PROJECTION PURSUIT THROUGH Φ-DIVERGENCE MINIMISATION

Jacques Touboul

▶ **To cite this version:**

Jacques Touboul. PROJECTION PURSUIT THROUGH Φ-DIVERGENCE MINIMISATION. 2008. hal-00432242v3

HAL Id: hal-00432242
https://hal.science/hal-00432242v3

Preprint submitted on 22 Mar 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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 solve this problem. In his seminal article, Huber (see "Projection pursuit", Annals of Statistics, 1985) demonstrates the interest of his method in a very simple given case. He considers the factorization of density through a Gaussian component and some residual density. Huber's work is based on maximizing relative entropy. Our proposal leads to a new algorithm. Furthermore, we will also consider the case when the density to be factorized is estimated from an i.i.d. sample. We will then propose a test for the factorization of the estimated density. Applications include a new test of fit pertaining to the Elliptical copulas.

Key words: Projection Pursuit; minimum Φ-divergence; Elliptical distribution; goodness-of-fit; copula; regression.
2000 MSC: 94A17 62F05 62J05 62G08.

1. Outline of the article

The objective of Projection Pursuit is to generate one or several projections providing as much information as possible about the structure of the data set regardless of its size:

Once a structure has been isolated, the corresponding data are eliminated from the data set. Through a recursive approach, this process is iterated to find another structure in the remaining data, until no futher structure can be evidenced in the data left at the end.

Friedman (1984 and 1987) and Huber (1985) count among the first authors to have introduced this type of approaches for evidencing structures. They each describe, with many examples, how to evidence such a structure and consequently how to estimate the density of such data through two different methodologies each. Their work is based on maximizing relative entropy.
For a very long time, the two methodologies exposed by each of the above authors were thought to be equivalent but Mu Zhu (2004) showed it was in fact not the case when the number of iterations in the algorithms exceeds the dimension of the space containing the data. In the present article, we will therefore only focus on Huber's study while taking into account Mu Zhu remarks.

At present, let us briefly introduce Huber's methodology. We will then expose our approach and objective.

1.1. Huber's analytic approach

Let f be a density on $\mathbb{R}^d$. We define an instrumental density g with same mean and variance as f. Huber's methodology requires us to start with performing the $K(f,g) = 0$ test - with K being the relative entropy. Should this test turn out to be positive, then $f = g$ and the algorithm stops. If the test were not to be verified, the first step of Huber's algorithm amounts to defining a vector a_1 and a density $f^{(1)}$ by

$$a_1 = arg \inf_{a\in\mathbb{R}^d_*} K(f\frac{g_a}{f_a}, g) \text{ and } f^{(1)} = f\frac{g_{a_1}}{f_{a_1}}, \tag{1.1}$$

where $\mathbb{R}^d_*$ is the set of non null vectors of $\mathbb{R}^d$, where f_a (resp. g_a) stands for the density of $a^\top X$ (resp. $a^\top Y$) when f (resp. g) is the density of X (resp. Y). More exactly, this results from the maximisation of $a \mapsto K(f_a, g_a)$ since $K(f,g) = K(f_a, g_a) + K(f\frac{g_a}{f_a}, g)$ and it is assumed that $K(f,g)$ is finite. In a second step, Huber replaces f with $f^{(1)}$ and goes through the first step again.
By iterating this process, Huber thus obtains a sequence $(a_1, a_2, ...)$ of vectors of $\mathbb{R}^d_*$ and a sequence of densities $f^{(i)}$.

Remark **1.1.** *Huber stops his algorithm when the relative entropy equals zero or when his algorithm reaches the d^{th} iteration, he then obtains an approximation of f from g :*
When there exists an integer j such that $K(f^{(j)}, g) = 0$ with $j \leq d$, he obtains $f^{(j)} = g$, i.e. $f = g\Pi_{i=1}^{j} \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$ since by induction $f^{(j)} = f\Pi_{i=1}^{j} \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$. Similarly, when, for all j, Huber gets $K(f^{(j)}, g) > 0$ with $j \leq d$, he assumes $g = f^{(d)}$ in order to derive $f = g\Pi_{i=1}^{d} \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$.
He can also stop his algorithm when the relative entropy equals zero without the condition $j \leq d$ is met. Therefore, since by induction we have $f^{(j)} = f\Pi_{i=1}^{j} \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$ with $f^{(0)} = f$, we obtain $g = f\Pi_{i=1}^{j} \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$. Consequently, we derive a representation of f as $f = g\Pi_{i=1}^{j} \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$.
Finally, he obtains $K(f^{(0)}, g) \geq K(f^{(1)}, g) \geq ..... \geq 0$ with $f^{(0)} = f$.

1.2. Huber's synthetic approach

Keeping the notations of the above section, we start with performing the $K(f,g) = 0$ test; should this test turn out to be positive, then $f = g$ and the algorithm stops, otherwise, the first step of his algorithm would consist in defining a vector a_1 and a density $g^{(1)}$ by

$$a_1 = arg \inf_{a\in\mathbb{R}^d_*} K(f, g\frac{f_a}{g_a}) \text{ and } g^{(1)} = g\frac{f_{a_1}}{g_{a_1}}. \tag{1.2}$$

More exactly, this optimisation results from the maximisation of $a \mapsto K(f_a, g_a)$ since $K(f,g) = K(f_a, g_a) + K(f, g\frac{f_a}{g_a})$ and it is assumed that $K(f,g)$ is finite. In a second step, Huber replaces g with $g^{(1)}$ and goes through the first step again. By iterating this process, Huber thus obtains a sequence $(a_1, a_2, ...)$ of vectors of $\mathbb{R}^d_*$ and a sequence of densities $g^{(i)}$.

Remark **1.2.** *First, in a similar manner to the analytic approach, this methodology enables us to approximate and even to represent f from g:*
To obtain an approximation of f, Huber either stops his algorithm when the relative entropy equals zero, i.e. $K(f, g^{(j)}) = 0$ implies $g^{(j)} = f$ with $j \leq d$, or when his algorithm reaches the d^{th}

iteration, i.e. he approximates f with $g^{(d)}$.
To obtain a representation of f, Huber stops his algorithm when the relative entropy equals zero, since $K(f, g^{(j)}) = 0$ implies $g^{(j)} = f$. Therefore, since by induction we have $g^{(j)} = g\Pi_{i=1}^{j} \frac{f_{a_i}}{g_{a_i}^{(i-1)}}$ with $g^{(0)} = g$, we then obtain $f = g\Pi_{i=1}^{j} \frac{f_{a_i}}{g_{a_i}^{(i-1)}}$.
Second, he gets $K(f, g^{(0)}) \geq K(f, g^{(1)}) \geq ..... \geq 0$ with $g^{(0)} = g$.

1.3. Proposal

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 that $\varphi(1) = 0$. 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 - see also Annex A.1 page 18.
Now, let us introduce our algorithm. We start with performing the $\Phi(g, f) = 0$ test; should this test turn out to be positive, then $f = g$ and the algorithm stops, otherwise, the first step of our algorithm would consist in defining a vector a_1 and a density $g^{(1)}$ by

$$a_1 = arg \inf_{a \in \mathbb{R}_*^d} \Phi(g\frac{f_a}{g_a}, f) \text{ and } g^{(1)} = g\frac{f_{a_1}}{g_{a_1}}. \tag{1.3}$$

Later on, we will prove that a_1 simultaneously optimises (1.1), (1.2) and (1.3).
In our second step, we will replace g with $g^{(1)}$, and we will repeat the first step.
And so on, by iterating this process, we will end up obtaining a sequence $(a_1, a_2, ...)$ of vectors in $\mathbb{R}_*^d$ and a sequence of densities $g^{(i)}$. We will thus prove that the underlying structures of f evidenced through this method are identical to the ones obtained through the Huber's method. We will also evidence the above structures, which will enable us to infer more information on f - see example below.

Remark **1.3.** *As in the previous algorithm, we first provide an approximate and even a represention of f from g:*
To obtain an approximation of f, we stop our algorithm when the divergence equals zero, i.e. $\Phi(g^{(j)}, f) = 0$ implies $g^{(j)} = f$ with $j \leq d$, or when our algorithm reaches the d^{th} iteration, i.e. we approximate f with $g^{(d)}$.
To obtain a representation of f, we stop our algorithm when the divergence equals zero. Therefore, since by induction we have $g^{(j)} = g\Pi_{i=1}^{j} \frac{f_{a_i}}{g_{a_i}^{(i-1)}}$ with $g^{(0)} = g$, we then obtain $f = g\Pi_{i=1}^{j} \frac{f_{a_i}}{g_{a_i}^{(i-1)}}$.
Second, he gets $\Phi(g^{(0)}, f) \geq \Phi(g^{(1)}, f) \geq ..... \geq 0$ with $g^{(0)} = g$.
Finally, the specific form of relationship (1.3) establishes that we deal with M-estimation. We can therefore state that our method is more robust than Huber's - see Yohai (2008), Toma (2009) as well as Huber (2004).

At present, let us study two examples:

Example **1.1.** *Let f be a density defined on $\mathbb{R}^3$ by $f(x_1, x_2, x_3) = n(x_1, x_2)h(x_3)$, with n being a bi-dimensional Gaussian density, and h being a non Gaussian density. Let us also consider g, a Gaussian density with same mean and variance as f.*
Since $g(x_1, x_2/x_3) = n(x_1, x_2)$, we then have $\Phi(g\frac{f_3}{g_3}, f) = \Phi(n.f_3, f) = \Phi(f, f) = 0$ as $f_3 = h$, i.e. the function $a \mapsto \Phi(g\frac{f_a}{g_a}, f)$ reaches zero for $e_3 = (0, 0, 1)'$.
We therefore obtain $g(x_1, x_2/x_3) = f(x_1, x_2/x_3)$.

Example **1.2.** *Assuming that the Φ-divergence is greater than the L^2 norm. Let us consider $(X_n)_{n\geq 0}$, the Markov chain with continuous state space E. Let f be the density of (X_0, X_1) and let g be the normal density with same mean and variance as f.*
Let us now assume that $\Phi(g^{(1)}, f) = 0$ with $g^{(1)}(x) = g(x)\frac{f_1}{g_1}$, i.e. let us assume that our algorithm stops for $a_1 = (1, 0)'$. Consequently, if (Y_0, Y_1) is a random vector with g density, then the distribution law of X_1 given X_0 is Gaussian and is equal to the distribution law of Y_1 given Y_0.
And then, for any sequence (A_i) - where $A_i \subset E$ - we have

$$\begin{aligned}
&\mathbf{P}\Big(X_{n+1} \in A_{n+1} \mid X_0 \in A_0, X_1 \in A_1, \ldots, X_{n-1} \in A_{n-1}, X_n \in A_n\Big) \\
&\quad = \mathbf{P}\left(X_{n+1} \in A_{n+1} \mid X_n \in A_n\right), \textit{ based on the very definition of a Markov chain,} \\
&\quad = \mathbf{P}\left(X_1 \in A_1 \mid X_0 \in A_0\right), \textit{ through the Markov property,} \\
&\quad = \mathbf{P}\left(Y_1 \in A_1 \mid Y_0 \in A_0\right), \textit{ as a consequence of the above nullity of the } \Phi\textit{-divergence.}
\end{aligned}$$

To recapitulate our method, if $\Phi(g, f) = 0$, we derive f from the relationship $f = g$; should a sequence $(a_i)_{i=1,\ldots j}$, $j < d$, of vectors in $\mathbb{R}^d_*$ defining $g^{(j)}$ and such that $\Phi(g^{(j)}, f) = 0$ exist, then $f(./a_i^\top x, 1 \leq i \leq j) = g(./a_i^\top x, 1 \leq i \leq j)$, i.e. f coincides with g on the complement of the vector subspace generated by the family $\{a_i\}_{i=1,\ldots,j}$ - see also section 2 for a more detailed explanation.

In this paper, after having clarified the choice of g, we will consider the statistical solution to the representation problem, assuming that f is unknown and X_1, X_2,... X_m are i.i.d. with density f. We will provide asymptotic results pertaining to the family of optimizing vectors $a_{k,m}$ - that we will define more precisely below - as m goes to infinity. Our results also prove that the empirical representation scheme converges towards the theoretical one. As an application, section 3.4 permits a new test of fit pertaining to the copula of an unknown density f, section 3.5 gives us an estimate of a density deconvoluted with a Gaussian component and section 3.6 presents some applications to the regression analysis. Finally, we will present simulations.

2. The algorithm

2.1. The model

As explained by Friedman (1984 and 1987) and Diaconis (1984), the choice of g depends on the family of distribution one wants to find in f. Until now, the choice has only been to use the class of Gaussian distributions. This can be extended to the class of elliptic distributions with almost all $\Phi-$divergences.

