NONPARAMETRIC ESTIMATION IN A MULTIPLICATIVE CENSORING MODEL WITH SYMMETRIC NOISE

F. COMTE[(1)] AND C. DION[(1),(2)]

Abstract. We consider the model $Y_i = X_i U_i$, $i = 1, \dots, n$, where the X_i, the U_i and thus the Y_i are all independent and identically distributed. The X_i have density f and are the variables of interest, the U_i are multiplicative noise with uniform density on $[1-a, 1+a]$, for some $0 < a < 1$, and the two sequences are independent. However, only the Y_i are observed. We study nonparametric estimation of both the density f and the corresponding survival function. In each context, a projection estimator of an auxiliary function is built, from which estimator of the function of interest is deduced. Risk bounds in term of integrated squared error are provided, showing that the dimension parameter associated with the projection step has to perform a compromise. Thus, a model selection strategy is proposed in both cases of density and survival function estimation. The resulting estimators are proven to reach the best possible risk bounds. Simulation experiments illustrate the good performances of the estimators and a real data example is described.

(1)*MAP5, UMR CNRS 8145, Université Paris Descartes, Sorbonne Paris Cité, 45 rue des Saints Pères, 75006 Paris*

(2)*LJK, UMR CNRS 5224, Université Joseph Fourier, 51 rue des Mathématiques, 38041 Grenoble*
email: fabienne.comte@parisdescartes.fr, charlotte.dion@imag.fr

AMS Subject Classification 62G07 − 62N01

Keywords. Censored data. Model selection. Multiplicative noise. Nonparametric estimator.

1. Introduction

We consider the following model

$$Y_i = X_i U_i,\ i = 1, \dots, n, \quad U_i \sim \mathcal{U}_{[1-a,1+a]},\ 0 < a < 1 \tag{1.1}$$

where $(X_i)_{\{i=1,\dots,n\}}$ and $(U_i)_{\{i=1,\dots,n\}}$ are two independent samples. The U_i's are independent and identically distributed (*i.i.d.*) random variables from uniform density on an interval $[1-a, 1+a]$ of $\mathbb{R}^+$ with $0 < 1-a < 1+a$ and a is assumed to be known. The X_i's are *i.i.d.* from an unknown density f on $\mathbb{R}^+$. Both sequences are unobserved. Only the Y_i's are observed. The model implies that they are *i.i.d.* and we denote by f_Y their density on $\mathbb{R}^+$. Our goal is to estimate nonparametrically the density f of the X_i's from the observations Y_i, $i = 1, \dots, n$.

Equation (1.1) can be obtained as follows. Classical models involving measurement errors are often additive and state that the variable of interest X_i is not directly observed because an additive noise hides it: only samples of $X_i + \xi_i$ are available, where ξ_i is an i.i.d. centred sequence. Then, in many contexts, this noise depends on the level of the signal and the simplest strategy is to consider that it is proportional to the signal. Thus, the model for the observations becomes $X_i + \alpha X_i \xi_i$, $\alpha \in \mathbb{R}$. Rewriting this $X_i(1 + \alpha\xi_i)$, we obtain a multiplicative noise model with noise $1 + \alpha\xi_i$ with mean 1. This corresponds to model (1.1) where we specified the final distribution of the noise as uniform, and symmetric around one[1].

[1]Extension to general $\mathcal{U}([a, b])$ distribution for the U_i's is possible.

In any case, Equation (1.1) models an approximate transmission of the information: the recorded values Y_i correspond to the value of interest X_i, up to an error of order of $\pm 100a\%$. This represents rather standard situations, when people have to give their height or the amount of money they devote to some specific expenses, i.e. quantities they may not know exactly with no intention to change them (for instance, weight or income may be intentionally biased). However, very few studies of this model have been conducted in the literature. We mainly found it in Sinha et al. [2011], who study "noise multiplied magnitude microdata" as a form of data masking in contexts where one needs to protect the privacy of survey respondents. The authors mainly study quantile estimation.

Nevertheless, multiplicative noise models can be found with other distributions for the noise U. The case of U following a uniform distribution on $[0,1]$ ($\mathcal{U}([0,1])$) has been introduced by Vardi [1989] who called it a "multiplicative censoring" model. This model was studied by Vardi and Zhang [1992], Asgharian et al. [2012], Abbaszadeh et al. [2013], Brunel et al. [2015], and is mostly applied in survival analysis, see van Es et al. [2000]. In these papers, nonparametric estimators of the density f or of the survival function $\overline{F} = 1 - F$, $F(x) = \int_0^x f(u)du$, of the unobserved random variable X are built and studied, but in different contexts. For instance, Asgharian et al. [2012] assume that part of the observations are directly observed and the proposed method is no longer valid if this proportion is null as in our model. In Brunel et al. [2015], kernel estimators are studied, while Abbaszadeh et al. [2013] build wavelet estimators of the density and its derivatives. The case of Gaussian U, for variables on $\mathbb{R}$, has also been considered in financial context and studied from statistical point of view by e.g. van Es et al. [2005].

In this paper, we build estimators of the density f and of the survival function $\overline{F} = 1 - F$. The operator linking the density of the observations and the density of interest is given by

$$f_Y(y) = \frac{1}{2a} \int_{\frac{y}{1+a}}^{\frac{y}{1-a}} \frac{f(x)}{x} dx, \quad y \in]0, +\infty[, \tag{1.2}$$

and the inversion of formula (1.2) is not obvious. This is why our strategy relies on two steps. Let us give here a sketch of the procedure. First, we approach an auxiliary function g expressed as a function of f and a. We prove that for an explicit transformation $t \in \mathbb{L}^2(\mathbb{R}^+) \mapsto \psi_t$ and this function g in $\mathbb{L}^2(\mathbb{R}^+)$, we have

$$\mathbb{E}[\psi_t(Y_1)] = \langle t, g \rangle, \tag{1.3}$$

where $\langle s, t \rangle = \int_{\mathbb{R}^+} s(x)t(x)dx$ denotes the scalar product of two functions of $\mathbb{L}^2(\mathbb{R}^+)$. The two functions g and ψ_t are given in Section 2. Relation (1.3) is used to build projection estimators of g. Indeed, considering the collection of spaces

$$\mathcal{S}_m = \mathrm{Vect}\{\varphi_0, \varphi_1, \ldots, \varphi_{m-1}\}$$

where $(\varphi_j)_{j \geqslant 0}$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}^+)$, the orthogonal projection g_m of g on S_m is given by $g_m = \sum_{j=0}^{m-1} a_j \varphi_j$, with $a_j = \langle g, \varphi_j \rangle$. From relation (1.3), we notice that $a_j = \mathbb{E}[\psi_{\varphi_j}(Y_1)]$ and replacing the expectation by its empirical counterpart $\widehat{a}_j$, we obtain the estimator $\widehat{g}_m = \sum_{j=0}^{m-1} \widehat{a}_j \varphi_j$. With similar ideas, we also define $\breve{\overline{G}}_m = \sum_{j=0}^{m-1} \breve{b}_j \varphi_j$ where $\breve{b}_j$ are also computed from the observations $Y_1, \ldots, Y_n$. Then we deduce, by inverting the relation between f and g (see Section 2.2) and between $\overline{F}$ and $\overline{G}$ (see Section 2.5), collections of estimators of f and $\overline{F}$, for which risk bounds are provided, in term of mean integrated squared error (MISE) on $\mathbb{R}^+$. Model selection criterion are proposed to automatically select m in both cases, and they are proven to make the adequate tradeoff between bias (m must be large enough for the projection bias to be small) and variance (estimating too many coefficients increases the estimation error), see Theorems 2.3 and 2.6.

Finally we illustrate our method on simulated and real data. Our purpose is to propose a new method of privacy protection by the mean of our multiplicative censoring model. On the data given in Sinha et al. [2011], knowing the level of noise a, we show how to recover the hidden information about the original data from the noisy observations.

The plan of the paper is the following. In Section 2, we describe our estimation method and the model selection procedure, for the density in Sections 2.2 to 2.4 and for the survival function in Section 2.5. We compute bounds on the integrated quadratic risk associated to the estimators and deduce rates of convergence. The strategy and the results are detailed for density estimation and then extended to the case of survival function estimation. In Section 3, we describe a deconvolution strategy based on the additive model obtained by taking the logarithm of (1.1): we compare our method to this one from theoretical point of view here and in practice in Section 4. Finally Section 4 illustrates the theoretical results, on simulated data (Section 4.1) and on real data (Section 4.2). Simulation experiments show the good performances of our method, and estimation on real data is presented through an example of application. Lastly, most proofs are gathered in Section 5.

2. Multiplicative denoising of density and survival function

2.1. **Notations.** The space $\mathbb{L}^2(\mathbb{R}^+)$ is the space of square integrable functions on the positive real line. The associated $\mathbb{L}^2$-norm is denoted $\|t\|^2 = \int_{\mathbb{R}^+} |t(x)|^2 dx$. The Fourier transform of $t \in \mathbb{L}^1$, for $x \in \mathbb{R}$ is: $t^\star(x) = \int t(u) e^{iux} du$. Finally, the supremum norm of a bounded function t is denoted by $\|t\|_\infty = \sup_{x \in \mathbb{R}^+} |t(x)|$. The Laguerre basis is defined by:

$$\varphi_0(x) = \sqrt{2} e^{-x}, \quad \varphi_k(x) = \sqrt{2} L_k(2x) e^{-x} \text{ for } k \geqslant 1, \quad x \geqslant 0, \tag{2.1}$$

with L_k the Laguerre polynomials

$$L_k(x) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \frac{x^j}{j!}. \tag{2.2}$$

It satisfies the orthonormality property $\langle \varphi_j, \varphi_k \rangle = \delta_{j,k}$ where $\delta_{j,k}$ is the Kronecker symbol equal to 1 if $j = k$ and to zero otherwise; and the following relations on the norms (see Abramowitz and Stegun [1964]):

$$\forall j \geqslant 0, \|\varphi_j\|_\infty \leqslant \sqrt{2}, \text{ and } \|\varphi_j'\|_\infty \leqslant 2\sqrt{2}(j+1), \tag{2.3}$$

where φ_j' is the derivative of φ_j. Any function of $\mathbb{L}^2(\mathbb{R}^+)$ can be decomposed on this basis.
Lastly, we state a useful lemma, proven in Section 5, relying on the fact that the density f_Y is given by (1.2).

Lemma 2.1. *The density f_Y defined in (1.2) satisfies $\lim_{y \to 0} y f_Y(y) = 0$ and $\lim_{y \to +\infty} y f_Y(y) = 0$.*

Lemma 2.1 is a useful property to justify the construction of the estimator.

2.2. **Estimation strategy.** Recall that f_Y is given by (1.2). Now let g be given by

$$g(x) := \frac{1}{2a} \left[f\left(\frac{x}{1+a}\right) - f\left(\frac{x}{1-a}\right) \right], \tag{2.4}$$

and consider a bounded function t, derivable and with derivative function t' in $\mathbb{L}^2(\mathbb{R}^+)$. Then, an integration by part and the Lemma 2.1 imply

$$\begin{aligned} \mathbb{E}[t(Y_1) + Y_1 t'(Y_1)] &= \frac{1}{2a} \int_0^{+\infty} t(y) \left[f\left(\frac{y}{1+a}\right) - f\left(\frac{y}{1-a}\right) \right] dy \\ &= \langle t, g \rangle. \end{aligned} \tag{2.5}$$

In other words

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

Our strategy is to use equation (2.5) to build a projection estimator of g, and then to look for an inversion of formula (2.4) to recover f. Precisely, it follows from (2.4) that

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

and iterating the relation (by changing x into $(1+a)x/(1-a)$, $x>0$), it yields

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

Thus a sequence of approximations of f, for $x>0$, is

$$f_N(x) = 2a\sum_{k=0}^{N-1} g\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right). \tag{2.6}$$

Besides, using that $f(x) - f_N(x) = f(((1+a)/(1-a))^N x)$, it is easy to check for $f \in \mathbb{L}^2(\mathbb{R}^+)$ that $\|f - f_N\|$ tends to 0 when N tends to infinity. Now, if f is square-integrable, so is g and therefore we can write its decomposition on the Laguerre basis:

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

Recall that $g_m := \sum_{j=0}^{m-1} a_j(g)\varphi_j$ is the orthogonal projection of g on $\mathcal{S}_m$. According to (2.5), we have $a_j(g) = \mathbb{E}[\varphi_j(Y_1) + Y_1\varphi_j'(Y_1)] = \langle \varphi_j, g\rangle$. Then the projection g_m of g on $\mathcal{S}_m$ is estimated by

$$\widehat{g}_m = \sum_{j=0}^{m-1} \widehat{a}_j\varphi_j, \quad \widehat{a}_j = \frac{1}{n}\sum_{i=1}^{n}[Y_i\varphi_j'(Y_i) + \varphi_j(Y_i)] = n^{-1}\sum_{i=1}^{n}\psi_{\varphi_j}(X_i), \tag{2.7}$$

with m in a finite collection $\mathcal{M}_n \subset \mathbb{N}$ that will be given later. Finally, plugging estimator (2.7) into (2.6), gives the collection of estimators of f, for $m \in \mathcal{M}_n$,

$$\widehat{f}_{N,m}(x) = 2a\sum_{k=0}^{N-1} \widehat{g}_m\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right). \tag{2.8}$$

2.3. **Risk bound for density estimator.** We first state a bound on the mean integrated squared error (MISE) of $\widehat{f}_{N,m}$ as an estimator of f.

Proposition 2.2. *Assume that $f \in \mathbb{L}^2(\mathbb{R}^+)$ and $\mathbb{E}[X_1^2] < +\infty$.*
(i) The estimator $\widehat{g}_m$ of g defined by (2.7) satisfies

$$\mathbb{E}[\|\widehat{g}_m - g\|^2] \leqslant \|g - g_m\|^2 + c_1\frac{m}{n} + c_2\frac{m^3}{n}, \qquad c_1 = 4, \quad c_2 = 16\mathbb{E}[Y_1^2]. \tag{2.9}$$

(ii) The estimator $\widehat{f}_{N,m}$ of f defined by (2.8) satisfies

$$\mathbb{E}[\|\widehat{f}_{N,m} - f\|^2] \leqslant \frac{8a^2}{(\sqrt{1+a} - \sqrt{1-a})^2}\left(\|g - g_m\|^2 + c_1\frac{m}{n} + c_2\frac{m^3}{n}\right) + 2\left(\frac{1-a}{1+a}\right)^N \|f\|^2. \tag{2.10}$$

Both risk bounds involve a bias term (proportional to $\|g - g_m\|^2$) which decreases when m increases, and a variance term with main order m^3/n, which increases with m. The last term of (2.10) is clearly exponentially decreasing with N. As the value of N is chosen by the statistician, taking $N \geqslant \log(n)/|\log((1-a)/(1+a))|$ makes this term negligible (if $a = 0.5$, and $n = 1000$, the condition is $N \geqslant 8$.)

