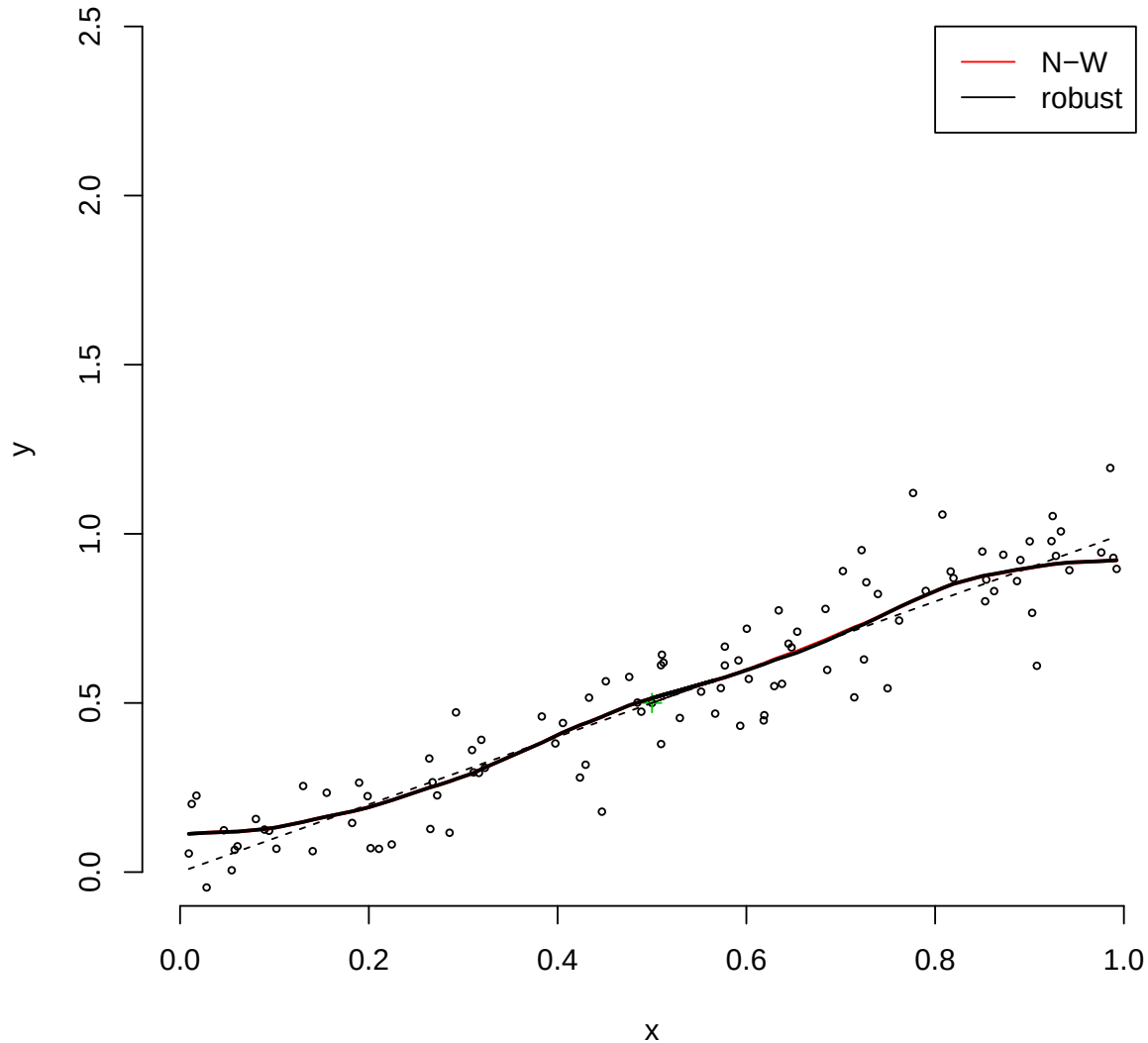
Under the homoscedastic regression model, **robust root$-n$ estimators for the scale σ** can be obtained.

Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the ordered statistics of the explanatory variable and denote as $(X_{(1)}, Y_{D_1}), \ldots, (X_{(n_j)}, Y_{D_n})$ the sample of observations ordered according to the values of the explanatory variable, that is, $X_{(\ell)} = X_{D_\ell}$.

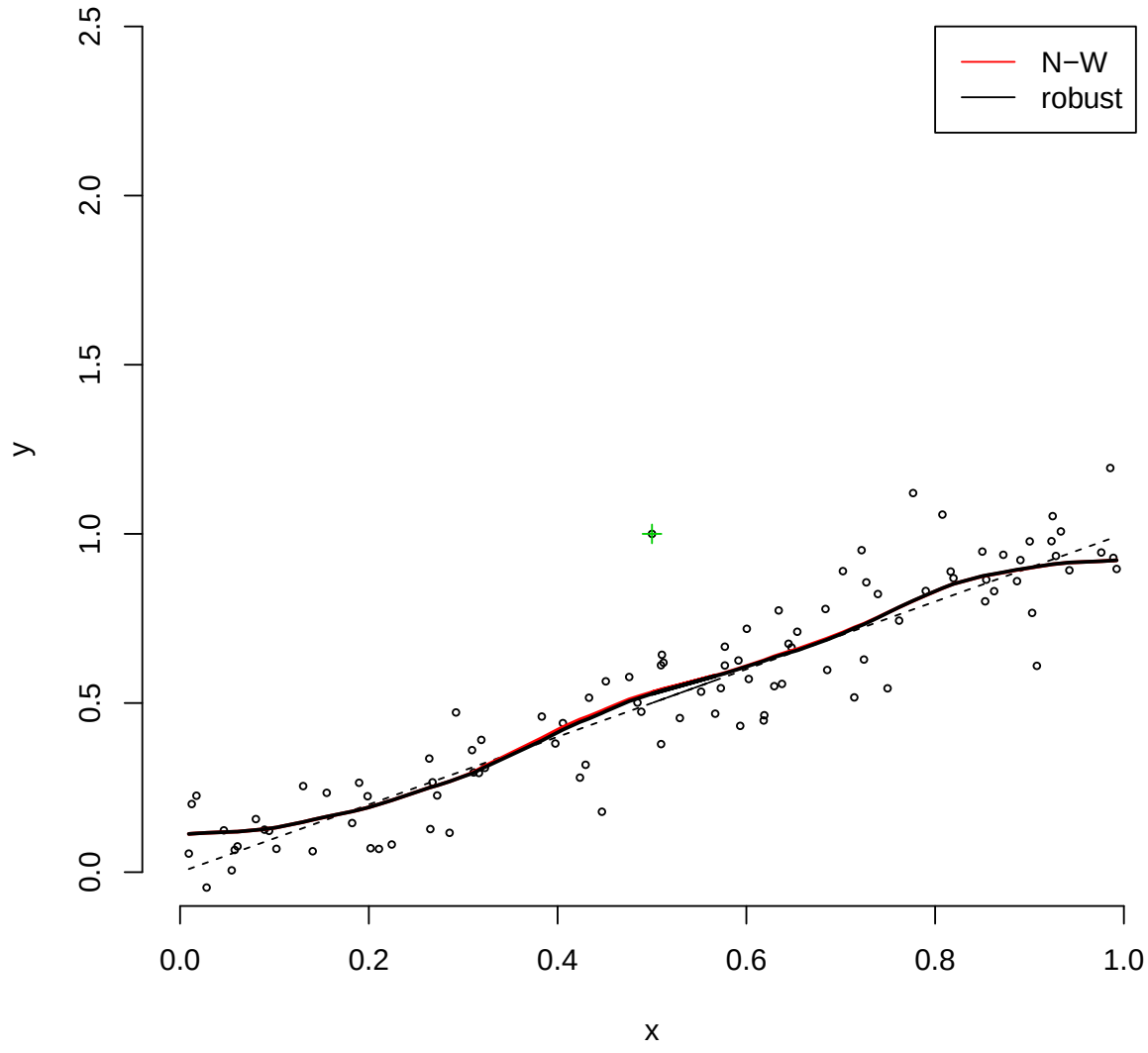
A robust consistent root$-n$ estimator of σ can be obtained as

$$\widehat{\sigma} = \frac{1}{\sqrt{2}\Phi^{-1}(3/4)} \underset{1 \leq \ell \leq n-1}{\text{median}} \left| Y_{D_{\ell+1}} - Y_{D_\ell} \right|.$$

