Nonparametric estimation in a multiplicative noise model

Charlotte Dion[(1),(2)]
Joint work with **Fabienne Comte**[(2)]

(1) *LJK, UMR CNRS 5224, Université Grenoble Alpes, Grenoble*
(2) *MAP5, UMR CNRS 8145, Université Paris Descartes, Paris Cité*

Lundi 18 avril 2016

Motivation: the model

Nonnegative random variable X

- height, weight...
- time between the first symptom of a disease and the death of the patient $\rightarrow$ survival data.

Interest: nonparametric estimation of

- the density function f
- the survival function $\overline{F}(x) = \mathbb{P}(X > x) = \int_x^\infty f(u)du$

  $(\mathbb{E}[X] = \int \overline{F}, \mathbb{E}[h(X)])$.

Motivation: the model

Classical noise model

$$Y_i = X_i + \varepsilon_i, i = 1, \ldots, n$$

with $\mathbb{E}[\varepsilon_i] = 0$. But often the noise depends on the level of the signal:

$$Y_i = X_i + \alpha X_i \varepsilon_i, \quad \alpha \in \mathbb{R}, \quad Y_i = X_i \underbrace{(1 + \alpha \varepsilon_i)}_{\Downarrow}$$

$$Y_i = X_i U_i, \ i = 1, \ldots, n, \quad U_i \sim \mathcal{U}_{[1-a,1+a]}, \ 0 < a < 1$$

with $\mathbb{E}[U_i] = 1$.

Motivation: the model

Classical noise model

$$Y_i = X_i + \varepsilon_i, i = 1, \ldots, n$$

with $\mathbb{E}[\varepsilon_i] = 0$. But often the noise depends on the level of the signal:

$$Y_i = X_i + \alpha X_i \varepsilon_i, \quad \alpha \in \mathbb{R}, \quad Y_i = X_i \underbrace{(1 + \alpha \varepsilon_i)}_{\Downarrow}$$

$$Y_i = X_i U_i, \ i = 1, \ldots, n, \quad U_i \sim \mathcal{U}_{[1-a,1+a]}, \ 0 < a < 1$$

with $\mathbb{E}[U_i] = 1$.

$\rightarrow$ **What does it represent?**

A partial transmission of the information X_i up to an error of order $\pm 100a\%$:

- unintentionally during a survey

What does it represent?

- deliberately to mask some data.

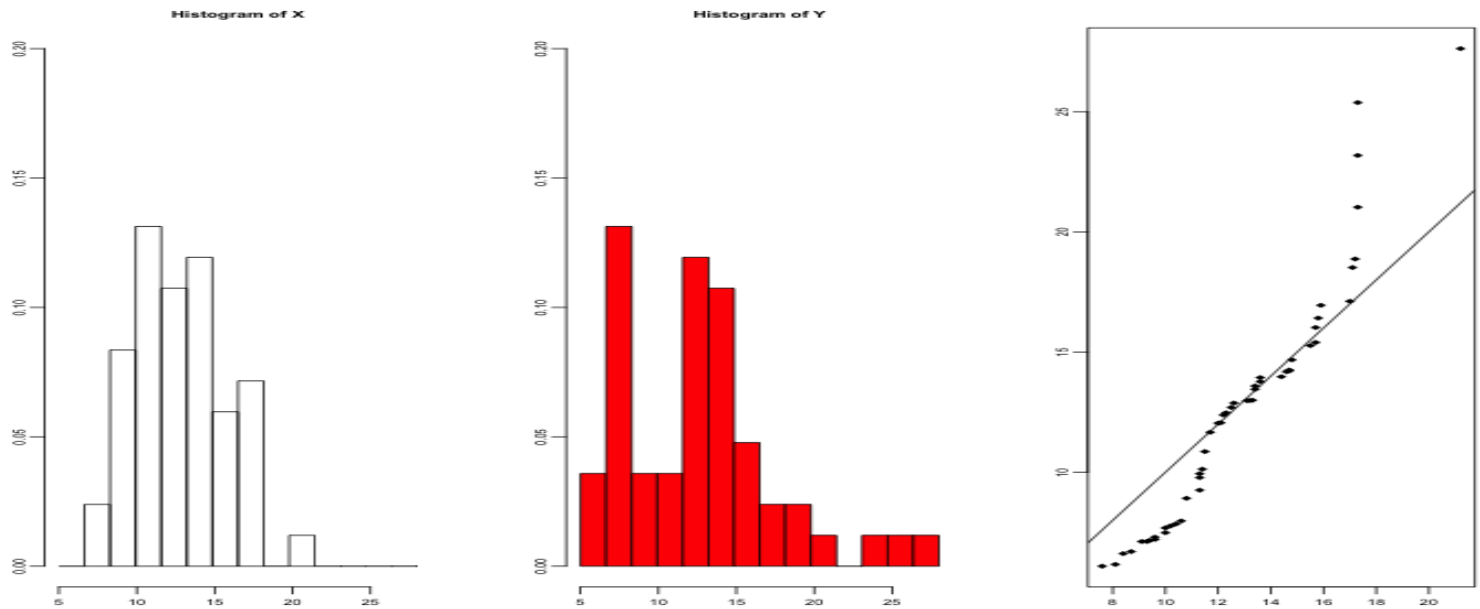


Figure : X, Y, X vs Y, $a = 0.5$

Motivation: literature case $U_i \sim U_{[0,1]}$

Vardi (1989), Vardi and Zhang (1992) asymptotic framework

- Link with deconvolution method.
- If $\varepsilon \sim \mathcal{E}(1)$, $\exp(-\varepsilon) \sim U_{[0,1]}$.
- The density function of a nonnegative random variable Y is decreasing $\Leftrightarrow Y = XU$ with $U \sim U_{[0,1]}$ independent of X.

