Seminario del

Instituto de Matemática Aplicada del Litoral
Mayo de 2010

Funcionales Robustos de Escala en Regresión No-paramétrica

por

Marcelo Ruiz

(UNRC)

en co–autoría con

Graciela Lina Boente (UBA-CONICET) y Ruben Zamar (University of British Columbia, Canadá)

(2008, Journal of Statiscal Planning and Inference)

Sección A

El marco general. Preliminares.

El modelo

Medimos x (posición) sin error en el intervalo $[0, 1]$ y obtenemos una respuesta, aleatoria, Y_x (velocidad de una partícula). Postulamos la relación

$$Y_x = g(x) + U_x\sigma(x),\ i = 1, \ldots, n,$$

donde $U : (\Omega, \mathcal{A}, P) \to \mathbb{R}$ es el componente aleatorio (error) con

$$\mathsf{E}[U] = \int udP(u) = 0 \text{ y } \mathsf{Var}[U] = \mathsf{E}[U - \mathsf{E}[U]]^2 = 1.$$

Así,

$$\mathsf{E}[Y_x] = g(x) \text{ y } \mathsf{Var}[Y_x] = \sigma^2(x)$$

g es la función de regresión y $\sigma(x)$ es la función de escala. Ambas desconocidas, satisfaciendo $g \in \mathcal{G}$ y $\sigma \in \Sigma$, con $\mathcal{G}$ y Σ clases de funciones suaves.

El problema de estimación

Dadas n observaciones $(x_1, Y_1) \ldots (x_n, Y_n)$ y $x \in (0, 1)$, queremos hallar estimadores

$$\begin{aligned} \widehat{g}_n(x, (x_1, Y_1) \ldots (x_n, Y_n)) &= \widehat{g}_n(x) \\ \widehat{\sigma}_n(x, (x_1, Y_1) \ldots (x_n, Y_n)) &= \widehat{\sigma}_n(x) \end{aligned}$$

de $g(x)$ y $\sigma(x)$, respectivamente.
Dos estrategias

- Estimación simultánea de regresión y escala, o
- Estimación preliminar de la escala

$\Downarrow$

Optamos por la seguda estrategia y cuando la función de escala es constante.

Comenzamos por lo más simple

La funciones de regresión y de escala constantes y $U_i \sim N(0,1)$:

$$Y_i = \mu_0 + U_i\sigma,\ i = 1, \ldots, n$$

Como

$$\frac{1}{(n-1)} \sum_{i=1}^{n-1} \left(\frac{U_{i+1} - U_i}{\sqrt{2}} \right)^2 \xrightarrow{a.s.} \mathsf{E} \left(\frac{U_2 - U_1}{\sqrt{2}} \right)^2 = 1$$

Entonces

$$\frac{1}{(n-1)} \sum_{i=1}^{n-1} \left(\frac{Y_{i+1} - Y_i}{\sqrt{2}} \right)^2 \approx \sigma^2$$

Estimador de Rice

El siguiente será el modelo del resto de la exposición

$$Y_i = g(x_i) + U_i\sigma,\ i = 1, \ldots, n$$

entonces

$$Y_{i+1} - Y_i = (g(x_{i+1}) - g(x_i)) + \sigma(U_{i+1} - U_i) \approx \sigma(U_{i+1} - U_i)$$

luego si $U_i \sim F_0 = N(0,1)$ el estimador de Rice de la varianza

$$\begin{aligned}
\widehat{\sigma}^2_{\mathsf{R},n} &= \frac{1}{(n-1)} \sum_{i=1}^{n-1} \left(\frac{Y_{i+1} - Y_i}{\sqrt{2}} \right)^2 \\
&\approx \sigma^2 \frac{1}{(n-1)} \sum_{i=1}^{n-1} \left(\frac{U_{i+1} - U_i}{\sqrt{2}} \right)^2 \\
&\approx \sigma^2
\end{aligned}$$

Falta de robustez de los estimadores tipo Rice

$$\text{Si } \exists i_0 : Y_{i_0} \to \infty \text{ entonces } \widehat{\sigma}^2_{\mathsf{R},n} \to \infty$$

Es decir, frente a observaciones atípicas el estimador basado en el cuadrado de las diferencias se quiebra.