2.1.1. Elliptical laws

The interest of this class lies in the fact that conditional densities with elliptical distributions are also elliptical - see Cambanis (1981), Landsman (2003). This very property allows us to use this class in our algorithm.

Definition 2.1. *X is said to abide by a multivariate elliptical distribution - noted $X \sim E_d(\mu, \Sigma, \xi_d)$ - if X presents the following density, for any x in $\mathbb{R}^d$:*

$$f_X(x) = \frac{c_d}{|\Sigma|^{1/2}}\xi_d\Big(\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\Big)$$

- *with Σ, being a $d \times d$ positive-definite matrix and with μ, being an d-column vector,*
- *with ξ_d, being referred as the "density generator",*
- *with c_d, being a normalisation constant, such that $c_d = \frac{\Gamma(d/2)}{(2\pi)^{d/2}}\Big(\int_0^\infty x^{d/2-1}\xi_d(x)dx\Big)^{-1}$,*

with $\int_0^\infty x^{d/2-1}\xi_d(x)dx < \infty$.

Property 2.1. *1/ For any $X \sim E_d(\mu, \Sigma, \xi_d)$, for any A, being a $m \times d$ matrix with rank $m \leq d$, and for any b, being an m-dimensional vector, we have $AX + b \sim E_m(A\mu + b, A\Sigma A', \xi_m)$.*
Therefore, any marginal density of multivarite elliptical distribution is elliptic, i.e.
$X = (X_1, X_2, ..., X_d) \sim E_d(\mu, \Sigma, \xi_d) \Rightarrow X_i \sim E_1(\mu_i, \sigma_i^2, \xi_1)$, $f_{X_i}(x) = \frac{c_1}{\sigma_i}\xi_1\Big(\frac{1}{2}(\frac{x-\mu_i}{\sigma})^2\Big)$, $1 \leq i \leq d$.
2/ Corollary 5 of Cambanis (1981) states that conditional densities with elliptical distributions are also elliptic. Indeed, if $X = (X_1, X_2)' \sim E_d(\mu, \Sigma, \xi_d)$, with X_1 (resp. X_2) being a size $d_1 < d$ (resp. $d_2 < d$), then $X_1/(X_2 = a) \sim E_{d_1}(\mu', \Sigma', \xi_{d_1})$ with $\mu' = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$ and $\Sigma' = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, with $\mu = (\mu_1, \mu_2)$ and $\Sigma = (\Sigma_{ij})_{1\leq i,j\leq 2}$.

Remark **2.1.** *Landsman (2003) shows that multivariate Gaussian distributions derive from $\xi_d(x) = e^{-x}$. He also shows that if $X = (X_1, ..., X_d)$ has an elliptical density such that its marginals verify $E(X_i) < \infty$ and $E(X_i^2) < \infty$ for $1 \leq i \leq d$, then μ is the mean of X and Σ is the covariance matrix of X. Consequently, from now on, we will assume that we are in this case.*

Definition 2.2. *Let t be an elliptical density on $\mathbb{R}^k$ and let q be an elliptical density on $\mathbb{R}^{k'}$. The elliptical densities t and q are said to belong to the same family - or class - of elliptical densities, if their generating densities are ξ_k and $\xi_{k'}$ respectively, which belong to a common given family of densities.*

Example **2.1.** *Consider two Gaussian densities $\mathcal{N}(0, 1)$ and $\mathcal{N}((0, 0), Id_2)$. They are said to belong to the same elliptical families as they both present $x \mapsto e^{-x}$ as generating density.*

2.1.2. Choice of g

Let us begin with studying the following case:
Let f be a density on $\mathbb{R}^d$. Let us assume there exists d not null independent vectors a_j, with $1 \leq j \leq d$, of $\mathbb{R}^d$, such that

$$f(x) = n(a_{j+1}^\top x, ..., a_d^\top x)h(a_1^\top x, ..., a_j^\top x), \tag{2.1}$$

with $j < d$, with n being an elliptical density on $\mathbb{R}^{d-j-1}$ and with h being a density on $\mathbb{R}^j$, which does not belong to the same family as n. Let $X = (X_1, ..., X_d)$ be a vector presenting f as density. Define g as an Elliptical distribution with same mean and variance as f.
For simplicity, let us assume that the family $\{a_j\}_{1\leq j\leq d}$ is the canonical basis of $\mathbb{R}^d$:
The very definition of f implies that $(X_{j+1}, ..., X_d)$ is independent from $(X_1, ..., X_j)$. Hence, the density of $(X_{j+1}, ..., X_d)$ given $(X_1, ..., X_j)$ is n.
Let us assume that $\Phi(g^{(j)}, f) = 0$, for some $j \leq d$. We then get $\frac{f(x)}{f_{a_1} f_{a_2} ... f_{a_j}} = \frac{g(x)}{g_{a_1}^{(1-1)} g_{a_2}^{(2-1)} ... g_{a_j}^{(j-1)}}$, since, by induction, we have $g^{(j)}(x) = g(x)\frac{f_{a_1}}{g_{a_1}^{(1-1)}}\frac{f_{a_2}}{g_{a_2}^{(2-1)}}...\frac{f_{a_j}}{g_{a_j}^{(j-1)}}$.
Consequently, the fact that conditional densities with elliptical distributions are also elliptical as well as the above relationship enable us to infer that

$$n(a_{j+1}^\top x, ., a_d^\top x) = f(./a_i^\top x, 1 \leq i \leq j) = g(./a_i^\top x, 1 \leq i \leq j).$$

In other words, f coincides with g on the complement of the vector subspace generated by the family $\{a_i\}_{i=1,...,j}$.

Now, if the family $\{a_j\}_{1\leq j\leq d}$ is no longer the canonical basis of $\mathbb{R}^d$, then this family is again a basis of $\mathbb{R}^d$. Hence, lemma F.1 - page 24 - implies that

$$g(./a_1^\top x, ..., a_j^\top x) = n(a_{j+1}^\top x, ..., a_d^\top x) = f(./a_1^\top x, ..., a_j^\top x), \tag{2.2}$$

which is equivalent to having $\Phi(g^{(j)}, f) = 0$ - since by induction $g^{(j)} = g\frac{f_{a_1}}{g_{a_1}^{(1-1)}}\frac{f_{a_2}}{g_{a_2}^{(2-1)}}...\frac{f_{a_j}}{g_{a_j}^{(j-1)}}$.
The end of our algorithm implies that f coincides with g on the complement of the vector subspace generated by the family $\{a_i\}_{i=1,...,j}$. Therefore, the nullity of the $\Phi-$divergence provides us with information on the density structure.
In summary, the following proposition clarifies our choice of g which depends on the family of distribution one wants to find in f :

Proposition 2.1. *With the above notations,* $\Phi(g^{(j)}, f) = 0$ *is equivalent to*

$$g(./a_1^\top x, ..., a_j^\top x) = f(./a_1^\top x, ..., a_j^\top x)$$

More generally, the above proposition leads us to defining the co-support of f as the vector space generated from vectors $a_1, ..., a_j$.

Definition 2.3. *Let f be a density on $\mathbb{R}^d$. We define the co-vectors of f as the sequence of vectors $a_1, ..., a_j$ which solves the problem $\Phi(g^{(j)}, f) = 0$ where g is an Elliptical distribution with same mean and variance as f. We define the co-support of f as the vector space generated from vectors $a_1, ..., a_j$.*

Remark **2.2.** *Any (a_i) family defining f as in (2.1), is an orthogonal basis of $\mathbb{R}^d$ - see lemma F.2*

2.2. Stochastic outline of our algorithm

Let X_1, X_2,..,X_m (resp. Y_1, Y_2,..,Y_m) be a sequence of m independent random vectors with same density f (resp. g). As customary in nonparametric $\Phi-$divergence optimizations, all estimates of f and f_a as well as all uses of Monté Carlo's methods are being performed using subsamples X_1, X_2,..,X_n and Y_1, Y_2,..,Y_n - extracted respectively from X_1, X_2,..,X_m and Y_1, Y_2,..,Y_m - since the estimates are bounded below by some positive deterministic sequence θ_m - see Annex B.
Let $\mathbb{P}_n$ be the empirical measure of the subsample X_1, X_2,.,X_n. Let f_n (resp. $f_{a,n}$ for any a in $\mathbb{R}_*^d$) be the kernel estimate of f (resp. f_a), which is built from X_1, X_2,..,X_n (resp. $a^\top X_1$, $a^\top X_2$,..,$a^\top X_n$). As defined in section 1.3, we introduce the following sequences $(a_k)_{k\geq 1}$ and $(g^{(k)})_{k\geq 1}$:

- a_k is a non null vector of $\mathbb{R}^d$ such that $a_k = arg\min_{a\in\mathbb{R}_*^d}\Phi(g^{(k-1)}\frac{f_a}{g_a^{(k-1)}}, f)$, (2.3)
- $g^{(k)}$ is the density such that $g^{(k)} = g^{(k-1)}\frac{f_{a_k}}{g_{a_k}^{(k-1)}}$ with $g^{(0)} = g$.

The stochastic setting up of the algorithm uses f_n and $g_n^{(0)} = g$ instead of f and $g^{(0)} = g$ - since g is known. Thus, at the first step, we build the vector $\check{a}_1$ which minimizes the $\Phi-$divergence between f_n and $g\frac{f_{a,n}}{g_a}$ and which estimates a_1 :
Proposition B.1 page 20 and lemma F.6 page 25 enable us to minimize the $\Phi-$divergence between f_n and $g\frac{f_{a,n}}{g_a}$. Defining $\check{a}_1$ as the argument of this minimization, proposition 3.3 page 8

shows us that this vector tends to a_1.
Finally, we define the density $\check{g}_m^{(1)}$ as $\check{g}_m^{(1)} = g\frac{f_{\check{a}_1,m}}{g_{\check{a}_1}}$ which estimates $g^{(1)}$ through theorem 3.1.

Now, from the second step and as defined in section 1.3, the density $g^{(k-1)}$ is unknown. Consequently, once again, we have to truncate the samples:
All estimates of f and f_a (resp. $g^{(1)}$ and $g_a^{(1)}$) are being performed using a subsample X_1, X_2,..,X_n (resp. $Y_1^{(1)}$, $Y_2^{(1)}$,..,$Y_n^{(1)}$) extracted from X_1, X_2,..,X_m (resp. $Y_1^{(1)}$, $Y_2^{(1)}$,..,$Y_m^{(1)}$ - which is a sequence of m independent random vectors with same density $g^{(1)}$) such that the estimates are bounded below by some positive deterministic sequence θ_m - see Annex B.
Let $\mathbb{P}_n$ be the empirical measure of the subsample X_1, X_2,..,X_n. Let f_n (resp. $g_n^{(1)}$, $f_{a,n}$, $g_{a,n}^{(1)}$ for any a in $\mathbb{R}_*^d$) be the kernel estimate of f (resp. $g^{(1)}$ and f_a as well as $g_a^{(1)}$) which is built from X_1, X_2,..,X_n (resp. $Y_1^{(1)}$, $Y_2^{(1)}$,..,$Y_n^{(1)}$ and $a^\top X_1$, $a^\top X_2$,..,$a^\top X_n$ as well as $a^\top Y_1^{(1)}$, $a^\top Y_2^{(1)}$,..,$a^\top Y_n^{(1)}$). The stochastic setting up of the algorithm uses f_n and $g_n^{(1)}$ instead of f and $g^{(1)}$. Thus, we build the vector $\check{a}_2$ which minimizes the $\Phi-$divergence between f_n and $g_n^{(1)}\frac{f_{a,n}}{g_{a,n}^{(1)}}$ - since $g^{(1)}$ and $g_a^{(1)}$ are unknown - and which estimates a_2. Proposition B.1 page 20 and lemma F.6 page 25 enable us to minimize the $\Phi-$divergence between f_n and $g_n^{(1)}\frac{f_{a,n}}{g_{a,n}^{(1)}}$. Defining $\check{a}_2$ as the argument of this minimization, proposition 3.3 page 8 shows us that this vector tends to a_2 in n. Finally, we define the density $\check{g}_n^{(2)}$ as $\check{g}_n^{(2)} = g_n^{(1)}\frac{f_{\check{a}_2,n}}{g_{\check{a}_2,n}^{(1)}}$ which estimates $g^{(2)}$ through theorem 3.1.

And so on, we will end up obtaining a sequence $(\check{a}_1, \check{a}_2, ...)$ of vectors in $\mathbb{R}_*^d$ estimating the covectors of f and a sequence of densities $(\check{g}_n^{(k)})_k$ such that $\check{g}_n^{(k)}$ estimates $g^{(k)}$ through theorem 3.1.

