ON THE PROJECTION PURSUIT METHODOLOGY - VERSION 2

Jacques Touboul

PROJECTION PURSUIT THROUGH Φ-DIVERGENCE MINIMISATION

Jacques Touboul
Université Pierre et Marie Curie
Laboratoire de Statistique Théorique et Appliquée
175 rue du Chevaleret, 75013 Paris, France
jack_touboul@hotmail.com

Abstract : 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 through divergence based methodologies to achieve the same results as Huber but this time through minimizations of almost any $\varphi-$divergence and then in examining the advantages gained.
Keywords : convexity; kernel estimator; maximum relative entropy; minimum Φ-divergence; projection pursuit; rate of convergence; uniform convergence.
MSC: 94A17 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 optimizes the relative entropy K. In a second step, Huber replaces f by $f^{(1)}$ and go 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 what we want to do.
Let us first introduce the concept of $\Phi-$divergence. Let φ be a strictly convex function defined by $\varphi : \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$, and such as $\varphi(1) = 0$. We define a $\Phi-$divergence of P from Q - where P and Q are two probability distributions over a space Ω such that Q is absolutely continuous with respect to P - by $\Phi(Q, P) = \int \varphi(\frac{dQ}{dP})dP$. Throughout this article, we will also assume that $\varphi(0) < \infty$, that φ' is continuous and that this divergence is greater than

the L^1 distance. 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 $g^{(a_1)} = \inf_{g^{(a)} \in \mathcal{G}} \Phi(g^{(a)}, f)$ (**), 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 the divergences of the $g^{(a_i)}$ to f - that we will call from now on $g^{(i)}$ - holds the following relationship: $\Phi(g^{(0)}, f) \geq \Phi(g^{(1)}, f) \geq ..... \geq 0$ with $g^{(0)} = g$.

The purpose of this article is to demonstrate the feasibility of extending Huber's method - which is based on maximizing the relative entropy - to our method - which is based on the minimization of a $\Phi-$divergence - and then to examine the advantages gained.

We will study the rate of convergence and laws of the different densities $g^{(i)}$ for all set i, as well as the tests on all parameters. Moreover, by changing the end of process tests, we will evidence the existence of convergences in i and we will perform new tests. In conclusion, we will present simulations. All demonstrations and reminders can be found in Annex.

1 Comparison of Huber's method to ours

First, let us expose the details of Huber method's.
Let f be a density defined on $\mathbb{R}^d$ and let g be a density belonging to a family of known and fixed law such that $K(f, g) < \infty$ and $K(g, f) < \infty$. We will assume that g still presents the same mean and variance as f.
We would like to build a sequence of density closer to f than g already is from a relative entropy standpoint. Since the first density of this sequence 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 [Huber, 1985], we derive the following lemma :

Lemma 1. The function g^* defined by $g^*(x) = g(x)h(x_1)$ is a density.
Moreover, $h = argin f\{K(f, gr);$ where r is such that $x \mapsto g(x)r(x_1)$ is a density$\}$.
Finally, we have $K(f, g) = K(f_1, g_1) + K(f, g^*)$.

Thus, and similarly, let us consider a, a vector set in $\mathbb{R}^d_*$. We define g^* by
$g^*(x) = g(x)h(a^\top x)$, then g^* is a density where h holds $h = \frac{f_a}{g_a}$,
$h = argin f\{K(f, gr);$ where r is such that $x \mapsto g(x)r(a^\top x)$ is a density$\}$, and
$K(f, g) = K(f_a, g_a) + K(f, g^*)$.
Hence, the vector a - that we will note a_1 - is a projection vector and g^* - that we will call $g^{(1)}$ - is the first element of the sequence defined at the beginning of this paragraph. Now, by iterating this process between f and $g^{(1)}$, instead of f and g, we obtain a new projection vector, that we will call a_2, as well as a new density $g^{(2)}$. And so on, this way, Huber gets the sequence of densities $(g^{(n)})_n$ he wanted. Finally, he shows, that under certain assumptions, there exist convergences between $(g^{(n)})_n$ and f. This concludes our reminder of Huber's method.
Let us now expose our method.
The fact that we consider a broader context than Huber's, leads us to an algorithm presenting a simpler end of process test.

Let Φ be a divergence such that $\varphi(0) < \infty$ and greater than the L^1 distance. Keeping the notation $h(x_1) = \frac{f_1(x_1)}{g_1(x_1)}$, let us introduce the following lemma :

Lemma 2. On the one hand, the function g^*, defined by $g^*(x) = g(x)h(x_1)$, is a density. On the other hand, we have
$h = arginf\{\Phi(gr, f);$ **where r is such that $x \mapsto g(x)r(x_1)$ is a density**$\}$.

Thus, similarly, let us consider a vector a set in $\mathbb{R}^d$ and let us define g^* by $g^*(x) = g(x)h(a^\top x)$, then we can say g^* is a density where h holds $h = \frac{f_a}{g_a}$ and $h = arginf\{\Phi(gr, f);$ where r is such that $x \mapsto g(x)r(a^\top x)$ is a density$\}$.
Hence, the vector a - that we will note a_1 - is a projection vector and g^* - that we will note $g^{(1)}$ - is the first element of the sequence defined at the beginning of this paragraph. Now, by iterating this process between f and $g^{(1)}$ (instead of between f and g), we get a new vector projection a_2 and a new density $g^{(2)}$. And so on, we obtain the sequence of the densities $(g^{(n)})_n$ we wanted. We will focus later on the selection of a_i. We find also that $\inf_{a\in\mathbb{R}^d_*} \Phi(g^*, f)$ is reached through lemma 10 (see page 15). We will therefore find these a_i and write their estimators.

2 First convergences

Based on the work of Broniatowski in [Broniatowski, 2003] and
[Broniatowski and Keziou, 2003], we derive estimators of the minimum expressions obtained above. Then, after introducing certain notations, we will produce almost sure uniform convergences of the transformed densities obtained.