Las observaciones son atípicas (outliers o inliers) cuando

$$U \sim G \neq F_0 \text{ y } G \in \mathcal{P}_\epsilon(F_0),$$

un entorno de contaminación de tamaño ϵ:

$$\mathcal{P}_\epsilon(F_0) = \{G : \ G = (1-\epsilon)F_0 + \epsilon H; \ H \in \mathcal{D}\}$$

con $\mathcal{D}$ la colección de distribuciones sobre $\mathbb{R}$, $\epsilon \in [0, 1/2)$ fijo.

Estimadores como solución de ecuaciones

$\widehat{\sigma}_{\mathrm{R},n}$ es solución de la ecuación no lineal

$$\frac{1}{n-1}\sum_{i=1}^{n-1}\chi\left(\frac{Y_{i+1}-Y_i}{a\widehat{\sigma}_{\mathrm{R},n}}\right)=b,$$

con $\chi(x)=x^2$ y las constantes $a=\sqrt{2}$ y $b=1/2$ satisfacen las condiciones

$$(C.A.)\;\; E\left[\chi\left(\frac{Z_2-Z_1}{a}\right)\right]=b \text{ y } E\left[\chi\left(Z_1\right)\right]=b,$$

Z_1, Z_2 v.a.'s i.i.d con $Z_1 \sim N(0,1)$.

$M-$estimadores

Para estimar robustamente la escala en el modelo homoscedástico

$$Y_i = g(x_i) + \sigma U_i,\ i = 1, \ldots, n.$$

con $U_i \sim G \in \mathcal{P}_\epsilon(F_0)$

☞ Nuestra propuesta:

$M-$estimadores de escala basados en las diferencias sucesivas de las respuestas

$$\widehat{\sigma}_{\mathsf{M},n} = \inf \left\{ s > 0 :\ \frac{1}{n-1} \sum_{i=1}^{n-1} \chi\left(\frac{Y_i^*}{as}\right) \leq b \right\}$$

con $\{Y_i^* = Y_{i+1} - Y_i\}_{i=1}^{n-1}$ ...

... $M-$estimadores

siendo

- χ una función de escores: par, $\chi(0) = 0$, no decreciente en $\mathbb{R}^+$, con $0 < \|\chi\|_\infty$;
- $a \in (0, \infty)$ y $b \in (0, 1)$ constantes tales que

$$E[\chi(Z_1)] = b \;\text{ y }\; E\left[\chi\left(\frac{Z_1^*}{a}\right)\right] = b,$$