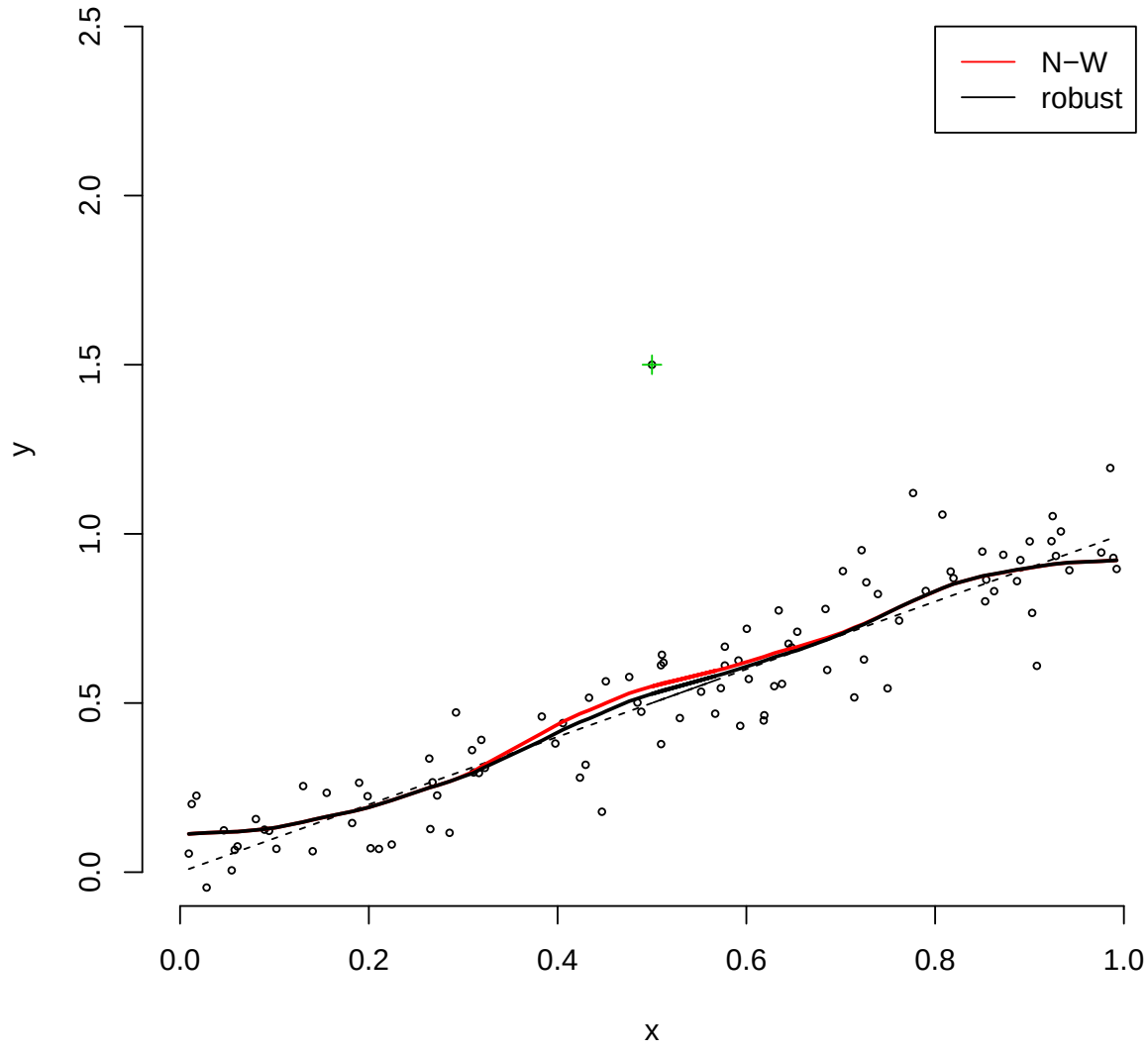
– Effect of outliers in estimation



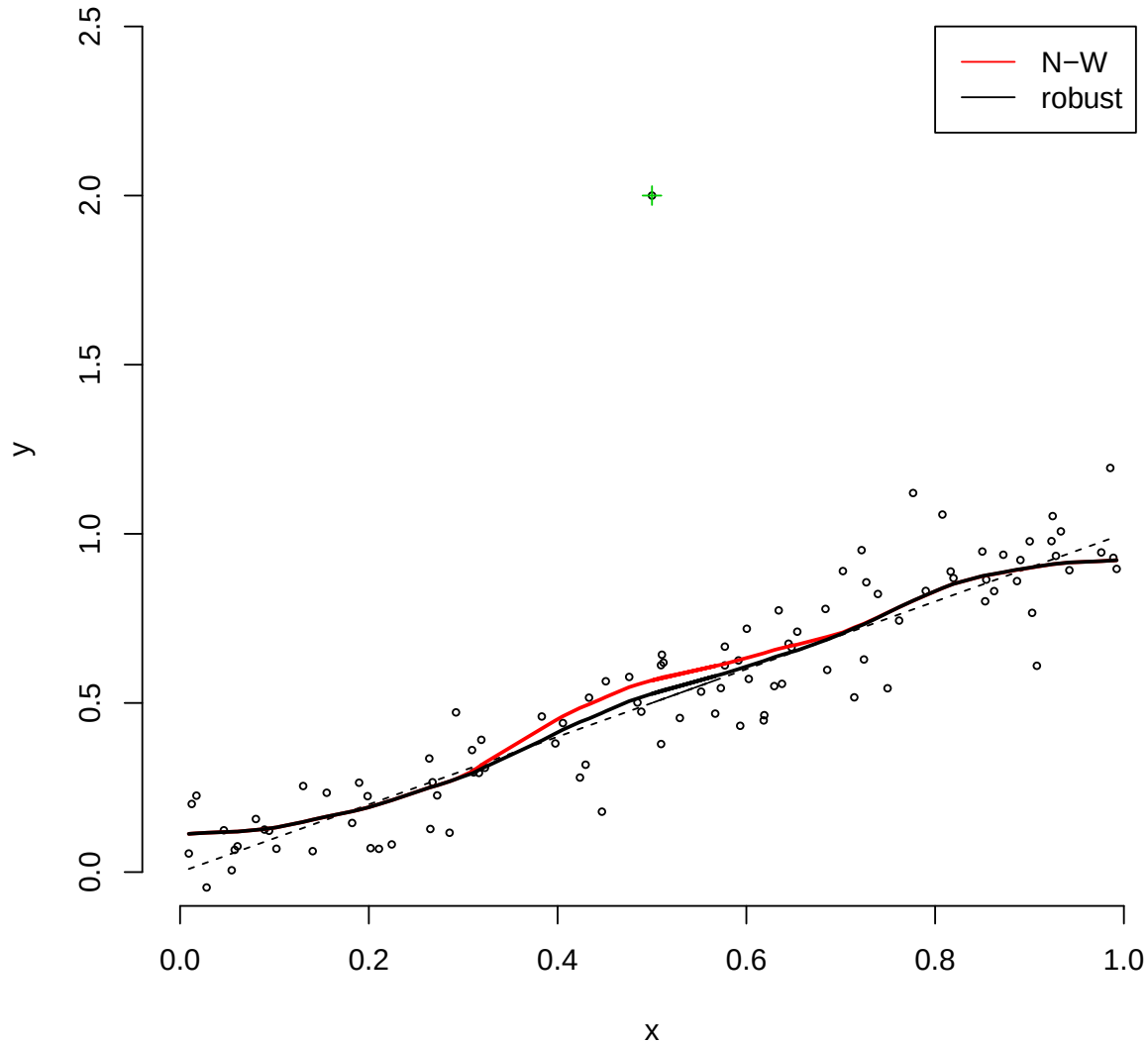
– Effect of outliers in estimation



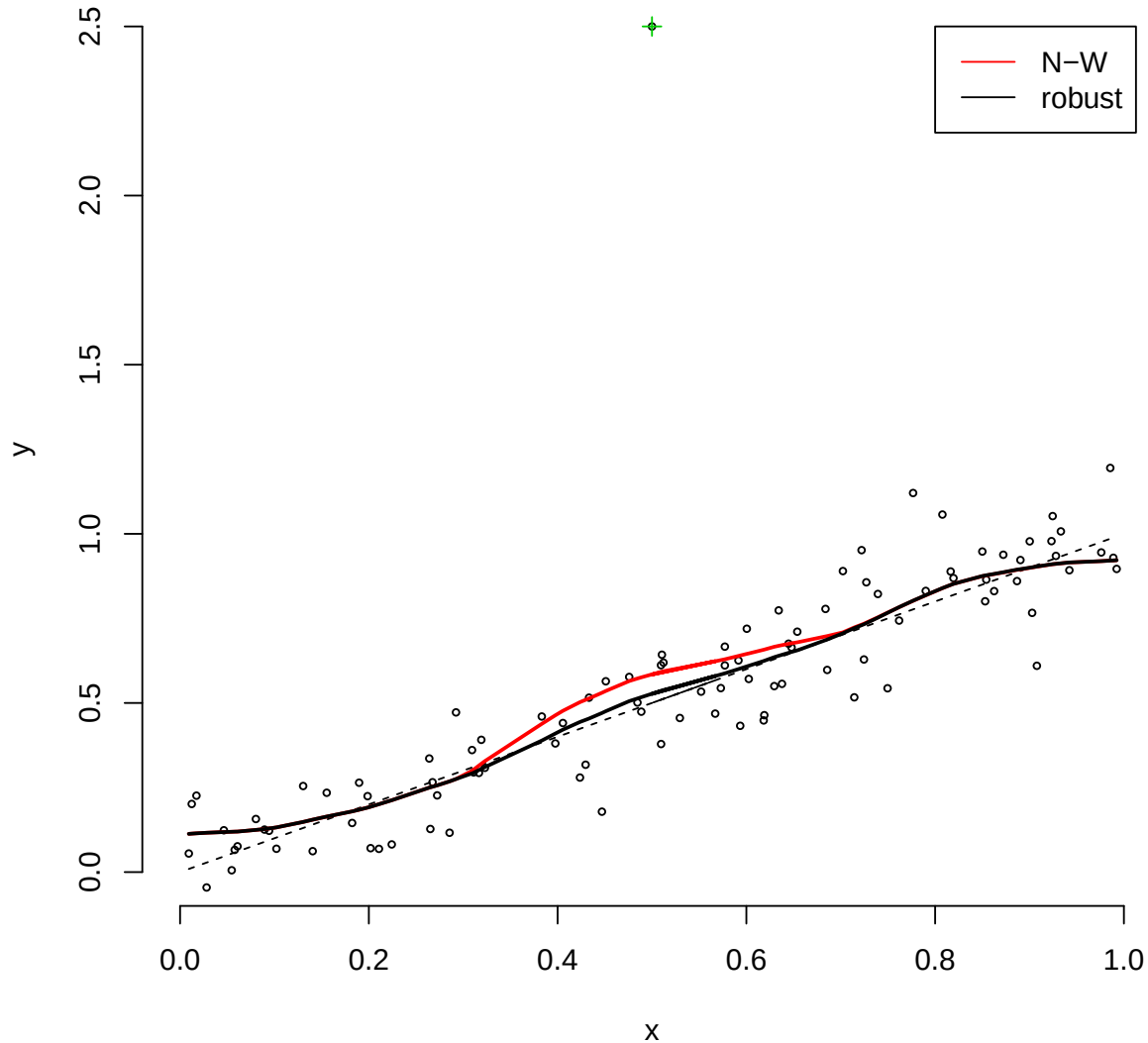
– Effect of outliers in estimation



– Effect of outliers in estimation



– Effect of outliers in estimation



– Effect of outliers in testing

