ON THE PROJECTION PURSUIT METHODOLOGY - VERSION 1

Jacques Touboul

PROJECTION PURSUIT THROUGH ENTROPY MINIMISATION

Jacques Touboul

Laboratoire de Statistique Théorique et Appliquée

Université Pierre et Marie Curie

jack_touboul@hotmail.com

Let us consider a defined density on a set of very large dimension. It is quite difficult to find an estimate of this density from a data set. However, it is possible through a projection pursuit methodology to achieve it. Over the last twenty years, many mathematicians have studied this approach, including Huber in 1985 (see "Projection pursuit", Annals of Statistics). In his article, Huber demonstrates the interest of his method in a very simple given case : considering two densities, a known one, and the other yet to be estimated, he manages through an algorithm to turn one into the other. He concludes, through a generalization of this process, by introducing a convergence in law. Huber's work is based on maximizing relative entropy.

Our work will consist in demonstrating that it is possible to achieve the same results as Huber's but this time through minimizations. We will then compare the different laws and tests obtained by Huber and us.

Keywords: projection pursuit; minimum relative entropy; maximum relative entropy; rate of convergence; kernel estimator; uniform convergence; convexity.

MSC(2000): 62H40(62H10,62H11,62H12,62H15).

Outline of the article

Let f be a density defined in $\mathbb{R}^d$. We define a first fairly flexible approximation of f (we will justify later this choice), which will simply be a density with same mean and variance as f and that we will name g .

Let's briefly consider Huber's findings :

Putting $\mathcal{F} = \{f^{(a)}; \text{for all } a \in \mathbb{R}^d_*,\ f^{(a)} = f\frac{g_a}{f_a} \text{ and } f^{(a)} \text{ is a density}\}$ - where generally h_u

is the density of $u^\top X$, if h is the density of X - then according to Huber, the first step of his algorithm amounts to defining a_1 and $f^{(a_1)}$ - that we will call from now on $f^{(1)}$ - by $f^{(a_1)} = \inf_{f^{(a)} \in \mathcal{F}} K(f^{(a)}, g)$ $(*)$,where a_1 is the vector in $\mathbb{R}^d$, which optimises this divergence.

In a second step, Huber replaces f by $f^{(1)}$ and goes through the first step again. By reiterating this process, Huber thus obtains a sequence $(a_1, a_2, ...)$ of vectors of $\mathbb{R}^d_*$. The sequence of the relative entropies of the $f^{(a_i)}$ to g - that we will call from now on $f^{(i)}$ - holds the relationship $K(f^{(0)}, g) \geq K(f^{(1)}, g) \geq ..... \geq 0$ with $f^{(0)} = f$.

Now, let us briefly describe that we want to do.

We also define $\mathcal{G} = \{g^{(a)}; \text{for all } a \in \mathbb{R}^d_*,\ g^{(a)} = g\frac{f_a}{g_a} \text{ and } g^{(a)} \text{ is a density}\}$. The first step of our algorithm consists in defining a_1 and $g^{(a_1)}$ - that we will call from now on $g^{(1)}$ - by $g^{(a_1)} = \inf_{g^{(a)} \in \mathcal{G}} K(g^{(a)}, f)$ $(**)$, where a_1 is the vector in $\mathbb{R}^d$, which optimises this divergence. Later on, we will demonstrate this very a_1 optimises simultaneously both $(*)$ and $(**)$. In our second step, we will replace g by $g^{(1)}$, and we will repeat the first step. And so on, by reiterating this process, we will end up obtaining a sequence $(a_1, a_2, ...)$ of vectors in $\mathbb{R}^d_*$. The sequence of relative entropies of the $g^{(a_i)}$ to f - that we will call from now on $g^{(i)}$ - holds the relationship below: $K(g^{(0)}, f) \geq K(g^{(1)}, f) \geq ..... \geq 0$ with $g^{(0)} = g$.

The purpose of this article is to demonstrate that these two methods are equivalent. We will study the rate of convergence and the distribution limit of the different densities $g^{(i)}$ and $f^{(i)}$ for all element i, as well as the tests on any of the parameters. By changing the end of process test, we will demonstrate that there is convergence in i. In conclusion, we will show some simulations.

Demonstrations and reminders can be found in annex.

1. Comparison of all the optimisation methods

In its section, we will expose the three lemmas forming the theoretical basis of our method. Firstly, let f be a density defined in $\mathbb{R}^d$ and let g be a density such that

$K(f, g) < \infty,\ K(g, f) < \infty$ and where g still presents the same mean and variance as f.

We would like to build a density closer to f than g already is from a relative entropy standpoint. Since this new density has to be derived only from f and g, we define h by

$h(x_1) = \frac{f_1(x_1)}{g_1(x_1)}$, where f_1 and g_1 are the marginal densities of f and g in the x_1 direction. Based on Huber's lemma 13.1 of [HUB85], we derive the following lemma :

Lemma 1. Let us consider g^* such that $g^*(x) = g(x)h(x_1)$ be a density. Moreover, $h = arginf\{K(f, gr);$ where r is such that $x \mapsto g(x)r(x_1)$ be a density$\}$. Finally, we have $K(f, g) = K(f_1, g_1) + K(f, g^*)$.

Thus, and similarly, let us consider a as a vector set in $\mathbb{R}^d_*$. We define g^* by $g^*(x) = g(x)h(a^\top x)$, then g^* is a density such that h verifies $h = \frac{f_a}{g_a}$, $h = arginf\{K(f, gr);$ where r is such that $x \mapsto g(x)r(a^\top x)$ be a density$\}$, and $K(f, g) = K(f_a, g_a) + K(f, g^*)$.
Vector a is a projection vector.We will focus later on the selection of a.
Moreover, by keeping the expression $h(x_1) = \frac{f_1(x_1)}{g_1(x_1)}$, we can introduce the following lemma,

Lemma 2. The function g^* defined by $g^*(x) = g(x)h(x_1)$ is a density and we have $h = arginf\{K(gr, f);$ where r is such that $x \mapsto g(x)r(x_1)$ be a density$\}$.

Considering a, a set vector in $\mathbb{R}^d_*$, let us define g^* by $g^*(x) = g(x)h(a^\top x)$, then we can say g^* is a density such that h verifies $h = \frac{f_a}{g_a}$ and $h = arginf\{K(gr, f);$ where r is such that $x \mapsto g(x)r(a^\top x)$ be a density$\}$. We find also that $\inf_{a\in\mathbb{R}^d_*} K(g^*, f)$ is reached through lemma 11 (see page 19).

Finally, we introduce a third lemma

Lemma 3. The expression $\int g\{ln\frac{g^*}{f}\}dx$, is positive and is minimized for $h = \frac{f_a}{g_a}$. Moreover, we have $K(g, f) = K(g_a, f_a) + \int g\{ln\frac{g^*}{f}\}dx$.

Basically, the above implies that the choice of a , which is equivalent to the choice of h, is such that $K(g_a, f_a)$ is being maximised since $K(g, f)$ is set.
<u>Conclusion :</u> **The choice of $h = \frac{f_a}{g_a}$ enables us to simultaneously solve the following three optimisation problems, for $a \in \mathbb{R}^d_*$,**

$$\begin{aligned}
\textbf{First, find } a \textbf{ such that } a &= \boldsymbol{arginf}_{\{a\in\mathbb{R}^d_*; g^{(a)}\in\ \mathcal{G}\}}\ K(g^{(a)}, f)\\
\textbf{Second, find } a \textbf{ such that } a &= \boldsymbol{arginf}_{\{a\in\mathbb{R}^d_*; g^{(a)}\in\ \mathcal{G}\}}\ K(f, g^{(a)})\\
\textbf{Third, find } a \textbf{ such that } a &= \boldsymbol{argsup}_{\{a\in\mathbb{R}^d_*; g^{(a)}\in\ \mathcal{G}\}}\ K(g_a, f_a)
\end{aligned}$$

2. First convergences - Main results

Based on the work of Broniatowski in [BRO03] and [BROKEZ], we derive estimators of the minimum and maximum expressions obtained above. Then, after introducing certain notations, we will produce almost sure uniform convergences of the transformed densities obtained.

2.1 *Writing of the estimators*

Based on the Broniatovski articles mentioned above, we deduct that the estimator of the maximum of the relative entropy, as derived from Huber's works, is $\hat{a} = arg\sup_{a\in\mathbb{R}_*^d} \hat{K}(g_a, f_a)$ where $\hat{K}(g_a, f_a) = \int ln(\frac{g_a(a^\top x)}{f_a(a^\top x)})g_a(a^\top x)dx - \int(\frac{g_a(a^\top x)}{f_a(a^\top x)} - 1)(\frac{f_a(a^\top x)}{f(x)})d\mathbb{P}_n$.
Similarly, the estimator of the minimal distance of Kullback-Lieber in our algorithm is
$\check{a} = arg\inf_{a\in\mathbb{R}_*^d} \check{K}(g\frac{f_a}{g_a}, f)$
where $\check{K}(g\frac{f_a}{g_a}, f) = \int ln(\frac{g(x)}{f(x)}\frac{f_a(a^\top x)}{g_a(a^\top x)})g(x)\frac{f_a(a^\top x)}{g_a(a^\top x)}dx - \int(\frac{g(x)}{f(x)}\frac{f_a(a^\top x)}{g_a(a^\top x)} - 1)d\mathbb{P}_n$.
These estimators implicitly suppose that f and g are known. Therefore, we introduce an estimate of the convolution kernel of these densities, which leads to the formulation of certain hypotheses as explained below. Let X_1, X_2,..,X_n be a sequence of independent random vectors with same law f. Let Y_1, Y_2,..,Y_n be a sequence of independent random vectors with same law g. Then the kernel estimators f_n, g_n, $f_{a,n}$ and $g_{a,n}$ of f, g, f_a and g_a, for all $a \in \mathbb{R}_*^d$, uniformly converge (see Deheuvels (1974) in [DEH74]). Let us consider now a sequence θ_n such that $\theta_n \to 0$, and $y_n/\theta_n^2 \to 0$, where y_n is the rate of convergence of the kernel estimator. Then, going forward, we will only consider the members of the sample $X_1, X_2, ..., X_n$ associated to f and the members of the sample $Y_1, Y_2, ..., Y_n$ associated to g verifying $f_n(X_i) \geq \theta_n$, $g_n(Y_i) \geq \theta_n$ and $g_{b,n}(b^\top Y_i) \geq \theta_n$, for all i and for all $b \in \mathbb{R}_*^d$. The vectors meeting these conditions will be once again called $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$.
So let us consider
$A_1(n,a) = \frac{1}{n}\Sigma_{i=1}^n ln\{\frac{g_{a,n}(a^\top Y_i)}{f_{a,n}(a^\top Y_i)}\}\frac{g_{a,n}(a^\top Y_i)}{g_n(Y_i)}$, $A_2(n,a) = \frac{1}{n}\Sigma_{i=1}^n(\frac{g_{a,n}(a^\top X_i)}{f_{a,n}(a^\top X_i)} - 1)\frac{f_{a,n}(a^\top X_i)}{f_n(X_i)}$,
$B_1(n,a) = \frac{1}{n}\Sigma_{i=1}^n ln\{\frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)}\frac{g_n(Y_i)}{f_n(Y_i}\}$ $B_2(n,a) = \frac{1}{n}\Sigma_{i=1}^n(1 - \{\frac{f_{a,n}(a^\top X_i)}{g_{a,n}(a^\top X_i)}\frac{g_n(X_i)}{f_n(X_i}\})$.
Assuming the number of random vectors thus discarded is negligible compared to n, the uniform convergence mentioned above still holds and the definition of θ_n enables us to estimate the maximum of $K(g_a, f_a)$ (and respectively of the minimum of $K(g\frac{f_a}{g_a}, f)$) by the following