2.1 Writing the estimators

Let φ^* be a function defined by, $\forall t \in \mathbb{R}$, $\varphi^*(t) = t\varphi'^{-1}(t) - \varphi(\varphi'^{-1}(t))$, where φ' is the derivate function of φ, φ'^{-1} being the reciprocal function of φ'. Let $\mathcal{F}$ be the class of the function defined by $\mathcal{F} = \{x \mapsto \varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)});\ b \in \mathbb{R}^d_*\}$, then Broniatowski in [Broniatowski, 2003] and [Broniatowski and Keziou, 2003] shows that the estimator of $\Phi(g\frac{f_a}{g_a}, f)$ is:
$\hat{\Phi}(g\frac{f_a}{g_a}, f) = \sup_{b\in\mathbb{R}^d}\{\int\ \varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)})\ g(x)\frac{f_a(a^\top x)}{g_a(a^\top x)}\ dx\ -\ \int\ \varphi^*(\varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)}))\ d\mathbb{P}_n\}$
where $\mathbb{P}_n = \frac{1}{n}\Sigma\delta_{X_i}$ and thus

<u>Conclusion :</u> The estimator of the minimum divergence argument is

$$\begin{aligned}\hat{a} &= arg\inf_{a\in\mathbb{R}^d} \hat{\Phi}(g\frac{f_a}{g_a}, f)\\ &= arg\inf_{a\in\mathbb{R}^d}\sup_{b\in\mathbb{R}^d}\{\int \varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)})g(x)\frac{f_a(a^\top x)}{g_a(a^\top x)}dx - \int \varphi^*(\varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)}))d\mathbb{P}_n\}.\end{aligned}$$

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 [Deheuvels, 1974]). Let us consider now a positive sequence θ_n such that $\theta_n \to 0$, $y_n/\theta_n^2 \to 0$, where y_n is the rate of convergence of the

kernel estimator, $y_n^{(1)}/\theta_n^2 \to 0$, where $y_n^{(1)}$ is defined by $|\varphi(\frac{g_n(x)}{f_n(x)}\frac{f_{b,n}(b^\top x)}{g_{b,n}(b^\top x)}) - \varphi(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)})| \leq y_n^{(1)}$ for all b in $\mathbb{R}_*^d$ and all x in $\mathbb{R}^d$, and finally $\frac{y_n^{(2)}}{\theta_n^2} \to 0$, where $y_n^{(2)}$ is defined by $|\varphi'(\frac{g_n(x)}{f_n(x)}\frac{f_{b,n}(b^\top x)}{g_{b,n}(b^\top x)}) - \varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)})| \leq y_n^{(2)}$ for all b in $\mathbb{R}_*^d$ and all x in $\mathbb{R}^d$. 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 $b \in \mathbb{R}_*^d$, and the vectors meeting these conditions will be once again called $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$.
Moreover let us consider $B_1(n,a) = \frac{1}{n}\Sigma_{i=1}^n \varphi'\{\frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)}\frac{g_n(Y_i)}{f_n(Y_i}\}\frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)}$ and $B_2(n,a) = \frac{1}{n}\Sigma_{i=1}^n \varphi^*\{\varphi'\{\frac{f_{a,n}(a^\top X_i)}{g_{a,n}(a^\top X_i)}\frac{g_n(X_i)}{f_n(X_i}\}\}$. 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 minimum of $\Phi(g\frac{f_a}{g_a}, f))$ by the following limit $\lim_{n\to\infty} \sup_{a\in\mathbb{R}_*^d} |(B_1(n,a) - B_2(n,a)) - \Phi(g\frac{f_a}{g_a}, f)| = 0$.

2.2 Notations

In this paragraph, we will formalize what we explained earlier in our "Outline of the Article" section, i.e. we will write the sequence of the transformed densities obtained.
Thus, let us define the following sequences: $\{g^{\{k\}}\}_{k=0..d}$, $\{a_k\}_{k=1..d}$, $\{\hat{a}_k\}_{k=1..d}$ 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 a new sequence of the transformed densities $\{\hat{g}^{(j)}\}_{j=1..d}$, where $\hat{g}^{(j)}(x) = \hat{g}^{(j-1)}(x)\frac{f_{\hat{a}_j}(\hat{a}_j^\top x)}{[\hat{g}^{(j-1)}]_{\hat{a}_j}(\hat{a}_j^\top x)} = g(x)\Pi_{k=1}^j \frac{f_{\hat{a}_k}(\hat{a}_k^\top x)}{[\hat{g}^{(k-1)}]_{\hat{a}_k}(\hat{a}_k^\top x)}$ for $j = 1..d$.

$\mathcal{N}$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 $\Phi(g, f) = 0$, it is necessary for g to have same mean and variance as f, since, for the relative entropy, $\frac{g}{f}.ln(\frac{g}{f}) + \frac{g}{f} - 1 = 0$ if $\frac{g}{f} = 1$. This explains why we choose g this way.

2.3 Convergence studies

In this section, we will concentrate on the different types of convergence.
If $\mathbf{P}$ and $\mathbf{P}^a$ are the densities of f and f_a respectively, let us consider
$\Theta = \mathbb{R}^d$, $\Theta^\Phi = \{b \in \Theta \mid \int \varphi^*(\varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)}))d\mathbf{P} < \infty\}$,
$M(b,a,x) = \int \varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)})g(x)\frac{f_a(a^\top x)}{g_a(a^\top x)}dx - \varphi^*(\varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)}))$,
$\mathbb{P}_n M(b,a) = \int \varphi'(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)})g(x)\frac{f_a(a^\top x)}{g_a(a^\top x)}dx \int \varphi^*(\varphi'(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)}))d\mathbb{P}_n$,
$\mathbf{P}M(b,a) = \int \varphi'(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)})g(x)\frac{f_a(a^\top x)}{g_a(a^\top x)}dx - \int \varphi^*(\varphi'(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)}))d\mathbf{P}$,
$\hat{c}_n(a) = arg\sup_{c\in\Theta} \mathbb{P}_n M(c,a)$, $\tilde{c}_n(a) = arg\sup_{c\in\Theta^\Phi} \mathbb{P}_n M(c,a)$,
$\hat{\gamma}_n = arg\inf_{a\in\Theta} \sup_{c\in\Theta} \mathbb{P}_n M(c,a)$ and $\tilde{\gamma}_n = arg\inf_{a\in\Theta} \sup_{c\in\Theta^\Phi} \mathbb{P}_n M(c,a)$.
We remark that $\hat{a}_k$ is a M-estimator for a_k, $k = 1..d$ and its rate of convergence is consequently in $O_{\mathbf{P}}(m^{-1/2})$. However, *Van der Vaart*, in chapter V of his work [van der Vaart, 1998],