Rates of convergence of estimators can be computed more precisely. To evaluate the order of $\|g - g_m\|^2$, the regularity of the function g has to be specified. Let us assume, in this paragraph, that g belongs to a Sobolev-Laguerre space (see Bongioanni and Torrea [2009]), defined by

$$W^s(\mathbb{R}^+, L) := \{f : \mathbb{R}^+ \to \mathbb{R}, f \in \mathbb{L}^2(\mathbb{R}^+), \sum_{j\geqslant 0} j^s\langle f, \varphi_j\rangle^2 \leqslant L < +\infty\}, \tag{2.11}$$

with $s > 0$ (see Comte and Genon-Catalot [2015] for equivalent definitions in case s is an integer). Then we get the following order for the squared bias term:

$$\|\widehat{g}_m - g\|^2 = \sum_{j=m}^{\infty} a_j^2(g) = \sum_{j=m}^{\infty} a_j^2(g) j^s j^{-s} \leqslant L m^{-s}.$$

Therefore we look for the choice $m = m_{\text{opt}}$ which minimizes $Lm^{-s} + c_2 m^3/n$. We obtain $m_{\text{opt}} = Cn^{1/(s+3)}$ with $C := (3c_2/(sL))^{-1/(s+3)}$, which implies $\mathbb{E}[\|\widehat{g}_{m_{\text{opt}}} - g\|^2] = O(n^{-s/(s+3)})$. This rate is the classical one in the multiplicative censoring model, and it is minimax optimal in case $U \sim \mathcal{U}([0,1])$, see Belomestny et al. (2016), Brunel *et al.* (2015).

2.4. **Model selection for density estimation.** As the regularity s of g in unknown, the choice $m = m_{\text{opt}}$ cannot be performed in practice. Therefore, a selection method must be set up to choose automatically the best m among the discrete collection $\mathcal{M}_n = \{m \in [\![1, n]\!],\ m^3 \leqslant n\}$, realizing the bias-variance trade-off. We want to choose m minimizing the MISE of $\widehat{f}_{N,m}$. Considering bound (2.10), the theoretical value is

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

as $\|g - g_m\|^2 = \|g\|^2 - \|g_m\|^2$ and $\|g\|^2$ does not depend on m. But functions g_m are unknown, thus we replace them by estimators. Therefore, we may select m as the minimizer of the sum $-\|\widehat{g}_m\|^2 + \text{pen}(m)$ with

$$\text{pen}(m) := \kappa_1 \frac{m}{n} + \kappa_2 \mathbb{E}[Y_1^2] \frac{m^3}{n} =: \text{pen}_1(m) + \text{pen}_2(m). \tag{2.12}$$

The penalty terms have the order of the variance term in (2.9). Note that the definition of $\mathcal{M}_n$ ensures that it is bounded. As $\mathbb{E}[Y_1^2]$ is unknown, we finally propose to replace it by its empirical counterpart and we get:

$$\widehat{m} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} \{-\|\widehat{g}_m\|^2 + \widehat{\text{pen}}(m)\}, \tag{2.13}$$

where

$$\widehat{\text{pen}}(m) = 2\kappa_1 \frac{m}{n} + 2\kappa_2 \widehat{C}_2 \frac{m^3}{n} := 2\text{pen}_1(m) + 2\widehat{\text{pen}}_2(m), \quad \widehat{C}_2 = \frac{1}{n}\sum_{k=1}^{n} Y_k^2. \tag{2.14}$$

The constants κ_1 and κ_2 are numerical constants which are calibrated in the simulations. Note that $\|\widehat{g}_m\|^2 = \sum_{j=0}^{m-1} \widehat{a}_j^2$ with $\widehat{a}_j$ given in (2.7) is easy to compute. Our final estimator is

$$\widehat{f}_{N,\widehat{m}}(x) = 2a \sum_{k=0}^{N-1} \widehat{g}_{\widehat{m}} \left(\left(\frac{1+a}{1-a} \right)^k (1+a)x \right). \tag{2.15}$$

We can prove the following result.

Theorem 2.3. *Assume that $f \in \mathbb{L}^2(\mathbb{R}^+)$, that f is bounded and that $\mathbb{E}[X_1^8] < +\infty$. For the final estimator $\widehat{f}_{N,\widehat{m}}$ defined by (2.7), (2.13) and (2.15), there exists κ_0 such that for $\kappa_1, \kappa_2 \geqslant \kappa_0$,*

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

where pen *is given by (2.12), and C_a is a positive constant depending on a and $\|f\|_\infty$.*

The theoretical study gives the bounds: $\kappa_1 \geqslant 32$ and $\kappa_2 \geqslant 288$. But it is well known that these theoretical constants are too large in practice: this is why the calibration step for choosing the values of the constants is done through simulations. Theorem 2.3 is a non-asymptotic bound for the MISE of the adaptive estimator $\widehat{f}_{N,\widehat{m}}$. It shows that the selection method leads to an estimator with smallest possible risk among all the estimators in the collection. Note that as previously, the choice $N = \log(n)/|\log((1-a)/(1+a))|$ is suitable for the last term to be negligible.

2.5. **Survival function estimation.** In this section, we extend the previous procedure to provide an estimator of the survival function of X, defined on $\mathbb{R}^+$ by

$$\overline{F}(x) = 1 - F(x) = \int_x^{+\infty} f(u)du. \tag{2.16}$$

We denote by $\overline{F}_Y$ the survival function of Y, defined accordingly. We also define a similar function $\overline{G}$ associated with g (which is not a density). We can prove the following Lemma.

Lemma 2.4. *For all x in $\mathbb{R}^+$,*

$$\overline{G}(x) := \int_x^{\infty} g(u)du = \frac{1}{2a}\left[(1+a)\overline{F}\left(\frac{x}{1+a}\right) - (1-a)\overline{F}\left(\frac{x}{1-a}\right)\right] = xf_Y(x) + \overline{F}_Y(x). \tag{2.17}$$

By integrating relation (2.6), we also get a relation between $\overline{F}$ and $\overline{G}$: for $x > 0$, let

$$\overline{F}_N(x) := \frac{2a}{1+a}\sum_{k=0}^{N-1}\left(\frac{1-a}{1+a}\right)^k \overline{G}\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right), \tag{2.18}$$

then

$$\overline{F}(x) - \overline{F}_N(x) = \left(\frac{1-a}{1+a}\right)^N \overline{F}\left(\left(\frac{1+a}{1-a}\right)^N x\right).$$

Note that $\overline{G}(0) = 1$ and thus $\lim_{N\to\infty}\overline{F}_N(0) = 1$, which is coherent with $\overline{F}(0) = 1$. Moreover, if $\mathbb{E}[X_1] < +\infty$ the function $\overline{F}$, and thus $\overline{G}$, is square integrable on $\mathbb{R}^+$. Denoting by $\overline{G}_m$ the orthogonal projection of $\overline{G}$ on $\mathcal{S}_m$, we have

$$\overline{G}_m = \sum_{j=0}^{m-1} b_j(\overline{G})\varphi_j, \quad \text{with } b_j(\overline{G}) := <\overline{G}, \varphi_j> .$$

According to relation (2.17), the coefficients $b_j(\overline{G})$ can also be written as follows:$b_j(\overline{G}) = \mathbb{E}[Y\varphi_j(Y)] + <\overline{F}_Y, \varphi_j>$. Thus we estimate the projection $\overline{G}_m$ of $\overline{G}$ on $\mathcal{S}_m$ by

$$\breve{\overline{G}}_m = \sum_{j=0}^{m-1}\breve{b}_j\varphi_j, \quad \breve{b}_j = \frac{1}{n}\sum_{i=1}^{n}\left[\int_{\mathbb{R}^+}\varphi_j(x)\mathbb{1}_{Y_i\geqslant x}dx + Y_i\varphi_j(Y_i)\right]. \tag{2.19}$$

Finally, plugging (2.19) into (2.18), an estimator of $\overline{F}$ is given by

$$\breve{\overline{F}}_{N,m} = \frac{2a}{1+a}\sum_{k=0}^{N-1}\left(\frac{1-a}{1+a}\right)^k \breve{\overline{G}}_m\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right). \tag{2.20}$$

We can prove the following bound.

Proposition 2.5. *Assume that $\mathbb{E}[X_1^2] < +\infty$. Then, $\overline{F}$ is square integrable and the estimator $\breve{\overline{F}}_{N,m}$ of $\overline{F}$ given by (2.20) satisfies*

$$\mathbb{E}[\|\breve{\overline{F}}_{N,m} - \overline{F}\|^2] \leqslant \mathfrak{C}(a)\left(\|\overline{G} - \overline{G}_m\|^2 + 4\mathbb{E}[Y_1^2]\frac{m}{n} + \frac{2\mathbb{E}[Y_1]}{n}\right) + \left(\frac{1-a}{1+a}\right)^{3N}\|\overline{F}\|^2, \tag{2.21}$$

where $\mathfrak{C}(a) = 8a^2/((1+a)^{3/2} - (1-a)^{3/2})^2$.

Inequality (2.21) provides a squared-bias/variance decomposition with bias proportional to $\|\overline{G} - \overline{G}_m\|^2$ and variance proportional to $\mathbb{E}[Y_1^2]m/n$. The term of order $\mathbb{E}[Y_1]/n$ is negligible, as well as the last one, for $N \geqslant \log(n)/[3\log((1+a)/(1-a))]$ (if $a = 0.5$, and $n = 1000$, the condition is $N \geqslant 3$). If $\overline{G}$ belongs to $W^s(\mathbb{R}^+, L)$ defined by (2.11), then choosing $m^\star_{\text{opt}}$ proportional to $n^{1/(s+1)}$ yields $\mathbb{E}[\|\breve{\overline{F}}_{N,m^\star_{\text{opt}}} - \overline{F}\|^2] = O(n^{s/(s+1)})$. The rate is better than the one obtained for density estimation.

However, it remains a nonparametric rate while cumulative distribution functions are estimated with parametric rates in direct problems.
Then we proceed as in the density case for selecting m and set:

$$\breve{m} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} \{-\|\breve{G}_m\|^2 + \widetilde{\operatorname{pen}}(m)\}, \quad \widetilde{\operatorname{pen}}(m) = 2\breve{\kappa}\widehat{C}_2\frac{m}{n} \tag{2.22}$$

where $\widehat{C}_2$ is given by (2.14). The constant $\breve{\kappa}$ is calibrated in the simulation part. We can prove the following oracle-type inequality of the final estimator $\breve{\overline{F}}_{N,\breve{m}}$.

Theorem 2.6. *If $\overline{F} \in \mathbb{L}^2(\mathbb{R}^+)$ and $\mathbb{E}[X_1^4] < \infty$, the final estimator $\breve{\overline{F}}_{N,\breve{m}}$ defined by (2.20) and (2.22) satisfies*

$$\mathbb{E}[\|\breve{\overline{F}}_{N,\breve{m}} - \overline{F}\|^2] \leqslant \mathfrak{C}(a)\left(6\inf_{m\in\mathcal{M}_n}\{\|\overline{G} - \overline{G}_m\|^2 + \operatorname{pen}(m)\} + \frac{D_a}{n}\right) + \left(\frac{1-a}{1+a}\right)^{3N}\|\overline{F}\|^2 \tag{2.23}$$

where $\mathfrak{C}(a)$ is defined in Proposition 2.5 and D_a is a constant depending on a.

Only a sketch of proof of the Theorem is given in Section 5.8, and we find $\breve{\kappa} \geqslant 192$.

2.6. **Case of unknown a.** Parameter a is not identifiable, unless additional information is available. Two cases can be considered. First, if an additional K-sample is available, where the signal is a deterministic known constant, then we have a set of observations of U, say $U_1^{(1)}, \dots, U_K^{(1)}$. In this case, we can use the maximum likelihood estimator $\max_{1\leqslant i\leqslant K}(|U_i^{(1)} - 1|)$ as an estimator of a with rate of convergence K (i.e. the mean square risk is of order $1/K^2$). Secondly, we can consider the model of repeated observations, where the variable X_i can be observed repeatedly, with independent errors:

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

where $(U_{i,1})_i$ and $(U_{i,2})_i$ are independent i.i.d. samples with distribution $\mathcal{U}([1-a, 1+a])$. Then we have

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

which yields $\mathbb{E}[Y_{i,1}^2/Y_{i,2}^2] = (1 + a^2/3)/(1 - a^2)$. Therefore, we make the proposal

$$\widehat{a}_n = \sqrt{\frac{\overline{W}_n - 1}{\overline{W}_n + 1/3}}, \quad \text{with } \overline{W}_n = \frac{1}{n}\sum_{i=1}^n W_i, \quad W_i := \frac{Y_{i,1}^2}{Y_{i,2}^2}. \tag{2.24}$$

Clearly, $\widehat{a}_n$ is a consistent estimator of a and by the limit central Theorem and the delta-method, we obtain the convergence in distribution

$$\sqrt{n}(\widehat{a}_n - a) \xrightarrow{\mathcal{L}} Z, \quad Z \sim \mathcal{N}\left(0, \sigma^2(a)\right), \sigma^2(a) = \frac{1-a^2}{40}(15 + 8a^2 + a^4) \in (0, 0.375).$$

This estimator can be plugged into the previous estimation procedure.

3. Model transformation and deconvolution approach

We present now another estimation strategy, to which ours may be compared. The idea is to rewrite the model under an additive form by taking logarithm of (1.1) (see van Es et al. [2005]). We obtain

$$Z_j := \log(Y_j) = \log(X_j) + \log(U_j) =: T_j + \varepsilon_j, \quad j = 1\dots, n. \tag{3.1}$$

Estimating the density of T_1 in model (3.1) is a classical deconvolution problem on $\mathbb{R}$ (see for example Comte et al. [2006]). Each sample $(Z_j)_j$, $(T_j)_j$, $(\varepsilon_j)_j$ is *i.i.d.* from density f_Z, f_T, f_ε respectively, and $(T_j)_j$, $(\varepsilon_j)_j$ are independent. They satisfy $f_Z = f_T \star f_\varepsilon$ where $\star$ denotes the convolution product.