limit: $\lim_{n\to\infty}\sup_{a\in\mathbb{R}^d_*}|(A_1(n,a)-A_2(n,a))-K(g_a,f_a)|=0$,
(resp. $\lim_{n\to\infty}\sup_{a\in\mathbb{R}^d_*}|(B_1(n,a)-B_2(n,a))-K(g\frac{f_a}{g_a},f)|=0$.)

2.2 *Notations*

Let us define the following sequences $\{g^{\{k\}}\}_{k=0..d}$, $\{a_k\}_{k=1..d}$, $\{\hat{a}_k\}_{k=1..d}$, and $\{\check{a}_k\}_{k=1..d}$ where through an immediate induction, we have $g^0=g$, $g^{\{1\}}(x)=g(x)\frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)}$ - because the optimal h is $h=\frac{f_a}{g_a}$ - and $g^{\{j\}}(x)=g^{\{j-1\}}(x)\frac{f_{a_j}(a_j^\top x)}{g_{a_j}^{\{j-1\}}(a_j^\top x)}$ for $j=1..d$, i.e
$g^{\{j\}}(x)=g(x)\Pi_{k=1}^j\frac{f_{a_k}(a_k^\top x)}{[g^{\{k-1\}}]_{a_k}(a_k^\top x)}$. We define this way two new sequences, following the two different optimisation methods. Indeed, if $g_n^{\{k\}}$ represents $\hat{g}^{\{k\}}$ or $\check{g}^{\{k\}}$ and if $\tilde{a}_k$ represents $\check{a}_k$ or $\hat{a}_k$, we get $\{g_n^{\{j\}}\}_{j=1..d}$ where $g_n^{\{j\}}(x)=g_n^{\{j-1\}}(x)\frac{f_{\tilde{a}_j}(\tilde{a}_j^\top x)}{[g_n^{\{j-1\}}]_{\tilde{a}_j}(\tilde{a}_j^\top x)}$, i.e.
$g_n^{\{j\}}(x)=g(x)\Pi_{k=1}^j\frac{f_{\tilde{a}_k}(\tilde{a}_k^\top x)}{[g_n^{\{k-1\}}]_{\tilde{a}_k}(\tilde{a}_k^\top x)}$.

$\mathcal{N}$ota $\mathcal{B}$ene

In between each transformed density, we carry out a test of Kolmogorov-Smirnov to check if it is close to the real law. Many other adjustment tests can be carried out such that Stephens', Anderson-Darling's and Cramer-Von Mises'. Moreover, if f and g are gaussian, then in order to get $K(g,f)=0$, it is necessary for g to have same mean and variance as f, since $\frac{g}{f}.ln(\frac{g}{f})+\frac{g}{f}-1=0$ in $\frac{g}{f}=1$. This explains why we choose g this way.

2.3 *Convergence studies*

In this section, we will concentrate on the different types of convergence as a function of the two types of optimisation methodologies. These two methodologies are not symmetrical but lead to the same results. Although it is not obvious at first, they constitute an alternative to Huber's methodology [HUB85].
First, let us consider $\varphi(x)=x\ ln(x)-x+1$, ie $\varphi'(x)=ln(x)$ and
$\varphi^*(x)=x\varphi'^{-1}(x)-\varphi(\varphi'^{-1}(x))=e^x-1$, ie again $\varphi^*(\varphi'(x))=x-1$, and let us introduce some notations. If $\mathbf{P}$ and $\mathbf{P}^a$ are the densities of f and f_a respectively, let us consider

$$\begin{aligned}\Theta &= \mathbb{R}^d_*,\ \Theta^1_a=\{b\in\Theta\mid \int \varphi^*(\varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)}))f_a(a^\top x)\,dx<\infty\},\\ \Theta^2 &= \{b\in\Theta\mid \int\varphi^*(\varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)}))d\mathbf{P}<\infty\},\end{aligned}$$

$$
\begin{aligned}
m(b,a,x) &= \int \varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)}) g_a(a^\top x)\ dx \ - \ \varphi^*(\varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)})) \\
\mathbf{P}^a m(b,a) &= \int \varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)}) g_a(a^\top x)\ dx - \int \varphi^*(\varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)})) f_a(a^\top x)\ dx \\
\mathbb{P}_n m(b,a) &= \int \varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)}) g_a(a^\top x)\ dx - \int \varphi^*(\varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)}))\ \frac{f_a(a^\top x)}{f(x)} d\mathbb{P}_n \\
M(b,a,x) &= \int \varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)}) g(x) \frac{f_a(a^\top x)}{g_a(a^\top x)} dx - \ \varphi^*(\varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)})) \\
\mathbb{P}_n M(b,a) &= \int \varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)}) g(x) \frac{f_a(a^\top x)}{g_a(a^\top x)} dx - \int \varphi^*(\varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)})) d\mathbb{P}_n \\
\mathbf{P} M(b,a) &= \int \varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)}) g(x) \frac{f_a(a^\top x)}{g_a(a^\top x)} dx - \int \varphi^*(\varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)})) d\mathbf{P}
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{b}_n(a) &= arg \sup_{b\in\Theta} \ \mathbb{P}_n^a m(b,a) & \qquad \check{c}_n(a) &= arg \sup_{c\in\Theta} \ \mathbb{P}_n M(c,a) \\
\tilde{b}_n(a) &= arg \sup_{b\in\Theta_a^1} \ \mathbb{P}_n^a m(b,a) & \qquad \tilde{c}_n(a) &= arg \sup_{c\in\Theta^2} \ \mathbb{P}_n M(c,a) \\
\hat{\beta}_n &= arg \sup_{a\in\Theta} \ \sup_{b\in\Theta} \ \mathbb{P}_n^a m(b,a) & \qquad \check{\gamma}_n &= arg \inf_{a\in\Theta} \ \sup_{c\in\Theta} \ \mathbb{P}_n M(c,a) \\
\tilde{\beta}_n &= arg \sup_{a\in\Theta} \ \sup_{b\in\Theta_a^1} \ \mathbb{P}_n^a m(b,a) & \qquad \tilde{\gamma}_n &= arg \inf_{a\in\Theta} \ \sup_{c\in\Theta^2} \ \mathbb{P}_n M(c,a)
\end{aligned}
$$

We remark that $\hat{a}_k$ and $\check{a}_k$ are M-estimators for a_k for $k = 1..d$.

However, *Van der Vaart*, in chapter V of his work [VDW], thoroughly studies M-estimators and formulates hypotheses that we will use here in our context and for all set a_k, as defined in the above section **(2.2)**:

$(H1)$: $\sup_{a\in\Theta;\ b\in\Theta_a^1} |\mathbb{P}_n m(b,a) - \mathbf{P}^a m(b,a)| \to 0$ a.s. (respectively in probability)

$(H2)$: For all $\varepsilon > 0$, there is $\eta > 0$ such that, for all $b \in \Theta_a^1$ verifying

$\|b - a_k\| \geq \varepsilon$ for all $a \in \Theta$, we have $\mathbf{P}^a m(b,a) < \mathbf{P}^a m(a_k,a) - \eta$,

$(H3)$: There is a neighbourhood of a_k, V, and a positive function H, such that,

for all $b \in V$, we have $|m(b,a_k,x)| \leq H(x)$ $(\mathbf{P}^a - a.s.)$ *with* $\mathbf{P}^a H < \infty$,

$(H4)$: There is a neighbourhood V of a_k, such that for all ε, there is a η

such that for all $b \in V$ and $a \in \Theta$, verifying

$\|a - a_k\| \geq \varepsilon$, we have $\mathbf{P}^a m(b,a_k) - \eta > \mathbf{P}^a m(b,a)$.

and

$(H'1)$: $\sup_{a\in\Theta;\ c\in\Theta^2} |\mathbb{P}_n M(c,a) - \mathbf{P}M(c,a)| \to 0$ a.s. (respectively in probability)

$(H'2)$: For all $\varepsilon > 0$, there is $\eta > 0$, such that for all $c \in \Theta^2$ verifying $\|c - a_k\| \geq \varepsilon$

we have $\mathbf{P}M(c,a) - \eta > \mathbf{P}M(a_k,a)$, with $a \in \Theta$.

$(H'3)$: There is a neighbourhood of a_k, V, and a positive function H,

such that, for all $c \in V$ we have $|M(c,a_k,x)| \leq H(x)$ $(\mathbf{P} - a.s.)$ *with* $\mathbf{P}H < \infty$,

$(H'4)$: There is a neighbourhood V of a_k, such that for all ε, there is a η

such that for all $c \in V$ and $a \in \Theta$, verifying

$\|a - a_k\| \geq \varepsilon$, we have $\mathbf{P}M(c,a_k) < \mathbf{P}M(c,a) - \eta$.