thoroughly studies M-estimators and formulates hypotheses that we will use here in our context and for all set a_k, as defined in section **(2.2)**:

$(H1)$: $\sup_{a\in\Theta;\ c\in\Theta^\Phi} |\mathbb{P}_n M(c,a) - \mathbf{P}M(c,a)| \to 0$ a.s. (respectively in probability),

$(H2)$: For all $\varepsilon > 0$, there is $\eta > 0$, such that for all $c \in \Theta^\Phi$ verifying $\|c - a_k\| \geq \varepsilon$, we have $\mathbf{P}M(c,a) - \eta > \mathbf{P}M(a_k,a)$, with $a \in \Theta$.

$(H3)$: $\exists\ Z < 0$, $n_0 > 0$ such that$(n \geq n_0 \ \Rightarrow \sup_{a\in\Theta} \sup_{c\in\{\Theta^\Phi\}^c} \mathbb{P}_n M(c,a) < Z)$

$(H4)$: 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 $\mathbf{P}H < \infty$,

$(H5)$: 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 $\mathbf{P}M(c,a_k) < \mathbf{P}M(c,a) - \eta$.

We will thus demonstrate that:

Proposition 1. Assuming conditions $(H1)$ to $(H5)$ hold, we have
(1) $\sup_{a\in\Theta} \|\hat{c}_n(a) - a_k\|$ **tends to 0 a.s. (respectively in probability)**
(2) $\hat{\gamma}_n$ **tends to** a_k **a.s. (respectively in probability).**

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

Theorem 1. For a set $j = 1..d$, we have almost everywhere and even uniformly almost everywhere, the following convergence: $\hat{g}^{\{j\}} \to g^{\{j\}}$, *when* $n \to \infty$.

3 Rate of convergence

In this section, we will expose results on the rate of convergence of our estimator. If m is the size of the sample and under the following hypothesis

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

- which thanks to lemma 5 (see page 13) implies that $\hat{g}^{(k)}$ is strictly positive and bounded - we have:

Theorem 2. For all $k = 1, ..., d$, we have

$$|\hat{g}^{(k)} - g^{(k)}| = O_{\mathbf{P}}(m^{-k/2}), \tag{3.1}$$

$$\int |\hat{g}^{(k)}(x) - g^{(k)}(x)|dx = O_{\mathbf{P}}(m^{-k/2}), \tag{3.2}$$

$$\Phi(\hat{g}^{(k)}, f) - \Phi(g^{(k)}, f) = O_{\mathbf{P}}(m^{-k/2}). \tag{3.3}$$

4 Estimator laws

Putting $I_{a_k} = \frac{\partial^2}{\partial a^2}\Phi(g\frac{f_{a_k}}{g_{a_k}}, f)$, 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:
(H6) : Estimators $\hat{\gamma}_n$ and $\hat{c}_n(a_k)$ converge towards a_k in probability.
(H7) : 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 λ-integrable functions.
(H8) : 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 the partial derivatives of $(b,a) \mapsto M(b,a)$ are all dominated in V_k by a $\mathbf{P}$_integrable function $H(x)$.
(H9) : $\mathbf{P}\|\frac{\partial}{\partial b}M(a_k,a_k)\|^2$ 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 $H6$ to $H9$ 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** $\mathcal{A} = (\mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k,a_k)\frac{\partial^2}{\partial a \partial a}\Phi(g\frac{f_{a_k}}{g_{a_k}},f))$, $\mathcal{C} = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k,a_k)$ **and** $\mathcal{B} = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k,a_k) + \frac{\partial^2}{\partial a \partial a}\Phi(g\frac{f_{a_k}}{g_{a_k}},f)$.

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. In this section, we will establish first the existence of 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 $g^{(j)}(x) = g(x)\Pi_{k=1}^{j}\frac{f_{\tilde{a}_k}(\tilde{a}_k^\top x)}{[g^{(j-1)}]_{\tilde{a}_k}(\tilde{a}_k^\top x)}$, with $g^{(0)} = g$. As a reminder, for all divergence setting an upper bound for the L^1 distance, we also have, $\Phi(g^{(0)},f) \geq \Phi(g^{(k)},f) \geq \Phi(g^{(k-1)},f) \geq 0$. Thus under hypothesis $(H0)$ -

$(H0)$: f and g are strictly positive and bounded -
lemma 5 (see page 13) implies that, for all k, $g^{(k)}$ is a strictly positive and bounded density. We then get:

Theorem 4. Given that Φ is greater than the L^1 distance, if $[\min_a \Phi(g^{(k)}\frac{f_a}{[g^{(k)}]_a},f)] \to 0$, **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** $\lim_k g^{(k)} = f$.

We thus infer the two following corollaries

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

Corollary 2. Given that Φ is greater than the L^1 distance, then if $\lim_n \lim_k [\min_a \Phi(\hat{g}^{(k)} \frac{f_a}{[\hat{g}^{(k)}]_a}, f)] = 0$, we have $\lim_n \lim_k \hat{g}^{(k)} = f$.