Test: $H_0 : m_1 = m_2$ vs $H_1 : m_1 < m_2$

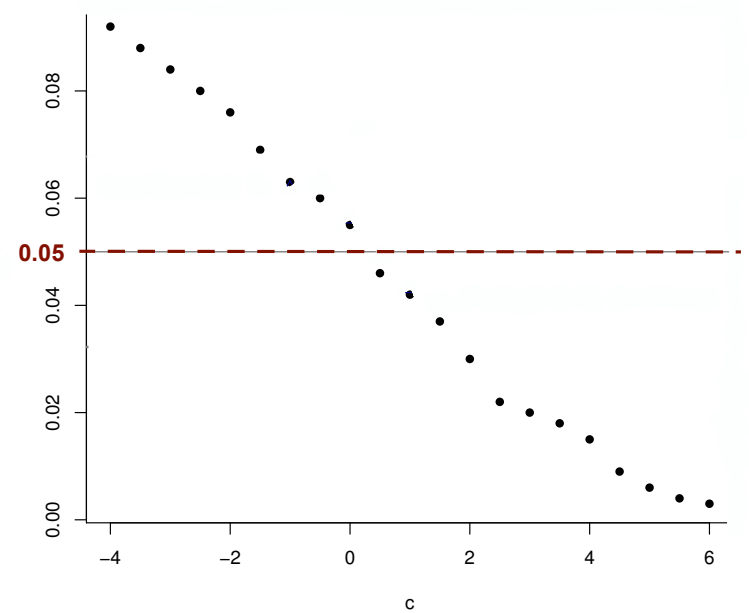
Method in Neumeyer and PF (2009, *JSPI*)

– Effect of outliers in testing

Test: $H_0 : m_1 = m_2$ vs $H_1 : m_1 < m_2$

Method in Neumeyer and PF (2009, *JSPI*)

The graph shows the **empirical size** of a test depending on **one** outlier.