Thus we will demonstrate that

Proposition 1. Assuming conditions $(H1 \to 4)$ are true, we have

(1) $\sup_{a\in\Theta} \|\hat{b}_n(a) - a_k\|$ **tends to 0 a.s.(respectively in probability)**

(2) $\hat{\beta}_n$ **tends to** a_k **a.s.(respectively in probability).**

Proposition 2. Assuming conditions $(H'1 \to 4)$ are true, we have

(1) $\sup_{a\in\Theta} \|\check{c}_n(a) - a_k\|$ **tends to 0 a.s. (respectively in probability)**

(2) $\check{\gamma}_n$ **tends to** a_k **a.s. (respectively in probability).**

Finally, if n is the number of vectors in the sample, then we have

<u>Theorem</u> 1. For a set $j = 1..d$, we have almost everywhere and even uniformly almost everywhere, the following convergences :

$$\check{g}^{\{j\}} \to g^{\{j\}},\ \textit{when}\ n \to \infty \ \textbf{and}\ \hat{g}^{\{j\}} \to g^{\{j\}},\ \textit{when}\ n \to \infty.$$

3. Rate of Convergence

In this section, we will show results that evidence the fact that these two optimisation methodologies are equivalent.

If m is the size of the sample and under the hypothesis:

$(H0)$: f and g are assumed to be strictly positive and bounded

(for which, we already shown in lemma 6 (see page 17) that $\check{g}^{(k)}$ was strictly positive and bounded),

Theorem 2. For all $k = 0..d$, we have

$$|\hat{g}^{(k)} - g^{(k)}| = O_{\mathbf{P}}(m^{-k/2}) \text{ and } |\check{g}^{(k)} - g^{(k)}| = O_{\mathbf{P}}(m^{-k/2}), \quad (1)$$

$$\int |\hat{g}^{(k)}(x) - g^{(k)}(x)|dx = O_{\mathbf{P}}(m^{-k/2}) \text{ and } \int |\check{g}^{(k)}(x) - g^{(k)}(x)|dx = O_{\mathbf{P}}(m^{-k/2}), \quad (2)$$

$$K(\check{g}^{(k)}, f) - K(g^{(k)}, f) = O_{\mathbf{P}}(m^{-k/2}) \text{ and } K(\hat{g}^{(k)}, f) - K(g^{(k)}, f) = O_{\mathbf{P}}(m^{-k/2}). \quad (3)$$

4. Estimator laws

We have $I_{a_k} = \frac{\partial^2}{\partial a^2} K(g\frac{f_{a_k}}{g_{a_k}}, f)$, and $x \to g(b, a, x) = \varphi'(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)})\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)}$.

Let us consider now four new hypotheses:

(H'5) : Estimators $\hat{\gamma}_n$ and $\hat{c}_n(a_k)$ converge towards a_k in probability.

(H'6) : The function φ is $\mathcal{C}^3$ in $(0, +\infty)$ and there is a neighbourhood of (a_k, a_k), that we will note V'_k, such that, for all (b, a) of V'_k, the gradient $\nabla(\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)})$ and the Hessian $\mathcal{H}(\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)})$ exist ($\lambda\_a.s.$), and the first order partial derivative $\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)}$ and the first and second order derivative of $(b, a) \mapsto g(b, a, x)$ are dominated (λ_a.s.) by integrable functions.

(H'7) : The function $(b, a) \mapsto M(b, a, x)$ is $\mathcal{C}^3$ in a neighbourhood V'_k of (a_k, a_k) for all x; and all the partial derivatives of $(b, a) \mapsto M(b, a, x)$ are dominated in V'_k by a $\mathbf{P}$_integrable function $H(x)$.

(H'8) : $\mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2$ and $\mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2$ are finite and the expressions $\mathbf{P}\frac{\partial^2}{\partial b_i \partial b_j}M(a_k, a_k)$ and I_{a_k} exist and are invertible.

We then have:

Theorem 3. Assuming that conditions $H'5$ to $H'8$ hold, then

$\sqrt{n}\mathcal{A}.(\hat{c}_n(a_k) - a_k) \overset{\mathcal{L}aw}{\to} \mathcal{B}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2) + \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2)$ **and**

$\sqrt{n}\mathcal{A}.(\hat{\gamma}_n - a_k) \overset{\mathcal{L}aw}{\to} \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2) + \mathcal{C}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2)$

where $\mathcal{A} = (\mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k)\frac{\partial^2}{\partial a \partial a}K(g\frac{f_{a_k}}{g_{a_k}}, f))$, $\mathcal{C} = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k)$ **and**

$\mathcal{B} = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k) + \frac{\partial^2}{\partial a \partial a}K(g\frac{f_{a_k}}{g_{a_k}}, f)$.

We also note $x \to g(b, a, x) = \varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)})g_a(a^\top x)$. Let us consider the following hypotheses:

(H5) : The estimators $\hat{\beta}_n$ and $\hat{b}_n(a_k)$ converge towards a_k in probability.

(H6) : The function φ is $\mathcal{C}^3$ in $(0, +\infty)$ and there exists a neibourghhood of (a_k, a_k), that we will name V_k, such that, for all (b, a) of V_k, the gradient $\nabla(\frac{g_a(a^\top x)}{f_a(a^\top x)})$ and the Hessian $\mathcal{H}(\frac{g_a(a^\top x)}{f_a(a^\top x)})$ exist $(\lambda - a.s.)$ and the first order partial derivative $\frac{g_a(a^\top x)}{f_a(a^\top x)}$ and the first and second order

derivative of $(b,a) \mapsto g(b,a,x)$ are dominated (λ_a.s.) by integrable functions.

(H7) : The function $(b,a) \mapsto m(b,a)$ is $\mathcal{C}^3$ in a neighbourhood V_k of (a_k,a_k) for all x and all the partial derivatives of $(b,a) \mapsto m(b,a)$ are dominated in V_k by a $\mathbf{P}$_integrable function $H(x)$.

(H8) : $\mathbf{P}\|\frac{\partial}{\partial b}m(a_k,a_k)\|^2$ and $\mathbf{P}\|\frac{\partial}{\partial a}m(a_k,a_k)\|^2$ are finite

and the quantities $\mathbf{P}\frac{\partial^2}{\partial b_i \partial b_j}m(a_k,a_k)$ and $\mathbf{P}\frac{\partial^2}{\partial a_i \partial a_j}m(a_k,a_k)$ are invertible. Then we have

Theorem 4. Assuming that conditions $H5$ to $H8$ hold, then

$\sqrt{n}\mathcal{D}.(\hat{b}_n(a_k) - a_k) \overset{\mathcal{L}aw}{\rightarrow} \mathcal{E}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}m(a_k,a_k)\|^2) + \mathcal{F}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}m(a_k,a_k)\|^2)$ **and**

$\sqrt{n}\mathcal{D}.(\hat{\beta}_n - a_k) \overset{\mathcal{L}aw}{\rightarrow} \mathcal{G}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}m(a_k,a_k)\|^2) + \mathcal{F}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}m(a_k,a_k)\|^2)$

where $\mathcal{E} = \mathbf{P}\frac{\partial^2}{\partial a^2}m(a_k,a_k)$, $\mathcal{F} = \mathbf{P}\frac{\partial^2}{\partial a \partial b}m(a_k,a_k)$, $\mathcal{G} = \mathbf{P}\frac{\partial^2}{\partial b^2}m(a_k,a_k)$ **and**

$\mathcal{D} = (\mathbf{P}\frac{\partial^2}{\partial b^2}m(a_k,a_k)\mathbf{P}\frac{\partial^2}{\partial a^2}m(a_k,a_k) - \mathbf{P}\frac{\partial^2}{\partial a \partial b}m(a_k,a_k)\mathbf{P}\frac{\partial^2}{\partial b \partial a}m(a_k,a_k))$ **with** $\mathcal{D} \in \mathbb{R}_*^+$.

5. New evolution in the process

The idea is simple: let us assume the algorithm does not stop after d iterations but only when the end of process test permits. Thus in this section, we will first be able to write a convergence between $g^{(j)}$ and f in j, then second, a new end of process test will provide us with an alternative to the Kolmogorov test.

5.1 *New convergence*

In this paragraph, we will evidence the fact there is a convergence between the law generated by $g^{(k)}$ and f.

First, a simple induction shows that the sequence of the transformed densities always holds $g_n^{\{j\}}(x) = g(x)\Pi_{k=1}^{j}\frac{f_{\tilde{a}_k}(\tilde{a}_k^\top x)}{[g_n^{\{k-1\}}]_{\tilde{a}_k}(\tilde{a}_k^\top x)}$, with $g^{(0)} = g$. As a reminder, the relative entropy is greater than the L^1 distance and we also have $K(g^{(0)},f) \geq K(g^{(k)},f) \geq K(g^{(k-1)},f) \geq 0$ and $K(g_a^{(0)},f_a) \geq K(g_a^{(k)},f_a) \geq K(g_a^{(k-1)},f_a) \geq 0$, where $g^{(0)} = g$.

Thus under hypothesis $(H0)$

$(H0)$: f and g are strictly positive and bounded,

lemma 6 (see page 17) implies that, for all k, $g^{(k)}$ is a strictly positive and bounded density. We then get

Theorem 5. Since K is greater than the L^1 distance, if

$[\min_a K(g^{(k)}\frac{f_a}{[g^{(k)}]_a}, f)] \to 0$, **when** $k \to \infty$ **(ie when the number of iterations is not finite), then the law generated by** $g^{(k)}$**, when** $k \to \infty$**, will be the same law as the one generated by** f**, ie** $\lim_k g^{(k)} = f$.
Similarly, if $[\max_a K(g_a^{(k)}, f_a)] \to 0$, **when** $k \to \infty$ **(ie the number of iterations is not finite) then** $\lim_k g^{(k)} = f$.

We thus infer the two following corollaries

Corollary 1. Based on theorem 1 and since K **is greater than the** L^1 **distance, then if** $[\min_a K(\hat{g}^{(k)}\frac{f_a}{[\hat{g}^{(k)}]_a}, f)] \to 0$, **when** $k \to \infty$, **(ie when the number of iterations is not finite), we have** $\lim_k \lim_n \hat{g}^{(k)} = f$.
Similarly, if $[\max_a K(\check{g}_a^{(k)}, f_a)] \to 0$, **when** $k \to \infty$ **(ie when the number of iterations is not finite), then** $\lim_k \lim_n \check{g}^{(k)} = f$.