5.2 Testing of criteria

Theorem 5. The law of the criteria writes

$$\sqrt{n}(Var_{\mathbf{P}}(M(a_k, a_k)))^{-1/2}(\mathbb{P}_n M(\hat{c}_n(a_k), \hat{\gamma}_n) - \mathbb{P}_n M(a_k, a_k)) \overset{\mathcal{L}aw}{\rightarrow} \mathcal{N}(0, I), \tag{5.1}$$

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

Thus, making the following hypothesis:
$(H10)$*: there is a k such that* $[\Phi(g^{(k)} \frac{f_{a_k}}{[g^{(k)}]_{a_k}}, f)] = 0$, then

Theorem 6. The law of the end of algorithm states:

$$\sqrt{n}(Var_{\mathbf{P}}(M(a_k, a_k)))^{-1/2}(\mathbb{P}_n M(\hat{c}_n(a_k), \hat{\gamma}_n)) \overset{\mathcal{L}aw}{\rightarrow} \mathcal{N}(0, I), \tag{5.2}$$

where k represents the last iteration of the algorithm.

We can then build a confidence ellipsoid 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 α level reduced centered normal distribution, then, expression (5.2) implies that
$\{b \in \mathbb{R}^d;\ \sqrt{n}(Var_{\mathbf{P}}(M(\hat{c}_n(a_k), \hat{\gamma}_n)))^{-1/2}(\mathbb{P}_n M(\hat{c}_n(a_k), \hat{\gamma}_n) \leq q_{1-level}^{\mathcal{N}(0,1)}\}$
is a α level confidence ellipsoid of a_k according to our algorithm.

6 Simulation

Let us study three examples:

The first will be with relative entropy, the second with the χ^2-divergence and the third with a Cressie-Read divergence (still with $\gamma = 1.25$). We recall the definition of divergences from annex A (see page 12).

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 $\Phi(g^{(k)}, f) = 0$, where Φ is a divergence and $a_j = arg\inf_b \Phi(g^{(j-1)} f_b / [g^{(j-1)}]_b, f)$, 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 $\Phi(g^{(k)}, f) = 0$, where Φ is a divergence and $a_j = argsup_b \Phi([g^{(j-1)}]_b, f_b)$, 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 : With the relative entropy

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 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

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

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

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

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

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

enables us to estimate our sequence of (a_i), reduced to (a_1, a_2), through the following $0.9(=\alpha)$ level confidence ellipsoid

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

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 n° 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 n° 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 : With the χ^2 divergence

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,

$f(x) = Gaussian(x1 + x2) * Gaussian(x0 + x2) * Gumbel(x0 + x1)$,

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

In the first part of our program, we theoretically obtain $k = 1$ and $a_1 = (1, 1, 0)$. To get this result, we perform the following test: $H0 : \ a_1 = (1, 1, 0)$ versus $H1 : \ a_1 \neq (1, 1, 0)$. Moreover, using the same reasoning as in Example 1, we estimate a_1 by the following $0.9(=\alpha)$ level confidence ellipsoid $\mathcal{E}_i = \{b \in \mathbb{R}^3; \ Var_{\mathbf{P}}(M(b, b))\chi^2(gf_b/g_b, f) \leq \frac{q_{1-\alpha}^{\mathcal{N}(0,1)}}{\sqrt{n}} = 0.182434\}$.

Now, if we follow Huber's method, we also theoretically obtain $k = 1$ and $a_1 = (1, 1, 0)$. To get this result, we perform the following test: $H0 : \ a_1 = (1, 1, 0)$ versus $H1 : \ a_1 \neq (1, 1, 0)$. Hence, using the same reasoning as in Example 1, we are able to estimate our sequence of (a_i), reduced to a_1, through the following $0.9(=\alpha)$ level confidence ellipsoid

$\mathcal{E}_i' = \{b \in \mathbb{R}^3; \ Var_{\mathbf{P}}(m(b, b))\chi^2([g^{(1)}]_b, f_b) \leq q_{1-\alpha}^{\mathcal{N}(0,1)}/\sqrt{n} = 0.182434\}$.

And, we obtain

	Our Algorithm	Huber's Algorithm
Kolmogorov Smirnov test, $H_0 : \ f = g$	H_0 False	H_0 False
Projection Study n° 0 :	minimum : 0.0445199	maximum : 0.00960693
	at point : (1.0,1,0.0)	at point : (1.0,0,1.0)
	P-Value : 0.997535	P-Value : 0.99975
Test :	$H_0 : \ a_1 \in \mathcal{E}_0$: True	$H_0 : \ a_1 \in \mathcal{E}_0'$: True
K(Kernel Estimation of $g^{(1)}$, $g^{(1)}$)	6.99742	9.59275

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

Example 3 : With the Cressie-Read divergence (Φ)

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 $(0, 1)$. 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 $k = 1$ and $a_1 = (1, 0)$. To get this result, we perform the following test: $H0 : \ a_1 = (1, 0)$ versus $H1 : \ a_1 \neq (1, 0)$. Moreover, using the same reasoning as in Example 1, we estimate a_1 by the following $0.9(=\alpha)$ level confidence ellipsoid :

$\mathcal{E}_i = \{b \in \mathbb{R}^2; \ Var_{\mathbf{P}}(M(b, b))\Phi(gf_b/g_b, f) \leq q_{1-\alpha}^{\mathcal{N}(0,1)}/\sqrt{n} = 0.182434\}$.

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 are able to estimate our sequence of (a_i), reduced to a_1, through the following $0.9(=\alpha)$ level confidence ellipsoid

$\mathcal{E}_i' = \{b \in \mathbb{R}^2; \ Var_{\mathbf{P}}(m(b, b))\Phi([g^{(1)}]_b, f_b) \leq q_{1-\alpha}^{\mathcal{N}(0,1)}/\sqrt{n} = 0.182434\}$.