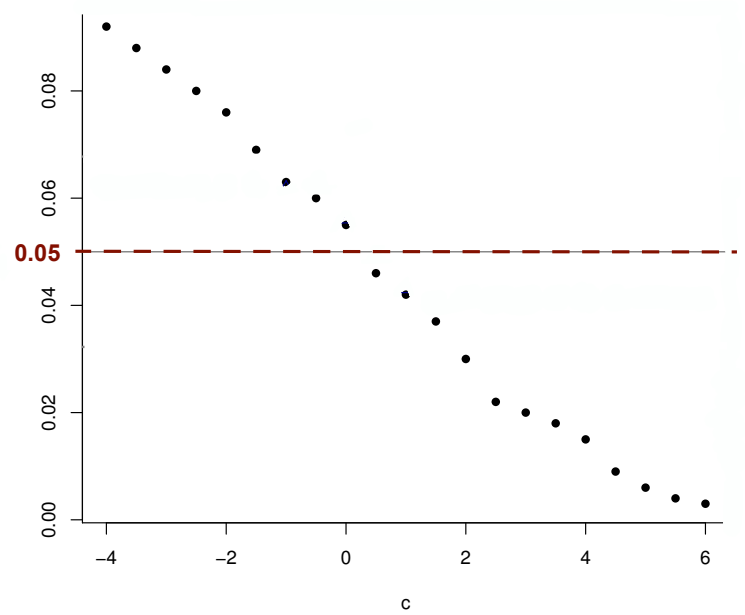


– Effect of outliers in testing

Test: $H_0 : m_1 = m_2$ vs $H_1 : m_1 < m_2$

Method in Neumeyer and PF (2009, *JSPI*)

The graph shows the **empirical size** of a test depending on **one** outlier.



Robust version of this test $\longrightarrow$ Boente and PF (2016, *CSDA*)

– Comparison of regression curves

A very relevant problem in Statistics is the **comparison of means**:

- Testing the equality of the means of k populations $\longrightarrow$ **ANOVA**.
- Conditional setting (covariate - response) $\longrightarrow$ **comparison of regression curves**.

– Comparison of regression curves

A very relevant problem in Statistics is the **comparison of means**:

- Testing the equality of the means of k populations $\longrightarrow$ **ANOVA**.
- Conditional setting (covariate - response) $\longrightarrow$ **comparison of regression curves**.

Consider two independent vectors (X_1, Y_1) and (X_2, Y_2) and suppose that

$$Y_1 = m_1(X_1) + \sigma_1(X_1)\varepsilon_1$$

$$Y_2 = m_2(X_2) + \sigma_2(X_2)\varepsilon_2,$$

where ε_1 and ε_2 have common distribution F_ε.
We like to test the hypothesis:

$$H_0 : m_1 \equiv m_2.$$

– Comparison of regression curves

Let (X_j, Y_j), $1 \leq j \leq k$, be k independent random vectors satisfying **fully nonparametric regression models**

$$Y_j = m_j(X_j) + \sigma_j(X_j)\varepsilon_j,$$

where

- $m_j(x) = \mathbb{E}(Y_j \mid X_j = x)$ regression function,
- $\sigma_j^2(x) = \textsc{Var}(Y_j \mid X_j = x)$ conditional variance function,
- ε_j is the regression error.

The hypothesis of **equality of means** is stated in terms of the regression functions:

$$H_0 : m_1(\cdot) = m_2(\cdot) = \cdots = m_k(\cdot).$$

– Comparison of regression curves

- **Comparison of 2 regression curves:**

  - Delgado (1993, *JASA*)

  - Kulasekera (1995, *JASA*)

  - Neumeyer and Dette (2003, *Ann. Statist.*)

  - Neumeyer and PF (2009, *JSPI*)

  - Srihera and Stute (2010, *JMVA*)

  - Boente and PF (2015, *CSDA*) [robust version of Neumeyer and PF (2009)]

  - ...

- **Comparison of $k \geq 2$ regression curves:**

  - PF, Van Keilegom and González-Manteiga (2007, *Stat. Sinica*)

  - PF, Jiménez-Gamero and El Ghouch (2015, *Scand. J. Stat.*)

  - ...

– Comparison of regression curves

Regression errors in population j:

$$\varepsilon_j = \frac{Y_j - m_j(X_j)}{\sigma_j(X_j)}.$$

– Comparison of regression curves

Regression errors in population j:

$$\varepsilon_j = \frac{Y_j - m_j(X_j)}{\sigma_j(X_j)}.$$

Let m_0 be the common regression curve under the null hypothesis. Define the regression errors under H_0 in population j:

$$\varepsilon_{0j} = \frac{Y_j - m_0(X_j)}{\sigma_j(X_j)} = \varepsilon_j + \frac{m_j(X_j) - m_0(X_j)}{\sigma_j(X_j)}.$$

– Comparison of regression curves

Regression errors in population j:

$$\varepsilon_j = \frac{Y_j - m_j(X_j)}{\sigma_j(X_j)}.$$

Let m_0 be the common regression curve under the null hypothesis. Define the regression errors under H_0 in population j:

$$\varepsilon_{0j} = \frac{Y_j - m_0(X_j)}{\sigma_j(X_j)} = \varepsilon_j + \frac{m_j(X_j) - m_0(X_j)}{\sigma_j(X_j)}.$$

Theorem (PF, Van Keilegom and González-Manteiga, 2007)

H_0 *is true* $\iff$ ε_j *and* ε_{0j} *have the same distribution*

– Comparison of regression curves

Regression errors in population j:

$$\varepsilon_j = \frac{Y_j - m_j(X_j)}{\sigma_j(X_j)}.$$

Let m_0 be the common regression curve under the null hypothesis. Define the regression errors under H_0 in population j:

$$\varepsilon_{0j} = \frac{Y_j - m_0(X_j)}{\sigma_j(X_j)} = \varepsilon_j + \frac{m_j(X_j) - m_0(X_j)}{\sigma_j(X_j)}.$$

Theorem (PF, Van Keilegom and González-Manteiga, 2007)

H_0 *is true* $\iff$ ε_j *and* ε_{0j} *have the same distribution*

The previous theorem can be interpreted in terms of . . .

- . . . cumulative distr. functions $\longrightarrow$ PF, VK, GM (2007, *Stat. Sinica*).
- . . . **characteristic functions** $\longrightarrow$ PF, Jiménez-Gamero and El Ghouch (2015, *Scand. J. Stat.*).

– Reminder: Characteristic function

The **characteristic function** of a random variable X is given by

$$\begin{aligned} \varphi_X \quad : \quad & \mathbb{R} \longrightarrow \mathbb{C} \\ & t \longrightarrow \varphi_X(t) = \mathbb{E}(\exp\{\,\mathrm{i}\,tX\}) = \mathbb{E}(\cos(tX) + \mathrm{i}\,\sin(tX)) \end{aligned}$$

– Reminder: Characteristic function

The **characteristic function** of a random variable X is given by

$$\begin{aligned} \varphi_X \quad : \quad & \mathbb{R} \longrightarrow \mathbb{C} \\ & t \longrightarrow \varphi_X(t) = \mathbb{E}(\exp\{\,\mathrm{i}\,tX\}) = \mathbb{E}(\cos(tX) + \mathrm{i}\,\sin(tX)) \end{aligned}$$

The characteristic function ...