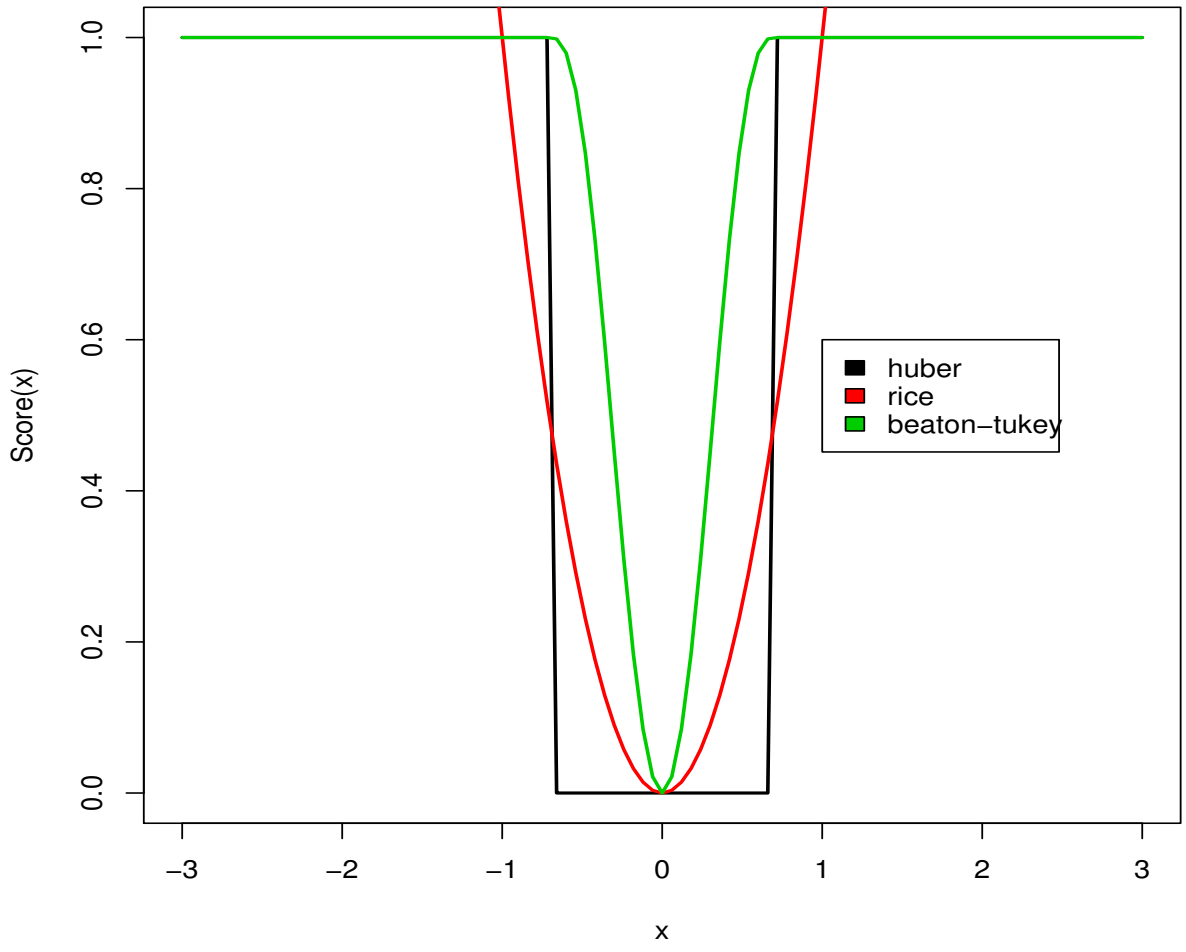
$Z_1^* = Z_2 - Z_1$, $\{Z_i\}_{i=1,2}$ v.a.'s i.i.d. con $Z_1 \sim F_0$.

Nota: Bajo ciertas condiciones sobre χ, $\widehat{\sigma}_{\mathrm{M},n}$ es es la única solución positiva de

$$\frac{1}{n-1}\sum_{i=1}^{n-1}\chi\left(\frac{Y_i^*}{as}\right) = b.$$

Escores

Ejemplos típicos de funciones de escores



Ejemplos

- "Estimador de Rice".
  $\chi(x) = x^2$, $b = 1$ y $a = \sqrt{2}$

$$\widehat{\sigma}_{\mathsf{R},n} = \left(\frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_i^*)^2 \right)^{1/2},$$

- "Estimador de Boente, Fraiman y Meloche".
  $\chi(x) = I_{\{y:\ |y| > \Phi^{-1}(3/4)\}}(x)$, $b = 1/2$ y $a = \sqrt{2}$

$$\widehat{\sigma}_{\mathsf{MAD},n} = \frac{q_{1/2}}{\sqrt{2}\Phi^{-1}(3/4)}$$

... Ejemplos

- "Estimador primer cuartil".
  $\chi(x) = I_{\{y:\ |y|>\Phi^{-1}(5/8)\}}(x)$, $b = 3/4$ y $a = \sqrt{2}$

$$\widehat{\sigma}_{\mathsf{PC},n} = \frac{q_{1/4}}{\sqrt{2}\Phi^{-1}(5/8)},$$

- "$M-$estimador con función de Beaton-Tukey", $\widehat{\sigma}_{\mathsf{BT},n}$.

$$\chi(x) = \chi_c(x) = \begin{cases} 3\left(x/c\right)^2 - 3\left(x/c\right)^4 + \left(x/c\right)^6 & \text{si} \quad |x| \leq c \\ 1 & \text{si} \quad |x| > c \end{cases}$$

con $c = 0.70417$, $b = 3/4$ y $a = \sqrt{2}$.

$M-$funcionales de escala

Huber (en el 64) define el $M-$ funcional de escala,

$$S : \mathcal{D} \longrightarrow I\!R_0^+,$$

como

$$S(F) = \inf \left\{ s > 0 : \; E\left[\chi\left(\frac{X}{as}\right)\right] \leq b \right\}.$$

Bajo ciertas condiciones sobre la función de escores,χ,

$S(F)$ es la única solución positiva de $E\left[\chi\left(\frac{X}{as}\right)\right] = b.$

...$M-$funcionales de escala

En nuestro problema, S opera sobre convoluciones de las distribuciones de los errores:

si $U_i \sim G$ y $G \in \mathcal{P}_\epsilon(F_0)$ consideramos $S(G^*_\sigma)$ donde G^*_σ es la f.d. de $\sigma U_1^* = \sigma(U_2 - U_1)$.