And, we obtain

	Our Algorithm	Huber's Algorithm
Kolmogorov Smirnov test, $H_0 : f = g$	H_0 False	H_0 False
Projection Study n° 0 :	minimum : 0.0210058	maximum : 0.0619417
	at point : (1.001,0)	at point : (1.0,0.0)
	P-Value : 0.989552	P-Value : 0.969504
Test :	$H_0 : a_1 \in \mathcal{E}_0$: True	$H_0 : a_1 \in \mathcal{E}'_0$: True
K(Kernel Estimation of $g^{(1)}$, $g^{(1)}$)	6.47617	2.09937

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

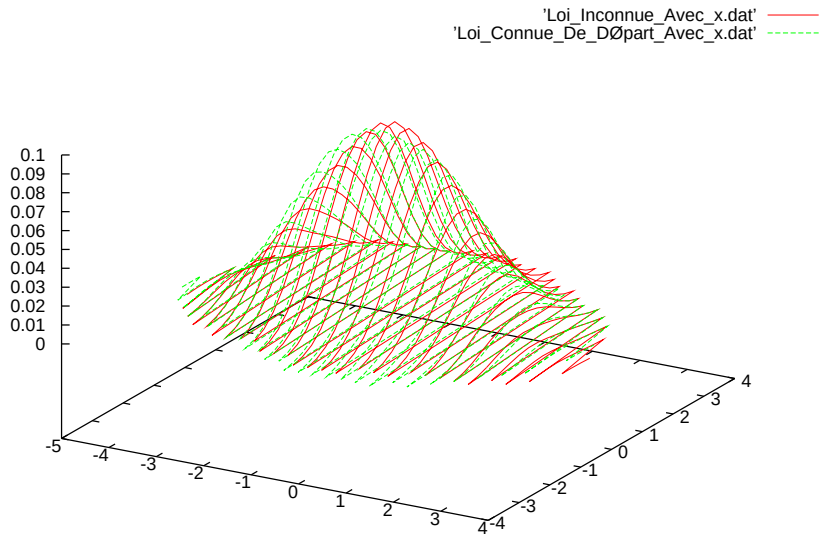

Figure 1: *Graph of the distribution to estimate and of the starting Gaussian.*

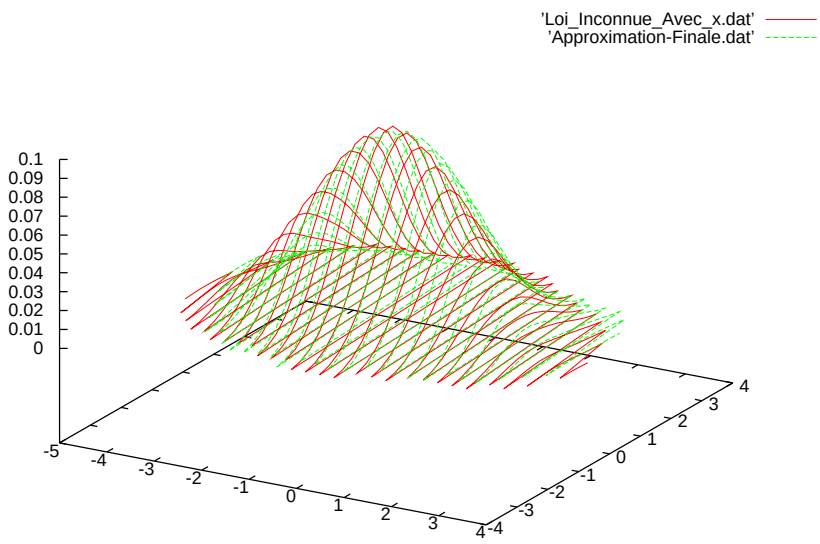

Figure 2: *Graph of the distribution to estimate and of our own estimate.*

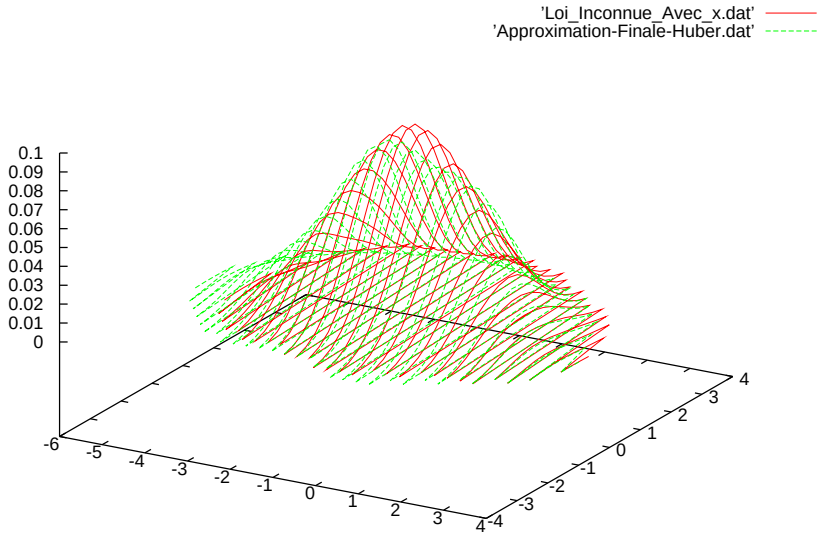

Figure 3: *Graph of the distribution to estimate and of Huber's estimate.*

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 Φ-divergence method constitutes a better alternative to Huber's. Indeed, the convergence results and simulations we carried out, convincingly fulfilled our expectations regarding our methodology. One of the key advantage of our method over Huber's lies in the fact that - since there exist divergences smaller than the relative entropy - our method requires a considerably shorter computation time.

A Annex - Reminders

A.1 Φ-Divergence

Let us call h_a the density of $a^\top Z$ if h is the density of Z. Let φ be a strictly convex function defined by $\varphi : \overline{\mathbb{R}^+} \to \overline{\mathbb{R}^+}$, and such that $\varphi(1) = 0$.

Definition 1. *We define $\Phi-$divergence of P from Q, where P and Q are two probability distributions over a space Ω such that Q is absolutely continuous with respect to P, by*

$$\Phi(Q, P) = \int \varphi(\frac{dQ}{dP}) dP. \tag{A.1}$$

The above expression (A.1) is also valid if P and Q are both dominated by the same probability.