3. Results

3.1. Convergence results

3.1.1. Hypotheses on f

In this paragraph, we define the set of hypotheses on f which could possibly be of use in our work. Discussion on several of these hypotheses can be found in Annex E.
In this section, to be more legible we replace g with $g^{(k-1)}$. Let

$$\Theta = \mathbb{R}^d,\ \Theta^\Phi = \{b \in \Theta \mid \int \varphi^*(\varphi'(\tfrac{g(x)}{f(x)}\tfrac{f_b(b^\top x)}{g_b(b^\top x)}))d\mathbf{P} < \infty\},$$
$$M(b,a,x) = \int \varphi'(\tfrac{g(x)}{f(x)}\tfrac{f_b(b^\top x)}{g_b(b^\top x)})g(x)\tfrac{f_a(a^\top x)}{g_a(a^\top x)}dx - \ \varphi^*(\varphi'(\tfrac{g(x)}{f(x)}\tfrac{f_b(b^\top x)}{g_b(b^\top x)})),$$
$$\mathbb{P}_n M(b,a) = \int M(b,a,x)d\mathbb{P}_n,\ \mathbf{P}M(b,a) = \int M(b,a,x)d\mathbf{P},$$

where $\mathbf{P}$ is the probability measure presenting f as density.
Similarly as in chapter V of Van der Vaart (1998), let us define :
$(H1)$: 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 $a \in \Theta$.
$(H2)$: $\exists\, Z < 0, n_0 > 0$ such that $(n \geq n_0 \ \Rightarrow\ \sup_{a\in\Theta}\sup_{c\in\{\Theta^\Phi\}^c}\mathbb{P}_n M(c,a) < Z)$
$(H3)$: There is a neighbourhood V of a_k, 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 $\mathbf{P}H < \infty$,
$(H4)$: 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 $\mathbf{P}M(c,a_k) < \mathbf{P}M(c,a) - \eta$.

Putting $I_{a_k} = \frac{\partial^2}{\partial a^2}\Phi(g\frac{f_{a_k}}{g_{a_k}}, f)$, and $x \to \rho(b,a,x) = \varphi'(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)})\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)}$, let us now consider three new hypotheses:
$(H5)$: The function φ is C^3 in $(0,+\infty)$ and there is a neighbourhood V_k' of (a_k,a_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 derivatives $\frac{g(x)f_a(a^\top x)}{g_a(a^\top x)}$ and the first and second order derivatives of $(b,a) \mapsto \rho(b,a,x)$ are dominated (λ_a.s.) by λ-integrable functions.

$(H6)$: The function $(b,a) \mapsto M(b,a)$ is 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 $H(x)$.

$(H7)$: $\mathbf{P}\|\frac{\partial}{\partial b}M(a_k,a_k)\|^2$ 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.

Finally, we define

$(H8)$: There exists k such that $\mathbf{P}M(a_k,a_k) = 0$.

$(H9)$: $(Var_{\mathbf{P}}(M(a_k,a_k)))^{1/2}$ exists and is invertible.

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

3.1.2. Estimation of the first co-vector of f

Let $\mathcal{R}$ be the class of all positive functions r defined on $\mathbb{R}$ and such that $g(x)r(a^\top x)$ is a density on $\mathbb{R}^d$ for all a belonging to $\mathbb{R}^d_*$. The following proposition shows that there exists a vector a such that $\frac{f_a}{g_a}$ minimizes $\Phi(gr, f)$ in r:

Proposition 3.1. *There exists a vector a belonging to $\mathbb{R}^d_*$ such that*

$$arg\min_{r\in\mathcal{R}} \Phi(gr, f) = \frac{f_a}{g_a} \text{ and } r(a^\top x) = \frac{f_a(a^\top x)}{g_a(a^\top x)}.$$

Remark **3.1.** *This proposition proves that a_1 simultaneously optimises (1.1), (1.2) and (1.3). In other words, it proves that the underlying structures of f evidenced through our method are identical to the ones obtained through Huber's methods - see also Annex D.*

Following Broniatowski (2009), let us introduce the estimate of $\Phi(g\frac{f_{a,n}}{g_a}, f_n)$, through

$$\check{\Phi}(g\frac{f_{a,n}}{g_a}, f_n) = \int M(a,a,x)d\mathbb{P}_n(x)$$

Proposition 3.2. *Let $\check{a}$ be such that $\check{a} := arg\inf_{a\in\mathbb{R}^d_*} \check{\Phi}(g\frac{f_{a,n}}{g_a}, f_n)$.*
Then, $\check{a}$ is a strongly convergent estimate of a, as defined in proposition 3.1.

Let us also introduce the following sequences $(\check{a}_k)_{k\geq 1}$ and $(\check{g}_n^{(k)})_{k\geq 1}$, for any given n - see section 2.2.:

- $\check{a}_k$ is an estimate of a_k as defined in proposition 3.2 with $\check{g}_n^{(k-1)}$ instead of g,
- $\check{g}_n^{(k)}$ is such that $\check{g}_n^{(0)} = g$, $\check{g}_n^{(k)}(x) = \check{g}_n^{(k-1)}(x)\frac{f_{\check{a}_k,n}(\check{a}_k^\top x)}{[\check{g}^{(k-1)}]_{\check{a}_k,n}(\check{a}_k^\top x)}$, i.e. $\check{g}_n^{(k)}(x) = g(x)\Pi_{j=1}^k \frac{f_{\check{a}_j,n}(\check{a}_j^\top x)}{[\check{g}^{(j-1)}]_{\check{a}_j,n}(\check{a}_j^\top x)}$.

We also note that $\check{g}_n^{(k)}$ is a density.

3.1.3. Convergence study at the k^{th} step of the algorithm:

In this paragraph, we will show that the sequence $(\check{a}_k)_n$ converges towards a_k and that the sequence $(\check{g}_n^{(k)})_n$ converges towards $g^{(k)}$.
Let $\check{c}_n(a) = arg\sup_{c\in\Theta} \mathbb{P}_n M(c,a)$, with $a \in \Theta$, and $\check{\gamma}_n = arg\inf_{a\in\Theta} \sup_{c\in\Theta} \mathbb{P}_n M(c,a)$. We state

Proposition 3.3. *Both $\sup_{a\in\Theta}\|\check{c}_n(a) - a_k\|$ and $\check{\gamma}_n$ converge toward a_k a.s.*

Finally, the following theorem shows that $\check{g}_n^{(k)}$ converges almost everywhere towards $g^{(k)}$:

Theorem 3.1. *It holds $\check{g}_n^{(k)} \to_n g^{(k)}$ a.s.*

3.2. Asymptotic Inference at the k^{th} step of the algorithm

The following theorem shows that $\check{g}_n^{(k)}$ converges towards $g^{(k)}$ at the rate $O_{\mathbf{P}}(n^{-\frac{2}{2+d}})$ in three differents cases, namely for any given x, with the L^1 distance and with the relative entropy:

Theorem 3.2. *It holds* $|\check{g}_n^{(k)}(x) - g^{(k)}(x)| = O_{\mathbf{P}}(n^{-\frac{2}{2+d}})$, $\int |\check{g}_n^{(k)}(x) - g^{(k)}(x)|dx = O_{\mathbf{P}}(n^{-\frac{2}{2+d}})$ *and* $|K(\check{g}_n^{(k)}, f) - K(g^{(k)}, f)| = O_{\mathbf{P}}(n^{-\frac{2}{2+d}})$.

Remark **3.2.** *With the relative entropy, we have $n = O(m^{1/2})$ - see lemma F.13. The above rates consequently become $O_{\mathbf{P}}(m^{-\frac{1}{2+d}})$.*

The following theorem shows that the laws of our estimators of a_k, namely $\check{c}_n(a_k)$ and $\check{\gamma}_n$, converge towards a linear combination of Gaussian variables.

Theorem 3.3. *It holds*
$\sqrt{n}\mathcal{A}.(\check{c}_n(a_k) - a_k) \overset{\mathcal{L}aw}{\to} \mathcal{B}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2) + C.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2)$ *and*
$\sqrt{n}\mathcal{A}.(\check{\gamma}_n - a_k) \overset{\mathcal{L}aw}{\to} C.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k, a_k)\|^2) + C.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k, a_k)\|^2)$
where $\mathcal{A} = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k)(\mathbf{P}\frac{\partial^2}{\partial a_i \partial a_j}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i \partial b_j}M(a_k, a_k))$, $C = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k)$ *and*
$\mathcal{B} = \mathbf{P}\frac{\partial^2}{\partial b \partial b}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i \partial a_j}M(a_k, a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i \partial b_j}M(a_k, a_k)$.

3.3. A stopping rule for the procedure

In this paragraph, we will call $\check{g}_n^{(k)}$ (resp. $\check{g}_{a,n}^{(k)}$) the kernel estimator of $\check{g}^{(k)}$ (resp. $\check{g}_a^{(k)}$). We will first show that $g_n^{(k)}$ converges towards f in k and n. Then, we will provide a stopping rule for this identification procedure.

3.3.1. Estimation of f

The following proposition provides us with an estimate of f:

Theorem 3.4. *We have* $\lim_n \lim_k \check{g}_n^{(k)} = f$ *a.s.*

Consequently, the following corollary shows that $\Phi(g_n^{(k-1)}\frac{f_{a_k,n}}{g_{a_k,n}^{(k-1)}}, f_{a_k,n})$ converges towards zero as k and then as n go to infinity:

Corollary 3.1. *We have* $\lim_n \lim_k \Phi(\check{g}_n^{(k)}\frac{f_{a_k,n}}{[\check{g}^{(k)}]_{a_k,n}}, f_n) = 0$ *a.s.*

3.3.2. Testing of the criteria

In this paragraph, through a test of our criteria, namely $a \mapsto \Phi(\check{g}_n^{(k)}\frac{f_{a,n}}{[\check{g}^{(k)}]_{a,n}}, f_n)$, we will build a stopping rule for this procedure.
First, the next theorem enables us to derive the law of our criteria:

Theorem 3.5. *For a fixed k, we have*
$\sqrt{n}(Var_{\mathbf{P}}(M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)))^{-1/2}(\mathbb{P}_n M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n) - \mathbb{P}_n M(a_k, a_k)) \overset{\mathcal{L}aw}{\to} \mathcal{N}(0, I)$,
where k represents the k^{th} step of our algorithm and where I is the identity matrix in $\mathbb{R}^d$.

Note that k is fixed in theorem 3.5 since $\check{\gamma}_n = arg\inf_{a\in\Theta}\sup_{c\in\Theta}\mathbb{P}_n M(c, a)$ where M is a known function of k - see section 3.1.1. Thus, in the case when $\Phi(g^{(k-1)}\frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) = 0$, we obtain

Corollary 3.2. *We have* $\sqrt{n}(Var_{\mathbf{P}}(M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)))^{-1/2}\mathbb{P}_n M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n) \overset{\mathcal{L}aw}{\to} \mathcal{N}(0, I)$.

Hence, we propose the test of the null hypothesis

$$(H_0) : \Phi(g^{(k-1)}\frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) = 0 \text{ versus the alternative } (H_1) : \Phi(g^{(k-1)}\frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) \neq 0.$$

Based on this result, we stop the algorithm, then, defining a_k as the last vector generated, we derive from corollary 3.2 a α-level confidence ellipsoid around a_k, namely

$$\mathcal{E}_k = \{b \in \mathbb{R}^d;\ \sqrt{n}(Var_{\mathbf{P}}(M(b,b)))^{-1/2}\mathbb{P}_n M(b,b) \leq q_\alpha^{\mathcal{N}(0,1)}\}$$

where $q_\alpha^{\mathcal{N}(0,1)}$ is the quantile of a α-level reduced centered normal distribution and where $\mathbb{P}_n$ is the empirical measure araising from a realization of the sequences $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$. Consequently, the following corollary provides us with a confidence region for the above test:

Corollary 3.3. *$\mathcal{E}_k$ is a confidence region for the test of the null hypothesis (H_0) versus (H_1).*