Motivation: literature case $U_i \sim U_{[0,1]}$

Vardi (1989), Vardi and Zhang (1992) asymptotic framework

- Link with deconvolution method.
- If $\varepsilon \sim \mathcal{E}(1)$, $\exp(-\varepsilon) \sim U_{[0,1]}$.
- The density function of a nonnegative random variable Y is decreasing $\Leftrightarrow Y = XU$ with $U \sim U_{[0,1]}$ independent of X.

Asgharian *et. al.* (2012) asymptotic nonparametric estimation of F.

Brunel *et. al.* (2015) (non-asymptotic) adaptive estimator of f and of survival function $\overline{F}$, optimal rates of convergence.

Model

$$Y_i = X_i U_i,\ i = 1, \ldots, n, \quad U_i \sim \mathcal{U}_{[1-a,1+a]},\ 0 < a < 1$$

- the $(X_i)_{\{i=1,\ldots,n\}}$ and $(U_i)_{\{i=1,\ldots,n\}}$ are independent
- the X_i are *i.i.d.* with density f
- the U_i are *i.i.d.* with density $\mathcal{U}_{[1-a,1+a]}$, a known
- the Y_i are observed, *i.i.d.* with density f_Y on $\mathbb{R}^+$.

Issues: how can we estimate the density f and the associated survival function $\overline{F}$?

Notations

$$\mathbb{L}^2(\mathbb{R}^+) = \{t : \mathbb{R}^+ \to \mathbb{R}, \int_0^\infty |t(x)|^2 dx < \infty\}$$

and the associated scalar product $\langle t, v\rangle = \int_0^{+\infty} t(x)v(x)dx$ and norm $\|t\|^2 = \int_{\mathbb{R}^+} |t(x)|^2 dx$.

If t is bounded: $\|t\|_\infty = \sup_{x\in\mathbb{R}^+} |t(x)|$.

Assumption $f \in \mathbb{L}^2(\mathbb{R}^+)$

Density f_Y

$$f_Y(y) = \frac{1}{2a} \int_{\frac{y}{1+a}}^{\frac{y}{1-a}} \frac{f(x)}{x} dx, \quad y \in]0, +\infty[,$$

- If $\|f\|_\infty < +\infty$, then $\|f_Y\|_\infty < \infty$
- $y f_Y(y) \underset{y \to 0}{\to} 0$ and $y f_Y(y) \underset{y \to +\infty}{\to} 0$

Auxiliary function

For a bounded t, derivable and $t' \in \mathbb{L}^2(\mathbb{R}^+)$,

$$\begin{aligned}\mathbb{E}[t(Y_1) + Y_1 t'(Y_1)] &= \frac{1}{2a}\int_0^{+\infty} t(y)\left[f\left(\frac{y}{1+a}\right) - f\left(\frac{y}{1-a}\right)\right] dy \\ &= \langle t, g\rangle.\end{aligned}$$

$$g(x) := \frac{1}{2a}\left[f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right)\right],\ x \in \mathbb{R}^+$$

$$\mathbb{E}[\psi_t(Y_1)] = \langle t, g\rangle \quad \text{with} \quad \psi_t(y) := t(y) + yt'(y).$$

Auxiliary function

For a bounded t, derivable and $t' \in \mathbb{L}^2(\mathbb{R}^+)$,

$$\begin{aligned}\mathbb{E}[t(Y_1) + Y_1 t'(Y_1)] &= \frac{1}{2a}\int_0^{+\infty} t(y)\left[f\left(\frac{y}{1+a}\right) - f\left(\frac{y}{1-a}\right)\right] dy \\ &= \langle t, g\rangle.\end{aligned}$$

$$g(x) := \frac{1}{2a}\left[f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right)\right],\ x \in \mathbb{R}^+$$

$$\mathbb{E}[\psi_t(Y_1)] = \langle t, g\rangle \quad \text{with} \quad \psi_t(y) := t(y) + yt'(y).$$

$\rightarrow$ Strategy (different from $U \sim \mathcal{U}_{[0,1]}$):

- Build a projection estimator of g.
- Look for an inversion formula to get f.

$$f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right) \quad = \quad 2ag(x)$$

$$
\begin{aligned}
f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right) &= 2ag(x) \\
f(x) - f\left(\frac{1+a}{1-a}x\right) &= 2ag((1+a)x)
\end{aligned}
$$

$$
\begin{aligned}
f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right) &= 2ag(x) \\
f(x) - f\left(\frac{1+a}{1-a}x\right) &= 2ag((1+a)x) \\
f\left(\frac{1+a}{1-a}x\right) - f\left(\left(\frac{1+a}{1-a}\right)^2 x\right) &= 2ag\left(\frac{1+a}{1-a}(1+a)x\right) \\
&\vdots \\
f\left(\left(\frac{1+a}{1-a}\right)^{N-1} x\right) - f\left(\left(\frac{1+a}{1-a}\right)^{N} x\right) &= 2ag\left(\left(\frac{1+a}{1-a}\right)^{N-1}(1+a)x\right)
\end{aligned}
$$