Taking the Fourier transform of the equality implies $f_Z^* = f_T^* f_\varepsilon^*$. Then using the Fourier inversion formula, we get the following closed form for the density f_T,

$$f_T(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-iux}\frac{f_Z^*(u)}{f_\varepsilon^*(u)}du, \quad x \in \mathbb{R}. \tag{3.2}$$

An estimator of f_T is obtained by replacing f_Z^* by its empirical counterpart, $\widehat{f}_Z^*(u) = (1/n)\sum_{j=1}^n e^{iuZ_j}$. However, although formula (3.2) is well defined, the ratio $\widehat{f}_Z/f_\varepsilon^*$ is not integrable on the whole real line, since f_ε^* tends to zero near infinity. Therefore, we do not only plug $\widehat{f}_Z^*$ in equation (3.2) but we also introduce a cut-off which avoids integrability problems. Finally the estimator is defined by:

$$\widetilde{f}_{T,\ell}(x) = \frac{1}{2\pi}\int_{-\pi\ell}^{\pi\ell} e^{-iux}\frac{\widehat{f}_Z^*(u)}{f_\varepsilon^*(u)}du = \frac{1}{2\pi}\int_{-\pi\ell}^{\pi\ell} e^{-iux}\frac{1}{n}\sum_{j=1}^n \frac{e^{iuZ_j}}{f_\varepsilon^*(u)}du\ . \tag{3.3}$$

Clearly $\mathbb{E}[\widetilde{f}_{T,\ell}(x)] = f_{T,\ell}(x)$ with

$$f_{T,\ell}(x) := \frac{1}{2\pi}\int_{-\pi\ell}^{\pi\ell} e^{-iux} f_T^*(u)du.$$

We can remark that, by Plancherel-Parseval formula, $\|f_{T,\ell} - f_T\|^2 = (2\pi)^{-1}\int_{|u|\geqslant\pi\ell}|f_T^*(u)|^2du$. Then, with an additional bound on the variance, we recall the following result.

Proposition 3.1. *If $f_\varepsilon(u) \neq 0$, for all $u \in \mathbb{R}$, the estimator $\widetilde{f}_{T,\ell}$ defined by (3.3), satisfies*

$$\mathbb{E}[\|\widetilde{f}_{T,\ell} - f_T\|^2] \leqslant \frac{1}{2\pi}\int_{|u|\geqslant\pi\ell}|f_T^*(u)|^2du + \frac{1}{2\pi n}\int_{-\pi\ell}^{\pi\ell}\frac{du}{|f_\varepsilon^*(u)|^2}\ .$$

Several proofs of this bound can be found in the literature, see for example Comte and Lacour [2011], Dion [2014]. Using $\varepsilon_j = \log(U_j)$, we have

$$f_\varepsilon^*(u) = \frac{1}{2a}\int e^{iu\log(t)}\mathbb{1}_{[1-a,1+a]}(t)dt = \frac{(1+a)e^{\log(1+a)iu} - (a-1)e^{\log(1-a)iu}}{2a(1+iu)},$$

and

$$|f_\varepsilon^*(u)|^2 = \frac{1+a^2-(1-a^2)\cos(u\log((1+a)/(1-a)))}{2a^2(1+u^2)} \tag{3.4}$$

which never reaches zero, as $0 < a < 1$. Besides, $1/|f_\varepsilon^*(u)|^2 \leqslant 2a^2(1+u^2)/(2a^2) = 1+u^2$, for $u \in \mathbb{R}$. Therefore, Proposition 3.1 writes in the present case

$$\mathbb{E}[\|\widetilde{f}_{T,\ell} - f_T\|^2] \leqslant \frac{1}{2\pi}\int_{|u|\geqslant\pi\ell}|f_T^*(u)|^2du + \frac{\ell}{n} + \frac{\pi^2\ell^3}{3n}. \tag{3.5}$$

We can see that here ℓ plays the role of m previously, and we have to choose it in order to make a compromise between the squared bias term $(2\pi)^{-1}\int_{|u|\geqslant\pi\ell}|f_T^*(u)|^2du$ which decreases when ℓ increases and the variance term (with main term $\pi^2\ell^3/(3n)$) which increases when ℓ increases. Thus as previously, writing that $(2\pi)^{-1}\int_{|u|\geqslant\pi\ell}|f_T^*(u)|^2du = \|f_T\|^2 - \|f_{T,\ell}\|^2$, we omit the constant term $\|f_T\|^2$ and estimate the second term by $-\|\widetilde{f}_{T,\ell}\|^2$; then we replace the variance by its upper bound, up to a multiplicative constant. Finally, we set

$$\widetilde{\ell} = \underset{\ell\in\mathcal{M}_n}{\operatorname{argmin}}\ \{-\|\widetilde{f}_{T,\ell}\|^2 + \widetilde{\operatorname{pen}}(\ell)\},\ \text{with}\ \widetilde{\operatorname{pen}}(\ell) := \widetilde{\kappa}\left(\frac{\ell}{n} + \frac{\pi^2}{3}\frac{\ell^3}{n}\right), \tag{3.6}$$

where $\widetilde{\kappa}$ is a numerical constant calibrated in Section 4.1. We can prove for the estimator $\widetilde{f}_{T,\widetilde{\ell}}$ a non-asymptotic oracle-type inequality.

Theorem 3.2. *The estimator $\widetilde{f}_{T,\widetilde{\ell}}$ defined by (3.3) and (3.6) satisfies*

$$\mathbb{E}[\|\widetilde{f}_{T,\widetilde{\ell}} - f_T\|^2] \leqslant 4 \inf_{\ell \in \mathcal{M}_n} \{\|f_{T,\ell} - f_T\|^2 + \widetilde{\text{pen}}(\ell)\} + \frac{K}{n}$$

with K a numerical constant and $\widetilde{\kappa} \geqslant 4$.

Finally to estimate f (the density of X) we have to apply the following relations:

$$f(v) = f_T(\log(v))/v, \qquad f_T(v) = f_{\log(X)}(v) = f(e^v)e^v.$$

We define the estimator of f by

$$\widetilde{f}_{\widetilde{\ell}}(x) := \widetilde{f}_{T,\widetilde{\ell}}(\log(x))/x. \tag{3.7}$$

We can see on this definition that the estimator is not defined near zero, thus we have to consider the truncated integral

$$\int_\alpha^{+\infty} \left(\widetilde{f}_{\widetilde{\ell}}(x) - f(x)\right)^2 dx \leqslant \frac{1}{\alpha}\|\widetilde{f}_{T,\widetilde{\ell}} - f_T\|^2$$

to obtain a bound on the risk: for any $\alpha > 0$

$$\mathbb{E}\left[\int_\alpha^{+\infty} (\widetilde{f}_{\widetilde{\ell}}(x) - f(x))^2 dx\right] \leqslant \frac{4}{\alpha} \inf_{\ell \in \mathcal{M}_n} \{\|f_{T,\ell} - f_T\|^2 + \widetilde{\text{pen}}(\ell)\} + \frac{K}{\alpha n}.$$

We can see on these bounds that, the smaller α, the larger the bound. This is clearly confirmed by the simulations hereafter.

4. Numerical study

4.1. **Simulated data.** In this Section we evaluate our estimators of the density and the survival function on simulated data. We compute three estimators: the estimators of f, $\widehat{f}_{N,\widehat{m}}$ given by (2.8) and $\widetilde{f}_{\widetilde{\ell}}$ given by (3.7) and the estimator of $\overline{F}$, $\check{F}_{N,\check{m}}$, given by (2.20). For each estimator, there is a preliminary step before estimating the target function. Indeed, we first compute the collection of projection estimators of function g: $\widehat{g}_m$. Then we implement the selection procedure for the dimension parameter m. We obtain the final estimator of g: $\widehat{g}_{\widehat{m}}$. Finally, applying formula (2.15) with $N = 30$, we obtain our final estimator $\widehat{f}_{30,\widehat{m}}$ of f. The estimation procedure is implemented similarly for the survival function $\overline{F}$.

For the deconvolution density estimator we first estimate the density $f_T = f_{\log(X)}$ with the collection $\widetilde{f}_{T,\ell}$ as given by (3.3). The integrals are computed using Riemann approximations with thin discretisations. We select the best cut-off parameter ℓ among the collection, according to the criterion given in Section 3. Finally we use formula (3.7) to obtain $\widetilde{f}_{\widetilde{\ell}}$.

Each selection procedure depends on a parameter which has to be calibrated, namely κ_1, κ_2 in (2.14), $\check{\kappa}$ in (2.22), $\widetilde{\kappa}$ in (3.6). They are chosen from preliminary simulation experiments. Different cases of density f have been investigated with different parameter values, and a large number of repetitions. Comparing the MISE obtained as functions of the constants of interest, yields to select values making a good compromise over all experiences. We choose: $\kappa_1 = 0.5$, $\kappa_2 = 0.01$, $\check{\kappa} = 0.3$, $\widetilde{\kappa} = 4$. In the following we investigate 3 densities for X:

- $\Gamma(4, 0.5)$.
- $\mathcal{E}(1)$
- $0.5\Gamma(2, 0.4) + 0.5\Gamma(11, 0.4)$

The first one is uni-modal and 0 in 0, the second one is decreasing and is 1 in 0: we are indeed interested in the behaviour near 0 of the estimators. The last one is bi-modal. For each density, we could start the estimation procedure near $x = 0$ for our estimator $\widehat{f}_{N,\widehat{m}}$. But in order to compare our estimator with the deconvolution estimator $\widetilde{f}_{\widetilde{\ell}}$ which is not defined in 0, we start in $x = 0.1$ for all the grids of density estimation. Figure 1 illustrates the kind of data generated by the model and the

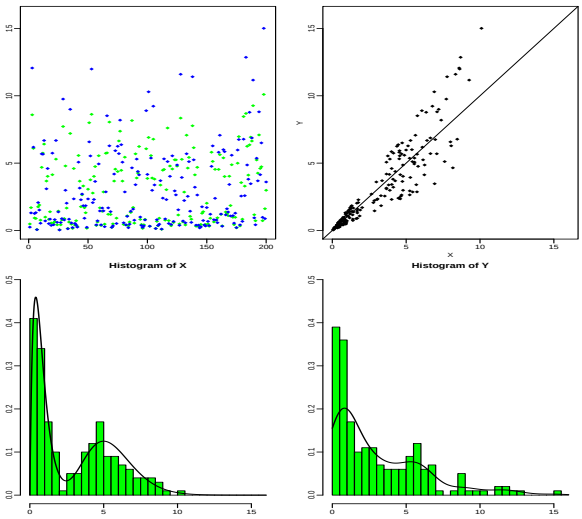

FIGURE 1. Example of database when $X \sim 0.5\Gamma(2, 0.4) + 0.5\Gamma(11, 0.5)$, $a = 0.5$, $n = 200$. Top left: plot of X (green or grey) and Y (blue or black). Top right: Y as a function of X. Bottom left: histogram of X with the true density f, bottom right: histogram of Y with a projection Laguerre estimator of f_Y applied on the (Y_i)'s.

effect of the censoring variable. It is a real issue to successfully reconstruct the density of X from the censored data Y.

Let us first comment the density estimation procedure. For the projection estimator $\widehat{f}_{N,\widehat{m}}$ we choose $m_{max} = 10$ or 15 because the selected m are small most of the time. For the deconvolution estimator: $\ell_{max} = 10$ and the selected ℓ are often small (1,2,3).
Figure 2 illustrates the good performances of our estimation procedure by projection. We represent 20 estimators $\widehat{f}_{N,\widehat{m}}$ of f (for 20 simulated samples) in the exponential case and the mixed-gamma case, and the beam of estimators are very close and close to the true density. On Figure 3, we can see both estimators $\widehat{f}_{N,\widehat{m}}$, $\widetilde{f}_{\widetilde{\ell}}$ and the true density. We also plot on this graph the projection estimator of density f_Y from the observations $(Y_i)_i$. It is defined for observations $(Z_i)_i$,

$$\widehat{f}_{Z,m} = \sum_{j=0}^{m-1} \widehat{b}_j \varphi_j \quad \text{with} \quad \widehat{b}_j = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(Z_i), \tag{4.1}$$

and $\widehat{\widehat{m}} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}} \{-\|\widehat{f}_{Z,m}\|^2 + m/n\}$ (the calibration constant has been chosen equal to 1 here). We can notice from the graph that estimator $\widetilde{f}_{\widetilde{\ell}}$ is closer to $\widehat{f}_{Y,\widehat{\widehat{m}}}$ (4.1) and f_Y than to f, the target function. However, estimator $\widehat{f}_{N,\widehat{m}}$ catches the difference between f and f_Y which is the aim here, and fits well the true density f of sample $(X_i)_i$.

Then we compute approximation of the MISE from 100 or 200 Monte-Carlo simulations. The number of repetitions has been checked to be large enough to insure the stability of the MISEs. They are multiplied by 100 and summed up in Table 1 for different values of parameter a and of the number of observations n, to complete the illustration. When a goes from 0.25 to 0.5 the estimation is more difficult, and this increases the value of the errors. Likewise, when n increases from 200 to 10000 the estimation is easier and the MISEs are smaller. We can see again that the results are specifically good for exponential densities. For the mixed-gamma case function g is hard to estimate because it has 2 modes, thus the estimation of f is also difficult and requires more observations, see the third line of Table 1. Still according to Table 1, the projection estimator performs better than the deconvolution method. All along, it has been seen that the projection method is computationally faster than the

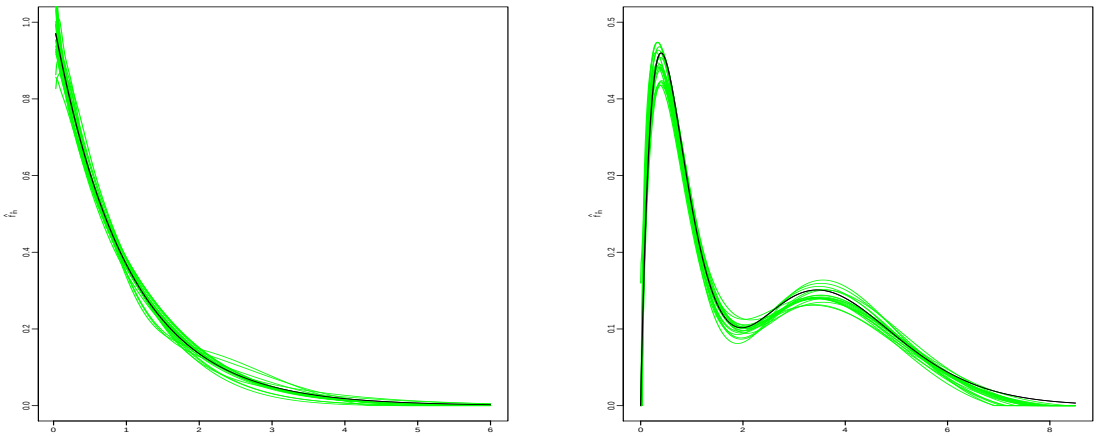

FIGURE 2. 20 estimator $\widehat{f}_{N,\widehat{m}}$ of f in plain grey line (green) versus the true density f in black bold plain line: on the left when $X \sim \mathcal{E}(1)$ with $a = 0.5$, $n = 1000$; on the right when $X \sim 0.5\Gamma(4, 0.25) + 0.5\Gamma(20, 0.5)$, with $a = 0.25$, $n = 1000$.

deconvolution strategy. Besides, the deconvolution estimator is very unstable around zero.