3.4. Goodness-of-fit test for copulas

Let us begin with studying the following case:
Let f be a density defined on $\mathbb{R}^2$ and let g be an Elliptical distribution with same mean and variance as f. Assuming first that our algorithm leads us to having $\Phi(g^{(2)}, f) = 0$ where family (a_i) is the canonical basis of $\mathbb{R}^2$. Hence, we have $g^{(2)}(x) = g(x)\frac{f_1}{g_1}\frac{f_2}{g_2^{(1)}} = g(x)\frac{f_1}{g_1}\frac{f_2}{g_2}$ - through lemma F.7 page 26 - and $g^{(2)} = f$. Therefore, $f = g(x)\frac{f_1}{g_1}\frac{f_2}{g_2}$, i.e. $\frac{f}{f_1 f_2} = \frac{g}{g_1 g_2}$, and then

$$\frac{\partial^2}{\partial x \partial y} C_f = \frac{\partial^2}{\partial x \partial y} C_g$$

where C_f (resp. C_g) is the copula of f (resp. g).
At present, let f be a density on $\mathbb{R}^d$ and let g be the density defined in section 2.1.2.
Let us assume that our algorithm implies that $\Phi(g^{(d)}, f) = 0$.
Hence, we have, for any $x \in \mathbb{R}^d$, $g(x)\Pi_{k=1}^d \frac{f_{a_k}(a_k^\top x)}{[g^{(k-1)}]_{a_k}(a_k^\top x)} = f(x)$, i.e. $\frac{g(x)}{\Pi_{k=1}^d g_{a_k}(a_k^\top x)} = \frac{f(x)}{\Pi_{k=1}^d f_{a_k}(a_k^\top x)}$, since lemma F.7 page 26 implies that $g_{a_k}^{(k-1)} = g_{a_k}$ if $k \leq d$.
Moreover, the family $(a_i)_{i=1\ldots d}$ is a basis of $\mathbb{R}^d$ - see lemma F.8 page 26. Hence, putting $A = (a_1, ..., a_d)$ and defining vector y (resp. density $\tilde{f}$, copula $\tilde{C}_f$ of $\tilde{f}$, density $\tilde{g}$, copula $\tilde{C}_g$ of $\tilde{g}$) as the expression of vector x (resp. density f, copula C_f of f, density g, copula C_g of g) in basis A, the above equality implies

$$\frac{\partial^d}{\partial y_1 \ldots \partial y_d} \tilde{C}_f = \frac{\partial^d}{\partial y_1 \ldots \partial y_d} \tilde{C}_g.$$

Finally, we perform a statistical test of the null hypothesis $(H_0) : \frac{\partial^d}{\partial y_1 \ldots \partial y_d}\tilde{C}_f = \frac{\partial^d}{\partial y_1 \ldots \partial y_d}\tilde{C}_g$ versus the alternative $(H_1) : \frac{\partial^d}{\partial y_1 \ldots \partial y_d}\tilde{C}_f \neq \frac{\partial^d}{\partial y_1 \ldots \partial y_d}\tilde{C}_g$. Since, under (H_0), we have $\Phi(g^{(d)}, f) = 0$, then, as explained in section 3.3.2, corollary 3.3 provides us with a confidence region for our test.

Theorem 3.6. *Keeping the notations of corollary 3.3, we infer that $\mathcal{E}_d$ is a confidence region for the test of the null hypothesis (H_0) versus the alternative hypothesis (H_1).*

3.5. Rewriting of the convolution product

In the present paper, we first elaborated an algorithm aiming at isolating several known structures from initial datas. Our objective was to verify if for a known density on $\mathbb{R}^d$, a known density n on $\mathbb{R}^{d-j-1}$ such that, for $d > 1$,

$$f(x) = n(a_{j+1}^\top x, ..., a_d^\top x)h(a_1^\top x, ..., a_j^\top x), \tag{3.1}$$

did indeed exist, with $j < d$, with $(a_1, \ldots, a_d)$ being a basis of $\mathbb{R}^d$ and with h being a density on $\mathbb{R}^j$.

Secondly, our next step consisted in building an estimate (resp. a representation) of f without necessarily assuming that f meets relationship (3.1) - see theorem 3.4.

Consequently, let us consider Z_1 and Z_2, two random vectors with respective densities h_1 and h_2 - which is Elliptical - on $\mathbb{R}^d$. Let us consider a random vector X such that $X = Z_1 + Z_2$ and let f be its density. This density can then be written as :

$$f(x) = h_1 * h_2(x) = \int_{\mathbb{R}^d} h_1(x)h_2(t - x)dt.$$

Then, the following property enables us to represent f under the form of a product and without the integral sign

Proposition 3.4. *Let ϕ be a centered Elliptical density with $\sigma^2.I_d$, $\sigma^2 > 0$, as covariance matrix, such that it is a product density in all orthogonal coordinate systems and such that its characteristic function $s \mapsto \Psi(\frac{1}{2}|s|^2\sigma^2)$ is integrable - see Landsman (2003).*
Let f be a density on $\mathbb{R}^d$ which can be deconvoluted with ϕ, i.e.

$$f = \overline{f} * \phi = \int_{\mathbb{R}^d} \overline{f}(x)\phi(t - x)dt,$$

where $\overline{f}$ is some density on $\mathbb{R}^d$.
Let $g^{(0)}$ be the Elliptical density belonging to the same Elliptical family as f and having same mean and variance as f.
Then, the sequence $(g^{(k)})_k$ converges uniformly a.s. and in L^1 towards f in k, i.e.

$$\lim_{k\to\infty} \sup_{x\in\mathbb{R}^d} |g^{(k)}(x) - f(x)| = 0, \text{ and } \lim_{k\to\infty} \int_{\mathbb{R}^d} |g^{(k)}(x) - f(x)|dx = 0.$$

Finally, with the notations of section 3.3 and of proposition 3.4, the following theorem enables us to estimate any convolution product of a multivariate Elliptical density ϕ with a continuous density $\overline{f}$:

Theorem 3.7. *It holds* $\lim_n \lim_k \check{g}_n^{(k)} = \overline{f} * \phi$ *a.s.*

3.6. On the regression

In this section, we will study several applications of our algorithm pertaining to the regression analysis. We define $(X_1, ..., X_d)$ (resp. $(Y_1, ..., Y_d)$) as a vector with density f (resp. g - see section 2.1.2).

Remark **3.3.** *In this paragraph, we will work in the L^2 space. Then, we will first only consider the $\Phi-$divergences which are greater than or equal to the L^2 distance - see Vajda (1973). Note also that the co-vectors of f can be obtained in the L^2 space - see lemma F.6 and proposition B.1.*

3.6.1. The basic idea

In this paragraph, we will assume that $\Theta = \mathbb{R}_*^2$ and that our algorithm stops for $j = 1$ and $a_1 = (0, 1)'$. The following theorem provides us with the regression of X_1 on X_2 :

Theorem 3.8. *The probability measure of X_1 given X_2 is the same as the probability measure of Y_1 given Y_2. Moreover, the regression between X_1 and X_2 is*

$$X_1 = E(Y_1/Y_2) + \varepsilon,$$

where ε is a centered random variable orthogonal to $E(X_1/X_2)$.

Remark **3.4.** *This theorem implies that $E(X_1/X_2) = E(Y_1/Y_2)$. This equation can be used in many fields of research. The Markov chain theory has been used for instance in example 1.2. Moreover, if g is a Gaussian density with same mean and variance as f, then Saporta (2006) implies that $E(Y_1/Y_2) = E(Y_1) + \frac{Cov(Y_1,Y_2)}{Var(Y_2)}(Y_2 - E(Y_2))$ and then*

$$X_1 = E(Y_1) + \frac{Cov(Y_1, Y_2)}{Var(Y_2)}(Y_2 - E(Y_2)) + \varepsilon.$$

3.6.2. General case

In this paragraph, we will assume that $\Theta = \mathbb{R}_*^d$ and that our algorithm stops with j for $j < d$. Lemma F.9 implies the existence of an orthogonal and free family $(b_i)_{i=j+1,...,d}$ of $\mathbb{R}_*^d$ such that $\mathbb{R}^d = Vect\{a_i\} \overset{\perp}{\oplus} Vect\{b_k\}$ and such that

$$g(b_{j+1}^\top x, ..., b_d^\top x/a_1^\top x, ..., a_j^\top x) = f(b_{j+1}^\top x, ..., b_d^\top x/a_1^\top x, ..., a_j^\top x). \tag{3.2}$$

Hence, the following theorem provides us with the regression of $b_k^\top X$, $k = 1, ..., d$, on $(a_1^\top X, ..., a_j^\top X)$:

Theorem 3.9. *The probability measure of $(b_{j+1}^\top X, ..., b_d^\top X)$ given $(a_1^\top X, ..., a_j^\top X)$ is the same as the probability measure of $(b_{j+1}^\top Y, ..., b_d^\top Y)$ given $(a_1^\top Y, ..., a_j^\top Y)$. Moreover, the regression of $b_k^\top X$, $k = 1, ..., d$, on $(a_1^\top X, ..., a_j^\top X)$ is $b_k^\top X = E(b_k^\top Y/a_1^\top Y_1, ..., a_j^\top Y) + b_k^\top \varepsilon$, where ε is a centered random vector such that $b_k^\top \varepsilon$ is orthogonal to $E(b_k^\top X/a_1^\top X, ..., a_j^\top X)$.*

Corollary 3.4. *If g is a Gaussian density with same mean and variance as f, and if $Cov(X_i, X_j) = 0$ for any $i \neq j$, then, the regression of $b_k^\top X$, $k = 1, ..., d$, on $(a_1^\top X, ..., a_j^\top X)$ is $b_k^\top X = E(b_k^\top Y) + b_k^\top \varepsilon$, where ε is a centered random vector such that $b_k^\top \varepsilon$ is orthogonal to $E(b_k^\top X/a_1^\top X, ..., a_j^\top X)$.*

4. Simulations

Let us study four examples. The first involves a χ^2-divergence, the second a Hellinger distance, the third a Cressie-Read divergence (still with $\gamma = 1.25$) and the fourth a Kullback Leibler divergence.

In each example, 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 $\Phi(g^{(k)}, f) = 0$, with Φ being 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 the second example, we will study the robustness of our method with two ouliers. In the third example, defining (X_1, X_2) as a vector with f as density, we will study the regression of X_1 on X_2. And finally, in the fourth example, we will perform our goodness-of-fit test for copulas.

Simulation **4.1** (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.*
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* $(H_1) : \ a_1 \neq (1, 1, 0)$.

Then, corollary 3.3 enables us to estimate a_1 *by the following 0.9(=*α*) level confidence ellipsoid*

$$\mathcal{E}_1 = \{b \in \mathbb{R}^3; \ (Var_{\mathbf{P}}(M(b,b)))^{(-1/2)}\mathbb{P}_n M(b,b) \leq q_\alpha^{\mathcal{N}(0,1)}/\sqrt{n} \simeq 0,2533/7.0710678 = 0.03582203\}.$$

And, we obtain

Our Algorithm	
Projection Study 0 :	*minimum : 0.0201741*
	at point : (1.00912,1.09453,0.01893)
	P-Value : 0.81131
Test :	$H_0 : a_1 \in \mathcal{E}_1$ *: True*
χ^2*(Kernel Estimation of* $g^{(1)}$*,* $g^{(1)}$*)*	*6.1726*

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

Simulation **4.2** (With the Hellinger distance *H*)**.**
We are in dimension 20*(=d). We first generate a sample with* 100*(=n) observations, namely two outliers* $x = (2, 0, \ldots, 0)$ *and* 98 *values of a random variable X with a density law f defined by* $f(x) = Gumbel(x_0).Normal(x_1, \ldots, x_9)$*, where the Gumbel law parameters are -5 and 1 and where the normal distribution is reduced and centered.*
Our reasoning is the same as in Simulation 4.1.
In the first part of the program, we theoretically obtain $k = 1$ *and* $a_1 = (1, 0, \ldots, 0)$*. To get this result, we perform the following test* $(H_0) : \ a_1 = (1, 0, \ldots, 0)$ *versus* $(H_1) : \ a_1 \neq (1, 0, \ldots, 0)$.
We estimate a_1 *by the following 0.9(=*α*) level confidence ellipsoid*

$$\mathcal{E}_i = \{b \in \mathbb{R}^2; \ (Var_{\mathbf{P}}(M(b,b)))^{-1/2}\mathbb{P}_n M(b,b) \leq q_\alpha^{\mathcal{N}(0,1)}/\sqrt{n} \simeq 0.02533\}.$$

And, we obtain

Our Algorithm	
Projection Study 0	*minimum : 0.002692*
	at point : (1.01326, 0.0657, 0.0628, 0.1011, 0.0509, 0.1083, 0.1261, 0.0573, 0.0377, 0.0794, 0.0906, 0.0356, 0.0012, 0.0292, 0.0737, 0.0934, 0.0286, 0.1057, 0.0697, 0.0771)
	P-Value : 0.80554
Test :	$H_0 : a_1 \in \mathcal{E}_1$ *: True*
H(Estimate of $g^{(1)}$*,* $g^{(1)}$*)*	*3.042174*

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

Simulation **4.3** (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* $X = (X_1, X_2)$ *with a density law f defined by* $f(x) = Gumbel(x_0).Normal(x_1)$*, where the Gumbel*

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 presenting same mean and variance as* f.

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* $(H_1): \ a_1 \neq (1,0)$.

Then, corollary 3.3 enables us to estimate a_1 *by the following 0.9(=*α*) level confidence ellipsoid*

$\mathcal{E}_1 = \{b \in \mathbb{R}^2; \ (Var_{\mathbf{P}}(M(b,b)))^{(-1/2)}\mathbb{P}_n M(b,b) \leq q_\alpha^{N(0,1)}/\sqrt{n} \simeq 0.03582203\}$.