Observar que:

✓ Consistencia Fisher: $\forall \sigma > 0,\ S(F^*_{0,\sigma}) = \sigma$.

✓ Equivariancia a transformaciones de escala:

$$\forall \sigma > 0 :\ S(G^*_\sigma) = \sigma S(G^*).$$

✓ Si $F_n(y) = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n-1} I_{(-\infty,y]}(Y_i^*)$, $S(F_n) = \widehat{\sigma}_{\mathsf{M},n}$.

Consistencia fuerte y normalidad asintótica

Teorema: Bajo las hipótesis

- χ es continua, acotada con $\|\chi\|_\infty = 1$, par, $\chi(0) = 0$ y estrictamente creciente sobre el conjunto $\{x : 0 < \chi(x) < 1\}$.
- $M_n = \max\limits_{1 \le i \le n-1} (x_{i+1} - x_i) \to 0.$
- La función de regresión, g, es continua.

la sucesión de $M-$estimadores converge casi seguramente al valor del funcional en G_σ^*, es decir

$$\forall G \in \mathcal{P}_\epsilon(F_0): \ S(F_n) = \widehat{\sigma}_{\mathsf{M},n} \xrightarrow{c.s.} S(G_\sigma^*).$$

Normalidad asintótica

Teorema: Si son válidas las hipótesis H.1.,
H.4. Continuidad Lipschitz de g.
H.6. $M_n = \max_{1 \le i \le n-1} |x_{i+1} - x_i| = O(n^{-1})$, y
H.7. χ es de clase C^2 y las funciones $\chi_1(u) = u\chi'(u)$ y $\chi_2(u) = u^2\chi''(u)$ están acotadas.
entonces

$$n^{1/2} \left(\widehat{\sigma}_{\mathsf{M},n} - S(G_\sigma^*) \right) \xrightarrow{d} N(0, v),$$

donde...

... Normalidad asintótica

... la varianza asintótica es $v = v_1/v_2^2$, siendo

$$v_1 = Var\left[\chi\left(\frac{\sigma U_1^*}{aS(G_\sigma^*)}\right)\right] + 2\,Cov\left[\chi\left(\frac{\sigma U_1^*}{aS(G_\sigma^*)}\right), \chi\left(\frac{\sigma U_2^*}{aS(G_\sigma^*)}\right)\right]$$

$$v_2 = E\left[\chi'\left(\frac{\sigma U_1^*}{aS(G_\sigma^*)}\right)\left(\frac{\sigma U_1^*}{a(S(G_\sigma^*))^2}\right)\right].$$

Sección B

Sesgo Máximo y Punto de Ruptura

Robustez: Sesgo Máximo

En ausencia de contaminación ($\epsilon = 0$) estimamos asintóticamente sin sesgo:

$$S(G_\sigma^*) = \sigma \text{ si } G = F_0$$

Cuando hay contaminación en los datos ($\epsilon > 0$), en general, $\{\widehat{\sigma}_{\mathsf{M},n}\}_{n\geq 1}$ es asintóticamente sesgado

$$S(G_\sigma^*) \neq \sigma \text{ si } G \neq F_0$$

Robustez: Sesgo Máximo

¿Cómo cuantificamos el sesgo?

☛ Sesgo Asintótico:

$$B\left(S(G_\sigma^*)\right) = \left(S(G_\sigma^*)/\sigma\right) - 1.$$

☛ Sesgo Asintótico Generalizado (podemos pesar inliers y outliers de forma diferencial):

$$B_g\left(S(G_\sigma^*)\right) = \begin{cases} L_1\left(\dfrac{S(G_\sigma^*)}{\sigma}\right), & \text{si } 0 < S(G_\sigma^*) \leq \sigma, \\ L_2\left(\dfrac{S(G_\sigma^*)}{\sigma}\right), & \text{si } \ \sigma < S(G_\sigma^*) < +\infty. \end{cases}$$