Remark. Estimation at $x = 0$. Note that by definition of function g (2.4) $g(0) = 0$, then $f_N(0) = 0$ and the projection estimator $\widehat{f}_{N,\widehat{m}}$ may have to be corrected in point zero if the true density is non-zero in zero. But, we can see that, if the function f is continuous in 0^+, then $\lim_{y\to 0} f_Y(y) = f(0)\log((1+a)/(1-a))/(2a)$. This implies that estimating f_Y in zero by a direct the projection estimator of f_Y relying on the Laguerre basis $(\widehat{f}_{Y,\widehat{\widehat{m}}})$ and applying the multiplicative correction factor $2a/\log((1+a)/(1-a))$ should be an adequate approximation of f near zero. On Figure 2 the grid begins in 0.03 for the exponential density and in 0 for the mixed-gamma density. If the statistician wants to start the estimation in 0, the plugging of corrected $\widehat{f}_{Y,\widehat{\widehat{m}}}(0)$ for the first value of estimator $\widehat{f}_{N,\widehat{m}}$ is a good strategy.

For the estimation of the survival function the grid of estimation begins in 0. We choose for the maximal dimension $m_{max} = 10, 15, 20$ ($n = 200, 1000, 10000$) and the selected m are small most of the time. The left graph of Figure 4 illustrates the good estimation of the survival function of X when it has an exponential distribution with parameter 1 from observations $(Y_i)_i$. On the right, the second graph shows the mixed-gamma case: our estimator $\overline{\widehat{F}}_{N,\widehat{\widehat{m}}}$ (plain grey line) detects well the bimodal character of the density (true $\overline{F}$ in plain black line). We also represent the empirical distribution function $\overline{F}_{Y,n}$ in dotted grey line, given for a sample $(Z_i)_i$ by:

$$\overline{F}_{Z,n}(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Z_i \leqslant t}. \tag{4.2}$$

We can see that this function is not a good approximation of $\overline{F}$ when $a = 0.5$. To confirm this fact, Table 2 provides the MISEs (times 100) for estimator $\overline{\widehat{F}}_{N,\widehat{\widehat{m}}}$ of $\overline{F}$. They can be compared to the MISEs of estimators of $\overline{F}$: $\overline{F}_{Y,n}$ (available in practice) and $\overline{F}_{X,n}$ (not available in practice). This table highlights the quality of our estimator when $a = 0.5$ (results in bold black). When $a = 0.25$ as expected $\overline{F}_{Y,n}$ can be considered as a satisfying approximation of the survival function of X, except for the exponential case, where our estimator is the best. Again, when a increases the MISEs are higher and with n increases the MISEs are smaller.

4.2. **Application.** Klein et al. [2013] detail the problem of confidential protection of data. The issue it how to alter the data before releasing it to the public in order to minimize the risk of disclosure and

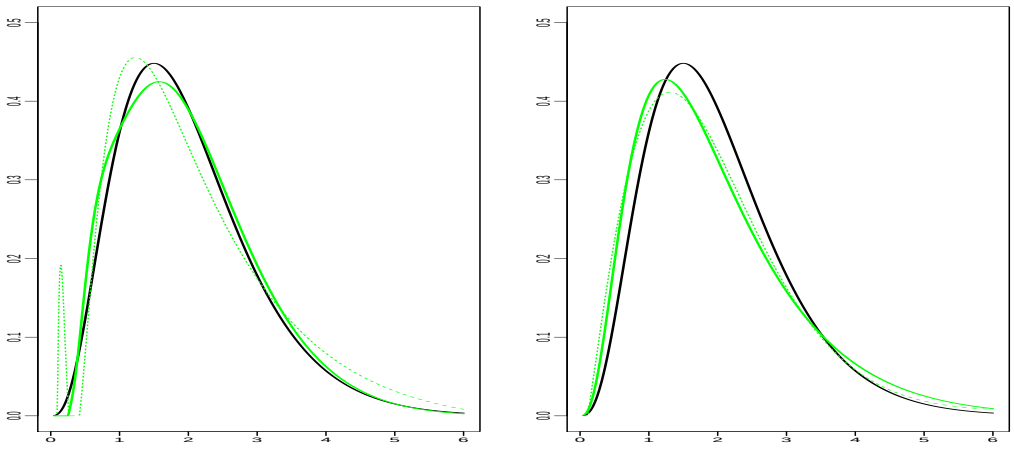

FIGURE 3. Gamma case: $X \sim \Gamma(4, 0.5)$, $a = 0.5$, $n = 1000$. Left graph: f in bold black line, estimator $\widehat{f}_{N,\widehat{m}}$ of f in plain bold grey line (green), estimator $\widetilde{f}_{\widetilde{\ell}}$ of f in thin dotted grey line (green). Right graph: f in bold black line, f_Y in bold grey line (green), estimator of f_Y by projection $\widehat{f}_{Y,\widehat{\widehat{m}}}$ in dotted grey line (green).

Distribution of f	a	Estimator $\widehat{f}_{N,\widehat{m}}$			Estimator $\widetilde{f}_{\widetilde{\ell}}$		
		$n = 200$	$n = 1000$	$n = 10000$	$n = 200$	$n = 1000$	$n = 10000$
Exponential	0.25	0.386	0.075	0.006	0.703	0.153	0.024
	0.5	0.470	0.095	0.009	0.964	0.231	0.030
Gamma	0.25	0.538	0.110	0.014	1.122	0.987	0.017
	0.5	0.972	0.394	0.154	1.589	1.851	0.217
Mixed-gamma	0.25	1.070	0.146	0.015	1.603	0.346	0.048
	0.5	1.441	0.563	0.208	2.703	2.833	0.337

TABLE 1. MISE for the estimators of f: $\widehat{f}_{N,\widehat{m}}$ and $\widetilde{f}_{\widetilde{\ell}}$, times 100, with 200 repetitions for $n = 200, 1000$ and 100 for $n = 10000$.

Distribution f	a	$\breve{\overline{F}}_{N,\breve{m}}$			$\overline{F}_{Y,n}$		$\overline{F}_{X,n}$	
		$n = 200$	$n = 1000$	$n = 10000$	$n = 200$	$n = 1000$	$n = 200$	$n = 1000$
Exponential	0.25	**0.194**	**0.043**	0.026	0.253	0.055	0.248	0.054
	0.5	**0.269**	**0.072**	0.027	0.277	0.106	0.234	0.054
Gamma	0.25	0.269	0.151	0.121	**0.260**	**0.081**	0.245	0.054
	0.5	**0.500**	**0.133**	0.121	0.756	0.497	0.281	0.057
Mixed-gamma	0.25	0.677	0.175	0.098	**0.610**	**0.126**	0.557	0.097
	0.5	0.888	**0.225**	0.126	**0.855**	0.430	0.517	0.102

TABLE 2. MISE for the estimators of $\overline{F}$: $\breve{\overline{F}}_{N,\breve{m}}$, $\overline{F}_{Y,n}$, $\overline{F}_{X,n}$, times 100, with 200 repetitions for $n = 200, 1000$ and 100 for $n = 10000$.

at the same time to remain able to find the main characteristics of the original dataset when the level of noise is known.

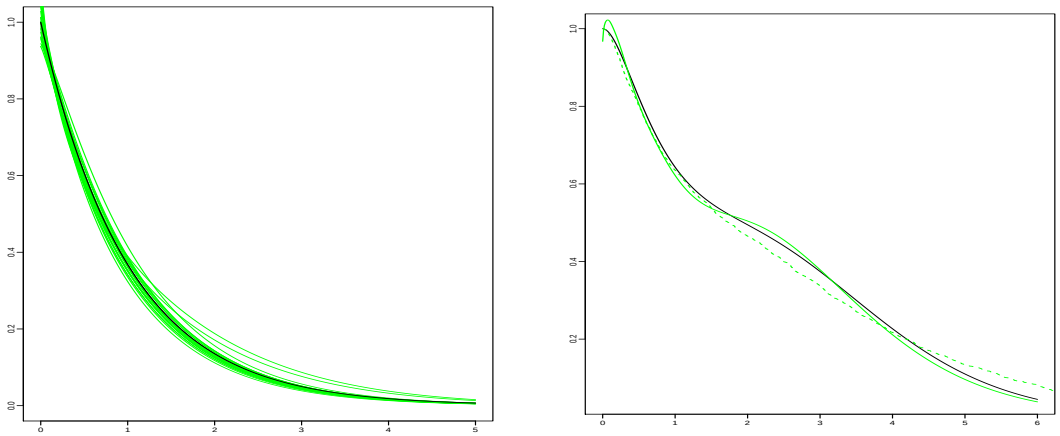

FIGURE 4. Left: 20 estimators $\breve{\overline{F}}_{N,\breve{m}}$ of $\overline{F}$ in plain grey line (green) versus the true function $\overline{F}$ in black bold plain line: when $X \sim \mathcal{E}(1)$, $a = 0.25$, $n = 200$. Right: estimator $\breve{\overline{F}}_{N,\breve{m}}$ in bold plain grey line (green) and empirical distribution $\overline{F}_n$ of Y in dotted bold grey line (green), when $X \sim 0.5\Gamma(4, 0.25) + 0.5\Gamma(20, 0.5)$ ($\overline{F}$ in black plain bold line), $a = 0.5$, $n = 1000$.

The multiplicative noise perturbation can be proposed in this context. Sinha et al. [2011] investigate this method on $n = 51$ magnitude data, different noise distributions, among which a uniform density $\mathcal{U}_{[1-a,1+a]}$ for $a = 0.1$ (the data set is publicly available from the American Community Survey (ACS) via `http://factfinder.census.gov`). The question is: how can the moments, the quantiles, the minimal value, maximal value of the sample X be estimated from the observations Y_i. They propose a strategy which delivers good results. But, looking at the noisy data Y_i one can see that they are very close from the true ones and thus in that case the privacy may be not insured. We illustrate this fact on Figure 5: it represents the multiplicative noise scenario, with $a = 0.1$ on the left and $a = 0.5$ on the right, for the original data $(X_i)_{i=1,\dots,n=51}$ from Sinha et al. [2011]. The three graphs are: top left the histogram of the $(X_i)_i$ the real data, top right an histogram of $(Y_i = X_iU_i)_i$ and on the bottom a plot of Y versus X.

Thus here we choose to illustrate the second choice: $a = 0.5$. What is the estimated density of X from these observations $(Y_i = X_iU_i)_i$? Are we capable of giving predictions of the data from this estimated density? What are the mean, the min, the max, the main quantiles of our new sample?

Figure 6 shows the estimator $\widehat{f}_{30,\widehat{m}}$ of f from the $(Y_i)_i$, the projection estimator of f on the sample $(X_i)_i$: $\widehat{f}_{X,\widehat{\widehat{m}}}$ (a benchmark, not available in practice) and $\widehat{f}_{Y,\widehat{\widehat{m}}}$ the projection estimator of f_Y on the $(Y_i)_i$. It seems that the two densities are very different. The quality of the method is asserted by the fact that $\widehat{f}_{X,\widehat{\widehat{m}}}$ and $\widehat{f}_{30,\widehat{m}}$ are very close. Then, from the estimator $\widehat{f}_{30,\widehat{m}}$ we simulate a new sample $(X\text{pred}_i)_i$ of length $n = 51$. To do so, we generate a "discrete variable" because we have a discrete version of the estimator of the density function f. The graph of the sorted new sample versus the sorted original sample is presented of Figure 7. The lining up of the values confirms the goodness of our estimator $\widehat{f}_{N,\widehat{m}}$ from the noisy observations $(Y_i)_i$. Finally we can compare the quantities of interest of $(X_i)_i$ (not available), $(Y_i)_i$ (noisy sample) and $(X\text{pred}_i)_i$, see Table 3. Except for the third quantile Q3 at (75 %), the information we get from our new sample is very close from the information from X.

The proposed procedure allows to correctly mask the data and to recover the main information from the original sample, as soon as the level of noise (given by a) is known. The method is easy to use in practice and insures the privacy protection of the data.

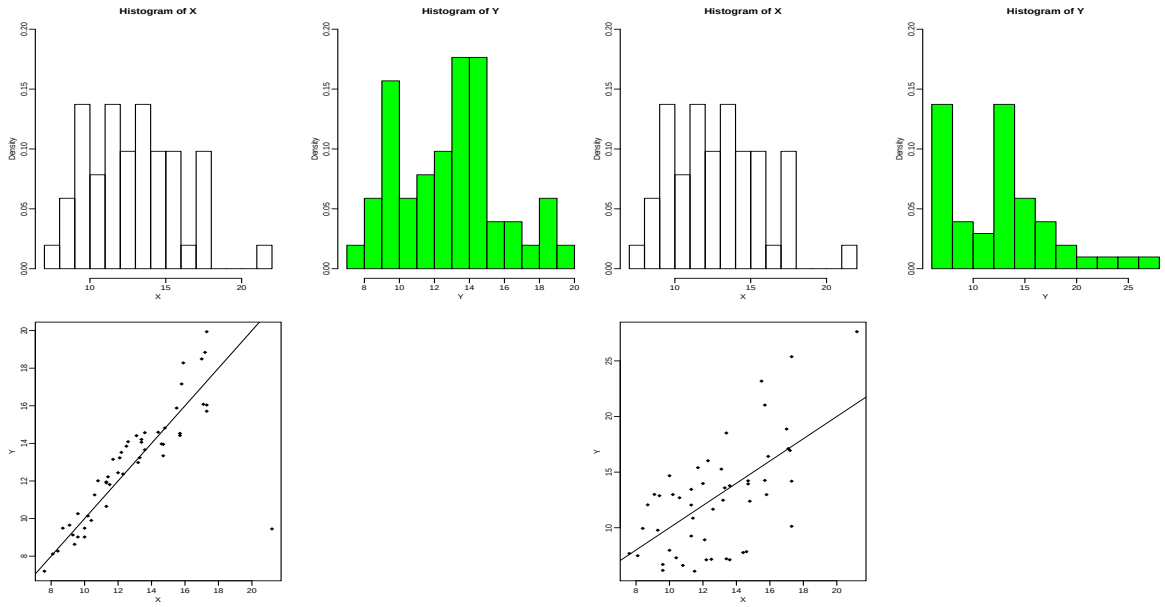

FIGURE 5. Illustration of uniform noise multiplication on real data. Three graphs for $a = 0.1$ on the left and $a = 0.5$ on the right. Top left histogram of $(X_i)_i$, top right histogram of $(Y_i)_i$, bottom plot of Y_i versus X_i.