- exists for any random variable.
- is uniformly continuous and bounded for any random variable.
- does not vanish around zero: $\varphi_X(0) = 1$.
- is real-valued when X is symmetric around the origin.
- **characterizes the distribution of the random variable**: there exists a bijective (and continuous) application between the set of cumulative distribution functions and the set of characteristic functions.
- If X and Y are independent random variables, then $\varphi_{X+Y} = \varphi_X \varphi_Y$

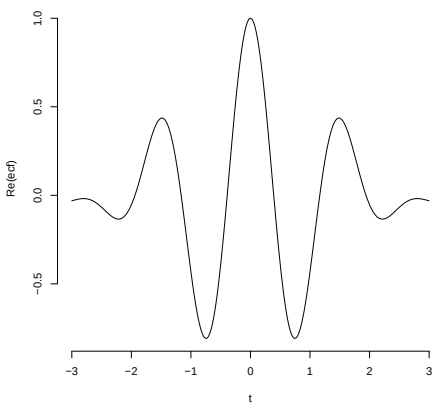
– Reminder: Empirical characteristic function

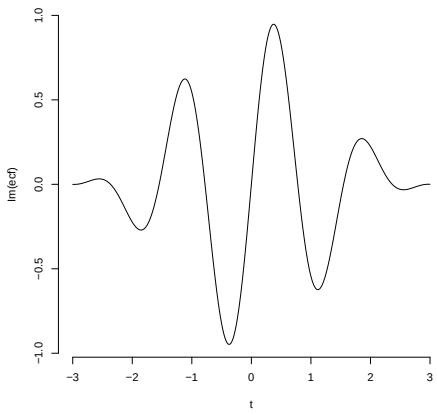
Given an i.i.d. sample $X_1, \ldots, X_n$ of X, the **empirical characteristic function**

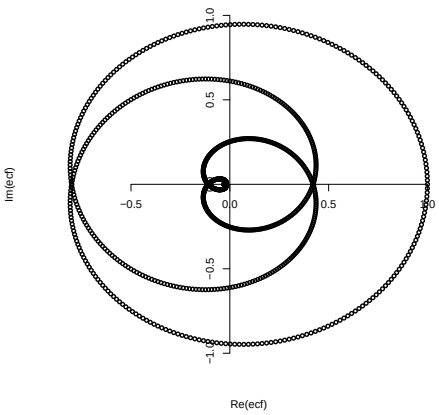
$$\hat{\varphi}_X(t) = \frac{1}{n}\sum_{l=1}^{n} \exp(\mathrm{i}\, tX_i) = \frac{1}{n}\sum_{l=1}^{n} \cos(tX_i) + \mathrm{i}\,\frac{1}{n}\sum_{l=1}^{n} \sin(tX_i).$$

Example: Sample of size $n = 100$ from $N(4, 1)$

$\mathrm{Re}(\hat{\varphi}_X(t))$ $\qquad$ $\mathrm{Im}(\hat{\varphi}_X(t))$ $\qquad$ $(\mathrm{Re}(\hat{\varphi}_X(t)), \mathrm{Im}(\hat{\varphi}_X(t)))$







– A test based on characteristic functions

Why using characteristic functions?

- Usually requires less stringent assumptions.
- Simulation results show that tests based on characteristic functions compete very satisfactorily with those based on the empirical cumulative distribution functions.

– A test based on characteristic functions

Why using characteristic functions?

- Usually requires less stringent assumptions.
- Simulation results show that tests based on characteristic functions compete very satisfactorily with those based on the empirical cumulative distribution functions.

Proposals for hypothesis testing based on characteristic functions:

- Two-sample problem: Alba-Fernández et al. (2008, *CSDA*).
- Goodness-of-fit for the errors in regression models: Jiménez-Gamero et al. (2005, *JMVA*) and Hušková and Meintanis (2007 *Statistics*, 2010 *TEST*).
- Goodness-of-fit for parametric regression functions: Hušková and Meintanis (2009, *Kybernetika*).
- Comparison of regression curves: PF, Jiménez-Gamero and El Ghouch (2015, *Scand. J. Stat.*).
- Comparison of conditional variance functions: PF, Jiménez-Gamero and El Ghouch (2015, *Electr. J. Statist.*).
- Goodness-of-fit for the parametric form of the conditional variance: PF and Jiménez-Gamero (2019, *AStA*)
- ...and many others

– A test based on characteristic functions

Data: $(X_{j\ell}, Y_{j\ell})$, $1 \le \ell \le n_j$, be iid observations from (X_j, Y_j), $1 \le j \le k$

For each population j, $1 \le j \le k$, we construct two samples of estimated residuals:

$$\hat{\varepsilon}_{j\ell} = \frac{Y_{j\ell} - \hat{m}_j(X_{j\ell})}{\hat{\sigma}_j(X_{j\ell})} \qquad \text{and} \qquad \hat{\varepsilon}_{0j\ell} = \frac{Y_{j\ell} - \hat{m}_0(X_{j\ell})}{\hat{\sigma}_j(X_{j\ell})},$$

$1 \le \ell \le n_j$, whose empirical characteristics functions are

$$\hat{\varphi}_j(t) = \frac{1}{n_j} \sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\hat{\varepsilon}_{j\ell}) \qquad \text{and} \qquad \hat{\varphi}_{0j}(t) = \frac{1}{n_j} \sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\hat{\varepsilon}_{0j\ell}).$$

– A test based on characteristic functions

Data: $(X_{j\ell}, Y_{j\ell})$, $1 \leq \ell \leq n_j$, be iid observations from (X_j, Y_j), $1 \leq j \leq k$

For each population j, $1 \leq j \leq k$, we construct two samples of estimated residuals:

$$\hat{\varepsilon}_{j\ell} = \frac{Y_{j\ell} - \hat{m}_j(X_{j\ell})}{\hat{\sigma}_j(X_{j\ell})} \qquad \text{and} \qquad \hat{\varepsilon}_{0j\ell} = \frac{Y_{j\ell} - \hat{m}_0(X_{j\ell})}{\hat{\sigma}_j(X_{j\ell})},$$

$1 \leq \ell \leq n_j$, whose empirical characteristics functions are

$$\hat{\varphi}_j(t) = \frac{1}{n_j} \sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\hat{\varepsilon}_{j\ell}) \qquad \text{and} \qquad \hat{\varphi}_{0j}(t) = \frac{1}{n_j} \sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\hat{\varepsilon}_{0j\ell}).$$

The testing procedure consists of comparing $\hat{\varphi}_j(t)$ and $\hat{\varphi}_{0j}(t)$, $1 \leq j \leq k$, using a weighted L_2-distance.

$$T = \sum_{j=1}^{k} \frac{n_j}{n} \int |\hat{\varphi}_j(t) - \hat{\varphi}_{0j}(t)|^2 \, w(t)dt,$$

where w is a non-negative weight function to ensure the finiteness of the integral.

– A test based on characteristic functions

In PF, Jiménez-Gamero and El Ghouch (2015) a detailed study of this statistic was performed. In particular, the asymptotic distribution under H_0 was obtained:

$$nT \xrightarrow{\mathcal{D}} \mathbf{Z}^{\mathrm{T}} \mathcal{A} \mathbf{Z}$$

- $\mathbf{Z} \sim N_k(0, \Sigma)$. The elements of Σ depend on population characteristics (basically densities of the covariates and conditional variance functions).
- $\mathcal{A}$ is a diagonal matrix whose elements depend on the characteristic funcion of the regression errors.

– A test based on characteristic functions

In PF, Jiménez-Gamero and El Ghouch (2015) a detailed study of this statistic was performed. In particular, the asymptotic distribution under H_0 was obtained:

$$nT \xrightarrow{\mathcal{D}} \mathbf{Z}^{\mathrm{T}} \mathcal{A} \mathbf{Z}$$

- $\mathbf{Z} \sim N_k(0, \Sigma)$. The elements of Σ depend on population characteristics (basically densities of the covariates and conditional variance functions).
- $\mathcal{A}$ is a diagonal matrix whose elements depend on the characteristic funcion of the regression errors.
- The limiting distribution of nT under H_0 is a **linear combination of independent chi-square variables**: $\sum_{j=1}^{k} \beta_j \chi^2_{1,j}$, where

  - $\chi^2_{1,1}, \ldots, \chi^2_{1,k}$ are independent chi-square random variates with 1 d.f.

  - $\beta_1, \ldots, \beta_k$ are the **eigenvalues** of $\mathcal{A}\Sigma$.

- All the quantities can be $\mathcal{A}$ and Σ can be (*easily*) estimated.
- The practical performance of the test based on the asymptotic distribution was shown to be good.