The most used distances (Kullback, Hellinger or χ^2) belong to the Cressie-Read family (see Csiszar 1967 and Cressie - Read 1984). They are defined by a specific φ. Indeed,
- with the relative entropy, we associate $\varphi(x) = xln(x) - x + 1$
- with the Hellinger distance, we associate $\varphi(x) = 2(\sqrt{x} - 1)^2$
- with the χ^2 distance, we associate $\varphi(x) = \frac{1}{2}(x-1)^2$
- more generally, with power divergences, we associate $\varphi(x) = \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma-1)}$, where $\gamma \in \mathbb{R} \setminus (0, 1)$
- and, finally, with the L^1 norm, which is also a divergence, we associate $\varphi(x) = |x - 1|$.

We will notice that we have, in particular, the following inequalities:

$$d_{L^1}(g, f) \leq K(g, f) \leq \chi^2(g, f).$$

Let us now present some well-known properties of divergences.

Proposition 2. *A fundamental property of $\Phi-$divergences is the fact that there is a unique case of nullity. We have $\Phi(P, Q) = 0 \Leftrightarrow P = Q$.*

Property 1. *The application $Q \mapsto \Phi(Q, P)$ is*

$$(i) \quad \textit{convex}, \tag{A.2}$$

(ii) *lower semi -continuous, for the topology that makes all the applications of the form* $Q \mapsto \int f dQ$ *continuuous where* f *is bounded and continued, and* (A.3)

(iii) *lower semi-continuous for the topology of the uniform convergence.*

Finally, we will also use the following property derived from the first part of corollary (1.29) page 19 of [Friedrich and Igor, 1987],

Property 2.
If $T : (X, A) \to (Y, B)$ is measurable and if $\Phi(P, Q) < \infty$, then $\Phi(P, Q) \geq \Phi(PT^{-1}, QT^{-1})$.

A.2 Useful properties and lemmas

We introduce several theorems and properties related to convexity.

Property 3 (Characterization of convex functions)**.**
Let f be a function of I in $\mathbb{R}$.
f is convex if and only if one of the assertions below holds:
(i) Any arc of the graph of f is above its chord,
(ii) The epigraph of f is convex (in the meaning of the convex part of an affine space),
(iii) For any $(x_1, .., x_n) \in I^n$ and for any $(\lambda_1, .., \lambda_n) \in \mathbb{R}_+^n$ such that $\sum_1^n \lambda_i = 1$, we have $f(\sum_1^n \lambda_i x_i) \leq \sum_1^n \lambda_i f(x_i)$.
(iv) For any $x \in I$, the function $t \mapsto \frac{f(t)-f(x)}{t-x}$ is increasing in $I \setminus x$.

Then, according to theorem III.4 of [AZE, 1997], we have

Theorem 7. Let $f : I \to \mathbb{R}$ be a convex function. Then f is a Lipschitz function in all compact interval $[a, b] \subset int\{I\}$. In particular, f is continuous on $int\{I\}$.

Now, we introduce useful lemmas.

Lemma 3. *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 4. *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 3 and 4 can be evidenced by a reductio ad absurdum argument. Moreover, by induction and through lemmas 3 and 4, we have

Lemma 5. *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 6. *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. □

B Annex - 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^* present 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,

$$\begin{aligned} K(f,g^*) &= \int f\{ln(f) - ln(g^*)\}dx, & \text{(B.1)} \\ &= \int f\{ln(f(.|x_1)) - ln(g^*(.|x_1)) + ln(f_1(x_1)) - ln(g_1(x_1)h(x_1))\}dx, \\ &= \int f\{ln(f(.|x_1)) - ln(g(.|x_1)) + ln(f_1(x_1)) - ln(g_1(x_1)h(x_1))\}dx, & \text{(B.2)} \end{aligned}$$

as $g^*(.|x_1) = g(.|x_1)$. Since the minimum of this last equation (B.2) 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 2 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
We remark that g and g^* present 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,

$$\begin{aligned}\Phi(g^*, f) &= \int f\varphi(\frac{g^*}{f})dx, \\ &= \int f(x).\varphi(\frac{g^*(./x_1)}{f(./x_1)}\frac{g_1(x_1)h(x_1)}{f_1(x_1)})dx.\end{aligned} \tag{B.3}$$

Thus, the minimum in h of (B.3) is reached through the minimization of $\varphi(\frac{g^*(./x_1)}{f(./x_1)}\frac{g_1(x_1)h(x_1)}{f_1(x_1)})$, in h. And since $h = h(x_1)$, this minimisation is obtained by choosing $h = \frac{f_1}{g_1}$, which completes the demonstration of this lemma.

Proof of lemma 7

Lemma 7. *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 8

Lemma 8. *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 Φ is greater than the L^1 distance, we get the result.

Proof of lemma 9

Lemma 9. *G is closed in L^1 for the topology of the uniform convergence.*

By definition of the closure of a set and lemma 6 (see page 13), we get the result.

Proof of lemma 10

Lemma 10. *We can say that $\inf_{a\in\mathbb{R}^d_*} \Phi(g^*, f)$ is reached.*

Indeed, let G be $\{g\frac{f_a}{g_a};\ a \in \mathbb{R}^d_*\}$ and Γ_c be $\Gamma_c = \{p;\ \Phi(p, f) \leq c\}$ for all $c>0$. From lemmas 7, 8 and 9 (see page 15), we get $\Gamma_c \cap G$ is a compact for the topology of the uniform convergence, if $\Gamma_c \cap G$ is not empty. Since proposition 1 (see page 13) implies $Q \mapsto \Phi(Q, P)$ is lower semi-continuous in L^1 for the topology of the uniform convergence, then the infimum is reached in L^1. And finally, taking for example $c = \Phi(g, f)$, Ω is necessarily not empty because we always have $\Phi(g^*, f) \leq \Phi(g, f)$. We therefore conclude.