	Mean	Standard deviation	Minimum	Maximum	Q1	Median	Q3
X	12.82	2.98	7.6	21.2	10.5	12.5	14.75
Y	12.59	4.93	6.10	27.6	7.91	12.70	14.46
Xpred	12.78	3.09	7.18	19.31	10.42	12.77	14.54

TABLE 3. Comparison of characteristic quantities from samples $(X_i)_i$, $(Y_i)_i$, $(X\text{pred}_i)_i$

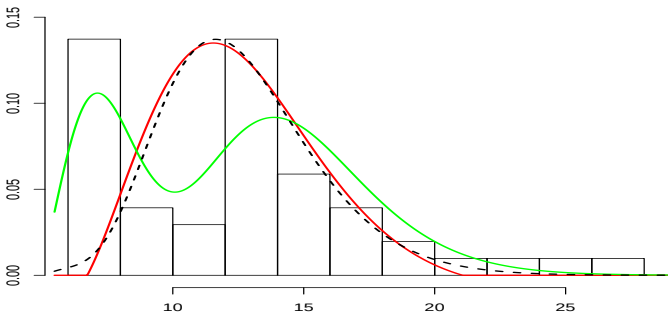

FIGURE 6. Histogram of the real data X_i's with full multiplicative noise, with $a = 0.5$, $Y_i = X_i U_i$. Dotted black line estimator $\widehat{f}_{X,\hat{\widehat{m}}}$ of f on the $(X_i)_i$, plain black line (red) $\widehat{f}_{N,\hat{m}}$ estimator of f on the $(Y_i)_i$, plain grey line (green) line estimator $\widehat{f}_{Y,\hat{\widehat{m}}}$ of f_Y on the $(Y_i)_i$.

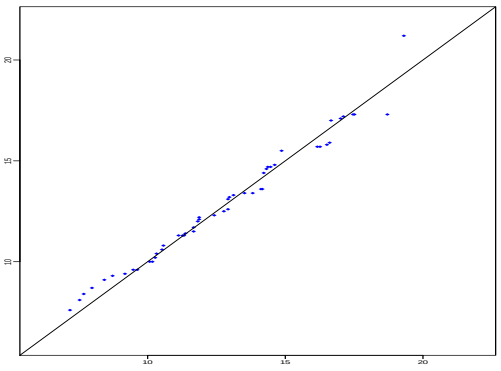

FIGURE 7. Plot of the new predictive sample $(X\text{pred}_i)_i$ versus original data $(X_i)_i$.

5. PROOFS

5.1. **Proof of Lemma 2.1.** Denote F the cumulative distribution function of X_1, it comes the bounds

$$\begin{aligned} \frac{1-a}{2a}\int_{\frac{y}{1+a}}^{\frac{y}{1-a}} f(x)dx &\leqslant yf_Y(y) \leqslant \frac{1+a}{2a}\int_{\frac{y}{1+a}}^{\frac{y}{1-a}} f(x)dx \\ \frac{1-a}{2a}\left[F\left(\frac{y}{1-a}\right)-F\left(\frac{y}{1+a}\right)\right] &\leqslant yf_Y(y) \leqslant \frac{1+a}{2a}\left[F\left(\frac{y}{1-a}\right)-F\left(\frac{y}{1+a}\right)\right]. \end{aligned} \tag{5.1}$$

Equation (5.1) shows that $yf_Y(y) \underset{y\to 0}{\to} 0$ and $yf_Y(y) \underset{y\to+\infty}{\to} 0$. □

5.2. **Useful properties of the Laguerre basis.**

Property 5.1. *If $t \in \mathcal{S}_m$, (1) $\|t\|_\infty \leqslant \sqrt{2m}\|t\|$, (2) $\|t'\|_\infty \leqslant 2\sqrt{2}m^{3/2}\|t\|$ and (3) If $\|t\| = 1$, $\|t'\| \leqslant 1+\sqrt{2m(m-1)}$.*

The two first points are direct consequences of (2.3). The last point comes from the following Lemma.

Lemma 5.2. *For all $j \in \mathbb{N}$, the Laguerre basis function $(\varphi_j)_j$ satisfies:*

$$\varphi_0'(x) = -\varphi_0(x), \quad \varphi_j'(x) = -\varphi_j(x) - 2\sum_{k=0}^{j-1}\varphi_k(x), \quad j \geqslant 1. \tag{5.2}$$

Considering $t \in S_m$, such that $\|t\| = 1$, $t(x) = \sum_{j=0}^{m-1} a_j\varphi_j(x)$, then

$$\begin{aligned} t'(x) &= \sum_{j=0}^{m-1} a_j\varphi_j'(x) = \sum_{j=1}^{m-1} a_j\left(-\varphi_j(x) - 2\sum_{k=0}^{j-1}\varphi_k(x)\right) - a_0\varphi_0(x) \\ &= -\sum_{j=0}^{m-1} a_j\varphi_j(x) - 2\sum_{j=1}^{m-1} a_j\left(\sum_{k=0}^{j-1}\varphi_k(x)\right). \end{aligned}$$

Then, $\|t'\| \leqslant \|t\| + 2\|\sum_{j=1}^{m-1} a_j(\sum_{k=0}^{j-1}\varphi_k)\|$

$$\left(\sum_{j=1}^{m-1} a_j\left(\sum_{k=0}^{j-1}\varphi_k(x)\right)\right)^2 \leqslant \sum_{j=1}^{m-1} a_j^2\left(\sum_{j=1}^{m-1}(\varphi_0+\varphi_1+\cdots+\varphi_{j-1}\right)^2 = \|t\|^2\sum_{j=1}^{m-1}(\varphi_0+\varphi_1+\cdots+\varphi_{j-1})^2$$

thus using that $\|t\| = 1$ and integrating, as the (φ_j)'s form a b.o.n, it yields

$$\left\|\sum_{j=1}^{m-1} a_j\left(\sum_{k=0}^{j-1}\varphi_k\right)\right\|^2 \leqslant \sum_{j=1}^{m-1}(\|\varphi_0\|^2+\|\varphi_1\|^2+\cdots+\|\varphi_{j-1}\|^2) \leqslant \sum_{j=1}^{m-1} j = m(m-1)/2.$$

Finally $\|t'\| \leqslant 1 + \sqrt{2m(m-1)}$. □

Proof of Lemma 5.2.
The following equality holds $\varphi_j'(x) = -\varphi_j(x) + 2\sqrt{2}e^{-x}L_j'(2x)$ which is a polynomial function of degree j multiplied by e^{-x}. Thus, it could be decomposed as $\varphi_j'(x) = \sum_{k=0}^{j} a_k^{(j)}\varphi_k(x)$ with

$$\begin{aligned} a_k^{(j)} &= <\varphi_j', \varphi_k> = \int_0^{+\infty} \varphi_j'(x)\varphi_k(x)dx = [\varphi_j(x)\varphi_k(x)]_0^{+\infty} - \int_0^{+\infty} \varphi_j(x)\varphi_k'(x)dx \\ &= -\varphi_j(0)\varphi_k(0) - \int_0^{+\infty} \varphi_j(x)\varphi_k'(x)dx = -2 - 2 <\varphi_j, \varphi_k'> = -2 - 2a_j^{(k)} \end{aligned}$$

Notice that this formula is also true when $k = j$: $<\varphi_j', \varphi_j> = \int_0^{+\infty} \varphi_j'(x)\varphi_j(x)dx = -(1/2)\varphi_j^2(0) = -2/2 = -1$. Thus we obtain:

$$\begin{aligned} \varphi_j'(x) &= \sum_{k=0}^{j}(-2- <\varphi_j', \varphi_k>)\varphi_k(x) = -2\sum_{k=0}^{j}\varphi_k(x) - \sum_{k=0}^{j} <\varphi_j, \varphi_k'> \varphi_k(x) \\ &= -\varphi_j(x) - 2\sum_{k=0}^{j-1}\varphi_k(x) - \sum_{k=0}^{j-1} <\varphi_j, \varphi_k'> \varphi_k(x) \end{aligned}$$

Or the $<\varphi_j, \varphi_k'>$ are zero for $k \leqslant j-1$. Thus we obtain (5.2). □

5.3. **Proof of Proposition 2.2.**

Proof of (i). To compute $\mathbb{E}[\|g_m - \widehat{g}_m\|^2]$ we start by noting that $\|g_m - \widehat{g}_m\|^2 = \sum_{j=0}^{m-1}(\widehat{a}_j - a_j(g))^2$. This implies

$$\mathbb{E}[\|g_m - \widehat{g}_m\|^2] = \sum_{j=0}^{m-1} \mathrm{Var}(\widehat{a}_j) \leqslant \frac{1}{n}\sum_{j=0}^{m-1} \mathbb{E}[(Y_1\varphi_j'(Y_1) + \varphi_j(Y_1))^2].$$

Now, Equation (2.5) applied with $t = \varphi_j^2$ and (2.3) lead to

$$\begin{aligned} \mathbb{E}[(Y_1\varphi_j'(Y_1) + \varphi_j(Y_1))^2] &\leqslant \mathbb{E}[2Y_1^2\varphi_j'^2(Y_1) + 2\varphi_j(Y_1)^2] \leqslant 2\|\varphi_j'\|_\infty^2\mathbb{E}[Y_1^2] + 2\|\varphi_j\|_\infty^2 \\ &\leqslant 16(j+1)^2\mathbb{E}[Y_1^2] + 4. \end{aligned}$$

As $8\sum_{j=0}^{m-1}(j+1)^2 = 8\sum_{j=1}^{m} j^2 \leqslant 8m^3$, it yields

$$\mathbb{E}[\|g_m - \widehat{g}_m\|^2] \leqslant 16\mathbb{E}[Y_1^2]\ \frac{m^3}{n} + 4\ \frac{m}{n}, \tag{5.3}$$

which is the result (i). □

Proof of (ii). Let us study the mean of the estimator of f:

$$\begin{aligned} \mathbb{E}[\widehat{f}_{N,m}(x)] &= 2a\ \sum_{k=0}^{N-1}\mathbb{E}\left[\widehat{g}_m\left(\left(\frac{1+a}{1-a}\right)^k(1+a)x\right)\right] = 2a\ \sum_{k=0}^{N-1}\sum_{j=1}^{m} a_j\varphi_j\left(\left(\frac{1+a}{1-a}\right)^k(1+a)x\right) \\ &= 2a\ \sum_{k=0}^{N-1} g_m\left(\left(\frac{1+a}{1-a}\right)^k(1+a)x\right) := f_{N,m}(x). \end{aligned}$$

Thus the estimator $\widehat{f}_{N,m}$ is an unbiased estimator of $f_{N,m}$ and

$$\mathbb{E}[\|\widehat{f}_{N,m} - f\|^2] = \|f - f_{N,m}\|^2 + \mathbb{E}[\|f_{N,m} - \widehat{f}_{N,m}\|^2]. \tag{5.4}$$

In the following we denote by h_k the composition of h with the function

$$x \mapsto \left(\frac{1+a}{1-a}\right)^k (1+a)x.$$

We note that for any function $h \in \mathbb{L}^2(\mathbb{R}^+)$,

$$\|h_k\|^2 = \int h^2\left(\left(\frac{1+a}{1-a}\right)^k (1+a)x\right)dx = \frac{1}{1+a}\left(\frac{1-a}{1+a}\right)^k \int h^2(y)dy = \frac{1}{1+a}\left(\frac{1-a}{1+a}\right)^k \|h\|^2. \quad (5.5)$$

Let us study of the bias term $\|f - f_{N,m}\|^2$:

$$\begin{aligned}(f - f_{N,m})(x) &= 2a\sum_{k=0}^{N-1} g_k(x) + f\left(\left(\frac{1+a}{1-a}\right)^N x\right) - 2a\sum_{k=0}^{N-1} g_{m,k}(x) \\ &= 2a\sum_{k=0}^{N-1}(g_k - g_{m,k})(x) + f\left(\left(\frac{1+a}{1-a}\right)^N x\right).\end{aligned}$$

The triangular inequality gives

$$\|f - f_{N,m}\| \leqslant 2a\sum_{k=0}^{N-1}\|g_k - g_{m,k}\| + \left\|f\left(\left(\frac{1+a}{1-a}\right)^N \cdot\right)\right\|. \quad (5.6)$$

As a consequence, using (5.5), we get

$$\begin{aligned}\sum_{k=0}^{N-1}\|g_k - g_{m,k}\| &= \sum_{k=0}^{N-1}\frac{1}{\sqrt{1+a}}\left(\frac{1-a}{1+a}\right)^{k/2}\|g - g_m\| = \frac{1-\left(\frac{1-a}{1+a}\right)^{N/2}}{\sqrt{1+a}-\sqrt{1-a}}\|g - g_m\| \\ &\leqslant \frac{\|g - g_m\|}{\sqrt{1+a}-\sqrt{1-a}}. \quad (5.7)\end{aligned}$$

Furthermore, $\|f(((1+a)/(1-a))^N \cdot)\| = ((1-a)/(1+a))^{N/2}\|f\|$, and plugging this and (5.7) in (5.6), we obtain

$$\|f - f_{N,m}\|^2 \leqslant \frac{8a^2}{(\sqrt{1+a}-\sqrt{1-a})^2}\|g - g_m\|^2 + 2\left(\frac{1-a}{1+a}\right)^N \|f\|^2. \quad (5.8)$$