And, we obtain

Our Algorithm	
Projection Study 0 :	*minimum : 0.0210058*
	at point : (1.001,0.0014)
	P-Value : 0.989552
Test :	$H_0 : a_1 \in \mathcal{E}_1$ *: True*
Φ*(Kernel Estimation of* $g^{(1)}$, $g^{(1)}$*)*	*6.47617*

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

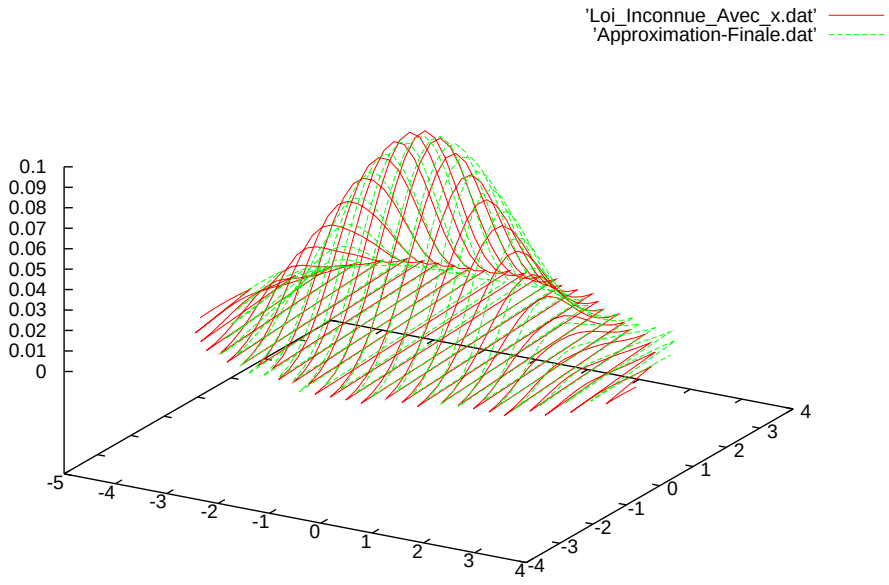

Figure 1: *Graph of the distribution to estimate (red) and of our own estimate (green).*

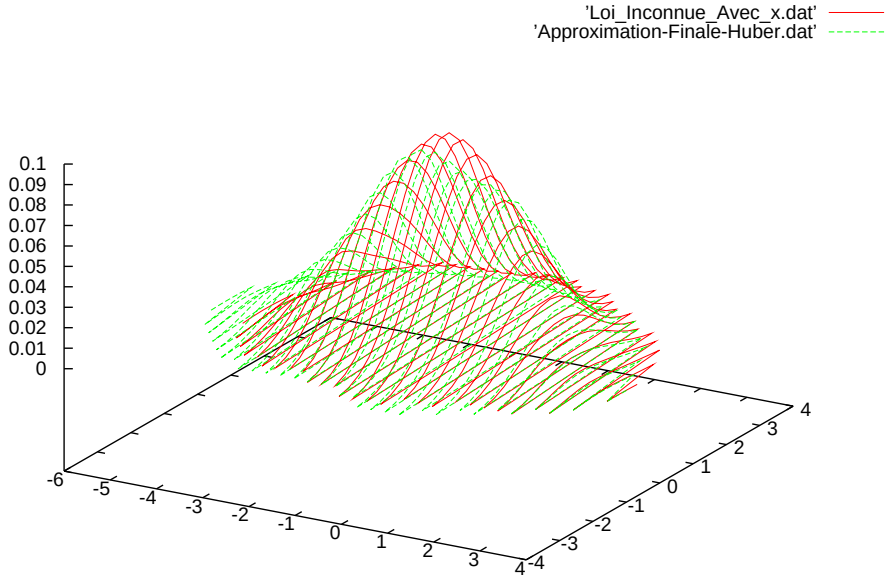

Figure 2: *Graph of the distribution to estimate (red) and of Huber's estimate (green).*

At present, keeping the notations of this simulation, let us study the regression of X_1 on X_2. Our algorithm leads us to infer that the density of X_1 given X_2 is the same as the density of y_1 given Y_2. Moreover, property A.1 implies that the co-factors of f are the same with all divergence. Consequently, we can use theorem 3.8, i.e. it implies that $X_1 = E(Y_1/Y_2) + \varepsilon$, where ε is a centered random variable orthogonal to $E(X_1/X_2)$. Thus, since g is a Gaussian density, remark 3.4 implies that

$$X_1 = E(Y_1) + \frac{Cov(Y_1, Y_2)}{Var(Y_2)}(Y_2 - E(Y_2)) + \varepsilon.$$

Now, using the least squares method, we estimate α_1 and α_2 such that $X_1 = a_1 + a_2 X_2 + \varepsilon$. Thus, the following table presents the results of our regression and of the least squares method if we assume that ε is Gaussian.

Our Regression	$E(Y_1)$	-4.545483
	$Cov(Y_1, Y_2)$	0.0380534
	$Var(Y_2)$	0.9190052
	$E(Y_2)$	0.3103752
	correlation coefficient (Y_1, Y_2)	0.02158213
Least squares method	a_1	-4.34159227
	a_2	0.06803317
	correlation coefficient (X_1, X_2)	0.04888484

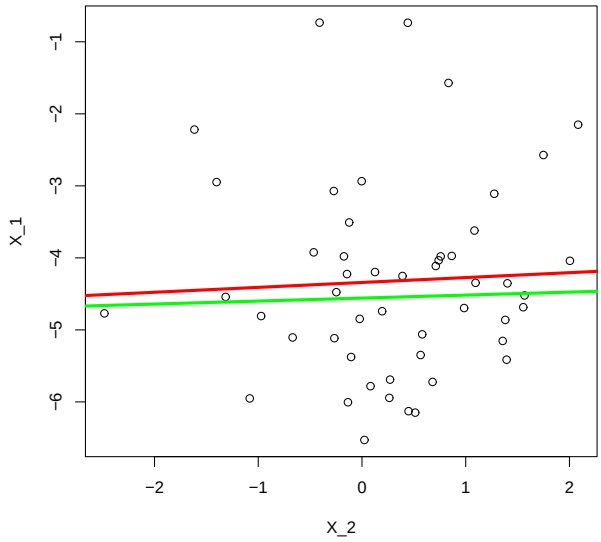

Figure 3: *Graph of the regression of X1 on X2 based on the least squares method (red) and based on our theory (green).*

Simulation **4.4** (With the relative entropy K)**.**
We are in dimension $2(=d)$*, and we use the relative entropy to perform our optimisations. Let us consider a sample of* $50(=n)$ *values of a random variable* X *with a density law* f *defined by :*

$$f(x) = c_\rho(F_{Gumbel}(x_0), F_{Exponential}(x_1)).Gumbel(x_0).Exponential(x_1),$$

where :

- c *is the Gaussian copula with correlation coefficient* $\rho = 0.5$*,*
- *the Gumbel distribution parameters are* -1 *and* 1 *and*
- *the Exponential density parameter is* 2*.*

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*.*
We theoretically obtain $k = 2$ *and* $(a_1, a_2) = ((1,0),(0,1))$*. To get this result, we perform the following test:*

$$(H_0) : \ (a_1, a_2) = ((1,0),(0,1)) \text{ versus } (H_1) : \ (a_1, a_2) \neq ((1,0),(0,1)).$$

Then, theorem 3.6 enables us to verify (H_0) *by the following* $0.9(=\alpha)$ *level confidence ellipsoid* $\mathcal{E}_2 = \{b \in \mathbb{R}^2; (Var_{\mathbf{P}}(M(b,b)))^{(-1/2)}\mathbb{P}_n M(b,b) \leq q_\alpha^{\mathcal{N}(0,1)}/\sqrt{n} \simeq 0,2533/7.0710678 = 0.0358220\}$.
And, we obtain

Our Algorithm	
Projection Study number 0 :	*minimum : 0.445199*
	at point : (1.0142,0.0026)
	P-Value : 0.94579
Test :	$H_0 : a_1 \in \mathcal{E}_1$ *: False*

Projection Study number 1 :	*minimum : 0.0263*
	at point : (0.0084,0.9006)
	P-Value : 0.97101
Test :	H_0 *: $a_2 \in \mathcal{E}_2$: True*
K(Kernel Estimation of $g^{(2)}$, $g^{(2)}$)	*4.0680*

Therefore, we can conclude that H_0 is verified.

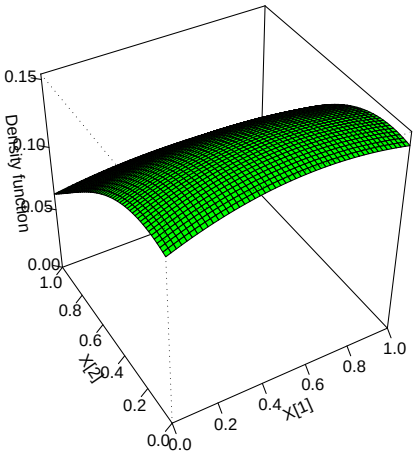

Figure 4: *Graph of the estimate of* $(x_0, x_1) \mapsto c_\rho(F_{Gumbel}(x_0), F_{Exponential}(x_1))$.

Critics of the simulations

In the case where f is unknown, we will never be sure to have reached the minimum of the Φ-divergence: 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 reach the minimum tends to 1. We also note that no theory on the optimal number of jumps to implement does exist, as this number depends on the specificities of each particular problem.

Moreover, we choose the $50^{-\frac{4}{4+d}}$ (resp. $100^{-\frac{4}{4+d}}$) for the AMISE of simulations 4.1, 4.2 and 4.3 (resp. simulation 4.4). This choice leads us to simulate 50 (resp. 100) random variables - see Scott (1992) page 151 -, none of which have been discarded to obtain the truncated sample.

Finally, we remark that some of the key advantages of our method over Huber's consist in the fact that - since there exist divergences smaller than the relative entropy - our method requires a considerably shorter computation time and also in the in the superiority in robustness of our method.

Conclusion

Projection Pursuit is useful in evidencing characteristic structures as well as one-dimensional projections and their associated distributions in multivariate data. Huber (1985) shows us how to achieve it through maximization of the relative entropy.
The present article shows that our Φ-divergence method constitutes a good alternative to Huber's particularly in terms of regression and robustness as well as in terms of copula's study. Indeed, the convergence results and simulations we carried out, convincingly fulfilled our expectations regarding our methodology.

A. 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 A.1. *We define the Φ−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 Cressie-Read (1984), Csiszár I. (1967) and the books of Friedrich and Igor (1987), Pardo Leandro (2006) and Zografos K. (1990)). 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 $\varphi(x) = |x - 1|$.

In particular we have 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.

Property A.1. *We have* $\Phi(P, Q) = 0 \Leftrightarrow P = Q$.

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

Property A.3 (corollary (1.29), page 19 of Friedrich and Igor (1987))**.** *If* $T : (X, A) \to (Y, B)$ *is measurable and if* $K(P, Q) < \infty$, *then* $K(P, Q) \geq K(PT^{-1}, QT^{-1})$, *with equality being reached when* T *is surjective for* (P, Q).

Theorem A.1 (theorem III.4 of Azé (1997))**.** *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* $int\{I\}$.

A.2. Useful lemmas

Through a reductio ad absurdum argument, we derive lemmas A.1 and A.2 :

lemme A.1. *Let f be a density in $\mathbb{R}^d$ bounded and positive. Then, any projection density of f - that we will name f_a, with $a \in \mathbb{R}^d_*$ - is also bounded and positive in $\mathbb{R}$.*

lemme A.2. *Let f be a density in $\mathbb{R}^d$ bounded and positive. Then any density $f(./a^\top x)$, for any $a \in \mathbb{R}^d_*$, is also bounded and positive.*

By induction and from lemmas A.1 and A.2, we have

lemme A.3. *If f and g are positive and bounded densities, then $g^{(k)}$ is positive and bounded.*

Finally we introduce a last lemma

lemme A.4. *Let f be an absolutely continuous density, then, for all sequences (a_n) tending to a in $\mathbb{R}^d_*$, 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 as a result of the corollary of Dini's second theorem - according to which *"A sequence of cumulative distribution functions which pointwise 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 sequences (a_n) converging towards a, ψ_{a_n} uniformly converges towards 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. Study of the sample

Let X_1, X_2,..,X_m be a sequence of independent random vectors with same density f. Let Y_1, Y_2,..,Y_m be a sequence of independent random vectors with same density g. Then, the kernel estimators f_m, g_m, $f_{a,m}$ and $g_{a,m}$ of f, g, f_a and g_a, for all $a \in \mathbb{R}^d_*$, almost surely and uniformly converge since we assume that the bandwidth h_m of these estimators meets the following conditions (see Bosq (1999)):

($\mathcal{H}yp$): $h_m \searrow_m 0$, $mh_m \nearrow_m \infty$, $mh_m/L(h_m^{-1}) \to_m \infty$ and $L(h_m^{-1})/LLm \to_m \infty$,
with $L(u) = ln(u \vee e)$.
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 $B_2(n,a) = \frac{1}{n}\Sigma_{i=1}^n \varphi^*\{\varphi'\{\frac{f_{a,n}(a^\top X_i)}{g_{a,n}(a^\top X_i)}\frac{g_n(X_i)}{f_n(X_i}\}\}$.