Robustez: Sesgo Máximo

Para cuantificar robustez necesitamos mirar en $\mathcal{P}_\epsilon(F_0)$

$$\Downarrow$$

Máximo sesgo asintótico generalizado

$$\overline{B}_g(\epsilon) = \sup_{G \in \mathcal{P}_\epsilon(F_0)} B_g\left(S(G^*_\sigma)\right).$$

Por la equivariancia del funcional $S(\cdot)$

$$\overline{B}_g(\epsilon) = \sup_{G \in \mathcal{P}_\epsilon(F_0)} B_g\left(S(G^*_\sigma)\right) = \sup_{G \in \mathcal{P}_\epsilon(F_0)} B_g\left(S(G^*)\right).$$

$$\Downarrow$$

Asumimos que $\sigma = 1$.

Robustez: Sesgo Máximo

¿Cuál es la relación entre robustez y máximo sesgo asintótico (generalizado)?

La sucesión de $M-$estimadores, $\{\widehat{\sigma}_{\mathsf{M},n}\}_{n\geq 1}$, es asintóticamente robusta si

$$\exists\, \epsilon \in (0, 1/2] : \ \overline{B}_g(\epsilon) < \infty$$

... Robustez: Sesgo Máximo

Por la equivariancia del $M-$funcional y por la monotonía de L_1 y L_2 se deduce que

$$\overline{B}_g(\epsilon) = \text{máx}\left\{ L_1\left(\inf_{G\in\mathcal{P}_\epsilon(F_0)} S(G^*)) \right), L_2\left(\sup_{G\in\mathcal{P}_\epsilon(F_0)} S\left(G^*\right) \right) \right\}$$

Se determina ϵ tal que

$$\inf_{G\in\mathcal{P}_\epsilon(F_0)} S(G^*) > 0 \text{ y } \sup_{G\in\mathcal{P}_\epsilon(F_0)} S\left(G^*\right) < \infty$$

Robustez: condiciones de finitud del $\overline{B}_g(\epsilon)$

La constante de ajuste b determina la robustez de los $M-$estimadores:

Teorema. Asumamos H.1.,

H.8. $H \in \mathcal{D}_0$ la colección de las f.d. absolutamente continuas.

H.9. F_0 posee densidad estrictamente positiva, unimodal y simétrica.

Entonces $\overline{B}_g(\epsilon) < \infty$

- ✓ para $b = 3/4$ si $\epsilon < 1/2$;
- ✓ para $b \in (0, 3/4)$ si $\epsilon < 1 - \sqrt{1-b}$;
- ✓ para $b \in (3/4, 1)$ si $\epsilon < \sqrt{1-b}$.

Robustez: punto de ruptura asintótico

El punto de ruptura asintótico de $\{\widehat{\sigma}_{\mathsf{M},n}\}_{n\geq 1}$

$$\epsilon^* = \epsilon^*(\{\widehat{\sigma}_{\mathsf{M},n}\}_{n\geq 1}) = \inf\left\{\epsilon \in (0, 1/2] : \ \overline{B}_g(\epsilon) = \infty\right\}.$$

Teorema. $\epsilon^*(\{\widehat{\sigma}_{\mathsf{M},n}\}_{n\geq 1})$ satisface

- ✓ Si $b = 3/4$, entonces $\epsilon^* = 1/2$.
- ✓ Si $b \in (0, 3/4)$, entonces $\epsilon^* = 1 - \sqrt{1-b}$.
- ✓ Si $b \in (3/4, 1)$, entonces $\epsilon^* = \sqrt{1-b}$.

$$\Downarrow$$

El máximo punto de ruptura asintótico que puede ser alcanzado, cuando b varía en el $(0, 1)$, es $\epsilon^*_{opt} = 1/2$ y se alcanza en $b = 3/4$.

Robustez: observaciones

✓ ϵ^* se considera bajo el modelo de contaminaciones independientes.

✓ La Hipótesis H.8. no es muy restrictiva.