For the variance term, we study $\|f_{N,m} - \widehat{f}_{N,m}\|^2$. We easily obtain

$$\|f_{N,m} - \widehat{f}_{N,m}\| = 2a\left\|\sum_{k=0}^{N-1}(\widehat{g}_{m,k} - g_{m,k})\right\| \leqslant 2a\sum_{k=0}^{N-1}\|\widehat{g}_m - g_m\|\frac{1}{\sqrt{1+a}}\left(\frac{1-a}{1+a}\right)^{k/2}$$

and finally

$$\|f_{N,m} - \widehat{f}_{N,m}\| \leqslant 2a\,\frac{1-\left(\frac{1-a}{1+a}\right)^{N/2}}{\sqrt{1+a}-\sqrt{1-a}}\|\widehat{g}_m - g_m\| \quad (5.9)$$

and $\mathbb{E}[\|\widehat{g}_m - g_m\|]$ has been evaluated in (5.3). Gathering (5.4), (5.8) and (5.9) implies (ii). □

5.4. **Proof of Theorem 2.3.** First, by the Cauchy-Schwarz inequality, we have

$$\mathbb{E}[\|\widehat{f}_{N,\widehat{m}} - f\|^2] \leqslant 2\mathbb{E}[\|f - f_{\widehat{m},N}\|^2] + 2\mathbb{E}[\|f_{N,\widehat{m}} - \widehat{f}_{N,\widehat{m}}\|^2]$$

Then we apply (5.8) and (5.9):

$$\begin{aligned}\mathbb{E}[\|\widehat{f}_{N,\widehat{m}} - f\|^2] &\leqslant 4\left(\frac{1-a}{1+a}\right)^N \|f\|^2 + \frac{16a^2}{(\sqrt{1+a}-\sqrt{1-a})^2}\left(\mathbb{E}[\|g - g_{\widehat{m}}\|^2] + \mathbb{E}[\|\widehat{g}_{\widehat{m}} - g_{\widehat{m}}\|^2]\right) \\ &= 4\left(\frac{1-a}{1+a}\right)^N \|f\|^2 + \frac{16a^2}{(\sqrt{1+a}-\sqrt{1-a})^2}\left(\mathbb{E}[\|g - \widehat{g}_{\widehat{m}}\|^2]\right). \quad (5.10)\end{aligned}$$

The last term is the MISE of the estimator $\widehat{g}_{\widehat{m}}$ which follows from the following Lemma.

Lemma 5.3. *Under the assumptions of Theorem 2.3, the estimator $\widehat{g}_{\widehat{m}}$ defined by (2.7) and (2.13), satisfies*

$$\mathbb{E}[\|\widehat{g}_{\widehat{m}} - g\|^2] \leqslant 6 \inf_{m\in\mathcal{M}}\{\|g - g_m\|^2 + \mathrm{pen}(m)\} + \frac{C'}{n}$$

with C' a positive constant depending on a and $\|f_Y\|_\infty$.

Gathering Lemma 5.3 and Inequality (5.10) ends the proof of Theorem 2.3. □

Proof of Lemma 5.3. Let us define the contrast

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^n [t(Y_i) + Y_i t'(Y_i)]. \tag{5.11}$$

It is easy to check that $\widehat{g}_m = \underset{t\in\mathcal{S}_m}{\mathrm{argmin}}\ \gamma_n(t)$, *i.e.* the estimator $\widehat{g}_m$ is also a minimum contrast estimator, and to compute that $\gamma_n(\widehat{g}_m) = -\|\widehat{g}_m\|^2$. We notice that

$$\gamma_n(t) - \gamma_n(s) = \|t - g\|^2 - \|s - g\|^2 - 2\nu_n(t - s) \tag{5.12}$$

with

$$\begin{aligned}\nu_n(t) &= \frac{1}{n}\sum_{i=1}^n t(Y_i) + Y_i t'(Y_i) - \langle t, g\rangle \\ &= \frac{1}{n}\sum_{i=1}^n t(Y_i) + Y_i t'(Y_i) - \mathbb{E}[t(Y_i) + Y_i t'(Y_i)] = \nu_{n,1}(t) + \nu_{n,2}(t) + \nu_{n,3}(t)\end{aligned}$$

where $\nu_{n,1}(t) := (1/n)\sum_{i=1}^n t(Y_i) - \mathbb{E}[t(Y_i)]$ and

$$\begin{aligned}\nu_{n,2}(t) &:= \frac{1}{n}\sum_{i=1}^n Y_i t'(Y_i)\mathbb{1}_{Y_i\leqslant c_n} - \mathbb{E}[Y_i t'(Y_i)\mathbb{1}_{Y_i\leqslant c_n}] \\ \nu_{n,3}(t) &:= \frac{1}{n}\sum_{i=1}^n Y_i t'(Y_i)\mathbb{1}_{Y_i> c_n} - \mathbb{E}[Y_i t'(Y_i)\mathbb{1}_{Y_i> c_n}]\end{aligned}$$

with

$$c_n := C_3\mathbb{E}[Y_1^2]\sqrt{n}/(\log(n)). \tag{5.13}$$

By definition of $\widehat{g}_{\widehat{m}}$, for all $m \in \mathcal{M}_n$, we have

$$\gamma_n(\widehat{g}_{\widehat{m}}) + \widehat{\mathrm{pen}}(\widehat{m}) \leqslant \gamma_n(g_m) + \widehat{\mathrm{pen}}(m).$$

Denoting $m \vee m' = m^*$,

$$\mathcal{B}_{m,m'} = \{t \in \mathcal{S}_{m\vee m'}, \|t\| = 1\}, \tag{5.14}$$

and using (5.12) we get

$$\begin{aligned}\|\widehat{g}_{\widehat{m}} - g\|^2 &\leqslant \|g - g_m\|^2 + \|\widehat{g}_{\widehat{m}} - g\|^2 - \|g - g_m\|^2 \\ &\leqslant \|g - g_m\|^2 + \widehat{\mathrm{pen}}(m) + 2\nu_n(\widehat{g}_{\widehat{m}} - g_m) - \widehat{\mathrm{pen}}(\widehat{m}) \\ &\leqslant \|g - g_m\|^2 + \frac{1}{4}\|\widehat{g}_{\widehat{m}} - g_m\|^2 + 4\sup_{t\in\mathcal{B}_{m,\widehat{m}}}\nu_n^2(t) + \widehat{\mathrm{pen}}(m) - \widehat{\mathrm{pen}}(\widehat{m}) \\ &\leqslant \|g - g_m\|^2 + \frac{1}{2}\|\widehat{g}_{\widehat{m}} - g\|^2 + \frac{1}{2}\|g_m - g\|^2 + 4\sup_{t\in\mathcal{B}_{m,\widehat{m}}}\nu_n^2(t) + \widehat{\mathrm{pen}}(m) - \widehat{\mathrm{pen}}(\widehat{m})\end{aligned}$$

Therefore we get

$$\begin{aligned}
\|\widehat{g}_{\widehat{m}} - g\|^2 &\leqslant 3\|g - g_m\|^2 + 8 \sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_n^2(t) + 2\widehat{\mathrm{pen}}(m) - 2\widehat{\mathrm{pen}}(\widehat{m}) \\
&\leqslant 3\|g - g_m\|^2 + 24 \sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,1}^2(t) + 24 \sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,2}^2(t) + 24 \sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,3}^2(t) \\
&\quad +2\widehat{\mathrm{pen}}(m) - 2\widehat{\mathrm{pen}}(\widehat{m}) + 2\mathrm{pen}(\widehat{m}) - 2\mathrm{pen}(\widehat{m}) + 2\mathrm{pen}(m) - 2\mathrm{pen}(m) \\
&\leqslant 3\|g - g_m\|^2 + 24(\sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,1}^2(t) - p_1(m,\widehat{m}))_+ + 24(\sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,2}^2(t) - p_2(m,\widehat{m}))_+ \\
&\quad +24 \sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,3}^2(t) + 2\widehat{\mathrm{pen}}(m) - 2\widehat{\mathrm{pen}}(\widehat{m}) + 2\mathrm{pen}(\widehat{m}) + 2\mathrm{pen}(m) \qquad (5.15)
\end{aligned}$$

with $p_1(m,m') = 6m^*/n$ satisfying $12p_1(m,m') \leqslant \mathrm{pen}_1(m) + \mathrm{pen}_1(m')$ for $\kappa_1 \geqslant 72$ and

$$p_2(m,m') = 24\mathbb{E}[Y_1^2]m^{3*}/n$$

$12\ p_2(m,m') \leqslant \mathrm{pen}_2(m) + \mathrm{pen}_2(m')$ for $\kappa_2 \geqslant 288$. Let us state intermediate results.

Lemma 5.4. *Under the assumption of Theorem 2.3,*

(i) $\mathbb{E}\left[(\sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,1}^2(t) - p_1(m,\widehat{m}))_+\right] \leqslant K_1/n,$

(ii) $\mathbb{E}\left[(\sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,2}^2(t) - p_2(m,\widehat{m}))_+\right] \leqslant K_2/n,$

(iii) $\mathbb{E}\left[\sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,3}^2(t)\right] \leqslant K_3/n,$

where K_1, K_2, K_3 are constants which do not depend on n.

(iv) *There exists a positive constant K_4 depending on a such that,*

$$\mathbb{E}[\{\ \mathrm{pen}(\widehat{m}) - \widehat{\mathrm{pen}}(\widehat{m})\}_+] \leqslant \frac{K_4}{n}.$$

Taking expectation of (5.15), using $\mathbb{E}[\widehat{\mathrm{pen}}(m)] = 2\mathrm{pen}(m)$, and plugging the results of Lemmas 5.4 implies Lemma 5.3. □

5.5. **Proof of Lemma 5.4.**
First notice that, for $i = 1, 2$,

$$\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}_{m,\widehat{m}}} \nu_{n,i}^2(t) - p_i(m,\widehat{m})\right)_+\right] \leqslant \sum_{m'\in\mathcal{M}_n} \left(\sup_{t\in\mathcal{B}_{m,m'}} \nu_{n,i}^2(t) - p_i(m,m')\right)_+ .$$

In the following we apply Talagrand's inequality to the two above terms. For that purpose, we compute the terms denoted by H^2, v and M in Theorem 5.7.

Proof of (i). We bound $\mathbb{E}[\sup_{t\in\mathcal{B}_{m,m'}} \nu_{n,1}^2(t)]$. For $t \in \mathcal{B}_{m,m'}$, using that $\sum_{j=0}^{m^*-1}\langle t,\varphi_j\rangle^2 = 1$, we get

$$\nu_{n,1}^2(t) = \left(\nu_{n,1}\left(\sum_{j=0}^{m^*-1}\langle t,\varphi_j\rangle\varphi_j\right)\right)^2 = \left(\sum_{j=0}^{m^*-1}\langle t,\varphi_j\rangle\nu_{n,1}(\varphi_j)\right)^2 \leqslant \sum_{j=0}^{m^*-1}\nu_{n,1}^2(\varphi_j)$$

$$\mathbb{E}[\sup_{t\in\mathcal{B}_{m,m'}} \nu_{n,1}^2(t)] \leqslant \sum_{j=0}^{m^*-1}\mathbb{E}[\nu_{n,1}(\varphi_j)^2] = \sum_{j=0}^{m^*-1}\frac{1}{n}\mathrm{Var}(\varphi_j(Y_1)) \leqslant \frac{2m^*}{n} =: H^2,$$

as $\varphi_j^2(x) \leqslant 2, \forall j, \forall x$. Now, $\|f\|_\infty = \sup_{x\in\mathbb{R}^+}|f(x)| < \infty$ implies that $|f_Y(y)| \leqslant (\|f\|_\infty/2a)\log((1+a)/(1-a))$ and $\|f_Y\|_\infty < \infty$. Thus

$$\mathrm{Var}(t(Y_1)) \leqslant \mathbb{E}[t(Y_1)^2] \leqslant \|f_Y\|_\infty\|t\|^2 = \|f_Y\|_\infty =: v$$

Finally, point (1) of Property 5.1 gives

$$\sup_{t\in\mathcal{B}_{m,m'}}\|t\|_\infty=\sqrt{2m^*}\sup_{t\in\mathcal{B}_{m,m'}}\|t\|=\sqrt{2m^*}=:M.$$

We obtain (for $\alpha=1/2$ in Theorem 5.7):

$$\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}_{m,m'}}\nu_{n,1}^2(t)-6\frac{m^*}{n}\right)_+\right]\leqslant 36\left(\frac{\|f_Y\|_\infty}{n}e^{-m^*/(6\|f_Y\|_\infty)}+C_1\frac{m^*}{n^2}e^{-C_2\sqrt{n}}\right)$$

with $C_1=\dfrac{588}{5/2-\sqrt{6}}$, $C_2=\dfrac{\sqrt{3/2}-1}{42}$. Consequently,

$$\begin{aligned}\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in\mathcal{B}_{m,m'}}\nu_{n,1}^2(t)-6\frac{m'}{n}\right)_+\right]&\leqslant\sum_{m'\in\mathcal{M}_n}36\left(\frac{\|f_Y\|_\infty}{n}e^{-m'/(6\|f_Y\|_\infty)}+C_1\frac{m'}{n^2}e^{-C_2\sqrt{n}}\right)\\&\leqslant\sum_{m'\in\mathcal{M}_n}36\left(\frac{\|f_Y\|_\infty}{n}e^{-m'/(6\|f_Y\|_\infty)}+C_1\frac{m'}{n^2}e^{-C_2\sqrt{m'}}\right)\leqslant\frac{K_1}{n}\end{aligned}$$

with C a positive constant depending on $\|f_Y\|_\infty$. This explains the choice $p_1(m,m')=6m^*/n$, and the constraint $\kappa_1\geqslant 12\times 6=72$.

Proof of (ii). As before

$$\begin{aligned}\mathbb{E}[\sup_{t\in\mathcal{B}_{m,m'}}\nu_{n,2}^2(t)]&\leqslant\sum_{j=0}^{m^*-1}\mathbb{E}[\nu_{n,2}^2(\varphi_j)]=\sum_{j=0}^{m^*-1}\frac{1}{n}\mathrm{Var}(Y_1\varphi_j'(Y_1)\mathbb{1}_{Y_1\leqslant c_n})\leqslant\sum_{j=0}^{m^*-1}\frac{1}{n}\mathbb{E}[Y_1^2(\varphi_j')^2(Y_1)]\\&\leqslant\sum_{j=0}^{m^*-1}\frac{1}{n}\|\varphi_j'^2\|_\infty\mathbb{E}[Y_1^2]\leqslant 8\mathbb{E}[Y_1^2]\frac{m^{*3}}{n}=:H^2\end{aligned}$$

We introduce the following result

Lemma 5.5. $\mathbb{E}[Y_i^2\psi^2(Y_i)]\leqslant\mathbb{E}[X_i^2\psi^2(X_iU_i)]\leqslant(1+a)^2\|\psi\|^2\mathbb{E}[X_i]$.