Corollary 2. Since K **is greater than the** L^1 **distance, then**
if $\lim_n \lim_k [\min_a K(\hat{g}^{(k)}\frac{f_a}{[\hat{g}^{(k)}]_a}, f)] = 0$, **we have** $\lim_n \lim_k \hat{g}^{(k)} = f$.
Similarly, if $\lim_n \lim_k [\max_a K(\check{g}_a^{(k)}, f_a)] = 0$, **we have** $\lim_n \lim_k \check{g}^{(k)} = f$.

5.2 *Testing of the criteria*

Theorem 6. The law of the criteria writes

$$\sqrt{n}(Var_{\mathbf{P}}(M(a_k, a_k)))^{-1/2}(\mathbb{P}_n M(\hat{c}_n(a_k), \hat{\gamma}_n) - \mathbb{P}_n M(a_k, a_k)) \overset{\mathcal{L}aw}{\to} \mathcal{N}(0, I),$$
$$\sqrt{n}(Var_{\mathbf{P}}(m(a_k, a_k)))^{-1/2}(\mathbb{P}_n m(\hat{b}_n(a_k), \hat{\beta}_n) - \mathbb{P}_n m(a_k, a_k)) \overset{\mathcal{L}aw}{\to} \mathcal{N}(0, I),$$

where k **represents the** k^{th} **step of the algorithm.**

Thus, making the following hypotheses:
$(H'10)$*: there is a* k *such that* $[K(g^{(k)}\frac{f_{a_k}}{[g^{(k)}]_{a_k}}, f)] = 0$,
$(H10)$*: there is a* k *such that* $[K([g^{(k)}]_{a_k}, f_{a_k})] = 0$,
We can say that

Theorem 7. The law of the end of algorithm states

$$\sqrt{n}(Var_{\mathbf{P}}(M(a_k, a_k)))^{-1/2}(\mathbb{P}_n M(\hat{c}_n(a_k), \hat{\gamma}_n)) \overset{\mathcal{L}aw}{\to} \mathcal{N}(0, I), \tag{4}$$
$$\sqrt{n}(Var_{\mathbf{P}}(m(a_k, a_k)))^{-1/2}(\mathbb{P}_n m(\hat{b}_n(a_k), \hat{\beta}_n)) \overset{\mathcal{L}aw}{\to} \mathcal{N}(0, I). \tag{5}$$

where k **represents the last iteration of the algorithm.**

We can then build confidence ellipsoids around the last a_k thanks also to the following corollary:

Corollary 3. If $q_{1-\alpha}^{\mathcal{N}(0,1)}$ is the quantile of a reduced centered normal distribution with level α, then, expression (4) implies that
$\{b \in \mathbb{R}^d;\ \sqrt{n}(Var_{\mathbf{P}}(M(\hat{c}_n(a_k), \hat{\gamma}_n)))^{-1/2}(\mathbb{P}_n M(\hat{c}_n(a_k), \hat{\gamma}_n) \leq q_{1-\alpha}^{\mathcal{N}(0,1)}\}$ is a confidence ellipsoid with a level α of a_k according to our algorithm and the expression (5) implies that $\{b \in \mathbb{R}^d;\ \sqrt{n}(Var_{\mathbf{P}}(m(\hat{b}_n(a_k), \hat{\beta}_n)))^{-1/2}(\mathbb{P}_n m(\hat{b}_n(a_k), \hat{\beta}_n)) \leq q_{1-\alpha}^{\mathcal{N}(0,1)}\}$ is a confidence ellipsoid with a level α of a_k based on Huber's algorithm.

6. Simulations

We will illustrate this section by detailing several examples.

In each example, the first part of our program will follow our algorithm and will aim at creating a sequence of densities $(g^{(j)})$, $j = 1, .., k$, $k < d$, such that $g(0) = g$, $g^{(j)} = g^{(j-1)} f_{a_j} / [g^{(j-1)}]_{a_j}$ *and* $K(g^{(k)}, f) = 0$, where K is the relative entropy and $a_j = arg\inf_b K(g^{(j-1)} f_b / [g^{(j-1)}]_b, f)$, for all $j = 1, ..., k$. Moreover, in a second step, our program will follow Huber's method and will create a sequence of densities $(g^{(j)})$, $j = 1, .., k$, $k < d$, such that $g(0) = g$, $g^{(j)} = g^{(j-1)} f_{a_j} / [g^{(j-1)}]_{a_j}$ and $K(g^{(k)}, f) = 0$, where K is the relative entropy and $a_j = argsup_b K([g^{(j-1)}]_b, f_b)$, for all $j = 1, ..., k$. Let us remark that we test upfront the hypothesis that f is gaussian through a Kolmogorov Smirnov test.

Example 1 :

We are in dimension 3(=d), and we consider a sample of 50(=n) values of a random variable X with a density law f defined by,

$f(x) = normal(x_1 + x_2) * Gumbel(x_0 + x_2) * Gumbel(x_0 + x_1)$,

where the Gumbel law parameters are $(-3, 4)$ and $(1, 1)$ and where the normal distribution parameters are $(-5, 2)$. Let us generate then a gaussian random variable Y - that we will name g - with a density which presents the same mean and variance as f.

In the first part of our program, we theoretically obtain $k = 2$, $a_1 = (1, 0, 1)$ and $a_2 = (1, 1, 0)$ (or $a_2 = (1, 0, 1)$ and $a_1 = (1, 1, 0)$ which leads us to the same conclusion). To get this result, we perform the following test

$H0 :\ (a1, a2) = ((1, 0, 1), (1, 1, 0))$ *versus* $H1 :\ (a1, a2) \neq ((1, 0, 1), (1, 1, 0))$.

Moreover, if i represents the last iteration of the algorithm, then

$\sqrt{n}(Var_{\mathbf{P}}(M(a_i,a_i)))^{(-1/2)}(\mathbb{P}_nM(c_n(a_i),\gamma_n)-\mathbb{P}_nM(a_i,a_i)) \overset{\mathcal{L}aw}{\to} \mathcal{N}(0,1)$,

and then we estimate (a_1,a_2) by the following $0.9(=\alpha)$ level confidence ellipsoid

$\mathcal{E}_i=\{b\in\mathbb{R}^3;\ Var_{\mathbf{P}}(M(b,b))K(g^{(i)}f_b/[g^{(i)}]_b,f)\leq q_{1-\alpha}^{\mathcal{N}(0,1)}/\sqrt{n}=0.182434\}$.

Indeed, if $i=1$ represents the last iteration of the algorithm, then $a_1\in\mathcal{E}_0$, and if $i=2$ represents the last iteration of the algorithm, then $a_2\in\mathcal{E}_1$, and so on, if i represents the last iteration of the algorithm, then $a_i\in\mathcal{E}_{i-1}$.

Now, if we follow Huber's method, we also theoretically obtain $k=2$, $a_1=(1,0,1)$ and $a_2=(1,1,0)$ (or $a_2=(1,0,1)$ and $a_1=(1,1,0)$ which leads us to the same conclusion). To get this result, we perform the following test:

$H0:\ (a_1,a_2)=((1,0,1),(1,1,0))\ versus\ H1:\ (a_1,a_2)\neq((1,0,1),(1,1,0))$.

The fact that, if i represents the last iteration of the algorithm, then

$\sqrt{n}(Var_{\mathbf{P}}(m(a_i,a_i)))^{(-1/2)}(\mathbb{P}_nm(b_n(a_i),\beta_n)-\mathbb{P}_nm(a_i,a_i)) \overset{\mathcal{L}aw}{\to} \mathcal{N}(0,1)$,

enables us to estimate our sequence of (a_i), reduced to (a_1,a_2), through the following $0.9(=\alpha)$ level confidence ellipsoid $\mathcal{E}'_i=\{b\in\mathbb{R}^3;\ Var_{\mathbf{P}}(m(b,b))K([g^{(i)}]_b,f_b)\leq q_{1-\alpha}^{\mathcal{N}(0,1)}/\sqrt{n}=0.182434\}$. Indeed, if $i=1$ represents the last iteration of the algorithm, then $a_1\in\mathcal{E}'_0$, and if $i=2$ represents the last iteration of the algorithm, then $a_2\in\mathcal{E}'_1$, and so on, if i represents the last iteration of the algorithm, then $a_i\in\mathcal{E}'_{i-1}$. Finally, we obtain

	Our Algorithm	Huber's Algorithm
Kolmogorov Smirnov test, $H_0: f=g$	H_0 False	H_0 False
Projection Study number 0 :	minimum : 0.0317505	maximum : 0.00715135
	at point : (1.0,1.0,0)	at point : (1.0,1.0,0)
	P-Value : 0.99851	P-Value : 0.999839
Test :	$H_0: a_1\in\mathcal{E}_0$: False	$H_0: a_1\in\mathcal{E}'_0$: False
Projection Study number 1 :	minimum : 0.0266514	maximum : 0.00727748
	at point : (1.0,0,1.0)	at point : (1,0.0,1.0)
	P-Value : 0.998852	P-Value : 0.999835
Test :	$H_0: a_2\in\mathcal{E}_1$: True	$H_0: a_2\in\mathcal{E}'_1$: True
K(Kernel Estimation of $g^{(2)}$, $g^{(2)}$)	0.444388	0.794124

Therefore, we conclude that $f = g^{(2)}$.

Example 2 :

We are in dimension 2(=d), and we consider a sample of 50(=n) values of a random variable X with a density law f defined by, $f(x) = Cauchy(x_0) * Normal(x_1)$, where the Cauchy law parameters are -5 and 1 and where the normal distribution parameters are $(0, 1)$.

Our reasoning is the same as in Example 1. In the first part of our program, we theoretically obtain $k = 1$ and $a_1 = (1, 0)$. To get this result, we perform the following test:

$H0 : \ a_1 = (1, 0)$ *versus* $H1 : \ a_1 \neq (1, 0)$.