– Robust version of the test

Let (X_j, Y_j), $j = 1, \ldots, k$, be k random vectors that follow the **homoscedastic nonparametric regression models** given by

$$Y_j = m_j(X_j) + \sigma_j \, \varepsilon_j,$$

where, for $j = 1, \ldots, k$,

- m_j is a nonparametric smooth function,
- the error ε_j is independent of the covariate X_j, has a **symmetric** distribution and has scale 1
- σ_j is the scale.

– Robust version of the test

Let (X_j, Y_j), $j = 1, \ldots, k$, be k random vectors that follow the **homoscedastic nonparametric regression models** given by

$$Y_j = m_j(X_j) + \sigma_j \, \varepsilon_j,$$

where, for $j = 1, \ldots, k$,

- m_j is a nonparametric smooth function,
- the error ε_j is independent of the covariate X_j, has a **symmetric** distribution and has scale 1
- σ_j is the scale.

Objective: testing the **null hypothesis** of equality of regression curves

$$H_0 : m_1(x) = \cdots = m_k(x) \text{ for all } x \in \mathcal{R},$$

where $\mathcal{R}$ is the common support of the covariates X_j, $j = 1, \ldots, k$, where the comparison will be performed, versus a **general alternative**

$$H_1 : H_0 \text{ is not true.}$$

– Robust version of the test

Notation:

- Samples: $\{(X_{j\ell}, Y_{j\ell}), \ell = 1, \ldots, n_j\}$ from (X_j, Y_j), $j = 1, \ldots, k$.
- $n = \sum_{i=1}^{k} n_j$
- $\widehat{m}_j(x)$ is the robust estimator of $m_j(x)$
- $\widehat{\sigma}_j$ a robust estimator of the error's scale σ_j
- Estimator of the common regression function under H_0:

$$\widehat{m}_0(x) = \sum_{j=1}^{k} \frac{n_j}{n} \frac{\widehat{f}_j(x)}{\widehat{f}(x)} \widehat{m}_j(x)$$

where $\widehat{f}_j(x)$ is the kernel estimator of the density of X_j, f_j.

– Robust version of the test

On the basis of these estimators, for each population j, we construct two samples of residuals based on the robust estimators given before

$$\hat{\varepsilon}_{j\ell} = \frac{Y_{j\ell} - \widehat{m}_j(X_{j\ell})}{\widehat{\sigma}_j} \qquad \text{and} \qquad \hat{\varepsilon}_{0j\ell} = \frac{Y_{j\ell} - \widehat{m}_0(X_{j\ell})}{\widehat{\sigma}_j}$$

and the corresponding characteristic functions

$$\widehat{\varphi}_j(t) = \frac{1}{n_j}\sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\, \hat{\varepsilon}_{j\ell}) \qquad \text{and} \qquad \widehat{\varphi}_{0j}(t) = \frac{1}{n_j}\sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\, \hat{\varepsilon}_{0j\ell})$$

– Robust version of the test

On the basis of these estimators, for each population j, we construct two samples of residuals based on the robust estimators given before

$$\hat{\varepsilon}_{j\ell} = \frac{Y_{j\ell} - \widehat{m}_j(X_{j\ell})}{\widehat{\sigma}_j} \qquad \text{and} \qquad \hat{\varepsilon}_{0j\ell} = \frac{Y_{j\ell} - \widehat{m}_0(X_{j\ell})}{\widehat{\sigma}_j}$$

and the corresponding characteristic functions

$$\widehat{\varphi}_j(t) = \frac{1}{n_j}\sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\, \hat{\varepsilon}_{j\ell}) \qquad \text{and} \qquad \widehat{\varphi}_{0j}(t) = \frac{1}{n_j}\sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\, \hat{\varepsilon}_{0j\ell})$$

The **test statistic** is

$$T = \sum_{j=1}^{k} \frac{n_j}{n} \int |\hat{\varphi}_j(t) - \hat{\varphi}_{0j}(t)|^2\, w(t) dt.$$

– Robust version of the test

On the basis of these estimators, for each population j, we construct two samples of residuals based on the robust estimators given before

$$\hat{\varepsilon}_{j\ell} = \frac{Y_{j\ell} - \widehat{m}_j(X_{j\ell})}{\widehat{\sigma}_j} \qquad \text{and} \qquad \hat{\varepsilon}_{0j\ell} = \frac{Y_{j\ell} - \widehat{m}_0(X_{j\ell})}{\widehat{\sigma}_j}$$

and the corresponding characteristic functions

$$\widehat{\varphi}_j(t) = \frac{1}{n_j}\sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\, \hat{\varepsilon}_{j\ell}) \qquad \text{and} \qquad \widehat{\varphi}_{0j}(t) = \frac{1}{n_j}\sum_{\ell=1}^{n_j} \exp(\mathrm{i}\, t\, \hat{\varepsilon}_{0j\ell})$$

The **test statistic** is

$$T = \sum_{j=1}^{k} \frac{n_j}{n} \int \left|\hat{\varphi}_j(t) - \hat{\varphi}_{0j}(t)\right|^2 w(t)dt.$$

- The null hypothesis will be rejected for large positive values of the test statistic T.
- To obtain the critical values, we need the (asymptotic) null distribution.

– Robust version of the test. Asymptotic null distribution

Theorem

Regularity assumptions (. . .) and $\mathbb{E}|\varepsilon_j| < \infty$. *Under* H_0,

$$n\,T \xrightarrow{\mathcal{D}} \mathbf{Z}^{\mathrm{T}}\mathbf{A}\mathbf{Z},$$

where $\mathbf{Z} = (Z_1, \ldots, Z_k)^{\mathrm{T}} \sim N(\mathbf{0}, \mathbf{\Sigma})$ *where* $\mathbf{\Sigma} = (\sigma_{j\ell})$ *with*

$$\sigma_{jj} = \sum_{s=1}^{k} \pi_j\, \pi_s\, e_s\, \alpha_j^{(s)}\, \frac{\sigma_s^2}{\sigma_j^2} + e_j\, \{1 - 2\pi_j \beta_j\}$$

$$\sigma_{j\ell} = \frac{\pi_\ell^{1/2}\, \pi_j^{1/2}}{\sigma_\ell\, \sigma_j} \sum_{s=1}^{k} e_s\, \pi_s\, \sigma_s^2 \alpha_{j,\ell}^{(s)} - \frac{\sigma_\ell}{\sigma_j} \pi_j^{1/2}\, \pi_\ell^{1/2}\, e_\ell\ \beta_j^{(\ell)} - \frac{\sigma_j}{\sigma_\ell} \pi_j^{1/2}\, \pi_\ell^{1/2}\, e_j\ \beta_\ell^{(j)}$$

and, for $j, \ell, s = 1, \ldots, k$, $\pi_j = \lim n_j/n > 0$, $e_j = \mathbb{E}[\psi_j^2(\varepsilon_j)]/\mathbb{E}[\psi_j'(\varepsilon_j)]^2$,

$$\beta_j^{(s)} = \mathbb{E}\left\{\frac{f_j(X_s)}{f(X_s)}\right\}, \quad \beta_j = \mathbb{E}\left\{\frac{f_j(X_j)}{f(X_j)}\right\}, \quad \alpha_{j,\ell}^{(s)} = \mathbb{E}\left(\frac{f_\ell(X_s)\, f_j(X_s)}{f^2(X_s)}\right), \quad \alpha_j^{(s)} = \mathbb{E}\left\{\frac{f_j^2(X_s)}{f^2(X_s)}\right\},$$

and $\mathbf{A} = \mathrm{DIAG}(a_1, \ldots, a_k)$ *with* $a_j = \int |t\, \varphi_j(t)|^2 w(t) dt$.