Using also Lemma 5.2, we obtain

$$\mathrm{Var}(Y_1t'(Y_1)\mathbb{1}_{Y_1\leqslant c_n})\leqslant\mathbb{E}[Y_1^2t'^2(Y_1)]\leqslant\|t'\|^2\mathbb{E}[X]\leqslant 3(1+a)^2m^{*2}\mathbb{E}[X]=:v$$

Finally: $\sup_{t\in\mathcal{B}_{m,m'}}(\sup_x|xt'(x)\mathbb{1}_{x\leqslant c_n}|\leqslant\sup_{t\in\mathcal{B}_{m,m'}}c_n\|t'\|_\infty\leqslant c_n2\sqrt{2}m^{*\,3/2}=:M$. We obtain, applying Theorem 5.7 with $\alpha=1/2$ again:

$$\begin{aligned}\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in B_{m,m'}}\nu_{n,2}^2(t)-24\mathbb{E}[Y_1^2]\frac{m'}{n}\right)_+\right]&\leqslant\sum_{m'\in\mathcal{M}_n}24\left(\frac{(3(1+a)^2m'^2\mathbb{E}[X]}{n}e^{-\frac{C_1\mathbb{E}[Y_1^2]m'}{(3(1+a)^2\mathbb{E}[X])}}\right.\\&+\left.C_2\frac{c_n^2m'^3}{n^2}e^{-C_3\mathbb{E}[Y_1^2]\sqrt{n}/c_n}\right)\\&\leqslant\frac{K_2}{n}\end{aligned}$$

with $C_1=8$, $C_2=2352/(5/2-\sqrt{6})$, $C_3=(\sqrt{3/2}-1)/42$ with c_n given by (5.13) and C' a constant depending on $\mathbb{E}[X_1]$ and $\mathbb{E}[Y_1^2]$. We choose

$$p_2(m,m')=24\mathbb{E}[Y_1^2]\frac{m^3}{n},\quad\mathrm{pen}_2(m)=\kappa_2\mathbb{E}[Y_1^2]\frac{m^{*3}}{n},\quad\kappa_2\geqslant 288$$

and obtain (ii). □

Proof of Lemma 5.5
We have, as $U \leqslant (1+a)$ a.s.,

$$\begin{aligned}\mathbb{E}[Y^2\psi^2(Y)] &\leqslant \mathbb{E}[(1+a)^2X^2\psi^2(XU)] = (1+a)^2\int_0^{+\infty}\int_{(1-a)}^{(1+a)} x^2\psi^2(xu)f(x)dudx \\ &\leqslant (1+a)^2\int_0^{+\infty}\psi^2(v)\,dv\int_0^{+\infty}xf(x)\,dx \\ &= (1+a)^2\|\psi\|^2\mathbb{E}[X]. \quad \square\end{aligned}$$

Proof of (iii). We use that $m^{*3} \leqslant n$, it yields

$$\begin{aligned}\mathbb{E}[\sup_{t\in\mathcal{B}_{m,m'}}\nu_{n,3}^2(t)] &\leqslant \frac{1}{n}\sum_{j=0}^{m^*-1}\mathrm{Var}(Y_1\varphi_j'(Y_1)\mathbb{1}_{Y_1>c_n}) \leqslant \frac{1}{n}\sum_{j=0}^{m^*-1}\|\varphi_j'^2\|_\infty\mathbb{E}[Y_1^2\mathbb{1}_{Y_1>c_n}] \\ &\leqslant 8\,\frac{m^{*3}}{n}\,\mathbb{E}[Y_1^2c_n^p\mathbb{1}_{Y_1>c_n}]c_n^{-1} \leqslant 8\,\frac{\mathbb{E}[Y_1^{2+p}]}{c_n^p}\end{aligned}$$

with the choice of c_n (5.13) we obtain

$$\mathbb{E}\left[\sup_{t\in\mathcal{B}_{m,m'}}\nu_{n,3}^2(t)\right] \leqslant 8\frac{\mathbb{E}[Y_1^{2+p}]\mathrm{Card}(\mathcal{M}_n)}{C_3^p\mathbb{E}[Y_1^2]^p\log(n)^pn^{p/2}} \leqslant \frac{K_3}{n}$$

for $p=4$, using that $\mathrm{card}(\mathcal{M}_n) \leqslant n^{1/3}$ and that the function $\log(n)^4/n^{2/3}$ is bounded with C" a positive constant depending on $\mathbb{E}[Y_1^4]$. $\square$
Proof of (iv). Let us study the difference

$$\mathbb{E}[\{\mathrm{pen}(\widehat{m}) - \widehat{\mathrm{pen}}(\widehat{m})\}_+] = \mathbb{E}\left[2\kappa_2\left\{\frac{\mathbb{E}[Y_1^2]}{2} - \widehat{C}_2\right\}_+\frac{\widehat{m}^3}{n}\right].$$

Denote $\Omega = \{|\mathbb{E}[Y_1^2] - \widehat{C}_2| \leqslant \mathbb{E}[Y_1^2]/2\}$. Then $\mathbb{E}[Y_1^2]/2 - \widehat{C}_2 \leqslant 0$ on Ω, thus

$$\mathbb{E}[\{\mathrm{pen}(\widehat{m}) - \widehat{\mathrm{pen}}(\widehat{m})\}_+] = \mathbb{E}\left[2\kappa_2\left(\frac{\mathbb{E}[Y_1^2]}{2} - \widehat{C}_2\right)\frac{\widehat{m}^3}{n}\mathbb{1}_{\Omega^c}\right] \leqslant \mathbb{E}\left[2\kappa_2\left(\mathbb{E}[Y_1^2] - \widehat{C}_2\right)\frac{\widehat{m}^3}{n}\mathbb{1}_{\Omega^c}\right].$$

By Cauchy-Schwarz we have

$$\mathbb{E}\left[\left|\mathbb{E}[Y_1^2] - \widehat{C}_2\right|\mathbb{1}_{\Omega^c}\right] \leqslant \mathbb{E}[|\mathbb{E}[Y_1^2] - \widehat{C}_2|^2]^{1/2}\mathbb{P}(\Omega^c)^{1/2}.$$

First, Markov's inequality implies

$$\mathbb{P}(\Omega^c) = \mathbb{P}\left(|\mathbb{E}[Y_1^2] - \widehat{C}_2| \geqslant \frac{E[Y_1^2]}{2}\right) \leqslant \frac{2^4}{\mathbb{E}[Y_1^2]^4}\mathbb{E}[|\mathbb{E}[Y_1^2] - \widehat{C}_2|^4].$$

Then the Rosenthal inequality implies that there exists a constant C, such that

$$\mathbb{E}[|\mathbb{E}[Y_1^2] - \widehat{C}_2|^4] \leqslant Cn^{-2}\mathbb{E}\left[\left(Y_1^2 - \mathbb{E}[Y_1^2]\right)^4\right].$$

Gathering the results we obtain:

$$\mathbb{E}\left[\left|\mathbb{E}[Y_1^2] - \widehat{C}_2\right|\mathbb{1}_{\Omega^c}\right] \leqslant \mathrm{Var}(\widehat{C}_2)^{1/2}\frac{4}{\mathbb{E}[Y_1^2]^2}\left(\frac{c_1}{n^3} + \frac{c_2}{n^2}\right)^{1/2}\left(\mathbb{E}[(Y_1^2 - \mathbb{E}[Y_1^2])^4]\right)^{1/2}.$$

Thus, as $\mathbb{E}[Y_1^k] = \mathbb{E}[X_1^k]\mathbb{E}[U_1^k]$ and the moments of $\mathbb{E}[U_1^k]$ are finite depending on a, if $\mathbb{E}[X_1^8] < \infty$ the quantities $m_4 := \mathbb{E}[(Y_1^2 - \mathbb{E}[Y_1^2])^4]$, $\mathrm{Var}(\widehat{C}_2)$ are bounded, we have the announced result. $\square$

5.6. **Proof of Lemma 2.4.** The survival function of Y satisfies

$$\begin{aligned}
2a\overline{F}_Y(y) &= \int_y^{+\infty} f_Y(z)dz = \int_y^{+\infty} \left(\int_{\frac{z}{1+a}}^{\frac{z}{1-a}} \frac{f(x)}{x} dx \right) dz \\
&= \int_{\frac{y}{1+a}}^{+\infty} \left(\int_{y\vee(1-a)x}^{x(1+a)} dz \right) \frac{f(x)}{x} dx \\
&= \int_{\frac{y}{1+a}}^{+\infty} \frac{f(x)}{x} [x(1+a) - y \vee (1-a)x] \, dx \\
&= (1+a) \int_{\frac{y}{1+a}}^{+\infty} f(x)dx - (1-a) \int_{\frac{y}{1+a}}^{+\infty} f(x) \mathbb{1}_{y<(1-a)x}(x)dx - y \int_{\frac{y}{1+a}}^{+\infty} \frac{f(x)}{x} \mathbb{1}_{y>(1-a)x}(x)dx.
\end{aligned}$$

with $x \vee y = \max(x, y)$. Finally it yields:

$$\overline{F}_Y(y) = \frac{1}{2a} \left[(1+a)\overline{F}\left(\frac{y}{1+a}\right) - (1-a)\overline{F}\left(\frac{y}{1-a}\right) \right] - y f_Y(y). \tag{5.16}$$

But, looking at the definition of g given in (2.4), we define analogously the function

$$\overline{G}(x) := \int_x^{\infty} g(y)dy = \frac{1}{2a} \left[(1+a)\overline{F}\left(\frac{x}{1+a}\right) - (1-a)\overline{F}\left(\frac{x}{1-a}\right) \right].$$

Thus relation (5.16) becomes:

$$\overline{G}(x) = x f_Y(x) + \overline{F}_Y(x). \quad \square$$

5.7. **Proof of Proposition 2.5.** First note that $\mathbb{E}[X_1^2] < +\infty$ implies that $\overline{F}$ integrable. Indeed

$$\int_0^{+\infty} \overline{F}^2(x)dx \leqslant \int_0^{+\infty} \overline{F}(x)dx = \mathbb{E}[X_1] \leqslant \mathbb{E}^{1/2}(X_1^2).$$

The result follows if we prove that

$$\mathbb{E}[\|\overline{G}_m - \breve{\overline{G}}_m\|^2] \leqslant \frac{2\mathbb{E}[Y_1]}{n} + 4\mathbb{E}[Y_1^2]\frac{m}{n}.$$

The MISE of estimator $\breve{\overline{G}}_m$ is:

$$\mathbb{E}[\|\breve{\overline{G}}_m - \overline{G}\|^2] = \|\mathbb{E}[\breve{\overline{G}}_m] - \overline{G}\|^2 + \mathbb{E}[\|\mathbb{E}[\breve{\overline{G}}_m] - \breve{\overline{G}}_m\|^2].$$

First, $\mathbb{E}[\overline{G}_m] = \sum_{j=0}^{m-1} \mathbb{E}[\breve{b}_j]\varphi_j = \sum_{j=0}^{m-1} b_j(\overline{G})\varphi_j = \overline{G}_m$. Then to compute the variance term $\mathbb{E}[\|\overline{G}_m - \breve{\overline{G}}_m\|^2]$ we start with the relation: $\|\overline{G}_m - \breve{\overline{G}}_m\|^2 = \sum_{j=0}^{m-1}(\breve{b}_j - b_j)^2$ and then

$$\begin{aligned}
\mathbb{E}[\|\overline{G}_m - \breve{\overline{G}}_m\|^2] &= \sum_{j=0}^{m-1} \mathrm{Var}(\breve{b}_j) \leqslant \frac{1}{n} \sum_{j=0}^{m-1} \mathbb{E}\left[\left(Y_1\varphi_j(Y_1) + \int_{\mathbb{R}^+} \varphi_j(x)\mathbb{1}_{Y_1\geqslant x}(x)dx \right)^2 \right] \\
&\leqslant \frac{2}{n} \sum_{j=0}^{m-1} \mathbb{E}[Y_1^2\varphi_j(Y_1)^2] + \frac{2}{n} \sum_{j=0}^{m-1} \mathbb{E}\left[\left(\int_{\mathbb{R}^+} \varphi_j(x)\mathbb{1}_{Y_1\geqslant x}(x)dx \right)^2 \right] \\
&\leqslant \frac{2m}{n}\|\varphi_j\|_\infty^2 \mathbb{E}[Y_1^2] + \frac{2}{n} \sum_{j=0}^{m-1} \mathbb{E}\left[\left(\int_{\mathbb{R}^+} \varphi_j(x)\mathbb{1}_{Y_1\geqslant x}(x)dx \right)^2 \right].
\end{aligned}$$

The last term:

$$\frac{1}{n}\mathbb{E}\left[\sum_{j=0}^{m-1} \left(\int_{\mathbb{R}^+} \varphi_j(x)\mathbb{1}_{Y_1\geqslant x}(x)dx \right)^2 \right] = \frac{1}{n}\mathbb{E}\left[\sum_{j=0}^{m-1} \langle \varphi_j \mathbb{1}_{Y_1\geqslant \cdot} \rangle^2 \right] \leqslant \frac{1}{n}\mathbb{E}\left[\|\mathbb{1}_{Y_1\geqslant \cdot}\|^2 \right] = \frac{\mathbb{E}[Y_1]}{n}.$$

Thus it comes

$$\mathbb{E}[\|\overline{G}_m - \breve{\overline{G}}_m\|^2] \leqslant \frac{2\mathbb{E}[Y_1]}{n} + 4\mathbb{E}[Y_1^2]\frac{m}{n}. \quad \square$$

5.8. **Proof of Theorem 2.6.** This proof follows the same line as the proof of Theorem 2.3. We define the contrast

$$\gamma_n^{(2)}(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^n \left[\int_{\mathbb{R}^+} t(x)\mathbb{1}_{Y_i \geqslant x} dx + Y_i t(Y_i)\right]. \tag{5.17}$$