✓ Para los estimadores introducidos en los ejemplos:

$$\epsilon^*(\{\widehat{\sigma}_{\mathsf{R},n}\}_{n\geq 1}) = 0,\ \epsilon^*(\{\widehat{\sigma}_{\mathsf{MAD},n}\}_{n\geq 1}) \approx 0.29$$

$$\epsilon^*(\{\widehat{\sigma}_{\mathsf{PC},n}\}_{n\geq 1}) \approx 1/2,\ \epsilon^*(\{\widehat{\sigma}_{\mathsf{BT},n}\}_{n\geq 1}) = 1/2.$$

Sección C

El funcional para muestras finitas

Simulación Monte Carlo: objetivos

Evaluación del comportamiento de los estimadores

$\widehat{\sigma}_{\mathrm{R},n}$, $\widehat{\sigma}_{\mathrm{MAD},n}$, $\widehat{\sigma}_{\mathrm{PC},n}$ y $\widehat{\sigma}_{\mathrm{BT},n}$

$\Downarrow$

Objetivos

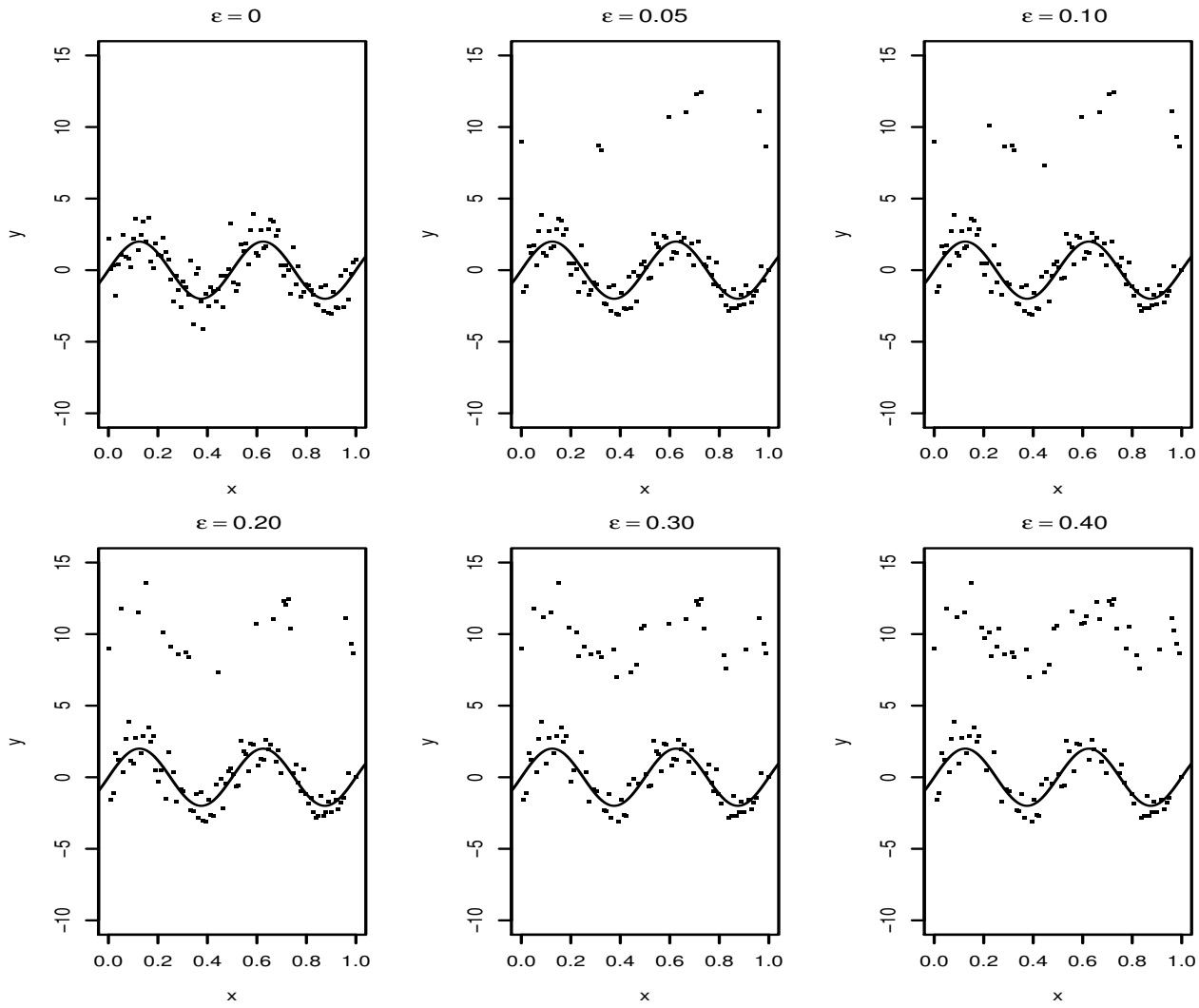
✓ Investigar las propiedades de eficiencia cuando $\epsilon = 0$.