Our goal is to estimate the minimum of $\Phi(g\frac{f_a}{g_a}, f)$. To do this, it is necessary for us to truncate our samples:
Let us consider now a positive sequence θ_m such that $\theta_m \to 0$, $y_m/\theta_n^2 \to 0$, where y_m is the almost sure convergence rate of the kernel density estimator - $y_m = O_{\mathbf{P}}(m^{-\frac{2}{4+d}})$, see lemma F.10 - $y_m^{(1)}/\theta_m^2 \to 0$, where $y_m^{(1)}$ is defined by

$$|\varphi(\frac{g_m(x)}{f_m(x)}\frac{f_{b,m}(b^\top x)}{g_{b,m}(b^\top x)}) - \varphi(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)})| \leq y_m^{(1)}$$

for all b in $\mathbb{R}_*^d$ and all x in $\mathbb{R}^d$, and finally $\frac{y_m^{(2)}}{\theta_m^2} \to 0$, where $y_n^{(2)}$ is defined by

$$|\varphi'(\frac{g_m(x)}{f_m(x)}\frac{f_{b,m}(b^\top x)}{g_{b,m}(b^\top x)}) - \varphi'(\frac{g(x)}{f(x)}\frac{f_b(b^\top x)}{g_b(b^\top x)})| \leq y_m^{(2)}$$

for all b in $\mathbb{R}_*^d$ and all x in $\mathbb{R}^d$.
We will generate f_m, g_m and $g_{b,m}$ from the starting sample and we will select the X_i and Y_i vectors such that $f_m(X_i) \geq \theta_m$ and $g_{b,m}(b^\top Y_i) \geq \theta_m$, for all i and for all $b \in \mathbb{R}_*^d$.
The vectors meeting these conditions will be called $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$.
Consequently, the next proposition provides us with the condition required for us to derive our estimations

Proposition B.1. *Using the notations introduced in Broniatowski (2009) and in section 3.1.1, it holds* $\lim_{n\to\infty} \sup_{a\in\mathbb{R}_*^d} |(B_1(n,a) - B_2(n,a)) - \Phi(g\frac{f_a}{g_a}, f)| = 0$.

Remark **B.1.** *With the relative entropy, we can take for θ_m the expression $m^{-\nu}$, with $0 < \nu < \frac{1}{4+d}$.*

C. Case study : f is known

In this Annex, we will study the case when f and g are known. We will then use the notations introduced in sections 3.1.1 and 3.1.2 with f and g, i.e. no longer with their kernel estimates.

C.1. Convergence study and Asymptotic Inference at the k^{th} step of the algorithm

In this paragraph, when k is less than or equal to d, we will show that the sequence $(\check{a}_k)_n$ converges towards a_k and that the sequence $(\check{g}^{(k)})_n$ converges towards $g^{(k)}$.
Both $\check{\gamma}_n$ and $\check{c}_n(a)$ are M-estimators and estimate a_k - see Broniatowski (2009). We state

Proposition C.1. *Assuming* $(H1)$ *to* $(H3)$ *hold. Both* $\sup_{a\in\Theta} \|\check{c}_n(a) - a_k\|$ *and* $\check{\gamma}_n$ *tends to* a_k *a.s.*

Finally, the following theorem shows us that $\check{g}^{(k)}$ converges uniformly almost everywhere towards $g^{(k)}$, for any $k = 1..d$.

Theorem C.1. *Assumimg* $(H1)$ *to* $(H3)$ *hold. Then,* $\check{g}^{(k)} \to_n g^{(k)}$ *a.s. and uniformly a.e.*

The following theorem shows that $\check{g}^{(k)}$ converges at the rate $O_{\mathbf{P}}(n^{-1/2})$ in three differents cases, namely for any given x, with the L^1 distance and with the $\Phi-$divergence:

Theorem C.2. *Assuming* $(H0)$ *to* $(H3)$ *hold, for any $k = 1, ..., d$ and any $x \in \mathbb{R}^d$, we have*

$$|\check{g}^{(k)}(x) - g^{(k)}(x)| = O_{\mathbf{P}}(n^{-1/2}), \tag{C.1}$$

$$\int |\check{g}^{(k)}(x) - g^{(k)}(x)|dx = O_{\mathbf{P}}(n^{-1/2}), \tag{C.2}$$

$$|K(\check{g}^{(k)}, f) - K(g^{(k)}, f)| = O_{\mathbf{P}}(n^{-1/2}). \tag{C.3}$$

The following theorem shows that the laws of our estimators of a_k, namely $\check{c}_n(a_k)$ and $\check{\gamma}_n$, converge towards a linear combination of Gaussian variables.

Theorem C.3. *Assuming that conditions* $(H1)$ *to* $(H6)$ *hold, then*
$\sqrt{n}\mathcal{A}.(\check{c}_n(a_k) - a_k) \overset{\mathcal{L}aw}{\to} \mathcal{B}.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k,a_k)\|^2) + C.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k,a_k)\|^2)$ *and*
$\sqrt{n}\mathcal{A}.(\check{\gamma}_n - a_k) \overset{\mathcal{L}aw}{\to} C.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial b}M(a_k,a_k)\|^2) + C.\mathcal{N}_d(0, \mathbf{P}\|\frac{\partial}{\partial a}M(a_k,a_k)\|^2)$
where $\mathcal{A} = (\mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k)(\mathbf{P}\frac{\partial^2}{\partial a_i\partial a_j}M(a_k,a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i\partial b_j}M(a_k,a_k)))$,
$C = \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k)$ *and* $\mathcal{B} = \mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i\partial a_j}M(a_k,a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i\partial b_j}M(a_k,a_k)$.

C.2. A stopping rule for the procedure

We now assume that the algorithm does not stop after d iterations. We then remark that, it still holds - for any $i > d$:

- $g^{(i)}(x) = g(x)\Pi_{k=1}^{i} \frac{f_{a_k}(a_k^\top x)}{[g_n^{(k-1)}]_{a_k}(a_k^\top x)}$, with $g^{(0)} = g$.
- $K(g^{(0)}, f) \geq K(g^{(1)}, f) \geq K(g^{(2)}, f)... \geq 0$.
- Theorems C.1, C.2 and C.3.

Moreover, as explained in section 14 of Huber (1985) for the relative entropy, the sequence $(\Phi(g^{(k-1)}\frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f))_{k\geq 1}$ converges towards zero. Then, in this paragraph, we will show that $g^{(i)}$ converges towards f in i. And finally, we will provide a stopping rule for this identification procedure.

C.2.1. Representation of f

Under $(H0)$, the following proposition shows us that the probability measure with density $g^{(k)}$ converges towards the probability measure with density f :

Proposition C.2 (Representation of f)**.** *We have* $\lim_k g^{(k)} = f$ *a.s.*

C.2.2. Testing of the criteria

Through a test of the criteria, namely $a \mapsto \Phi(g^{(k-1)}\frac{f_a}{g_a^{(k-1)}}, f)$, we will build a stopping rule for this procedure. First, the next theorem enables us to derive the law of the criteria.

Theorem C.4. *Assuming that* $(H1)$ *to* $(H3)$*,* $(H6)$ *and* $(H8)$ *hold. Then,*

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

where k represents the k^{th} *step of the algorithm and with I being the identity matrix in* $\mathbb{R}^d$.

Note that k is fixed in theorem C.4 since $\check{\gamma}_n = arg\inf_{a\in\Theta} \sup_{c\in\Theta} \mathbb{P}_n M(c, a)$ where M is a known function of k - see section 3.1.1. Thus, in the case where $\Phi(g^{(k-1)}\frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) = 0$, we obtain

Corollary C.1. *Assuming that* $(H1)$ *to* $(H3)$*,* $(H6)$*,* $(H7)$ *and* $(H8)$ *hold. Then,*

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

Hence, we propose the test of the null hypothesis

$(H_0) : K(g^{(k-1)}\frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) = 0$ versus $(H_1) : K(g^{(k-1)}\frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) \neq 0$.

Based on this result, we stop the algorithm, then, defining a_k as the last vector generated, we derive from corollary C.1 a α-level confidence ellipsoid around a_k, namely

$\mathcal{E}_k = \{b \in \mathbb{R}^d;\ \sqrt{n}(Var_{\mathbf{P}}(M(b, b)))^{-1/2}\mathbb{P}_n M(b, b) \leq q_\alpha^{\mathcal{N}(0,1)}\}$,

where $q_\alpha^{\mathcal{N}(0,1)}$ is the quantile of a α-level reduced centered normal distribution.
Consequently, the following corollary provides us with a confidence region for the above test:

Corollary C.2. $\mathcal{E}_k$ *is a confidence region for the test of the null hypothesis* (H_0) *versus* (H_1).

D. The first co-vector of f simultaneously optimizes four problems

Let us first study Huber's analytic approach.
Let $\mathcal{R}'$ be the class of all positive functions r defined on $\mathbb{R}$ and such that $f(x)r^{-1}(a^\top x)$ is a density on $\mathbb{R}^d$ for all a belonging to $\mathbb{R}_*^d$. The following proposition shows that there exists a vector a such that $\frac{f_a}{g_a}$ minimizes $K(fr^{-1}, g)$ in r:

Proposition D.1 (Analytic Approach)**.** *There exists a vector a belonging to $\mathbb{R}^d_*$ such that*
$arg\min_{r\in\mathcal{R}'} K(fr^{-1}, g) = \frac{f_a}{g_a}$, *and* $r(a^\top x) = \frac{f_a(a^\top x)}{g_a(a^\top x)}$ *as well as* $K(f,g) = K(f_a, g_a) + K(f\frac{g_a}{f_a}, g)$.

Let us also study Huber's synthetic approach:
Let $\mathcal{R}$ be the class of all positive functions r defined on $\mathbb{R}$ and such that $g(x)r(a^\top x)$ is a density on $\mathbb{R}^d$ for all a belonging to $\mathbb{R}^d_*$. The following proposition shows that there exists a vector a such that $\frac{f_a}{g_a}$ minimizes $K(gr, f)$ in r:

Proposition D.2 (Synthetic Approach)**.** *There exists a vector a belonging to $\mathbb{R}^d_*$ such that*
$arg\min_{r\in\mathcal{R}} K(f, gr) = \frac{f_a}{g_a}$, *and* $r(a^\top x) = \frac{f_a(a^\top x)}{g_a(a^\top x)}$ *as well as* $K(f,g) = K(f_a, g_a) + K(f, g\frac{f_a}{g_a})$.

In the meanwhile, the following proposition shows that there exists a vector a such that $\frac{f_a}{g_a}$ minimizes $K(g, fr^{-1})$ in r.

Proposition D.3. *There exists a vector a belonging to $\mathbb{R}^d_*$ such that*
$arg\min_{r\in\mathcal{R}'} K(g, fr^{-1}) = \frac{f_a}{g_a}$, *and* $r(a^\top x) = \frac{f_a(a^\top x)}{g_a(a^\top x)}$ *as well as* $K(g,f) = K(g_a, f_a) + K(g, f\frac{g_a}{f_a})$.

Remark **D.1.** *First, through property A.3 page 18, we get $K(f, g\frac{f_a}{g_a}) = K(g, f\frac{g_a}{f_a}) = K(f\frac{g_a}{f_a}, g)$ and $K(f_a, g_a) = K(g_a, f_a)$. Thus, proposition D.3 implies that finding the argument of the maximum of $K(g_a, f_a)$ amounts to finding the argument of the maximum $K(f_a, g_a)$. Consequently, the criteria of Huber's methodologies is $a \mapsto K(g_a, f_a)$. Second, if the Φ-divergence is the relative entropy, then our criteria is $a \mapsto K(g\frac{g_a}{f_a}, f)$ and property A.3 implies $K(g, f\frac{g_a}{f_a}) = K(g\frac{f_a}{g_a}, f)$.*

To recapitulate, the choice of $r = \frac{f_a}{g_a}$ enables us to simultaneously solve the following four optimisation problems, for $a \in \mathbb{R}^d_*$:
First, find a such that $a = arginf_{a\in\mathbb{R}^d_*} K(f\frac{g_a}{f_a}, g)$,
Second, find a such that $a = arginf_{a\in\mathbb{R}^d_*} K(f, g\frac{f_a}{g_a})$,
Third, find a such that $a = argsup_{a\in\mathbb{R}^d_*} K(g_a, f_a)$,
Fourth, find a such that $a = arginf_{a\in\mathbb{R}^d_*} K(g\frac{f_a}{g_a}, f)$.