We estimate a_1 by the following $0.75(=\alpha)$ level confidence ellipsoid

$\mathcal{E}_i = \{b \in \mathbb{R}^2; \ Var_{\mathbf{P}}(M(b, b))K(gf_b/g_b, f) \leq q_{1-\alpha}^{\mathcal{N}(0,1)}/\sqrt{n} = 0.0961665\}$.

Now, if we follow Huber's method, we also theoretically obtain $k = 1$ and $a_1 = (1, 0)$. To get this result, we perform the following test: $H0 : \ a_1 = (1, 0)$ *versus* $H1 : \ a_1 \neq (1, 0)$. Hence, using the same reasoning as in Example 1, we estimate a_1 through the following $0.75(=\alpha)$ level confidence ellipsoid

$\mathcal{E}_i' = \{b \in \mathbb{R}^2; \ Var_{\mathbf{P}}(m(b, b))K([g^{(1)}]_b, f_b) \leq q_{1-\alpha}^{\mathcal{N}(0,1)}/\sqrt{n} = 0.0961665\}$. And, we obtain

	Our Algorithm	Huber's Algorithm
Kolmogorov Smirnov test, H_0 :	H_0 False	H_0 False
$f = g$	minimum : 0.00263554	maximum : 0.00376235
Projection Study n° 0 :	at point : (1.0001,0)	at point : (1.0,0.0)
	P-Value : 0.998683	P-Value : 0.998121
Test :	$H_0 : a_1 \in \mathcal{E}_0$: True	$H_0 : a_1 \in \mathcal{E}_0'$: True
K(Kernel Estimation of $g^{(1)}$, $g^{(1)}$)	2.44546	2.32331

Therefore, we conclude that $f = g^{(1)}$.

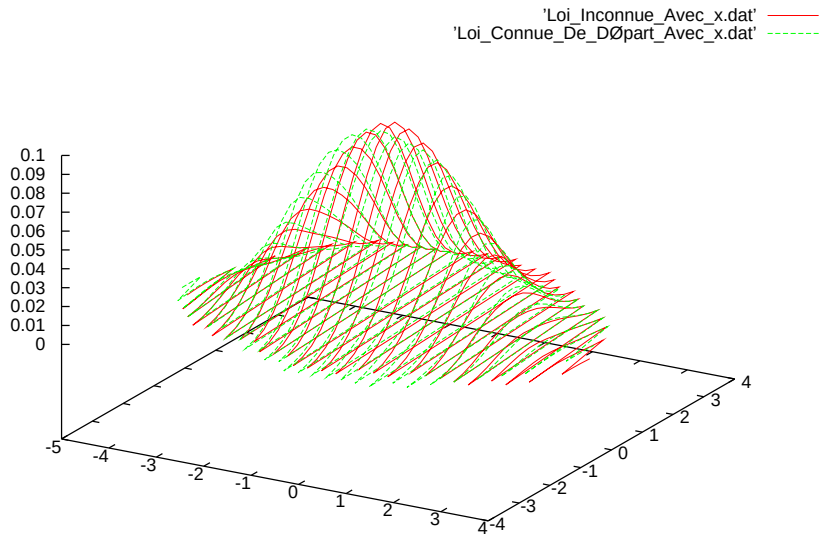

Figure 1: *Graph of the distribution to estimate and of the starting Gaussian.*

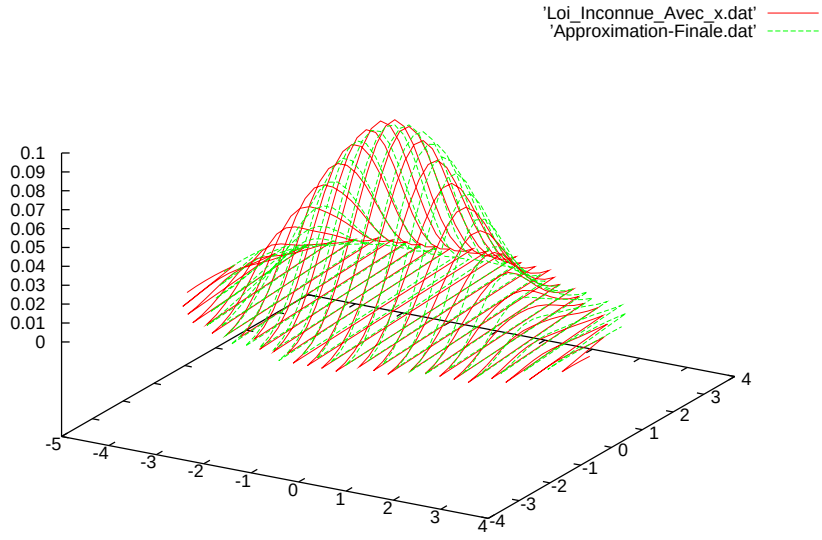

Figure 2: *Graph of the distribution to estimate and of our own estimate.*

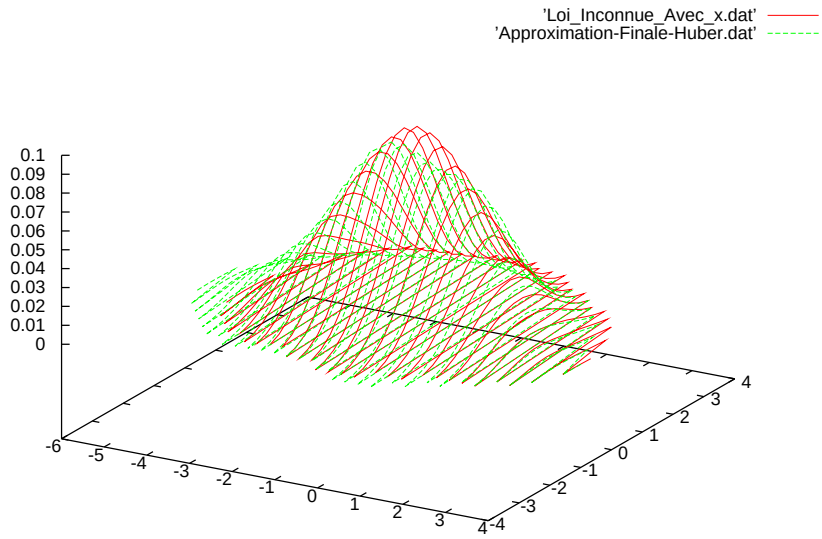

Figure 3: *Graph of the distribution to estimate and of the Huber estimate.*

Critics of the simulations

We note that as the approximations accumulate and according to the power of the calculators used, we might obtain results above or below the value of the thresholds of the different tests. Moreover, in the case where f is unknown, we will never be sure to have reached the minimum or the maximum of the relative entropy: we have indeed used the simulated annealing method to solve our optimisation problem, and therefore it is only when the number of random jumps tends in theory towards infinity that the probability to get the minimum or the maximum tends to 1. We note finally that no theory on the optimal number of jumps to implement does exist, as this number depends on the specificities of each particular problem.

Conclusion

The present article demonstrates that our relative entropy minimisation method constitutes a good alternative to Huber's. Indeed, the convergence results and simulations we carried out convincingly fulfilled our expectations regarding our methodology.

Annex I - Reminders

1 *The relative entropy*

Let us call h_a the density of $a^\top Z$ if h is the density of Z, and K the relative entropy or Kullback-Lieber distance, i.e. if P and Q are two probabilities then
$K(Q,P) = \int \varphi(\frac{\partial Q}{\partial P})\, dP$ if $P << Q$ and $K(Q,P) = +\infty$ otherwise, where
$\varphi : x \mapsto xln(x) - x + 1$ is strictly convex.
Let us present some well-known properties of the relative entropy.

Property 1. *A fundamental property of the relative entropy is the fact there is a unique case of nullity. We have* $K(P,Q) = 0 \Leftrightarrow P = Q$.

Property 2. *The application* $Q \mapsto K(Q,P)$ *is convex, lower semi-continuous for the topology that makes all the applications of the form* $Q \mapsto \int f dQ$ *continuous where* f *is bounded and continuous, lower semi-continuous for the topology of the uniform convergence, and greater than the* L^1 *distance.*

Moreover, the corollary (1.29) page 19 of [LIVAJ] enables us to say,

Property 3.
If $T : (X,A) \to (Y,B)$ *is measurable and if* $K(P,Q) < \infty$, *then* $K(P,Q) \geq K(PT^{-1}, QT^{-1})$, *and with equality being reached when* T *is surjective for* (P,Q).

And finally, according to the theorem III.4 of [AZE97], we have

Theorem 8. *Let* $f : I \to \mathbb{R}$ *be a convex function. Then* f *is a Lipschitz function in all compact intervals* $[a,b] \subset int\{I\}$. *In particular,* f *is continuous on* $int\{I\}$.

2 *Useful lemmas*

Lemma 4. *Let* f *be a density in* $\mathbb{R}^d$ *bounded and strictly positive. Then, any projection density of* f *, that we will name* f_a, $a \in \mathbb{R}^d_*$, *is also bounded and strictly positive in* $\mathbb{R}$.

Lemma 5. *Let* f *be a density in* $\mathbb{R}^d$ *bounded and strictly positive. Then all density* $f(./a^\top x)$, *for any* $a \in \mathbb{R}^d_*$, *is also bounded and strictly positive.*

The above lemmas 4 and 5 can be evidenced by a reductio ad absurdum argument. Moreover, by induction and lemmas 4 and 5, we have

Lemma 6. *If f and g are strictly positive and bounded densities, then $g^{(k)}$ is strictly positive and bounded.*

Finally we introduce a last lemma

Lemma 7. *Let f be an absolutely continuous density, then, for all sequence (a_n) tending to a in $\mathbb{R}^d_*$, the sequence f_{a_n} uniformly converges towards f_a.*

Proof :

For all a in $\mathbb{R}^d_*$, let F_a be the cumulative distribution function of $a^\top X$ and ψ_a be a complex function defined by $\psi_a(u,v) = F_a(\mathcal{R}e(u+iv)) + iF_a(\mathcal{R}e(v+iu))$, for all u and v in $\mathbb{R}$.
First, the function $\psi_a(u,v)$ is an analytic function, because $x \mapsto f_a(a^\top x)$ is continuous and since we have the corollary of Dini's second theorem - according to which *"A sequence of cumulative distribution functions which simply converges on $\mathbb{R}$ towards a continuous cumulative distribution function F on $\mathbb{R}$, uniformly converges towards F on $\mathbb{R}$"*- we deduct that, for all sequence (a_n) converging towards a, ψ_{a_n} uniformly converges toward ψ_a. Finally, the Weierstrass theorem, (see proposal (10.1) page 220 of the "Calcul infinitésimal" book of Jean Dieudonné), implies that all sequences $\psi'_{a,n}$ uniformly converge towards ψ'_a, for all a_n tending to a. We can therefore conclude. □

Annex II - Proofs

This last section includes the proofs of most of the lemmas, propositions, theorems and corollaries contained in the present article.