Proof of proposition 1

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

Proof :

Since $\tilde{c}_n(a) = arg\sup_{c\in\Theta^\Phi} \mathbb{P}_n M(c, a)$, we have $\mathbb{P}_n M(\tilde{c}_n(a), a) \geq \mathbb{P}_n M(a_k, a)$. And through condition $(H1)$, we get $\mathbb{P}_n M(\tilde{c}_n(a), a) \geq \mathbb{P}_n M(a_k, a) \geq \mathbf{P}M(a_k, a) - o_{\mathbb{P}}(1)$,
where $o_{\mathbb{P}}(1)$ does not depend on a. Thus, we get:

$$\begin{aligned}\mathbf{P}M(a_k, a) - \mathbb{P}M(\tilde{c}_n(a), a) &\leq \mathbb{P}_n M(\tilde{c}_n(a), a) - \mathbf{P}M(\tilde{c}_n(a), a) + o_{\mathbb{P}}(1) \\ &\leq \sup_{a\in\Theta;\ c\in\Theta^\Phi} |\mathbb{P}_n M(c, a) - \mathbf{P}M(c, a)| \to 0\ a.s.\ .\end{aligned} \tag{B.4}$$

Let $\varepsilon > 0$ be such that $\sup_{a\in\Theta} \|\tilde{c}_n(a) - a_k\| > \varepsilon$. 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 $\|\tilde{c}_n(a_n) - a_k\| > \varepsilon$, which implies thanks to $(H2)$ that there exists a η such that $\mathbf{P}M(\tilde{c}_n(a_n), a_n) - \mathbf{P}M(a_k, a_n) > \eta$. Thus, we can write:

$$\mathbf{P}(\sup_{a\in\mathbb{R}^d} \|\tilde{c}_n(a) - a_k\| > \varepsilon) \;\leq\; \mathbf{P}(\mathbf{P}M(\tilde{c}_n(a_n), a_n) - \mathbf{P}M(a_k, a_n) > \eta) \to 0 \; by \; (B.4).$$

Moreover, $(H1)$ and $(H3)$ imply that $\hat{c}_n(a) = \tilde{c}_n(a)$ for all $a \in \Theta$ and for n big enough. This results in $\sup_{a\in\Theta} \|\hat{c}_n(a) - a_k\| \to 0$ $a.s.$, which concludes our demonstration of the first part of the proposition.
For the second part, we remark that $(H1)$ and $(H3)$ also imply that $\hat{\gamma}_n = \tilde{\gamma}_n$ for all $a \in \Theta$. This explains why it is sufficient to demonstrate the result for $\tilde{\gamma}_n$ only.
Based on the first part of the demonstration and on condition $(H4)$, we can write:
$\mathbb{P}_n M(\tilde{c}_n(\tilde{\gamma}_n), \tilde{\gamma}_n) \geq \mathbb{P}_n M(\tilde{c}_n(a_k), a_k) \geq \mathbf{P}M(\tilde{c}_n(\tilde{\gamma}_n), a_k) - o_{\mathbb{P}}(1)$,
which implies:

$$\begin{aligned} \mathbf{P}M(\tilde{c}_n(\tilde{\gamma}_n), a_k) - \mathbf{P}M(\tilde{c}_n(\tilde{\gamma}_n), \tilde{\gamma}_n) &\leq \mathbb{P}_n M(\tilde{c}_n(\tilde{\gamma}_n), \tilde{\gamma}_n) - \mathbf{P}M(\tilde{c}_n(\tilde{\gamma}_n), \tilde{\gamma}_n) + o_{\mathbb{P}}(1) \\ &\leq \sup_{a\in\Theta;\; b\in\Theta^{\Phi}} |\mathbb{P}_n M(b, a) - \mathbf{P}M(b, a)| \to 0 \; a.s.. \text{(B.5)} \end{aligned}$$

Based on the first part of this demonstration and on $(H5)$, we infer the existence of η such that: $\mathbf{P}(\|\tilde{\gamma}_n - a_k\| \geq \varepsilon) \leq \mathbf{P}(\mathbf{P}M(\tilde{c}_n(\tilde{\gamma}_n), a_k) - \mathbf{P}M(\tilde{c}_n(\tilde{\gamma}_n), \tilde{\gamma}_n)) \to 0$ $a.s.$ by $(B.5)$. This concludes our demonstration. □

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 5 (see page 13), that the densities $\hat{g}^{(k)}$ and $g^{(k)}$ are also strictly positive and bounded. Using lemma 2, (see page 3), and lemma 6, (see page 13), we demonstrate the theorem by induction.

Proof of theorem 2

row 3.1: 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:

$$|\Psi_j| \leq O_{\mathbf{P}}(m^{-1/2}). \tag{B.6}$$

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 B.6, 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 3.2: This demonstration holds for the two types of optimisation.
Since f and g are assumed to be strictly positive and bounded, hence lemma 5 (see page 13) implies $g^{(k)}$ is also, for all k, strictly positive and bounded.
Moreover, theorem 3.1 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:

$$\begin{aligned}\int |\hat{g}^{(k)}(x) - g^{(k)}(x)|dx &= \int g^{(k)}(x)|\tfrac{\hat{g}^{(k)}(x)}{g^{(k)}(x)} - 1|dx, \text{ because } g^{(k)} > 0\\ &\leq \int g^{(k)}(x)C(x)m^{-k/2}dx,\end{aligned}$$

$$\begin{aligned}\text{Moreover, } \sup_{x\in\mathbb{R}^d} |\hat{g}^{(k)}(x) - g^{(k)}(x)| &= \sup_{x\in\mathbb{R}^d} g^{(k)}(x)|\tfrac{\hat{g}^{(k)}(x)}{g^{(k)}(x)} - 1|\\ &= \sup_{x\in\mathbb{R}^d} g^{(k)}(x)C(x)m^{-k/2} \to 0 \ a.s., \text{ by theorem 1.}\end{aligned}$$