✓ Comparar los errores cuadráticos medios (E.C.M.) estimados en presencia de outliers.

✓ Estudiar si el incremento de E.C.M. puede atribuirse al sesgo y/o la varianza.

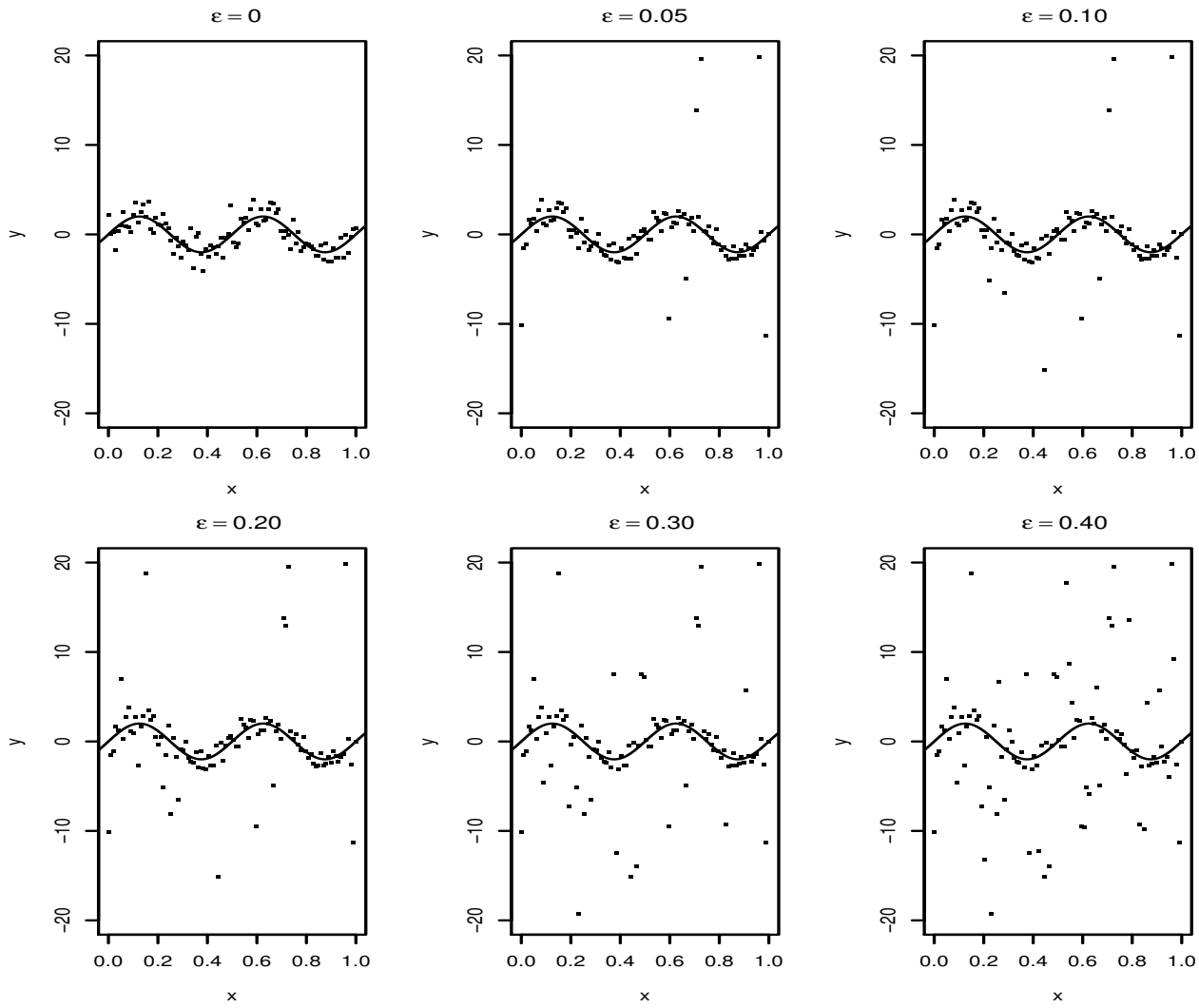
Simulación Monte Carlo: gráficos de modelos

Valores de las respuestas versus puntos del diseño para $g(x) = \text{sen}(4\pi x)$, $\sigma \equiv 1$. La curva corresponde al gráfico de la función g. Modelo central $F_0 = N(0, 1)$ y contaminación asimétrica $H = N(10, 1)$.