$$
\begin{aligned}
f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right) &= 2ag(x) \\
f(x) - f\left(\frac{1+a}{1-a}x\right) &= 2ag((1+a)x) \\
f\left(\frac{1+a}{1-a}x\right) - f\left(\left(\frac{1+a}{1-a}\right)^2 x\right) &= 2ag\left(\frac{1+a}{1-a}(1+a)x\right) \\
&\vdots \\
f\left(\left(\frac{1+a}{1-a}\right)^{N-1} x\right) - f\left(\left(\frac{1+a}{1-a}\right)^{N} x\right) &= 2ag\left(\left(\frac{1+a}{1-a}\right)^{N-1}(1+a)x\right) \\
f(x) - f\left(\left(\frac{1+a}{1-a}\right)^{N} x\right) &= 2a\sum_{k=0}^{N-1} g\left(\left(\frac{1+a}{1-a}\right)^{k}(1+a)x\right)
\end{aligned}
$$

$$
\begin{aligned}
f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right) &= 2ag(x) \\
f(x) - f\left(\frac{1+a}{1-a}x\right) &= 2ag((1+a)x) \\
f\left(\frac{1+a}{1-a}x\right) - f\left(\left(\frac{1+a}{1-a}\right)^2 x\right) &= 2ag\left(\frac{1+a}{1-a}(1+a)x\right) \\
&\vdots \\
f\left(\left(\frac{1+a}{1-a}\right)^{N-1} x\right) - f\left(\left(\frac{1+a}{1-a}\right)^{N} x\right) &= 2ag\left(\left(\frac{1+a}{1-a}\right)^{N-1}(1+a)x\right) \\
f(x) - f\left(\left(\frac{1+a}{1-a}\right)^{N} x\right) &= 2a\sum_{k=0}^{N-1} g\left(\left(\frac{1+a}{1-a}\right)^{k}(1+a)x\right)
\end{aligned}
$$

$$
f_N(x) := 2a\sum_{k=0}^{N-1} g\left(\left(\frac{1+a}{1-a}\right)^{k}(1+a)x\right)
$$

Projection

$f(x) - f_N(x) = f(((1+a)/(1-a))^N x)$, gives $\|f - f_N\| \underset{N\to\infty}{\to} 0$.

Notice that $f \in \mathbb{L}^2(\mathbb{R}^+) \Rightarrow g \in \mathbb{L}^2(\mathbb{R}^+)$.

Orthonormal basis of $\mathbb{L}^2(\mathbb{R}^+)$: $(\varphi_j)_{j\in\mathbb{N}}$,

$$g(x) = \sum_{j=0}^{\infty} a_j(g)\varphi_j(x), \quad \text{with } a_j(g) = \langle \varphi_j, g\rangle.$$

For $m \in \mathcal{M}_n \subset \mathbb{N}$,

$$g_m := \sum_{j=0}^{m-1} a_j(g)\varphi_j \text{ projection } \mathcal{S}_m = \text{Vect}\{\varphi_0, \varphi_1, \ldots, \varphi_{m-1}\}$$

with

$$a_j(g) = \langle \varphi_j, g\rangle = \mathbb{E}[\varphi_j(Y_1) + Y_1\varphi_j'(Y_1)] = \mathbb{E}[\psi_{\varphi_j}(Y_1)].$$

Estimator of g and f

$$\widehat{g}_m = \sum_{j=0}^{m-1} \widehat{a}_j \varphi_j, \quad \widehat{a}_j = \frac{1}{n} \sum_{i=1}^{n} [Y_i \varphi_j'(Y_i) + \varphi_j(Y_i)] = n^{-1} \sum_{i=1}^{n} \psi_{\varphi_j}(Y_i).$$

Estimator of g and f

$$\widehat{g}_m = \sum_{j=0}^{m-1} \widehat{a}_j \varphi_j, \quad \widehat{a}_j = \frac{1}{n} \sum_{i=1}^{n} [Y_i \varphi_j'(Y_i) + \varphi_j(Y_i)] = n^{-1} \sum_{i=1}^{n} \psi_{\varphi_j}(Y_i).$$

Then, as:

$$f_N(x) = 2a \sum_{k=0}^{N-1} g\left(\left(\frac{1+a}{1-a} \right)^k (1+a)x \right)$$

it comes

$$\widehat{f}_{N,m}(x) = 2a \sum_{k=0}^{N-1} \widehat{g}_m \left(\left(\frac{1+a}{1-a} \right)^k (1+a)x \right).$$

Choice: Laguerre basis

$$\varphi_0(x) = \sqrt{2}e^{-x}, \quad \varphi_k(x) = \sqrt{2}L_k(2x)e^{-x} \text{ for } k \geq 1, \quad x \geq 0$$

with L_k le $k^{\text{ème}}$ Laguerre polynomials

$$L_k(x) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \frac{x^j}{j!}.$$

$\rightarrow$ Orthonormal basis: $\langle \varphi_j, \varphi_k \rangle = \delta_{j,k}$.

$$\forall j \geq 0, \|\varphi_j\|_\infty \leq \sqrt{2}, \text{ and } \|\varphi_j'\|_\infty \leq 2\sqrt{2}(j+1)$$

where φ_j' is the derivative function of φ_j.

$\rightarrow$ All functions in $\mathbb{L}^2(\mathbb{R}^+)$ can be decomposed on this basis.