E. Hypotheses' discussion

E.1. Discussion of (*H*2).

Let us work with the relative entropy and with g and a_1.
For all $b \in \mathbb{R}^d_*$, we have $\int \varphi^*(\varphi'(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)}))f(x)dx = \int(\frac{g(x)f_b(b^\top x)}{f(x)g_b(b^\top x)} - 1)f(x)dx = 0$, since, for any b in $\mathbb{R}^d_*$, the function $x \mapsto g(x)\frac{f_b(b^\top x)}{g_b(b^\top x)}$ is a density. The complement of Θ^Φ in $\mathbb{R}^d_*$ is $\emptyset$ and then the supremum looked for in $\overline{\mathbb{R}}$ is $-\infty$. We can therefore conclude. It is interesting to note that we obtain the same verification with f, $g^{(k-1)}$ and a_k.

E.2. Dicussion of (*H*4).

This hypothesis consists in the following assumptions:
• *We work with the relative entropy, (0)*
• *We have $f(./a_1^\top x) = g(./a_1^\top x)$, i.e. $K(g\frac{f_1}{g_1}, f) = 0$ - we could also derive the same proof with f, $g^{(k-1)}$ and a_k - (1)*
Preliminary (*A*)*: Shows that* $A = \{(c,x) \in \mathbb{R}^d_*\backslash\{a_1\}\times R^d;\ \frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} > \frac{f_c(c^\top x)}{g_c(c^\top x)},\ g(x)\frac{f_c(c^\top x)}{g_c(c^\top x)} > f(x)\} = \emptyset$

through a reductio ad absurdum, i.e. if we assume $A \neq \emptyset$.
Thus, our hypothesis enables us to derive

$f(x) = f(./a_1^\top x) f_{a_1}(a_1^\top x) = g(./a_1^\top x) f_{a_1}(a_1^\top x) > g(./c^\top x) f_c(c^\top x) > f$

since $\frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} \geq \frac{f_c(c^\top x)}{g_c(c^\top x)}$ implies $g(./a_1^\top x) f_{a_1}(a_1^\top x) = g(x)\frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} \geq g(x)\frac{f_c(c^\top x)}{g_c(c^\top x)} = g(./c^\top x) f_c(c^\top x)$, i.e. $f > f$. We can therefore conclude.

Preliminary (B)*: Shows that* $B = \{(c, x) \in \mathbb{R}_*^d \backslash \{a_1\} \times R^d;\ \frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} < \frac{f_c(c^\top x)}{g_c(c^\top x)},\ g(x)\frac{f_c(c^\top x)}{g_c(c^\top x)} < f(x)\} = \emptyset$ *through a reductio ad absurdum, i.e. if we assume* $B \neq \emptyset$.
Thus, our hypothesis enables us to derive

$f(x) = f(./a_1^\top x) f_{a_1}(a_1^\top x) = g(./a_1^\top x) f_{a_1}(a_1^\top x) < g(./c^\top x) f_c(c^\top x) < f$

We can therefore conclude as above.
Let us now verify $(H4)$:
We have $PM(c, a_1) - PM(c, a) = \int ln(\frac{g(x) f_c(c^\top x)}{g_c(c^\top x) f(x)})\{\frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} - \frac{f_c(c^\top x)}{g_c(c^\top x)}\} g(x) dx$. Moreover, the logarithm ln is negative on $\{x \in \mathbb{R}_*^d;\ \frac{g(x) f_c(c^\top x)}{g_c(c^\top x) f(x)} < 1\}$ and is positive on $\{x \in \mathbb{R}_*^d;\ \frac{g(x) f_c(c^\top x)}{g_c(c^\top x) f(x)} \geq 1\}$.
Thus, the preliminary studies (A) and (B) show that $ln(\frac{g(x) f_c(c^\top x)}{g_c(c^\top x) f(x)})$ and $\{\frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} - \frac{f_c(c^\top x)}{g_c(c^\top x)}\}$ always present a negative product. We can therefore conclude, since $(c, a) \mapsto PM(c, a_1) - PM(c, a)$ is not null for all c and for all a - with $a \neq a_1$.

F. Proofs

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

Remark **F.1.** *1/* $(H0)$ *- according to which* f *and* g *are assumed to be positive and bounded - through lemma A.3 (see page 19) implies that* $\check{g}^{(k)}$ *and* $\hat{g}^{(k)}$ *are positive and bounded.*
2/ remark 2.1 page 5 implies that f_n, g_n, $\check{g}^{(k)}$ *and* $\hat{g}^{(k)}$ *are positive and bounded since we consider a Gaussian kernel.*

Proof of propositions D.1 and D.2. Let us first study proposition D.2.
Without loss of generality, we will prove this proposition with x_1 in lieu of $a^\top X$.
Let us define $g^* = gr$. We remark that g and g^* present the same density conditionally to x_1. Indeed, $g_1^*(x_1) = \int g^*(x) dx_2 ... dx_d = \int r(x_1) g(x) dx_2 ... dx_d = r(x_1) \int g(x) dx_2 ... dx_d = r(x_1) g_1(x_1)$.
Thus, we can demonstrate this proposition.
We have $g(.|x_1) = \frac{g(x_1,...,x_n)}{g_1(x_1)}$ and $g_1(x_1) r(x_1)$ is the marginal density of g^*. Hence,
$\int g^* dx = \int g_1(x_1) r(x_1) g(.|x_1) dx = \int g_1(x_1) \frac{f_1(x_1)}{g_1(x_1)} (\int g(.|x_1) dx_2 .. dx_d) dx_1 = \int f_1(x_1) dx_1 = 1$ and since g^* is positive, then g^* is a density. Moreover,

$$K(f, g^*) = \int f\{ln(f) - ln(g^*)\} dx, \tag{F.1}$$

$$= \int f\{ln(f(.|x_1)) - ln(g^*(.|x_1)) + ln(f_1(x_1)) - ln(g_1(x_1) r(x_1))\} dx,$$

$$= \int f\{ln(f(.|x_1)) - ln(g(.|x_1)) + ln(f_1(x_1)) - ln(g_1(x_1) r(x_1))\} dx, \tag{F.2}$$

as $g^*(.|x_1) = g(.|x_1)$. Since the minimum of this last equation (F.2) is reached through the minimization of $\int f\{ln(f_1(x_1)) - ln(g_1(x_1) r(x_1))\} dx = K(f_1, g_1 r)$, then property A.1 necessarily implies that $f_1 = g_1 r$, hence $r = f_1 / g_1$.

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 proposition D.2.
Similarly, if we replace $f^* = fr^{-1}$ with f and g with g^*, we obtain the proposition D.1. □

Proof of proposition D.3. The demonstration is very similar to the one for proposition D.2, save for the fact we now base our reasoning at row (F.1) on $\int g\{ln(g^*) - ln(f)\}dx$ instead of $K(f,g^*) = \int f\{ln(f) - ln(g^*)\}dx$. □

Proof of proposition 3.1.
Without loss of generality, we reason with x_1 in lieu of $a^\top x$.
Let us define $g^* = gr$. We remark that g and g^* present the same density conditionally to x_1. Indeed, $g_1^*(x_1) = \int g^*(x)dx_2...dx_d = \int h(x_1)g(x)dx_2...dx_d = h(x_1)\int g(x)dx_2...dx_d = h(x_1)g_1(x_1)$.
We can therefore prove this proposition.
First, since f and g are known, then, for any given function $h : x_1 \mapsto h(x_1)$, the application T, which is defined by:

$T : g(./x_1)\frac{h(x_1)f_1(x_1)}{g_1(x_1)} \mapsto g(./x_1)f_1(x_1)$,
$T : f(./x_1)f_1(x_1) \mapsto f(./x_1)f_1(x_1)$

is measurable.
Second, the above remark implies that

$\Phi(g^*, f) = \Phi(g^*(./x_1)\frac{g_1(x_1)h(x_1)}{f_1(x_1)}, f(./x_1)f_1(x_1)) = \Phi(g(./x_1)\frac{g_1(x_1)h(x_1)}{f_1(x_1)}, f(./x_1)f_1(x_1))$.

Consequently, property A.3 page 18 infers :

$$\begin{aligned}\Phi(g(./x_1)\tfrac{g_1(x_1)h(x_1)}{f_1(x_1)}, f(./x_1)f_1(x_1)) &\geq \Phi(T^{-1}(g(./x_1)\tfrac{g_1(x_1)h(x_1)}{f_1(x_1)}), T^{-1}(f(./x_1)f_1(x_1)))\\ &= \Phi(g(./x_1)f_1(x_1), f(./x_1)f_1(x_1)), \text{ by the very definition of } T.\\ &= \Phi(g\tfrac{f_1}{g_1}, f),\end{aligned}$$

which completes the proof of this proposition. □

Proof of lemma F.1.

lemme F.1. *We have* $g(./a_1^\top x, ..., a_j^\top x) = n(a_{j+1}^\top x, ..., a_d^\top x) = f(./a_1^\top x, ..., a_j^\top x)$.

Putting $A = (a_1, .., a_d)$, let us determine f in basis A. Let us first study the function defined by $\psi : \mathbb{R}^d \to \mathbb{R}^d$, $x \mapsto (a_1^\top x, .., a_d^\top x)$. We can immediately say that ψ is continuous and since A is a basis, its bijectivity is obvious. Moreover, let us study its Jacobian.
By definition, it is $J_\psi(x_1, \ldots, x_d) = \begin{vmatrix} \frac{\partial\psi_1}{\partial x_1} & \cdots & \frac{\partial\psi_1}{\partial x_d} \\ \cdots & \cdots & \cdots \\ \frac{\partial\psi_d}{\partial x_1} & \cdots & \frac{\partial\psi_d}{\partial x_d} \end{vmatrix} = \begin{vmatrix} a_{1,1} & \cdots & a_{1,d} \\ \cdots & \cdots & \cdots \\ a_{d,1} & \cdots & a_{d,d} \end{vmatrix} = |A| \neq 0$ since A is a basis. We can therefore infer : $\forall x \in \mathbb{R}^d$, $\exists! y \in \mathbb{R}^d$ such that $f(x) = |A|^{-1}\Psi(y)$, i.e. Ψ (resp. y) is the expression of f (resp of x) in basis A, namely $\Psi(y) = \tilde{n}(y_{j+1}, ..., y_d)\tilde{h}(y_1, ..., y_j)$, with $\tilde{n}$ and $\tilde{h}$ being the expressions of n and h in basis A. Consequently, our results in the case where the family $\{a_j\}_{1\leq j\leq d}$ is the canonical basis of $\mathbb{R}^d$, still hold for Ψ in basis A - see section 2.1.2. And then, if $\tilde{g}$ is the expression of g in basis A, we have $\tilde{g}(./y_1, ..., y_j) = \tilde{n}(y_{j+1}, ..., y_d) = \Psi(./y_1, ..., y_j)$, i.e. $g(./a_1^\top x, ..., a_j^\top x) = n(a_{j+1}^\top x, ..., a_d^\top x) = f(./a_1^\top x, ..., a_j^\top x)$. □

Proof of lemma F.2.

lemme F.2. *Should there exist a family* $(a_i)_{i=1...d}$ *such that* $f(x) = n(a_{j+1}^\top x, ..., a_d^\top x)h(a_1^\top x, ..., a_j^\top x)$, *with* $j < d$, *with* f, n *and* h *being densities, then this family is a orthogonal basis of* $\mathbb{R}^d$.