– Robust version of the test. Asymptotic null distribution

- The limiting distribution of nT under H_0 is a **linear combination of independent chi-square variables**: $\sum_{j=1}^{k} \beta_j \chi^2_{1,j}$, where

  - $\chi^2_{1,1}, \ldots, \chi^2_{1,k}$ are independent chi-square random variates with 1 d.f.

  - $\beta_1, \ldots, \beta_k$ are the estimated **eigenvalues** of $\mathbf{A}\Sigma$.

- All the quantities in $\mathbf{A}$ and Σ can be estimated by plug-in methods.

– Some simulations

Objective: compare the test in PF, JG and EG (2015) with the new proposal.

- $k = 2$
- Regression errors: $\varepsilon_1, \varepsilon_2 \sim N(0,1)$ + **some contamination with outliers**
- $\sigma_1 = \sqrt{0.25}$ and $\sigma_2 = \sqrt{0.50}$
- Covariates: $X_1, X_2 \sim Uniform[0,1]$
- Significance level: $\alpha = 0.05$
- **Score functions** for the robust estimation: Tukey's bisquare
- Cross-validation bandwidths

– Some simulations. Level approximation

Regression functions under the null hypothesis (level approximation):

$$\text{(L.1)} \quad m_1(x) = m_2(x) = x$$
$$\text{(L.2)} \quad m_1(x) = m_2(x) = \sin(2\pi x)$$
$$\text{(L.3)} \quad m_1(x) = m_2(x) = \exp(x)$$

No contamination:

	non robust			**robust**		
model	$(100, 100)$	$(200, 100)$	$(200, 200)$	$(100, 100)$	$(200, 100)$	$(200, 200)$
(L.1)	0.044	0.052	0.042	0.056	0.061	0.055
(L.2)	0.055	0.061	0.049	0.074	0.078	0.060
(L.3)	0.047	0.056	0.046	0.060	0.066	0.053

– Some simulations. Level approximation

Contamination #1:

$$\varepsilon_1 \sim \begin{cases} N(0,1) & \text{w.p. } 0.95 \\ N(5,0.1) & \text{w.p. } 0.05 \end{cases} \qquad \varepsilon_2 \sim \begin{cases} N(0,1) & \text{w.p. } 0.95 \\ N(10,0.1) & \text{w.p. } 0.05 \end{cases}$$

	non robust			robust		
model	$(100,100)$	$(200,100)$	$(200,200)$	$(100,100)$	$(200,100)$	$(200,200)$
(L.1)	0.153	0.159	0.376	0.059	0.055	0.055
(L.2)	0.158	0.173	0.384	0.076	0.062	0.060
(L.3)	0.150	0.161	0.374	0.065	0.055	0.059

– Some simulations. Level approximation

Contamination #1:

$$\varepsilon_1 \sim \begin{cases} N(0,1) & \text{w.p. } 0.95 \\ N(5,0.1) & \text{w.p. } 0.05 \end{cases} \qquad \varepsilon_2 \sim \begin{cases} N(0,1) & \text{w.p. } 0.95 \\ N(10,0.1) & \text{w.p. } 0.05 \end{cases}$$

	non robust			robust		
model	$(100,100)$	$(200,100)$	$(200,200)$	$(100,100)$	$(200,100)$	$(200,200)$
(L.1)	0.153	0.159	0.376	0.059	0.055	0.055
(L.2)	0.158	0.173	0.384	0.076	0.062	0.060
(L.3)	0.150	0.161	0.374	0.065	0.055	0.059

Contamination #2:

$$\varepsilon_1 \sim \begin{cases} N(0,1) & \text{w.p. } 0.90 \\ N(10,0.1) & \text{w.p. } 0.10 \end{cases} \qquad \varepsilon_2 \sim N(0,1)$$

	non robust			robust		
model	$(100,100)$	$(200,100)$	$(200,200)$	$(100,100)$	$(200,100)$	$(200,200)$
(L.1)	0.825	0.980	0.995	0.059	0.053	0.061
(L.2)	0.817	0.977	0.995	0.076	0.067	0.065
(L.3)	0.830	0.977	0.994	0.062	0.057	0.057

– Some simulations. Power

Regression functions under the alternative hypothesis:

(P.1) $m_1(x) = x$, $m_2(x) = x + 0.5x$

(P.2) $m_1(x) = \sin(2\pi x)$, $m_2(x) = \sin(2\pi x) + 0.5x$

(P.3) $m_1(x) = \exp(x)$, $m_2(x) = \exp(x) + 0.5x$

No contamination:

	non robust			**robust**		
model	$(100, 100)$	$(200, 100)$	$(200, 200)$	$(100, 100)$	$(200, 100)$	$(200, 200)$
(P.1)	0.796	0.874	0.985	0.788	0.868	0.985
(P.2)	0.806	0.885	0.987	0.807	0.880	0.987
(P.3)	0.796	0.873	0.985	0.791	0.867	0.985

– Some simulations. Power (robust test)

Contamination #1:

$$\varepsilon_1 \sim \begin{cases} N(0,1) & \text{w.p. } 0.95 \\ N(5,0.1) & \text{w.p. } 0.05 \end{cases} \qquad \varepsilon_2 \sim \begin{cases} N(0,1) & \text{w.p. } 0.95 \\ N(10,0.1) & \text{w.p. } 0.05 \end{cases}$$

	no contamination			**contamination**		
model	$(100,100)$	$(200,100)$	$(200,200)$	$(100,100)$	$(200,100)$	$(200,200)$
(P.1)	0.788	0.868	0.985	0.749	0.847	0.958
(P.2)	0.807	0.880	0.987	0.757	0.850	0.961
(P.3)	0.791	0.867	0.985	0.746	0.847	0.960

– Some simulations. Power (robust test)

Contamination #1:

$$\varepsilon_1 \sim \begin{cases} N(0,1) & \text{w.p. } 0.95 \\ N(5,0.1) & \text{w.p. } 0.05 \end{cases} \qquad \varepsilon_2 \sim \begin{cases} N(0,1) & \text{w.p. } 0.95 \\ N(10,0.1) & \text{w.p. } 0.05 \end{cases}$$

	no contamination			contamination		
model	$(100,100)$	$(200,100)$	$(200,200)$	$(100,100)$	$(200,100)$	$(200,200)$
(P.1)	0.788	0.868	0.985	0.749	0.847	0.958
(P.2)	0.807	0.880	0.987	0.757	0.850	0.961
(P.3)	0.791	0.867	0.985	0.746	0.847	0.960

Contamination #2:

$$\varepsilon_1 \sim \begin{cases} N(0,1) & \text{w.p. } 0.90 \\ N(10,0.1) & \text{w.p. } 0.10 \end{cases} \qquad \varepsilon_2 \sim N(0,1)$$

	no contamination			contamination		
model	$(100,100)$	$(200,100)$	$(200,200)$	$(100,100)$	$(200,100)$	$(200,200)$
(P.1)	0.788	0.868	0.985	0.804	0.876	0.968
(P.2)	0.807	0.880	0.987	0.813	0.888	0.974
(P.3)	0.791	0.867	0.985	0.800	0.879	0.966

– Illustration with real data: acid rain data

- Data from an environmental study about acid rain along a five-year period 1979-1983 in two locations of North Carolina, **Coweeta** and **Lewiston**.