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

$$\begin{aligned}\Phi(\check{g}^{(k)}, f) - \Phi(g^{(k)}, f) &= \int f\varphi(\tfrac{\check{g}^{(k)}}{f})dx - \varphi(\tfrac{g^{(k)}}{f})dx = \int f\{\varphi(\tfrac{\check{g}^{(k)}}{f}) - \varphi(\tfrac{g^{(k)}}{f}))\}dx\\ &\leq \int f\ R|\tfrac{\check{g}^{(k)}}{f} - \tfrac{g^{(k)}}{f}|dx = R\int |\check{g}^{(k)} - g^{(k)}|dx\end{aligned}$$

with the line before last being derived from theorem 7. Since we get the same expression as the one we found in our Proof of Theorem 3.2 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

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

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

since $(H6)$ enables us to reverse the derivative and integral signs.
Moreover, the central limit theorem implies: $\mathbb{P}_n\frac{\partial}{\partial b}m(a_k,a_k) \overset{\mathcal{L}aw}{\rightarrow} \mathcal{N}_d(0,\mathbf{P}\|\frac{\partial}{\partial b}m(a_k,a_k)\|^2)$, $\mathbb{P}_n\frac{\partial}{\partial a}m(a_k,a_k) \overset{\mathcal{L}aw}{\rightarrow} \mathcal{N}_d(0,\mathbf{P}\|\frac{\partial}{\partial a}m(a_k,a_k)\|^2)$, 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.

Proof of theorem 4
Let us consider ψ and $\psi^{(k)}$ the characteristic functions of the densities f and $g^{(k-1)}$, then we have, $|\psi(t) - \psi^{(k)}(t)| \leq \int |f(x) - g^{(k)}(x)|dx \leq \Phi(g^{(k)},f) = min_a\Phi(g^{(k-1)}\frac{f_a}{[g^{(k-1)}]_a},f)$, therefore the assumption that $\lim_k min_a\Phi(g^{(k-1)}\frac{f_a}{[g^{(k-1)}]_a},f) = 0$ implies $\lim_k g^{(k)} = f$.

Proof of theorem 5
Through a Taylor development of $\mathbb{P}_nM(\check{c}_n(a_k),\check{\gamma}_n)$ of rank 2, we get at point (a_k,a_k):
$\mathbb{P}_nM(\check{c}_n(a_k),\check{\gamma}_n)$
$= \mathbb{P}_nM(a_k,a_k) + \mathbb{P}_n\frac{\partial}{\partial a}M(a_k,a_k)(\check{\gamma}_n - a_k)^\top + \mathbb{P}_n\frac{\partial}{\partial b}M(a_k,a_k)(\check{c}_n(a_k) - a_k)^\top$
$+\frac{1}{2}\{(\check{\gamma}_n - a_k)^\top\mathbb{P}_n\frac{\partial^2}{\partial a \partial a}M(a_k,a_k)(\check{\gamma}_n - a_k) + (\check{c}_n(a_k) - a_k)^\top\mathbb{P}_n\frac{\partial^2}{\partial b \partial a}M(a_k,a_k)(\check{\gamma}_n - a_k)$
$+(\check{\gamma}_n - a_k)^\top\mathbb{P}_n\frac{\partial^2}{\partial a \partial b}M(a_k,a_k)(\check{c}_n(a_k) - a_k) + (\check{c}_n(a_k) - a_k)^\top\mathbb{P}_n\frac{\partial^2}{\partial b \partial b}M(a_k,a_k)(\check{c}_n(a_k) - a_k)\}$
The lemma below enables us to conclude.

Lemma 11. *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 $\mathbb{P}_nM(\check{c}_n(a_k),\check{\gamma}_n) = \mathbb{P}_nM(a_k,a_k) + O_{\mathbf{P}}(\frac{1}{n})$,
ie $\sqrt{n}(\mathbb{P}_nM(\check{c}_n(a_k),\check{\gamma}_n) - \mathbf{P}M(a_k,a_k)) = \sqrt{n}(\mathbb{P}_nM(a_k,a_k) - \mathbf{P}M(a_k,a_k)) + o_{\mathbf{P}}(1)$. Hence $\sqrt{n}(\mathbb{P}_nM(\check{c}_n(a_k),\check{\gamma}_n) - \mathbf{P}M(a_k,a_k))$ abides by the same limit distribution as $\sqrt{n}(\mathbb{P}_nM(a_k,a_k) - \mathbf{P}M(a_k,a_k))$, 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 of corollary 2.
Let us demonstrate that under hypothesis $\lim_{k\to\infty}\Phi(g^{(k)},f) = 0$, we have

$$\lim_{n\to\infty}\lim_{k\to\infty}\Phi(g_n^{(k)},f) = 0.$$

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

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

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

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

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 $\Phi(g_{\infty}^{\infty}, f) > \Phi(g^{(k)}, f) \geq 0$, since Φ is lower semi continuous, we have $\lim_n inf \Phi(g_n^{\infty}, f) \geq \Phi(g_{\infty}^{\infty}, f)$ and $\lim_n inf \Phi(g_n^{(k)}, f) \geq \Phi(g_{\infty}^{(k)}, f)$. Consequently, $\Phi(g_n^{\infty}, f) \geq \Phi(g_{\infty}^{\infty}, f) > \Phi(g_n^{(k)}, f)$, which leads to the contradiction we were looking for. Hence $(**)$ is true. We can therefore conclude that $(**)$ implies $\Phi(g_{\infty}^{\infty}, f) = 0$, ie $\lim_{n\to\infty} \lim_{k\to\infty} \Phi(g_n^{(k)}, f) = 0$, as a reductio ad absurdum argument would have led to $0 < \Phi(g_{\infty}^{\infty}, f) \leq \Phi(g^{(k)}, f)$, which would have contradicted the hypothesis according to which $\lim_{k\to\infty} \Phi(g^{(k)}, f) = 0$.

References

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

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

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

[Deheuvels, 1974] 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.*

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

[Friedrich and Igor, 1987] 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.*

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