Simulación Monte Carlo: gráficos de modelos

Valores de las respuestas versus puntos del diseño para $g(x) = \text{sen}(4\pi x)$, $\sigma \equiv 1$. La curva corresponde al gráfico de la función g. Modelo central $F_0 = N(0, 1)$ y contaminación simétrica $H = N(0, 10)$.



Simulación Monte Carlo: Eficiencia estimada y exacta

Eficiencias de $\widehat{\sigma}_{\mathrm{MAD},n}$, $\widehat{\sigma}_{\mathrm{PC},n}$ y $\widehat{\sigma}_{\mathrm{BT},n}$ relativas a $\widehat{\sigma}_{\mathrm{R},n}$ con $\epsilon = 0$.

n	E.R.$(\widehat{\sigma}_{\mathrm{MAD},n}, \widehat{\sigma}_{\mathrm{R},n})$	E.R.$(\widehat{\sigma}_{\mathrm{PC},n}, \widehat{\sigma}_{\mathrm{R},n})$	E.R.$(\widehat{\sigma}_{\mathrm{BT},n}, \widehat{\sigma}_{\mathrm{R},n})$
20	0.350	0.139	0.200
50	0.426	0.192	0.276
100	0.451	0.206	0.286
Asintótica	0.454	0.214	0.297

Simulación Monte Carlo: E.C.M.

	Estimador	$\epsilon = 0.05$	$\epsilon = 0.10$	$\epsilon = 0.20$	$\epsilon = 0.30$
$H = N(0, 10)$					
	$\widehat{\sigma}_{\mathsf{R},n}$	2.319	5.452	12.78	20.770
	$\widehat{\sigma}_{\mathsf{MAD},n}$	0.038	0.100	0.485	2.106
	$\widehat{\sigma}_{\mathsf{PC},n}$	0.063	0.118	0.382	1.050
	$\widehat{\sigma}_{\mathsf{BT},n}$	0.047	0.098	0.356	1.055
$H = N(10, 1)$					
	$\widehat{\sigma}_{\mathsf{R},n}$	2.037	4.703	9.778	13.670
	$\widehat{\sigma}_{\mathsf{MAD},n}$	0.045	0.126	0.668	5.461
	$\widehat{\sigma}_{\mathsf{PC},n}$	0.069	0.135	0.412	0.897
	$\widehat{\sigma}_{\mathsf{BT},n}$	0.052	0.115	0.393	0.913
$H(y) = \Delta_{10}(y)$					
	$\widehat{\sigma}_{\mathsf{R},n}$	1.813	5.032	12.450	20.460
	$\widehat{\sigma}_{\mathsf{MAD},n}$	0.039	0.126	1.130	71.250
	$\widehat{\sigma}_{\mathsf{PC},n}$	0.063	0.138	0.668	3.576
	$\widehat{\sigma}_{\mathsf{BT},n}$	0.047	0.117	0.657	4.165

Simulación Monte Carlo: análisis y conclusiones.

Conclusiones a partir del $\widehat{\text{E.C.M.}}$:

- ✓ Bajo contaminación, $\widehat{\sigma}_{\text{R},n}$ tiene el peor desempeño.
- ✓ Para bajos niveles de ϵ, $\widehat{\sigma}_{\text{MAD},n}$ es ligeramente mejor que sus competidores robustos.
- ✓ Si ϵ aumenta $\widehat{\sigma}_{\text{PC},n}$ y $\widehat{\sigma}_{\text{BT},n}$ se comportan
  - mejor que $\widehat{\sigma}_{\text{MAD},n}$ para contaminaciones simétricas.
  - sustancialmente mejor que $\widehat{\sigma}_{\text{MAD},n}$ para contaminaciones asimétricas e intercaladas.
- ✓ Para contaminaciones intercaladas todos los estimadores se deterioran a partir de $\epsilon = 0.25$.