- Variables:

  covariate: amount of rainfall per week

  response: logarithm of sulfate concentration

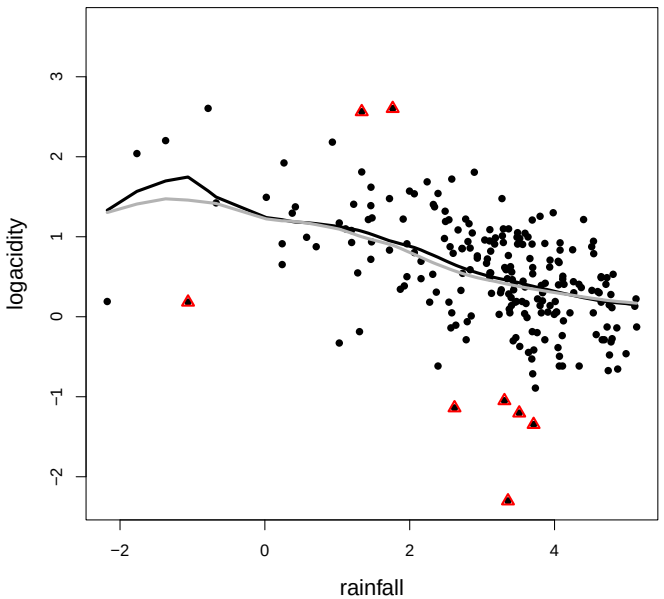
- Sample sizes:

  Coweeta: 220

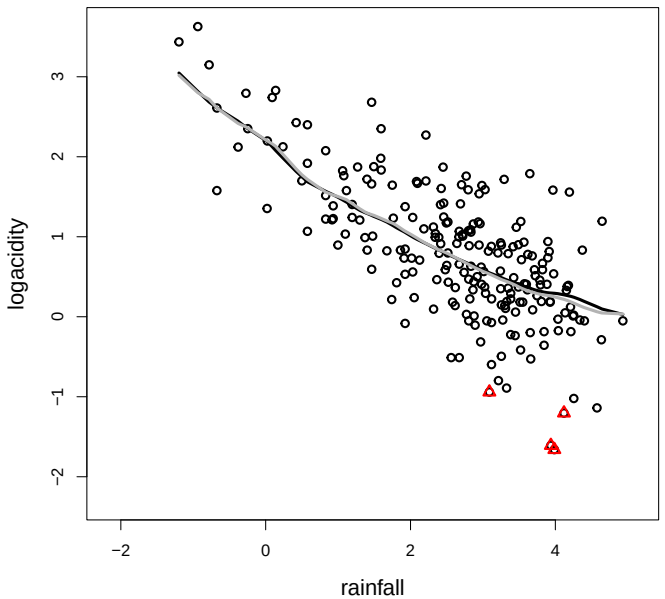
  Lewistion: 215

– Illustration with real data: acid rain data

Coweeta



Lewiston



Nadaraya–Watson (black lines), with c-v bandwidths $h_1 = 1.6$ and $h_2 = 0.8$

$\hookrightarrow$ non-robust test: $p-$value $= 0.04956$

Robust estimators (gray lines), with c-v bandwidths $h_1 = 1.3$ and $h_2 = 0.9$

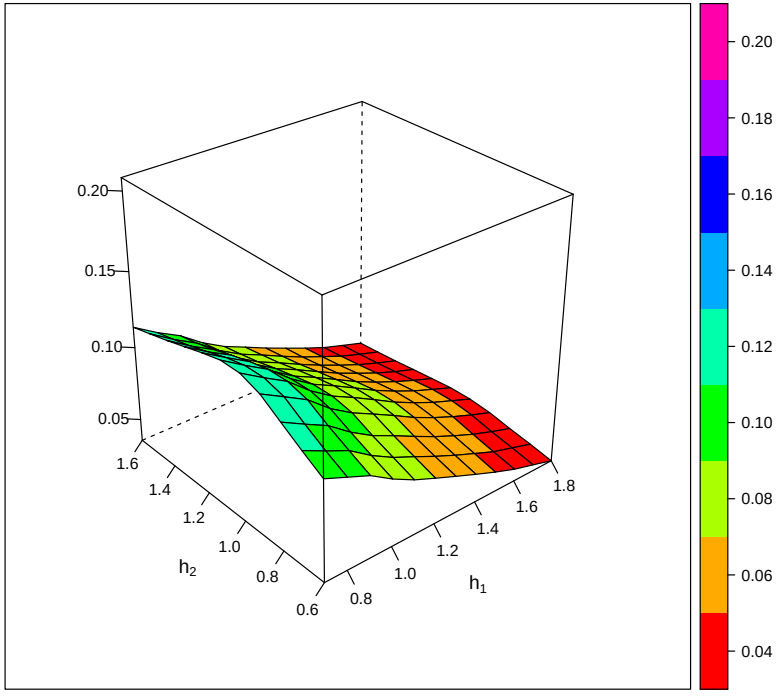
$\hookrightarrow$ robust test: $p-$value $= 0.11167$

– Illustration with real data: acid rain data

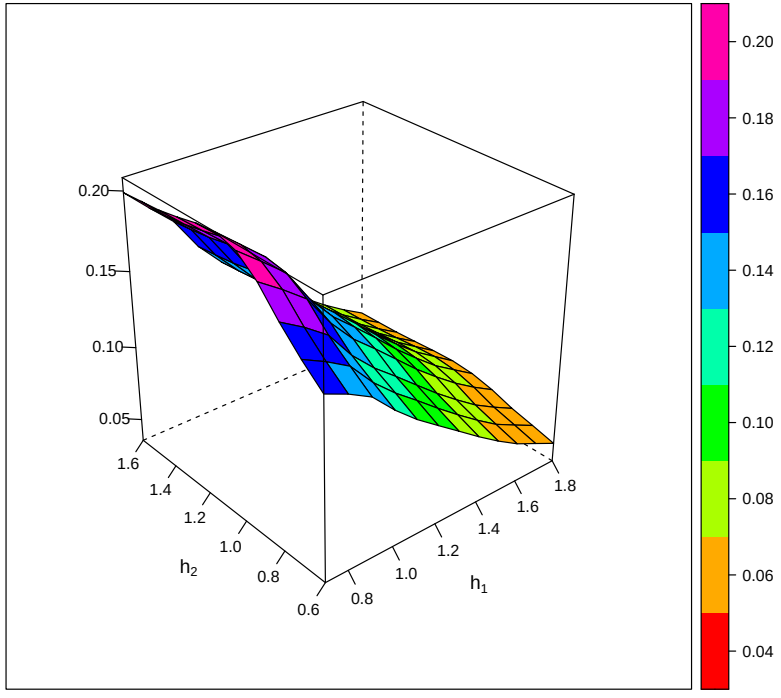
Analysis of the influence of the bandwidth choice on the $p-$value:

$\hookrightarrow$ grid of values for $(h_1, h_2) \in [0.7, 1.8] \times [0.6, 1.6]$

Classical Test



Robust Test



– Conclusions

- We have proposed and studied a new **robust** method to **test for the equality of regression curves** in a fully nonparametric setup.
- The new procedure adapts the procedure in PF, Jiménez-Gamero and El Ghouch (2015) to the robust context.
- **Critical values** can be obtained from the asymptotic null distribution of the test statistic (no bootstrap required).
- Consistency against root-n alternatives has also been proved.
- Gains: robust, avoids second moment.
- Price to pay: homoscedasticity, symmetry of the error, estimation process.
- Good practical performance in simulations.

Preprint available in arXiv: arxiv.org/abs/2205.12065

Robust tests for the comparison of regression curves

Juan Carlos Pardo-Fernández

Universidade de Vigo

Joint work with **Graciela Boente** (Universidad de Buenos Aires)

Symposium in honor to Ingrid Van Keilegom
A Coruña – June 15th, 2022