Proof of lemma 1

We remark that g and g^* determine the same density conditionally to x_1. Indeed,
$g_1^*(x_1) = \int g^*(x)dx_2...dx_d = \int h(x_1)g(x)dx_2...dx_d = h(x_1)\int g(x)dx_2...dx_d = h(x_1)g_1(x_1)$.
Thus, we can demonstrate this lemma.
We have $g(.|x_1) = \frac{g(x_1,...,x_n)}{g_1(x_1)}$ and $g_1(x_1)h(x_1)$ is the marginal density of g^*. Hence,
$\int g^*dx = \int g_1(x_1)h(x_1)g(.|x_1)dx = \int g_1(x_1)\frac{f_1(x_1)}{g_1(x_1)}(\int g(.|x_1)dx_2..dx_d)dx_1 = \int f_1(x_1)dx_1 = 1$

and since g^* is positive, then g^* is a density. Moreover,

$$\begin{aligned} K(f,g^*) &= \int f\{ln(f) - ln(g^*)\}dx, \quad (6) \\ &= \int f\{ln(f(.|x_1)) - ln(g^*(.|x_1)) + ln(f_1(x_1)) - ln(g_1(x_1)h(x_1))\}dx, \\ &= \int f\{ln(f(.|x_1)) - ln(g(.|x_1)) + ln(f_1(x_1)) - ln(g_1(x_1)h(x_1))\}dx, \quad (7) \end{aligned}$$

as $g^*(.|x_1) = g(.|x_1)$. Since the minimum of this last equation (7) is reached through the minimization of $\int f\{ln(f_1(x_1)) - ln(g_1(x_1)h(x_1))\}dx = K(f_1, g_1h)$, then proposition 1 necessarily implies that $f_1 = g_1h$, hence $h = f_1/g_1$.

Finally, we have $K(f,g) - K(f,g^*) = \int f\{ln(f_1(x_1)) - ln(g_1(x_1))\}dx = K(f_1,g_1)$, which completes the demonstration of the lemma.

Proof of lemma 2

The demonstration is very similar to lemma 1's save for the fact we now base our reasoning at row 6 on $K(g^*,f) = \int g^*\{ln(f) - ln(g^*)\}dx$ instead of $K(f,g^*) = \int f\{ln(f) - ln(g^*)\}dx$.

Proof of lemma 3

Without any loss of generalities, let us reason on the first component. In order to demonstrate this lemma, we will consider f^*, the density defined as $f^*(x) = f(x)t(x_1)$, where t is a function only depending on x_1. Since $f^*(.|x_1) = f(.|x_1)$, we get - as for lemma 1, and by reasoning with $K(g,f^*) = \int g\{ln(g)) - ln(f^*)\}dx$ and not with $K(f,g^*) = \int f\{ln(f) - ln(g^*)\}dx$ - the result.

Proof of lemma 8

Lemma 8. *The set Γ_c is closed in L^1 for the topology of the uniform convergence.*

By definition of the closure of a set, we have the result.

Proof of lemma 9

Lemma 9. *For all $c > 0$, we have $\Gamma_c \subset \overline{B}_{L^1}(f,c)$, where $B_{L^1}(f,c) = \{p \in L^1; \|f-p\|_1 \leq c\}$.*

Since K is greater than the L^1 distance, we get the result.

Proof of lemma 10

Lemma 10. *G is closed in L^1 for the topology of the uniform convergence.*

By definition of the closure of a set and lemma 7 (see page 17), we get the result.

Proof of lemma 11

Lemma 11. *We can say that* $\inf_{a\in\mathbb{R}^d_*} K(g^*, f)$ *is reached.*

Indeed, let G *be* $\{g\frac{f_a}{g_a};\ a \in \mathbb{R}^d_*\}$ *and* Γ_c *be* $\Gamma_c = \{p;\ K(p, f) \leq c\}$ *for all* $c>0$. *From lemmas 8, 9 and 10 (see page 18), we get* $\Gamma_c \cap G$ *is a compact for the topology of the uniform convergence, if* $\Gamma_c \cap G$ *is not empty. Hence, and since proposition 2 (see page 16) tells us that* $Q \mapsto K(Q, P)$ *is lower semi-continuous in* L^1 *for the topology of the uniform convergence, then the infimum is reached in* L^1.

Taking for example $c = K(g, f)$, Ω is necessarily not empty because we always have $K(g^*, f) \leq K(g, f)$.

Proof of propositions 1 and 2

First let us introduce the following lemma,

Lemma 12. *From* $(H1)$, *we derive the existence of* $C < 0$ *and* $n_0 > 0$ *such that*

$$(n \geq n_0) \ \Rightarrow\ \sup_{a\in\Theta} \sup_{b\in\{\Theta^1_a\}^c} \mathbb{P}_n m(b, a) < C$$

Proof :

This lemma comes from the fact that $\exists C > 0\ \sup_{a\in\Theta} \sup_{b\in\{\Theta^1_a\}^c} \mathbf{P}^a m(b, a) < -C$, which is true for any $b \in \{\Theta^1_a\}^c$. We have

$\infty = \int\ \varphi^*(\varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)}))f_a(a^\top x)\ dx = \int\ (\frac{g_b(b^\top x)}{f_b(b^\top x)} - 1)f_a(a^\top x)\ dx$ and since

$\int \varphi'(\frac{g_b(b^\top x)}{f_b(b^\top x)})g_a(a^\top x)dx \leq 1 + K(g_a, f_b) \leq 1 + K(g, f)$ as result of Property 3. Indeed, for all density h defined in $\mathbb{R}^d$ and for all u element in $\mathbb{R}^d_*$, considering $T : h_u \mapsto h$, we obtain the measurability of T and the following inequality

$\infty > K(g, f) \geq K(T^{-1}(g), T^{-1}(f)) = K(g_a, f_b)$, hence this conclusion. □

Given that $X_n \overset{\text{a.s.}}{\to} X$ if $\forall\varepsilon > 0,\ \mathbf{P}(\lim sup\{|X_n - X| > \varepsilon\}) = 0$, we prove proposition 1:

Proof :

Since $\tilde{b}_n(a) = arg\sup_{b\in\Theta^1_a}\ \mathbb{P}^a_n m(b, a)$, we have $\mathbb{P}_n m(\tilde{b}_n(a), a) \geq \mathbb{P}_n m(a_k, a)$.

And through condition $(H1)$, we get $\mathbb{P}_n m(\tilde{b}_n(a), a) \geq \mathbb{P}_n m(a_k, a) \geq \mathbf{P}^a m(a_k, a) - o_{\mathbb{P}}(1)$, where $o_{\mathbb{P}}(1)$ does not depend on a. Thus, we get:

$$\begin{aligned}\mathbf{P}^a m(a_k,a) - \mathbf{P}^a m(\tilde{b}_n(a),a) &\leq \mathbb{P}_n m(\tilde{b}_n(a),a) - \mathbf{P}^a m(\tilde{b}_n(a),a) + o_{\mathbb{P}}(1) \\ &\leq \sup_{a\in\Theta;\ b\in\Theta_a^1} |\mathbb{P}_n m(b,a) - \mathbf{P}^a m(b,a)| \to 0 \ a.s.\ . \end{aligned} \tag{8}$$

Let $\varepsilon > 0$ be such that $\sup_{a\in\Theta} \|\tilde{b}_n(a) - a_k\| > \varepsilon$. We notice that if such ε had failed to exist, the result would be obvious. Therefore, for this ε, there is $a_n \in \Theta$ such that $\|\tilde{b}_n(a_n) - a_k\| > \varepsilon$, which implies thanks to $(H2)$ that there exists a η such that $\mathbf{P}^{a_n} m(a_k,a_n) - \mathbf{P}^{a_n} m(\tilde{b}_n(a_n),a_n) > \eta$. Thus, we can write:

$$\mathbb{P}(\sup_{a\in\mathbb{R}^d} \|\tilde{b}_n(a) - a_k\| > \varepsilon) \leq \mathbb{P}(\mathbf{P}^{a_n} m(a_k,a_n) - \mathbf{P}^{a_n} m(\tilde{b}_n(a_n),a_n) > \eta) \to 0 \ by\ (8).$$

Moreover, $(H1)$, through the above lemma 12, implies that $\hat{b}_n(a) = \tilde{b}_n(a)$ for all $a \in \Theta$ and for n big enough. This results in $\sup_{a\in\Theta} \|\hat{b}_n(a) - a_k\| \to 0$ $a.s.$,
which concludes our demonstration of the first part of the proposition.
For the second part, we remark that $(H1)$, as a result of the above lemma 12, also implies that $\hat{\beta}_n = \tilde{\beta}_n$ for all $a \in \Theta$. This explains why it is sufficient to demonstrate the result for $\tilde{\beta}_n$ only. Based on the first part of the demonstration and on condition $(H3)$, we can write: $\mathbb{P}_n m(\tilde{b}_n(\tilde{\beta}_n),\tilde{\beta}_n) \geq \mathbb{P}_n m(\tilde{b}_n(a_k),a_k) \geq \mathbf{P}^a m(\tilde{b}_n(\tilde{\beta}_n),a_k) - o_{\mathbb{P}}(1)$, which implies:

$$\begin{aligned}\mathbf{P}^a m(\tilde{b}_n(\tilde{\beta}_n),a_k) - \mathbf{P}^a m(\tilde{b}_n(\tilde{\beta}_n),\tilde{\beta}_n) &\leq \mathbb{P}_n m(\tilde{b}_n(\tilde{\beta}_n),\tilde{\beta}_n) - \mathbf{P}^a m(\tilde{b}_n(\tilde{\beta}_n),\tilde{\beta}_n) + o_{\mathbb{P}}(1) \\ &\leq \sup_{a\in\Theta;\ b\in\Theta_a^1} |\mathbb{P}_n m(b,a) - \mathbf{P}^a m(b,a)| \to 0 \ a.s.. \end{aligned} \tag{9}$$

Based on the first part of this demonstration and on $(H4)$, we infer the existence of η such that: $\mathbb{P}(\|\tilde{\beta}_n - a_k\| \geq \varepsilon) \leq \mathbb{P}(\mathbf{P}^{\tilde{\beta}_n} m(\tilde{b}_n(\tilde{\beta}_n),a_k) - \mathbf{P}^{\tilde{\beta}_n} m(\tilde{b}_n(\tilde{\beta}_n),\tilde{\beta}_n) > \eta) \to 0$ $a.s.$ by (9), which concludes our demonstration. □

In a similar manner, we demonstrate proposition 2.