MISE: mean integrated squared error (1)

Definition

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] = \|g - \mathbb{E}[\widehat{g}_m]\|^2 + \mathbb{E}[\|\mathbb{E}[\widehat{g}_m] - \widehat{g}_m\|^2].$$

Here: $\mathbb{E}[\widehat{g}_m] = g_m$.

MISE: mean integrated squared error (1)

Definition

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] = \|g - \mathbb{E}[\widehat{g}_m]\|^2 + \mathbb{E}[\|\mathbb{E}[\widehat{g}_m] - \widehat{g}_m\|^2].$$

Here: $\mathbb{E}[\widehat{g}_m] = g_m$. Thus

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] = \underbrace{\|g - g_m\|^2}_{\text{Bias term}} +$$

MISE: mean integrated squared error (1)

Definition

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] = \|g - \mathbb{E}[\widehat{g}_m]\|^2 + \mathbb{E}[\|\mathbb{E}[\widehat{g}_m] - \widehat{g}_m\|^2].$$

Here: $\mathbb{E}[\widehat{g}_m] = g_m$. Thus

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] = \underbrace{\|g - g_m\|^2}_{\text{Bias term}} + \underbrace{\mathbb{E}[\|g_m - \widehat{g}_m\|^2]}_{\text{Variance term}}.$$

MISE: mean integrated squared error (2)

Proposition

Assume that $\mathbb{E}[X_1^2] < +\infty$.

(i) The estimator $\widehat{g}_m$ *of* g *satisfies*

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] \leq \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n}, \quad c_1 = 4, \quad c_2 = 16\mathbb{E}[Y_1^2].$$

MISE: mean integrated squared error (2)

Proposition

Assume that $\mathbb{E}[X_1^2] < +\infty$.

(i) The estimator $\widehat{g}_m$ *of* g *satisfies*

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] \leq \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n}, \quad c_1 = 4, \quad c_2 = 16\mathbb{E}[Y_1^2].$$

(ii) The estimator $\widehat{f}_{N,m}$ *of* f *satisfies*

$$\begin{aligned} \mathbb{E}[\|\widehat{f}_{N,m} - f\|^2] &\leq \frac{8a^2}{(\sqrt{1+a} - \sqrt{1-a})^2} \left(\|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right) \\ &+ 2 \left(\frac{1-a}{1+a} \right)^N \|f\|^2. \end{aligned}$$

Adaptive selection procedure

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] \leq \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n}, \qquad c_1 = 4,\ c_2 = 16\mathbb{E}[Y_1^2].$$

Discrete collection $\mathcal{M}_n = \{m \in [\![1, n]\!],\ m^3 \leq n\}$

$$m_{th} := \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\}$$

Adaptive selection procedure

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] \leq \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n}, \qquad c_1 = 4,\ c_2 = 16\mathbb{E}[Y_1^2].$$

Discrete collection $\mathcal{M}_n = \{m \in [\![1, n]\!],\ m^3 \leq n\}$

$$m_{th} := \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\}$$

But $\|g - g_m\|^2 = \|g\|^2 - \|g_m\|^2$, $m_{th} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ -\|g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\}$.

Adaptive selection procedure

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] \leq \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n}, \qquad c_1 = 4,\ c_2 = 16\mathbb{E}[Y_1^2].$$

Discrete collection $\mathcal{M}_n = \{m \in [\![1, n]\!],\ m^3 \leq n\}$

$$m_{th} := \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\}$$

But $\|g - g_m\|^2 = \|g\|^2 - \|g_m\|^2$, $m_{th} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ -\|g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\}$.

Penalty term

$$\operatorname{pen}(m) := \kappa_1 \frac{m}{n} + \kappa_2 \mathbb{E}[Y_1^2] \frac{m^3}{n} =: \operatorname{pen}_1(m) + \operatorname{pen}_2(m).$$

Adaptive selection procedure

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] \leq \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n}, \qquad c_1 = 4,\ c_2 = 16\mathbb{E}[Y_1^2].$$

Discrete collection $\mathcal{M}_n = \{m \in [\![1, n]\!],\ m^3 \leq n\}$

$$m_{th} := \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ \|g - g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\}$$

But $\|g - g_m\|^2 = \|g\|^2 - \|g_m\|^2$, $m_{th} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ -\|g_m\|^2 + c_1 \frac{m}{n} + c_2 \frac{m^3}{n} \right\}$.

Penalty term

$$\operatorname{pen}(m) := \kappa_1 \frac{m}{n} + \kappa_2 \mathbb{E}[Y_1^2] \frac{m^3}{n} =: \operatorname{pen}_1(m) + \operatorname{pen}_2(m).$$

But $\mathbb{E}[Y_1^2]$ is unknown $\rightarrow \widehat{C}_2 = \frac{1}{n} \sum_{k=1}^{n} Y_k^2$

$$\widehat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \{-\|\widehat{g}_m\|^2 + \widehat{\operatorname{pen}}(m)\}, \quad \widehat{\operatorname{pen}}(m) = 2\kappa_1 \frac{m}{n} + 2\kappa_2 \widehat{C}_2 \frac{m^3}{n}$$

Oracle-type inequality

Final estimator,

$$\widehat{f}_{N,\widehat{m}}(x) = 2a \sum_{k=0}^{N-1} \widehat{g}_{\widehat{m}} \left(\left(\frac{1+a}{1-a} \right)^k (1+a)x \right).$$

Theorem

Assume that f is bounded and that $\mathbb{E}[X_1^8] < +\infty$. For the final estimator $\widehat{f}_{N,\widehat{m}}$ there exists κ_0 such that for $\kappa_1, \kappa_2 \geq \kappa_0$,

$$\begin{aligned} \mathbb{E}[\|\widehat{f}_{N,\widehat{m}} - f\|^2] &\leq \frac{16a^2}{(\sqrt{1+a} - \sqrt{1-a})^2} \left(6 \inf_{m \in \mathcal{M}} \{\|g - g_m\|^2 + \text{pen}(m)\} + \frac{C_a}{n} \right) \\ &+ \left(\frac{1-a}{1+a} \right)^N \|f\|^2, \end{aligned}$$

where C_a is a positive constant depending on a and $\|f\|_\infty$.