It is such that $\gamma_n^{(2)}(\breve{\overline{G}}_m) = -\|\breve{\overline{G}}_m\|^2$ and $\breve{\overline{G}}_m = \underset{t\in\mathcal{S}_m}{\operatorname{argmin}}\, \gamma_n^{(2)}(t)$. Then, let

$$\nu_n^{(2)}(t) = \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^+} t(x)\mathbb{1}_{Y_i\geqslant x}dx + Y_i t(Y_i) - \mathbb{E}[\int_{\mathbb{R}^+} t(x)\mathbb{1}_{Y_i\geqslant x}dx + Y_i t(Y_i)] = \nu_{n,1}(t) + \nu_{n,2}(t) + \nu_{n,3}(t)$$

with

$$\begin{aligned}
\nu_{n,1}^{(2)}(t) &:= \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^+} t(x)\mathbb{1}_{Y_i\geqslant x}dx - \mathbb{E}\left[\int_{\mathbb{R}^+} t(x)\mathbb{1}_{Y_i\geqslant x}dx\right] \\
\nu_{n,2}^{(2)}(t) &:= \frac{1}{n}\sum_{i=1}^n Y_i t(Y_i)\mathbb{1}_{Y_i\leqslant c_n} - \mathbb{E}[Y_i t(Y_i)\mathbb{1}_{Y_i\leqslant c_n}] \\
\nu_{n,3}^{(2)}(t) &:= \frac{1}{n}\sum_{i=1}^n Y_i t(Y_i)\mathbb{1}_{Y_i> c_n} - \mathbb{E}[Y_i t(Y_i)\mathbb{1}_{Y_i> c_n}]
\end{aligned}$$

with c_n a numerical constant depending on n. Following the steps which lead to Equation (5.15), we choose $c_n := d\mathbb{E}[Y_1^2]\sqrt{n}/(\log(n))$ for numerical d a constant and we get the result with two applications of Talagrand inequality. $\square$

5.9. **Proof of Theorem 3.2.** Denote:

$$\phi_t(x) = \frac{1}{2\pi}\int t^*(-u)\frac{e^{iux}}{\overline{f_\varepsilon^*}(u)}du$$

and $\gamma(t) := \|t\|^2 - \frac{2}{n}\sum_{j=1}^n \phi_t(Z_j) = \|t\|^2 - 2\langle t, \widetilde{f}_{T,\ell}\rangle$. Let us define

$$\nu(t) := \frac{1}{2\pi}\langle t^*, \widetilde{f}_{T,\ell}^* - f_{T,\ell}^*\rangle = \frac{1}{n}\sum_{j=1}^n (\phi_t(Z_j) - \mathbb{E}[\phi_t(Z_j)]).$$

The two functions $\gamma(t)$ and $\nu(t)$ satisfy the following relation, for $t, s \in S_\ell$:

$$\begin{aligned}
&S_\ell = \{t \in \mathbb{L}^1(\mathbb{R}\cap\mathbb{L}^2(\mathbb{R}), \text{support}(t^*) \subset [-\pi\ell, \pi\ell]\}, \\
&\gamma(t) - \|t-f\|^2 - (\gamma(s) - \|s-f\|^2) = -2\nu(t-s).
\end{aligned} \tag{5.18}$$

Thus writing this relation with $\widetilde{f}_{T,\widetilde{\ell}}$ and $f_{T,\ell}$ and as, by definition, $\gamma(\widetilde{f}_{T,\widetilde{\ell}}) + \widetilde{\text{pen}}(\widetilde{\ell}) \leqslant \gamma(f_{T,\ell}) + \widetilde{\text{pen}}(\ell)$, it yields

$$\begin{aligned}
\|\widetilde{f}_{T,\widetilde{\ell}} - f_T\|^2 &= \|f_{T,\ell} - f_T\|^2 + \|\widetilde{f}_{T,\widetilde{\ell}} - f_T\|^2 - \|f_{T,\ell} - f_T\|^2 \\
&\leqslant \|f_{T,\ell} - f_T\|^2 + 2\nu(\widetilde{f}_{T,\widetilde{\ell}} - f_{T,\ell}) + \widetilde{\text{pen}}(\ell) - \widetilde{\text{pen}}(\widetilde{\ell}).
\end{aligned}$$

Let us remark that $\nu(\widetilde{f}_{T,\widetilde{\ell}} - f_{T,\ell}) = \|\widetilde{f}_{T,\widetilde{\ell}} - f_{T,\ell}\|\nu\left(\frac{\widetilde{f}_{T,\widetilde{\ell}} - f_{T,\ell}}{\|\widetilde{f}_{T,\widetilde{\ell}} - f_{T,\ell}\|}\right)$. This leads, as in the previous proofs, to

$$\|\widetilde{f}_{T,\widetilde{\ell}} - f_T\|^2 \leqslant 3\|f_{T,\ell} - f_T\|^2 + 4\widetilde{\text{pen}}(\ell) + 8\sum_{\ell'\in\mathcal{M}_n}\left(\sup_{t\in B_{\ell,\ell'}} \nu^2(t) - p(\ell,\ell')\right)_+$$

with a function p such that $\forall \ell, \ell'$, $4p(\ell,\ell') \leqslant \mathrm{pen}(\ell) + \mathrm{pen}(\ell')$.

Lemma 5.6. *There exists a constant $C>0$ such that*

$$\sum_{\ell' \in \mathcal{M}_n} \mathbb{E}\left[\left(\sup_{t \in B_{\ell,\ell'}} \nu^2(t) - p(\ell,\ell')\right)_+\right] \leqslant \frac{C}{N}.$$

We conclude that there exist two numerical constants $C_1, C_2 > 0$ such that

$$\mathbb{E}[\|\widetilde{f}_{\widetilde{\ell}} - f_T\|^2] \leqslant C_1 \inf_{\ell \in \mathcal{M}_n} \{\|f_{T,\ell} - f_T\|^2 + \widetilde{\mathrm{pen}}(\ell)\} + \frac{C_2}{N}. \quad \square$$

5.10. **Proof of the Lemma 5.6.** For $\ell \in \mathcal{M}_n$, we consider $t \in S_\ell$. We use Talagrand's inequality. We denote $B_{\ell,\ell'} = \{t \in S_{\ell \vee \ell'}, \|t\| = 1\}$ and $\ell^* = \ell \vee \ell'$. Using Proposition 3.1, we obtain

$$\mathbb{E}\left[\sup_{t \in B_{\ell,\ell'}} \nu^2(t)\right] = \mathbb{E}\|\widetilde{f}_{T,\ell^*} - f^*_{T,\ell}\|^2 \leqslant \frac{\ell^*}{n} + \frac{\pi^2(\ell^*)^3}{3n} \leqslant \frac{\ell^*}{n} + \frac{\pi^2(\ell^*)^3}{n} := H^2.$$

Then, by the Plancherel-Parseval inequality,it yields

$$\sup_{t \in B_{\ell,\ell'}} \|\phi_t\|_\infty \leqslant \frac{1}{\sqrt{2\pi}} \sqrt{2\pi\ell^* + \frac{2\pi^3 \ell^{*3}}{3}} \leqslant \sqrt{\ell^* + \pi^2 \ell^{*3}} := M.$$

Finally, for $t \in B_{\ell,\ell'}$, as we know the characteristic function f^*_ε, we have

$$\begin{aligned}
\mathrm{Var}(\phi_t(Z_1)) &\leqslant \mathbb{E}[|\phi_t(Z_1)|^2] = \frac{1}{4\pi^2}\mathbb{E}\left[\left|\int_{-\pi\ell^*}^{\pi\ell^*} t^*(-u)\frac{e^{iuZ_1}}{\overline{f^*_\varepsilon}(u)}du\right|^2\right] \\
&\leqslant \frac{\|f_Z\|_\infty}{4\pi^2}\int\left|\int_{-\pi\ell^*}^{\pi\ell^*} \frac{t^*(-u)e^{iuz}}{\overline{f^*_\varepsilon}(u)}du\right|^2 dz = \frac{\|f_Z\|_\infty}{2\pi}\int_{-\pi\ell^*}^{\pi\ell^*} \frac{|t^*(-u)|^2}{|\overline{f^*_\varepsilon}(u)|^2}du \\
&\leqslant \frac{\|f_Z\|_\infty}{2\pi}(1+(\pi\ell^*)^2)\int_{-\pi\ell^*}^{\pi\ell^*}|t^*(-u)|^2du \leqslant \frac{1+a}{2a}(1+(\pi\ell^*)^2) := v.
\end{aligned}$$

Indeed $\|f_Z\|_\infty \leqslant (1+a)/(2a) < +\infty$ since $\|f_Z\|_\infty = \|f_T \star f_\varepsilon\|_\infty \leqslant \|f_\varepsilon\|_\infty = (1+a)/(2a)$, with $f_\varepsilon(x) = e^x/(2a)\mathbf{1}_{[\log(1-a),\log(1+a)]}(x)$. According to Talagrand's inequality, for $\alpha = 1/2$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\sup_{t \in B_{\ell,\ell'}} \nu^2(t) - 4H^2\right)_+\right] \leqslant\ & 24 \quad \left(\frac{1+\pi^2\ell^{*2}}{n}\exp\left(-\ell^*/12\right)\right. \\
& + \quad \left.\frac{294}{(\frac{3}{2}-1)^2}\left[\frac{\ell^*}{n^2} + \frac{\pi^2\ell^{*3}}{n^2}\right]\exp\left(-\frac{(\sqrt{3/2}-1)}{42}\sqrt{\ell^* + \pi^2\ell^{*3}}\right)\right)
\end{aligned}$$

Then we use that $1 \leqslant \ell^{*3} \leqslant n$, thus for $\kappa \geqslant 4$ we obtain that there exist four numerical constants A_1, A_2, A_3, A_3 and a constant $C > 0$ such that

$$\begin{aligned}
\sum_{\ell' \in \mathcal{M}_n} \mathbb{E}\left[\left(\sup_{t \in B_{\ell,\ell'}} \nu^2(t) - p(\ell,\ell')\right)_+\right] &\leqslant \sum_{\ell' \in \mathcal{M}_n} A_1 \frac{1+\pi^2\ell'^2}{n}\exp\left(-A_2\ell'\right) + \frac{A_3}{n}\exp\left(-A_4\ell'^{3/2}\right) \\
&\leqslant \frac{C}{n}
\end{aligned}$$

with $p(\ell,\ell') = \dfrac{4\ \ell^*}{n} + \dfrac{4\pi^2\ \ell^{*3}}{3n}$. $\square$

Appendix

5.11. **Talagrand's inequality.** The following result follows from the Talagrand concentration inequality.

Theorem 5.7. *Consider $n \in \mathbb{N}^*$, $\mathcal{F}$ a class at most countable of measurable functions, and $(X_i)_{i\in\{1,\dots,n\}}$ a family of real independent random variables. One defines, for all $f \in \mathcal{F}$,*

$$\nu_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}[f(X_i)]).$$

Supposing there are three positive constants M, H and v such that $\sup_{f\in\mathcal{F}}\|f\|_\infty \leqslant M$, $\mathbb{E}[\sup_{f\in\mathcal{F}}|\nu_n(f)|] \leqslant H$, and $\sup_{f\in\mathcal{F}}(1/n)\sum_{i=1}^{n}\mathrm{Var}(f(X_i)) \leqslant v$, then for all $\alpha > 0$,

$$\begin{aligned}\mathbb{E}\left[\left(\sup_{f\in\mathcal{F}}|\nu_n(f)|^2 - 2(1+2\alpha)H^2\right)_+\right] &\leqslant \frac{4}{b}\left(\frac{v}{n}\exp\left(-b\alpha\frac{nH^2}{v}\right)\right.\\ &+ \left.\frac{49M^2}{bC^2(\alpha)n^2}\exp\left(-\frac{\sqrt{2}bC(\alpha)\sqrt{\alpha}}{7}\frac{nH}{M}\right)\right)\end{aligned}$$

with $C(\alpha) = (\sqrt{1+\alpha} - 1) \wedge 1$, and $b = \frac{1}{6}$.

References

M. Abbaszadeh, C. Chesneau, and H.I Doosti. Multiplicative censoring: Estimation of a density and its derivatives under the l-p-risk. *Revstat Statistical Journal*, 11:255–276, 2013.

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

M. Asgharian, M. Carone, and V. Fakoor. Large-sample study of the kernel density estimators under multiplicative censoring. *Ann. Statist.*, 40(1):159-187, 2012.

D. Belomestny, F. Comte, and V. Genon-Catalot. Laguerre estimation for k-monotone densities observed with noise. *Working paper* MAP5 2016-01, 2016

B. Bongioanni and J. L. Torrea. What is a Sobolev space for the Laguerre function systems? *Studia Math.*, 192(2):147–172, 2009.

E. Brunel, F. Comte, and V. Genon-Catalot. Nonparametric density and survival function estimation in the multiplicative censoring model. *Working paper* MAP5 2015-08 URL `https://hal.archives-ouvertes.fr/hal-01122847`, 2015

F. Comte and C. Lacour. Data-driven density estimation in the presence of additive noise with unknown distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(4):601–627, 2011.

F. Comte, Y. Rozenholc, and M-L Taupin. Penalized contrast estimator for adaptive density deconvolution. *Can. J. Stat.*, 34(3):431–452, 2006.

F. Comte and V. Genon-Catalot. Adaptive Laguerre density estimation for mixed Poisson models. *Electron. J. Stat.*, 9(1):1113–1149, 2015.

C. Dion. New adaptive strategies for nonparametric estimation in linear mixed models. *Journal of Statistical Planning and Inference*, 150:30-48, 2014.

B. van Es, C. A. J. Klaassen, and K. Oudshoorn. Survival analysis under cross-sectional sampling: length bias and multiplicative censoring. *J. Statist. Plann. Inference*, 91(2):295-312, 2000.

B. van Es, P. Spreij, and H. Van Zanten. Nonparametric volatility density estimation for discrete time models. *Journal of Nonparametric Statistics*, 17(2):237-249, 2005.

M. Klein, T. Mathew, and B. Sinha. A comparison of statistical disclosure control methods: Multiple imputation versus noise multiplication. *Research report series*, (02), 2013.

B. Sinha, T. K. Nayak, and L. Zayatz. Privacy protection and quantile estimation from noise multiplied data. *Sankhya B*, 73(2):297-315, 2011.

Y. Vardi. Multiplicative censoring, renewal processes, deconvolution and decreasing density: Non-parametric estimation. *Biometrika*, 76(4):pp. 751-761, 1989.

Y. Vardi and C.-H. Zhang. Large sample study of empirical distributions in a random-multiplicative censoring model. *Ann. Statist.*, 20(2):1022–1039, 1992.