Proof of theorem 1

The demonstration below holds for the two types of optimisation. Let us consider $g^{(0)} = g$, a density with same mean and variance as f. In this proof, we will assume f and g are strictly positive and bounded i.e. through lemma 6 (see page 17), the densities $\hat{g}^{(k)}$ and $g^{(k)}$ are also strictly positive and bounded. Using lemma 2, (see page 3), and lemma 7, (see page 17), we

demonstrate the theorem by induction.

Proof of theorem 2

row 1: Here let us consider m, the size of the sample and f and g two bounded densities. This demonstration holds for the two types of optimisation. Let us consider $\Psi_j = \{\frac{f_{\check{a_j}}(\check{a_j}^\top x)}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)} - \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}\}$. Since f and g are bounded , it is easy to prove that from a certain rank, we get $|\Psi_j| \leq max(\frac{1}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)}, \frac{1}{[g^{(j-1)}]_{a_j}(a_j^\top x)})|f_{\check{a_j}}(\check{a_j}^\top x) - f_{a_j}(a_j^\top x)|$. Moreover, we can remark the following:

First, based on what we stated earlier, for all set x and from a certain rank,

there is a constant $R>0$ independent from n, such that:

$$max(\frac{1}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)}, \frac{1}{[g^{(j-1)}]_{a_j}(a_j^\top x)}) \leq R = R(x) = O(1).$$

Second, since $\check{a}_k$ is an $M-$estimator of a_k for $k = 1..d$, its convergence rate is $O_{\mathbf{P}}(m^{-1/2})$. Thus using simple functions, we obtain an upper and lower bound for $f_{\check{a_j}}$ and for f_{a_j} and we reach the following conclusion:

$$|\Psi_j| \leq O_{\mathbf{P}}(m^{-1/2}). \tag{10}$$

We finally obtain:
$|\Pi_{j=1}^k \frac{f_{\check{a_j}}(\check{a_j}^\top x)}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)} - \Pi_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}| = \Pi_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)} |\Pi_{j=1}^k \frac{f_{\check{a_j}}(\check{a_j}^\top x)}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)} \frac{[g^{(j-1)}]_{a_j}(a_j^\top x)}{f_{a_j}(a_j^\top x)} - 1|$.
Based on relationship 10, the expression $\frac{f_{\check{a_j}}(\check{a_j}^\top x)}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)} \frac{[g^{(j-1)}]_{a_j}(a_j^\top x)}{f_{a_j}(a_j^\top x)}$ tends towards 1 at a rate of $O_{\mathbf{P}}(m^{-1/2})$ for all j. Consequently $\Pi_{j=1}^k \frac{f_{\check{a_j}}(\check{a_j}^\top x)}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)} \frac{[g^{(j-1)}]_{a_j}(a_j^\top x)}{f_{a_j}(a_j^\top x)}$ tends towards 1 at a rate of $O_{\mathbf{P}}(m^{-k/2})$. Thus from a certain rank, we get

$$|\Pi_{j=1}^k \frac{f_{\check{a_j}}(\check{a_j}^\top x)}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)} - \Pi_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}| = O_{\mathbf{P}}(m^{-k/2}) O_{\mathbf{P}}(1) = O_{\mathbf{P}}(m^{-k/2}).$$

In conclusion, we obtain

$$|\check{g}^{(k)} - g^{(k)}| = g(x)|\Pi_{j=1}^k \frac{f_{\check{a_j}}(\check{a_j}^\top x)}{[\check{g}^{(j-1)}]_{\check{a_j}}(\check{a_j}^\top x)} - \Pi_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}| \leq O_{\mathbf{P}}(m^{-k/2}).$$

row 2: This demonstration holds for the two types of optimisation.

Since f and g are assumed to be strictly positive and bounded, hence lemma 6 (see page 17) implies $g^{(k)}$ is also, for all k, strictly positive and bounded.

Moreover, theorem 1 implies that $|\frac{\hat{g}^{(k)}(x)}{g^{(k)}(x)} - 1| = O_{\mathbf{P}}(m^{-k/2})$ because $g^{(k)}(x)|\frac{\hat{g}^{(k)}(x)}{g^{(k)}(x)} - 1| = |\hat{g}^{(k)}(x) - g^{(k)}(x)|$. Hence, there exists a function C of $\mathbb{R}^d$ in $\mathbb{R}^+$ such that $\lim_{m\to\infty} m^{-k/2} C(x) = 0$ and $|\frac{\hat{g}^{(k)}(x)}{g^{(k)}(x)} - 1| \leq m^{-k/2} C(x)$, we have:

$\int |\hat{g}^{(k)}(x) - g^{(k)}(x)| dx = \int g^{(k)}(x) |\frac{\hat{g}^{(k)}(x)}{g^{(k)}(x)} - 1| dx$, because $g^{(k)} > 0$

$$\leq \int g^{(k)}(x)C(x)m^{-k/2}dx,$$

Moreover, $\sup_{x\in\mathbb{R}^d}|\hat{g}^{(k)}(x)-g^{(k)}(x)| = \sup_{x\in\mathbb{R}^d} g^{(k)}(x)|\frac{\hat{g}^{(k)}(x)}{g^{(k)}(x)}-1|$

$$= \sup_{x\in\mathbb{R}^d} g^{(k)}(x)C(x)m^{-k/2} \to 0 \ a.s.,\text{ by theorem 1.}$$

This implies that $\sup_{x\in\mathbb{R}^d} g^{(k)}(x)C(x) < \infty\ a.s.$, ie $\sup_{x\in\mathbb{R}^d} C(x) < \infty\ a.s.$ since $g^{(k)}$ has been assumed to be strictly positive and bounded.

Thus, $\int g^{(k)}(x)C(x)dx \leq \sup C.\int g^{(k)}(x)dx = \sup C < \infty$ since $g^{(k)}$ is a density, therefore we can conclude $\int|\hat{g}^{(k)}(x)-g^{(k)}(x)|dx \leq \sup C.m^{-k/2} = O_{\mathbf{P}}(m^{-k/2})$.

row 3: This demonstration holds for the two types of optimisation. We have

$$K(\check{g}^{(k)},f)-K(g^{(k)},f) = \int f\varphi(\tfrac{\check{g}^{(k)}}{f})dx - \varphi(\tfrac{g^{(k)}}{f})dx = \int f\{\varphi(\tfrac{\check{g}^{(k)}}{f})-\varphi(\tfrac{g^{(k)}}{f}))\}dx$$

$$\leq \int f\ R|\tfrac{\check{g}^{(k)}}{f}-\tfrac{g^{(k)}}{f}|dx = R\int|\check{g}^{(k)}-g^{(k)}|dx$$

with the line before last being derived from theorem 8. We get the same expression as the one we found in our Proof of Theorem 2 row 2, we then conclude in a similar manner.

Proof of theorem 3

By definition of the estimators $\hat{\gamma}_n$ and $\hat{c}_n(a_k)$, we have $\begin{cases} \mathbb{P}_n\frac{\partial}{\partial b}M(b,a)=0 \\ \mathbb{P}_n\frac{\partial}{\partial a}M(b(a),a)=0 \end{cases}$

ie $\begin{cases} \mathbb{P}_n\frac{\partial}{\partial b}M(\hat{c}_n(a_k),\hat{\gamma}_n)=0 \\ \mathbb{P}_n\frac{\partial}{\partial a}M(\hat{c}_n(a_k),\hat{\gamma}_n)+\mathbb{P}_n\frac{\partial}{\partial b}M(\hat{c}_n(a_k),\hat{\gamma}_n)\frac{\partial}{\partial a}\hat{c}_n(a_k)=0, \end{cases}$ which leads to the simplification

of the above system into $\begin{cases} \mathbb{P}_n\frac{\partial}{\partial b}M(\hat{c}_n(a_k),\hat{\gamma}_n)=0\ (E0) \\ \mathbb{P}_n\frac{\partial}{\partial a}M(\hat{c}_n(a_k),\hat{\gamma}_n)=0\ (E1) \end{cases}$.

Using a Taylor development of the $(E0)$ equation, we infer there exists $(\overline{c}_n,\overline{\gamma}_n)$ on the interval $[(\hat{c}_n(a_k),\hat{\gamma}_n),(a_k,a_k)]$ such that

$-\mathbb{P}_n\frac{\partial}{\partial b}M(a_k,a_k) = [(\mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k))^\top + o_{\mathbf{P}}(1),(\mathbf{P}\frac{\partial^2}{\partial a\partial b}M(a_k,a_k))^\top + o_{\mathbf{P}}(1)]a_n.$

with $a_n = ((\hat{c}_n(a_k)-a_k)^\top,(\hat{\gamma}_n-a_k)^\top)$.