Results on simulated data

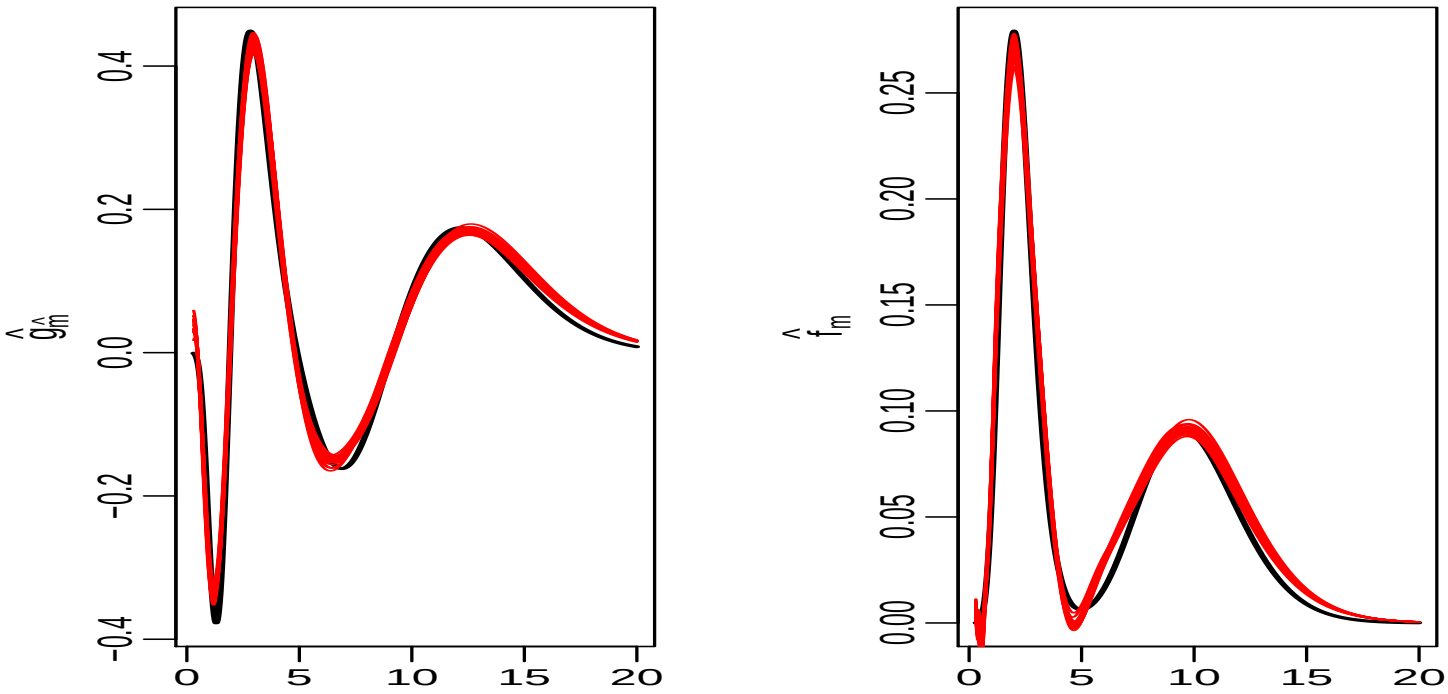


Figure : Mixed gamma case $n = 2000$, $a = 0.25$. 10 final estimators $\widehat{g}_{\widehat{m}}$ — of g — (left), 20 estimators $\widehat{f}_{N,\widehat{m}}$ — of f — (right)

And for the survival function?

Function of interest: $\overline{F}(x) = 1 - F(x) = \int_x^{+\infty} f(u)du$, $\overline{F}_Y$ for Y and:

$$\begin{aligned} \overline{G}(x) &:= \int_x^{\infty} g(u)du = \frac{1}{2a}\left[(1+a)\overline{F}\left(\frac{x}{1+a}\right) - (1-a)\overline{F}\left(\frac{x}{1-a}\right)\right] \\ &= x f_Y(x) + \overline{F}_Y(x). \end{aligned}$$

And for the survival function?

Function of interest: $\overline{F}(x) = 1 - F(x) = \int_x^{+\infty} f(u)du$, $\overline{F}_Y$ for Y and:

$$\begin{aligned} \overline{G}(x) &:= \int_x^{\infty} g(u)du = \frac{1}{2a}\left[(1+a)\overline{F}\left(\frac{x}{1+a}\right) - (1-a)\overline{F}\left(\frac{x}{1-a}\right)\right] \\ &= x f_Y(x) + \overline{F}_Y(x). \end{aligned}$$

$$\overline{F}_N(x) := \frac{2a}{1+a}\sum_{k=0}^{N-1}\left(\frac{1-a}{1+a}\right)^k \overline{G}\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right)$$

$\rightarrow$ $\overline{G}(0) = 1$ thus $\lim_{N\to\infty} \overline{F}_N(0) = 1$ which is coherent with $\overline{F}(0) = 1$.

And for the survival function?

Assumption $\mathbb{E}[X_1^2] < +\infty$ then $\overline{F} \in \mathbb{L}^2(\mathbb{R}^+)$, and $\overline{G} \in \mathbb{L}^2(\mathbb{R}^+)$.

Orthogonal projection of $\overline{G}$ on $\mathcal{S}_m$:

$$\overline{G}_m = \sum_{j=0}^{m-1} b_j(\overline{G})\varphi_j, \qquad \text{with } b_j(\overline{G}) := < \overline{G}, \varphi_j > = \mathbb{E}[Y\varphi_j(Y)] + < \overline{F}_Y, \varphi_j > .$$

And for the survival function?

Assumption $\mathbb{E}[X_1^2] < +\infty$ then $\overline{F} \in \mathbb{L}^2(\mathbb{R}^+)$, and $\overline{G} \in \mathbb{L}^2(\mathbb{R}^+)$.

Orthogonal projection of $\overline{G}$ on $\mathcal{S}_m$:

$$\overline{G}_m = \sum_{j=0}^{m-1} b_j(\overline{G})\varphi_j, \qquad \text{with } b_j(\overline{G}) := < \overline{G}, \varphi_j > = \mathbb{E}[Y\varphi_j(Y)] + < \overline{F}_Y, \varphi_j > .$$

Projection estimator

$$\widetilde{\overline{G}}_m = \sum_{j=0}^{m-1} \widetilde{b}_j \varphi_j, \quad \widetilde{b}_j = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i\varphi_j(Y_i) + \int_{\mathbb{R}^+} \varphi_j(x)\mathbb{1}_{Y_i \geq x} dx\right]$$

$$\widetilde{\overline{F}}_{N,m}(x) = \frac{2a}{1+a}\sum_{k=0}^{N-1}\left(\frac{1-a}{1+a}\right)^k \widetilde{\overline{G}}_m\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right).$$

- Better results than a classic deconvolution strategy.
- Transposition of the method for the estimation of the survival function.
- A kernel estimator instead of the projection estimator is possible.
- Application to data protection.

References

▶ Laguerre estimation for k-monotone densities observed with noise, **Belomestny, D., Comte, F. and Genon-Catalot, V.** (2016), *Preprint HAL*

▶ Nonparametric density and survival function estimation in the multiplicative censoring model **Brunel, E., Comte, F. and Genon-Catalot, V.**(2015) *to appear in TEST*

▶ Nonparametric estimation in a multiplicative censoring model with symmetric noise **Comte F. and Dion. C.**(2016), *Preprint HAL*

▶ Privacy protection and quantile estimation from noise multiplied data **Sinha, B. and Nayak, T. and Zayatz, L.** (2011) *Sankhya B*

▶ Large sample study of empirical distributions in a random-multiplicative censoring model **Vardi, Y. and Zhang, C-H.** (1992) *The Annals of Statistics*

Thank you for your attention !

If a is unknown ?

- A K-sample is available, where the signal is constant,
  $\rightarrow$ we have observations of U: $U_1^{(1)}, \ldots, U_K^{(1)}$
  $\rightarrow$ ML estimator: $\max_{1 \leq i \leq K}(|U_i^{(1)} - 1|)$.

If a is unknown ?

- A K-sample is available, where the signal is constant,
  $\rightarrow$ we have observations of U: $U_1^{(1)}, \ldots, U_K^{(1)}$
  $\rightarrow$ ML estimator: $\max_{1 \leq i \leq K}(|U_i^{(1)} - 1|)$.
- Repeated observations of X_i are available:

$$Y_{i,k} = X_i U_{i,k}, \quad k \in \{1, 2\}, \ i = 1, \ldots, n,$$

where $(U_{i,1})_i$ and $(U_{i,2})_i$ are independent samples from $\mathcal{U}_{[1-a,1+a]}$.

$$\mathbb{E}\left[Y_{i,1}^2 / Y_{i,2}^2\right] = \frac{1 + a^2/3}{1 - a^2}$$

then

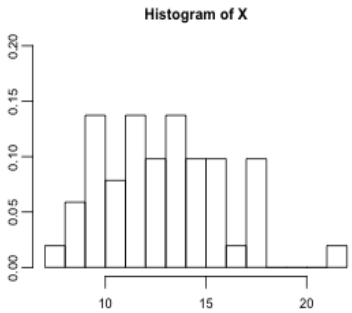
$$\widehat{a}_n = \sqrt{\frac{\overline{W}_n - 1}{\overline{W}_n + 1/3}}, \qquad \text{with } \overline{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i, \quad W_i := \frac{Y_{i,1}^2}{Y_{i,2}^2}.$$

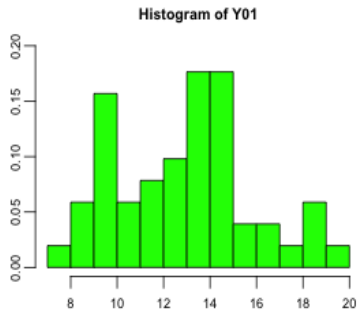
$\rightarrow$ we can plug this estimator.