Similarly, through a Taylor development of $(E1)$, we infer there exists $(\tilde{c}_n,\tilde{\gamma}_n)$ on the interval $[(\hat{c}_n(a_k),\hat{\gamma}_n),(a_k,a_k)]$ such that

$-\mathbb{P}_n\frac{\partial}{\partial a}M(a_k,a_k) = [(\mathbf{P}\frac{\partial^2}{\partial b\partial a}M(a_k,a_k))^\top + o_{\mathbf{P}}(1),(\mathbf{P}\frac{\partial^2}{\partial a^2}M(a_k,a_k))^\top + o_{\mathbf{P}}(1)]a_n.$

with $a_n = ((\hat{c}_n(a_k)-a_k)^\top,(\hat{\gamma}_n-a_k)^\top)$. Thus we get

$$\sqrt{n}a_n = \sqrt{n}\begin{bmatrix} \mathbf{P}\frac{\partial^2}{\partial b^2}M(a_k,a_k) & \mathbf{P}\frac{\partial^2}{\partial a\partial b}M(a_k,a_k) \\ \mathbf{P}\frac{\partial^2}{\partial b\partial a}M(a_k,a_k) & \mathbf{P}\frac{\partial^2}{\partial a^2}M(a_k,a_k) \end{bmatrix}^{-1}\begin{bmatrix} -\mathbb{P}_n\frac{\partial}{\partial b}M(a_k,a_k) \\ -\mathbb{P}_n\frac{\partial}{\partial a}M(a_k,a_k) \end{bmatrix} + o_{\mathbf{P}}(1)$$

$$= \sqrt{n}(\mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k)\frac{\partial^2}{\partial a\partial a}K(g\tfrac{f_{a_k}}{g_{a_k}},f))^{-1}$$

$$\cdot \begin{bmatrix} \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k)+\frac{\partial^2}{\partial a\partial a}K(g\frac{f_{a_k}}{g_{a_k}},f) & \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k) \\ \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k) & \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k) \end{bmatrix} \cdot \begin{bmatrix} -\mathbb{P}_n\frac{\partial}{\partial b}M(a_k,a_k) \\ -\mathbb{P}_n\frac{\partial}{\partial a}M(a_k,a_k) \end{bmatrix} + o_{\mathbf{P}}(1)$$

since $(H6)$ enables us to reverse the derivative and integral signs.

Moreover, the central limit theorem implies: $\mathbb{P}_n\frac{\partial}{\partial b}m(a_k,a_k) \overset{\mathcal{L}aw}{\to} \mathcal{N}_d(0,\mathbf{P}\|\frac{\partial}{\partial b}m(a_k,a_k)\|^2)$, $\mathbb{P}_n\frac{\partial}{\partial a}m(a_k,a_k) \overset{\mathcal{L}aw}{\to} \mathcal{N}_d(0,\mathbf{P}\|\frac{\partial}{\partial a}m(a_k,a_k)\|^2)$, since $\mathbf{P}\frac{\partial}{\partial b}m(a_k,a_k) = \mathbf{P}\frac{\partial}{\partial a}m(a_k,a_k) = 0$, which leads us to the result.

We similarly demonstrate theorem 4.

Proof of theorem 5

Let us consider ψ, ψ_a, $\psi^{(k)}$, $\psi_a^{(k)}$ the characteristic functions of the densities f, f_a, $g^{(k-1)}$ and $[g^{(k-1)}]_a$, then $|\psi(ta)-\psi^{(k-1)}(ta)| = |\psi_a(t)-\psi_a^{(k-1)}(t)| \leq \int |f_a(a^\top x)-[g^{(k-1)}]_a(a^\top x)|dx$, and then $\sup_a |\psi_a(t)-\psi_a^{(k-1)}(t)| \leq \sup_a \int |f_a(a^\top x)-[g^{(k-1)}]_a(a^\top x)|dx \leq \max_a K([g^{(k-1)}]_a, f_a)$ since $\psi(ta) = \mathbb{E}(e^{ita^\top x}) = \psi_a(t)$, where $t \in \mathbb{R}$ and $a \in \mathbb{R}^d_*$, and since the relative entropy is greater than the L^1 distance and since this maximum has been reached. Therefore, if $\max_a K([g^{(k-1)}]_a, f_a)$ tends to 0, we have $\lim_k g^{(k)} = f$. Moreover, we have $|\psi(t)-\psi^{(k)}(t)| \leq \int |f(x)-g^{(k)}(x)|dx \leq K(g^{(k)},f) = min_a K(g^{(k-1)}\frac{f_a}{[g^{(k-1)}]_a}, f)$, hence the hypothesis according to which $\lim_k min_a K(g^{(k-1)}\frac{f_a}{[g^{(k-1)}]_a}, f) = 0$ implies that $\lim_k g^{(k)} = f$.

Proof of theorem 6

Through a Taylor development of $\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n)$ of rank 2, we get at point (a_k, a_k):

$$\begin{aligned}
&\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) \\
&= \mathbb{P}_n M(a_k,a_k) + \mathbb{P}_n\tfrac{\partial}{\partial a}M(a_k,a_k)(\check{\gamma}_n - a_k)^\top + \mathbb{P}_n\tfrac{\partial}{\partial b}M(a_k,a_k)(\check{c}_n(a_k)-a_k)^\top \\
&+\tfrac{1}{2}\{(\check{\gamma}_n-a_k)^\top\mathbb{P}_n\tfrac{\partial^2}{\partial a\partial a}M(a_k,a_k)(\check{\gamma}_n-a_k) + (\check{c}_n(a_k)-a_k)^\top\mathbb{P}_n\tfrac{\partial^2}{\partial b\partial a}M(a_k,a_k)(\check{\gamma}_n-a_k) \\
&+(\check{\gamma}_n-a_k)^\top\mathbb{P}_n\tfrac{\partial^2}{\partial a\partial b}M(a_k,a_k)(\check{c}_n(a_k)-a_k) + (\check{c}_n(a_k)-a_k)^\top\mathbb{P}_n\tfrac{\partial^2}{\partial b\partial b}M(a_k,a_k)(\check{c}_n(a_k)-a_k)\}
\end{aligned}$$

The lemma below enables us to conclude.

Lemma 13. *Let H be an integrable function and let $C = \int H\, d\mathbf{P}$ and $C_n = \int H\, d\mathbb{P}_n$, then, $C_n - C = O_{\mathbf{P}}(\frac{1}{\sqrt{n}})$.*

Thus we get $\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) = \mathbb{P}_n M(a_k,a_k) + O_{\mathbf{P}}(\frac{1}{n})$, ie $\sqrt{n}(\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) - \mathbf{P}M(a_k,a_k)) = \sqrt{n}(\mathbb{P}_n M(a_k,a_k) - \mathbf{P}M(a_k,a_k)) + o_{\mathbf{P}}(1)$. Hence $\sqrt{n}(\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) - \mathbf{P}M(a_k,a_k))$ abides by the same limit distribution as $\sqrt{n}(\mathbb{P}_n M(a_k,a_k) - \mathbf{P}M(a_k,a_k))$, which is $\mathcal{N}(0, Var_{\mathbf{P}}(M(a_k,a_k)))$.

Proof of corollaries 1 and 2

Since they are both identical, we will only develop the proof for corollary 2. These demonstrations also apply to the two types of optimisation. We will name g_n the density of $\hat{g}^{(k)}$ or $\check{g}^{(k)}$. Now, let us demonstrate that, under hypothesis $\lim_{k\to\infty} K(g^{(k)}, f) = 0$, we have $\lim_{n\to\infty} \lim_{k\to\infty} K(g_n^{(k)}, f) = 0$.

If $g_n^\infty = \lim_{k\to\infty} g_n^{(k)}$, then we can say g_n^∞ is a density. Indeed, we infer

$\int g_n^\infty = \int \lim_{k\to\infty} g_n^{(k)} = \lim_{k\to\infty} \int g_n^{(k)} = 1$ from the Lebesgue theorem and by induction, we get $g_n^\infty = g.(\Pi_{i\geq 1} \frac{f_{\hat{a}_i}}{[g_n(i-1)]_{\hat{a}_i}}) \geq 0$. Moreover, we have

$$\forall k, 0 \leq K(g_n^\infty, f) \leq K(g_n^{(k)}, f) \leq K(g, f), \ (*)$$

since the sequence $(K(g_n^{(k)}, f))_k$ is decreasing. Taking the limit in n of $(*)$, we get $\forall k, 0 \leq \lim_{n\to\infty} K(g_n^\infty, f) \leq \lim_{n\to\infty} K(g_n^{(k)}, f) \leq K(g, f)$, ie

$$\forall k, 0 \leq K(g_\infty^\infty, f) \leq K(g^{(k)}, f) \leq K(g, f), \ (**)$$

where $g_\infty^\infty = \lim_{n\to\infty} g_n^\infty$ and $g^{(k)} = \lim_{n\to\infty} g_n^{(k)}$ thanks to theorem 1. Through a reductio ad absurdum, and assuming $K(g_\infty^\infty, f) > K(g^{(k)}, f) \geq 0$, since K is lower semi continuous, we have $\lim_n inf K(g_n^\infty, f) \geq K(g_\infty^\infty, f)$ and $\lim_n inf K(g_n^{(k)}, f) \geq K(g_\infty^{(k)}, f)$. Consequently, $K(g_n^\infty, f) \geq K(g_\infty^\infty, f) > K(g_n^{(k)}, f)$, which leads to the contradiction we were looking for. Hence $(**)$ is true. We can therefore conclude that $(**)$ implies $K(g_\infty^\infty, f) = 0$, ie

$\lim_{n\to\infty} \lim_{k\to\infty} K(g_n^{(k)}, f) = 0$, as a reductio ad absurdum argument would have led to $0 < K(g_\infty^\infty, f) \leq K(g^{(k)}, f)$, which would have contradicted the hypothesis according to which $\lim_{k\to\infty} K(g^{(k)}, f) = 0$.

References

[AZE97] AZE D.,*Eléments d'analyse convexe et variationnelle, Ellipse, 1997.*

[BRO03] Broniatowski M., *Estimation of the Kullback-Leibler divergence, Math. Methods Statist.,12(4):391–409(2004), 2003.*

[BROKEZ] Broniatowski M. and Amor Keziou, *Dual representation of ϕ-divergences and applications, C. R. Math. Acad. Sci. Paris,336(10):857–862, 2003.*

[DEH74] Deheuvels Paul, *Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité, C. R. Acad. Sci. Paris Sér. A,278:1217–1220, 1974.*

[HUB85] Huber Peter J., *Projection pursuit, Ann. Statist.,13(2):435–525, 1985*, With discussion.

[LIVAJ] Liese Friedrich and Vajda Igor, *Convex statistical distances, volume 95 of Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]. BSB B. G. Teubner Verlagsgesellschaft, 1987, with German, French and Russian summaries.*

[VDW] van der Vaart A. W., *Asymptotic statistics, volume 3 of Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 1998.*