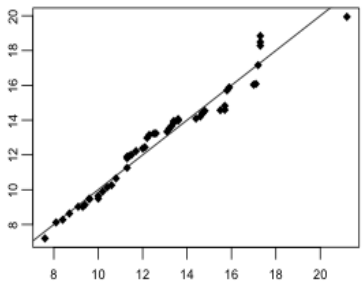
Real data: confidential protection

How to alter the data to minimize the risk of disclosure and to remain able to find the main characteristics of the original dataset?

% of people under the poverty level for the 51 states of USA: Sinha *et al.*(2011) $n = 51$ for $a = 0.1$ (from American Community Survey)







Real data: confidential protection

How to alter the data to minimize the risk of disclosure and to remain able to find the main characteristics of the original dataset?

% of people under the poverty level for the 51 states of USA: Sinha *et al.*(2011) $n = 51$ for $a = 0.1$ (from American Community Survey)

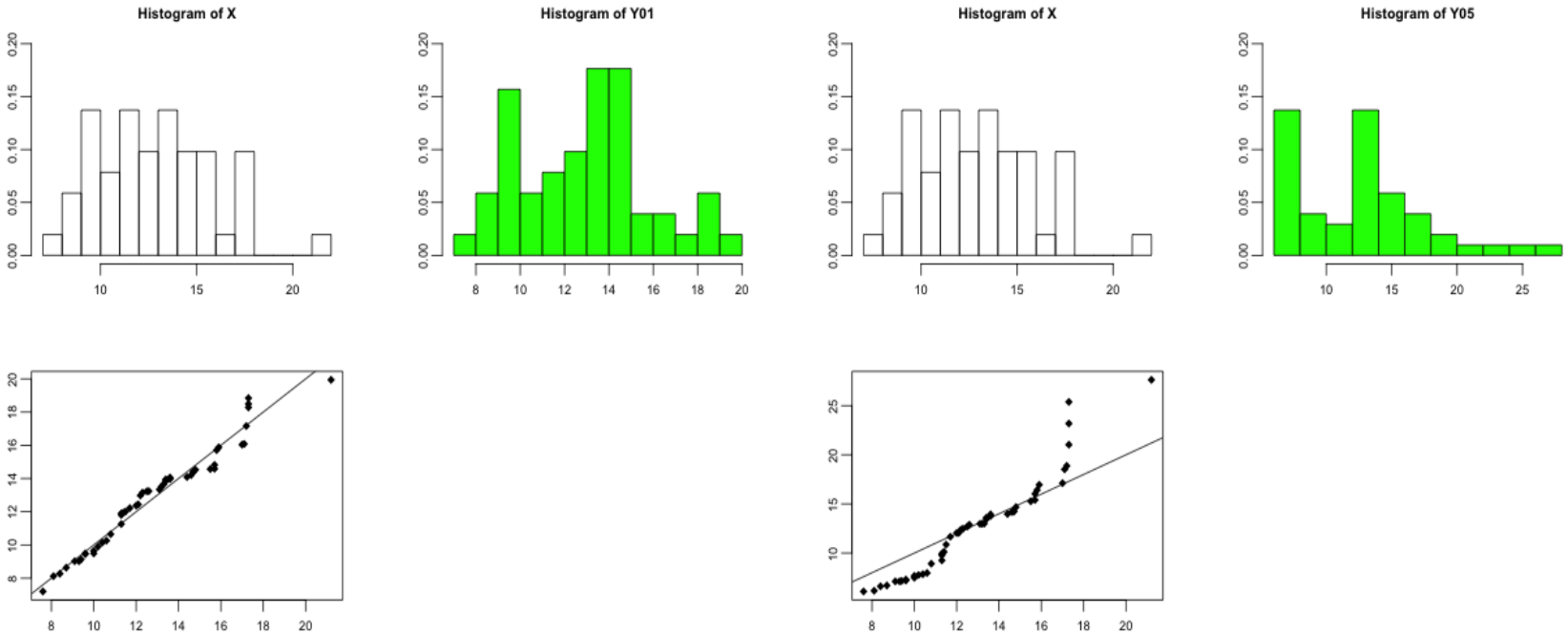


Figure : $a = 0.1$ and $a = 0.5$

Results on real data: $\widehat{f}_{N,\widehat{m}}$ better than $\widehat{f}_{Y,m}$?

▶ A Comparison of Statistical Disclosure Control Methods: Multiple Imputation Versus **Klein, M., Mathew, T. and Sinha, B.** (2013) *Research report series*

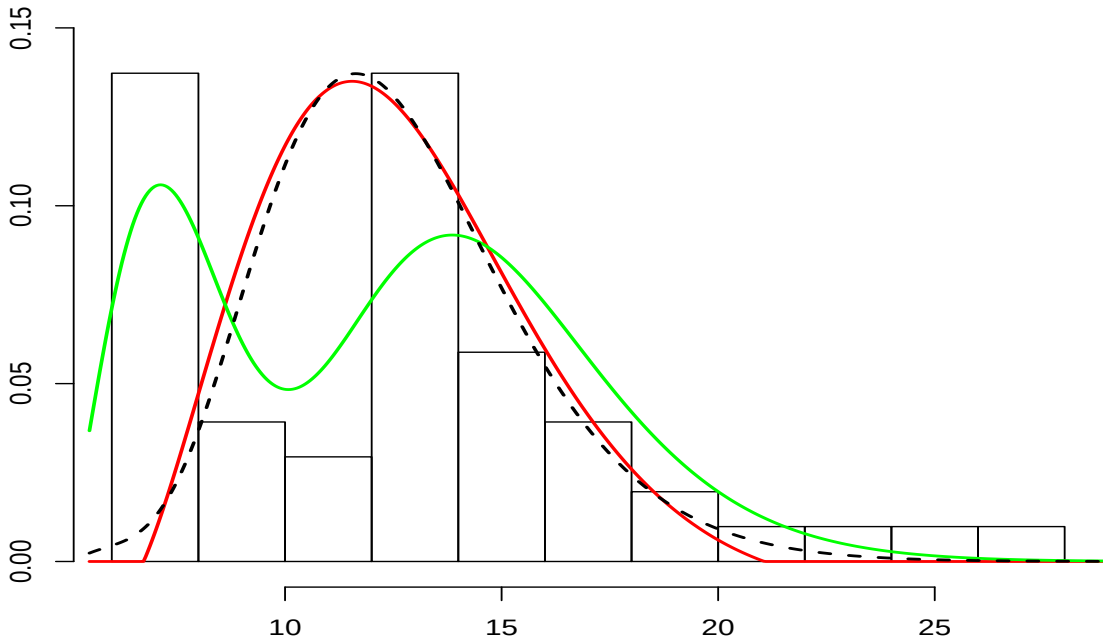


Figure : Histogram of the real data X_i's with full multiplicative noise, with $a = 0.5$, $Y_i = X_i U_i$.
- - projection estimator of f on the $(X_i)_i$, — estimator $\widehat{f}_{N,\widehat{m}}$,
— projection estimator of f_Y on the $(Y_i)_i$.

Results on real data

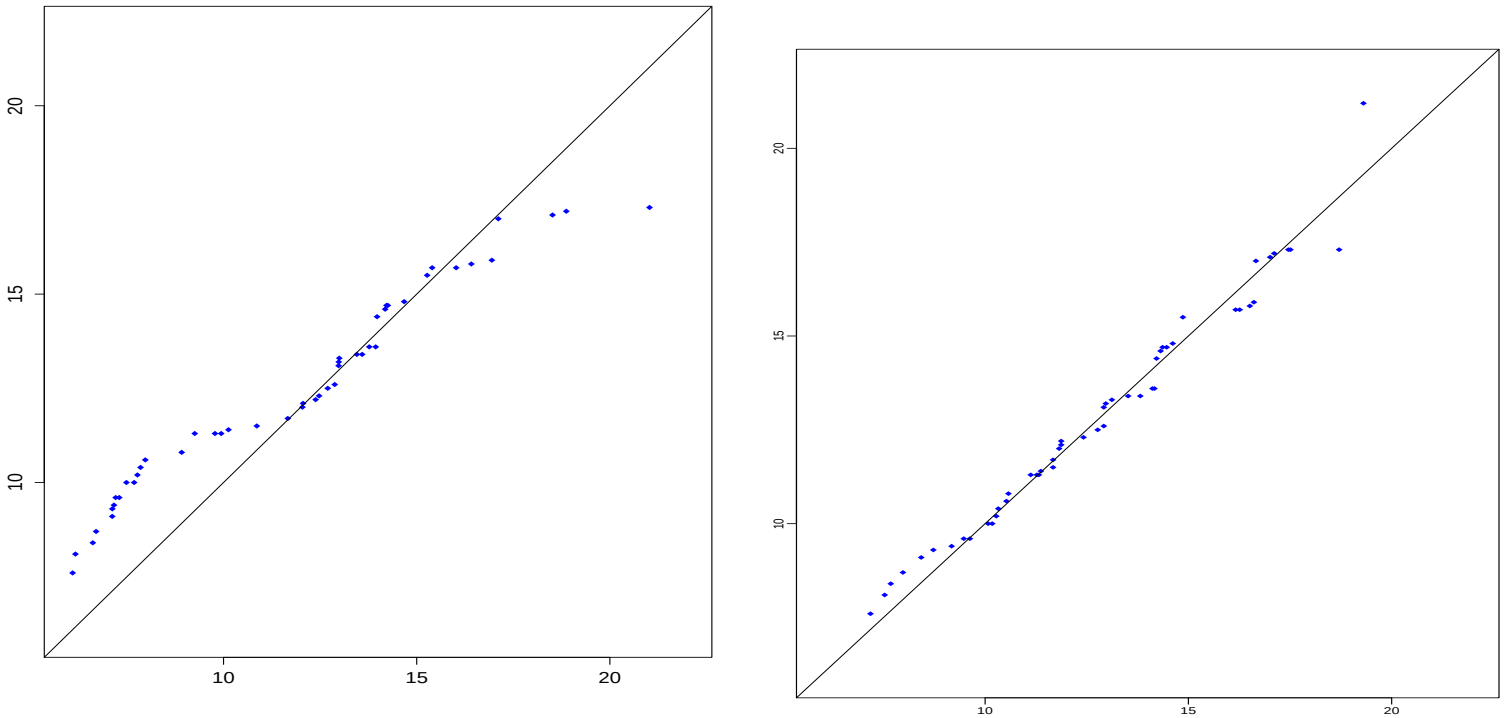


Figure : Before: Y vs X, after: Xnew vs X

$\rightarrow$ Our estimations of $Q1, Q3$, min, max are closer from the one of X as the values of Sinha *et al.*(2011)