Using a reductio ad absurdum, we have $\int f(x)dx = 1 \neq +\infty = \int n(a_{j+1}^\top x, ..., a_d^\top x)h(a_1^\top x, ..., a_j^\top x)dx$.
We can therefore conclude. □

Proof of proposition B.1.
Let us note first that we will prove this proposition for $k \geq 2$, i.e. in the case where $g^{(k-1)}$ is not known. The initial case using the known density $g^{(0)} = g$, will be an immediate consequence from the above.
Moreover, going forward, to be more legible, we will use g (resp. g_n) in lieu of $g^{(k-1)}$ (resp. $g_n^{(k-1)}$).
We can therefore remark that we have $f(X_i) \geq \theta_n - y_n$, $g(Y_i) \geq \theta_n - y_n$ and $g_b(b^\top Y_i) \geq \theta_n - y_n$, for all i and for all $b \in \mathbb{R}_*^d$, thanks to the uniform convergence of the kernel estimators. Indeed, we have $f(X_i) = f(X_i) - f_n(X_i) + f_n(X_i) \geq -y_n + f_n(X_i)$, by definition of y_n, and then $f(X_i) \geq -y_n + \theta_n$, by hypothesis on $f_n(X_i)$. This is also true for g_n and $g_{b,n}$.
This entails $\sup_{b\in\mathbb{R}_*^d} |\frac{1}{n}\Sigma_{i=1}^n \varphi'(\frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} \frac{g_n(Y_i)}{f_n(Y_i)}).\frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} - \int \varphi'(\frac{g(x)}{f(x)} \frac{f_b(b^\top x)}{g_b(b^\top x)}) g(x) \frac{f_a(a^\top x)}{g_a(a^\top x)} dx| \to 0$ *a.s.*
Indeed, let us remark that

$$
\begin{aligned}
&|\tfrac{1}{n}\Sigma_{i=1}^n \{\varphi'\{\tfrac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} \tfrac{g_n(Y_i)}{f_n(Y_i)}\} \tfrac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)}\} - \int \varphi'(\tfrac{g(x)}{f(x)} \tfrac{f_b(b^\top x)}{g_b(b^\top x)})\, g(x) \tfrac{f_a(a^\top x)}{g_a(a^\top x)}\, dx| \\
&= |\tfrac{1}{n}\Sigma_{i=1}^n \varphi'\{\tfrac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} \tfrac{g_n(Y_i)}{f_n(Y_i)}\} \tfrac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} - \tfrac{1}{n}\Sigma_{i=1}^n \varphi'\{\tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \tfrac{g(Y_i)}{f(Y_i)}\} \tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \\
&\quad + \tfrac{1}{n}\Sigma_{i=1}^n \varphi'\{\tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \tfrac{g(Y_i)}{f(Y_i)}\} \tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} - \int \varphi'(\tfrac{g(x)}{f(x)} \tfrac{f_b(b^\top x)}{g_b(b^\top x)})\, g(x) \tfrac{f_a(a^\top x)}{g_a(a^\top x)}\, dx| \\
&\leq |\tfrac{1}{n}\Sigma_{i=1}^n \varphi'\{\tfrac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} \tfrac{g_n(Y_i)}{f_n(Y_i)}\} \tfrac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} - \tfrac{1}{n}\Sigma_{i=1}^n \varphi'\{\tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \tfrac{g(Y_i)}{f(Y_i)}\} \tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)}| \\
&\quad + |\tfrac{1}{n}\Sigma_{i=1}^n \varphi'\{\tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \tfrac{g(Y_i)}{f(Y_i)}\} \tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} - \int \varphi'(\tfrac{g(x)}{f(x)} \tfrac{f_b(b^\top x)}{g_b(b^\top x)})\, g(x) \tfrac{f_a(a^\top x)}{g_a(a^\top x)}\, dx|
\end{aligned}
$$

Moreover, since $\int |\varphi'(\frac{g(x)}{f(x)} \frac{f_b(b^\top x)}{g_b(b^\top x)})\, g(x) \frac{f_a(a^\top x)}{g_a(a^\top x)}| dx < \infty$, as implied by lemma A.3, and since we assumed g such that $\Phi(g, f) < \infty$ and $\Phi(f, g) < \infty$ and since $b \in \Theta^\Phi$, the law of large numbers enables us to state that $|\frac{1}{n}\Sigma_{i=1}^n \varphi'\{\frac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \frac{g(Y_i)}{f(Y_i)}\} \frac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} - \int \varphi'(\frac{g(x)}{f(x)} \frac{f_b(b^\top x)}{g_b(b^\top x)})\, g(x) \frac{f_a(a^\top x)}{g_a(a^\top x)}\, dx| \to 0$ *a.s.*
Furthermore, $|\frac{1}{n}\Sigma_{i=1}^n \varphi'\{\frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} \frac{g_n(Y_i)}{f_n(Y_i)}\} \frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} - \frac{1}{n}\Sigma_{i=1}^n \varphi'\{\frac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \frac{g(Y_i)}{f(Y_i)}\} \frac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)}|$

$$
\leq \tfrac{1}{n}\Sigma_{i=1}^n |\varphi'\{\tfrac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} \tfrac{g_n(Y_i)}{f_n(Y_i)}\} \tfrac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} - \varphi'\{\tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \tfrac{g(Y_i)}{f(Y_i)}\} \tfrac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)}|
$$

and $|\varphi'\{\frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} \frac{g_n(Y_i)}{f_n(Y_i)}\} \frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} - \varphi'\{\frac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \frac{g(Y_i)}{f(Y_i)}\} \frac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)}| \to 0$ as a result of the hypotheses intially introduced on θ_n. Consequently, $\frac{1}{n}\Sigma_{i=1}^n |\varphi'\{\frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} \frac{g_n(Y_i)}{f_n(Y_i)}\} \frac{f_{a,n}(a^\top Y_i)}{g_{a,n}(a^\top Y_i)} - \varphi'\{\frac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)} \frac{g(Y_i)}{f(Y_i)}\} \frac{f_a(a^\top Y_i)}{g_a(a^\top Y_i)}| \to 0$, as it is a Cesàro mean. This enables us to conclude. Similarly, we obtain
$\sup_{b\in\mathbb{R}_*^d} |\frac{1}{n}\Sigma_{i=1}^n \varphi^*\{\varphi'\{\frac{f_{a,n}(a^\top X_i)}{g_{a,n}(a^\top X_i)} \frac{g_n(X_i)}{f_n(X_i)}\}\} - \int \varphi^*(\varphi'(\frac{g(x)}{f(x)} \frac{f_b(b^\top x)}{g_b(b^\top x)}))\, f(x) dx| \to 0$ *a.s.* □

Proof of lemma F.3. By definition of the closure of a set, we have

lemme F.3. *The set Γ_c is closed in L^1 for the topology of the uniform convergence.*

Proof of lemma F.4. Since Φ is greater than the L^1 distance, we have

lemme F.4. *For all $c > 0$, we have $\Gamma_c \subset \overline{B}_{L^1}(f, c)$, where $B_{L^1}(f, c) = \{p \in L^1; \|f - p\|_1 \leq c\}$.*

Proof of lemma F.5. The definition of the closure of a set and lemma A.4 (see page 19) imply

lemme F.5. *G is closed in L^1 for the topology of the uniform convergence.*

Proof of lemma F.6.

lemme F.6. $\inf_{a\in\mathbb{R}_*^d} \Phi(g^*, f)$ *is reached when the Φ-divergence is greater than the L^1 distance as well as the L^2 distance.*

Proof. 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 F.3, F.4 and F.5 (see page 25), 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 property A.2 (see page 18) implies that

$Q \mapsto \Phi(Q, P)$ is lower semi-continuous in L^1 for the topology of the uniform convergence, then the infimum is reached in L^1. (Taking for example $c = \Phi(g, f)$, Ω is necessarily not empty because we always have $\Phi(g\frac{f_a}{g_a}, f) \leq \Phi(g, f)$). Moreover, when the $\Phi-$divergence is greater than the L^2 distance, the very definition of the L^2 space enables us to provide the same proof as for the L^1 distance. □

Proof of lemma F.7.

lemme F.7. *For any $p \leq d$, we have $f_{a_p}^{(p-1)} = f_{a_p}$ - see Huber's analytic method -, $g_{a_p}^{(p-1)} = g_{a_p}$ - see Huber's synthetic method - and $g_{a_p}^{(p-1)} = g_{a_p}$ - see our algorithm.*

Proof. As it is equivalent to prove either our algorithm or Huber's, we will only develop here the proof for our algorithm. Assuming, without any loss of generality, that the a_i, $i = 1, .., p$, are the vectors of the canonical basis, since $g^{(p-1)}(x) = g(x)\frac{f_1(x_1)}{g_1(x_1)}\frac{f_2(x_2)}{g_2(x_2)}...\frac{f_{p-1}(x_{p-1})}{g_{p-1}(x_{p-1})}$ we derive immediately that $g_p^{(p-1)} = g_p$. We note that it is sufficient to operate a change in basis on the a_i to obtain the general case. □

Proof of lemma F.8.

lemme F.8. *If there exits p, $p \leq d$, such that $\Phi(g^{(p)}, f) = 0$, then the family of $(a_i)_{i=1,..,p}$ - derived from the construction of $g^{(p)}$ - is free and orthogonal.*

Proof. Without any loss of generality, let us assume that $p = 2$ and that the a_i are the vectors of the canonical basis. Using a reductio ad absurdum with the hypotheses $a_1 = (1, 0, ..., 0)$ and that $a_2 = (\alpha, 0, ..., 0)$, where $\alpha \in \mathbb{R}$, we get $g^{(1)}(x) = g(x_2, .., x_d/x_1)f_1(x_1)$ and $f = g^{(2)}(x) = g(x_2, .., x_d/x_1)f_1(x_1)\frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}$. Hence $f(x_2, .., x_d/x_1) = g(x_2, .., x_d/x_1)\frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}$.
It consequently implies that $f_{\alpha a_1}(\alpha x_1) = [g^{(1)}]_{\alpha a_1}(\alpha x_1)$ since
$1 = \int f(x_2, .., x_d/x_1)dx_2...dx_d = \int g(x_2, .., x_d/x_1)dx_2...dx_d\frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)} = \frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}$.
Therefore, $g^{(2)} = g^{(1)}$, i.e. $p = 1$ which leads to a contradiction. Hence, the family is free.
Moreover, using a reductio ad absurdum we get the orthogonality. Indeed, we have
$\int f(x)dx = 1 \neq +\infty = \int n(a_{j+1}^\top x, ..., a_d^\top x)h(a_1^\top x, ..., a_j^\top x)dx$. The use of the same argument as in the proof of lemma F.2, enables us to infer the orthogonality of $(a_i)_{i=1,..,p}$. □

Proof of lemma F.9.

lemme F.9. *If there exits p, $p \leq d$, such that $\Phi(g^{(p)}, f) = 0$, where $g^{(p)}$ is built from the free and orthogonal family a_1,...,a_j, then, there exists a free and orthogonal family $(b_k)_{k=j+1,...,d}$ of vectors of $\mathbb{R}_*^d$, such that $g^{(p)}(x) = g(b_{j+1}^\top x, ..., b_d^\top x/a_1^\top x, ..., a_j^\top x)f_{a_1}(a_1^\top x)...f_{a_j}(a_j^\top x)$*
and such that $\mathbb{R}^d = Vect\{a_i\} \overset{\perp}{\oplus} Vect\{b_k\}$.

Proof. Through the incomplete basis theorem and similarly as in lemma F.8, we obtain the result thanks to the Fubini's theorem. □

Proof of lemma F.10.

lemme F.10. *For any continuous density f, we have $y_m = |f_m(x) - f(x)| = O_{\mathbf{P}}(m^{-\frac{2}{4+d}})$.*

Defining $b_m(x)$ as $b_m(x) = |E(f_m(x)) - f(x)|$, we have $y_m \leq |f_m(x) - E(f_m(x))| + b_m(x)$. Moreover, from page 150 of Scott (1992), we derive that $b_m(x) = O_{\mathbf{P}}(\Sigma_{j=1}^{d} h_j^2)$ where $h_j = O_{\mathbf{P}}(m^{-\frac{1}{4+d}})$. Then, we obtain $b_m(x) = O_{\mathbf{P}}(m^{-\frac{2}{4+d}})$. Finally, since the central limit theorem rate is $O_{\mathbf{P}}(m^{-\frac{1}{2}})$, we infer that $y_m \leq O_{\mathbf{P}}(m^{-\frac{1}{2}}) + O_{\mathbf{P}}(m^{-\frac{2}{4+d}}) = O_{\mathbf{P}}(m^{-\frac{2}{4+d}})$. □

Proof of proposition 3.3. Proposition 3.3 comes immediately from proposition B.1 page 20 and lemma C.1 page 20. □

Proof of proposition 3.4. Let us first show by induction the following assertion

$\mathcal{P}(k) = \{g^{(k)}$ allows a deconvolution $g^{(k)} = \overline{g}^{(k)} * \phi\}$

Initialisation : For $k = 0$, we get the result since $g = g^{(0)}$ is elliptic.

Going from k to $k+1$: Let us assume $\mathcal{P}(k)$ is true, we then show that $\mathcal{P}(k+1)$.

Since the family of a_i, $i \leq k+1$ is free - see lemma F.8 - then, we define B as the basis of $\mathbb{R}^d$ such that its $k+1$ first vectors are the a_i, $i \leq k+1$ - see the incomplete basis theorem for its existence. Thus, in B and using the same procedure to prove lemma F.1 page 24, we have $\overline{g}^{(k)}(x) = \overline{g}^{(k)}(./x_{k+1})\overline{g}_{k+1}^{(k)}(x_{k+1})$. Consequently, the very definition of the convolution product, the Fubini's theorem and the hypothesis made on the Elliptical family imply that $g^{(k)}(x) = g^{(k)}(./x_{k+1})g_{k+1}^{(k)}(x_{k+1})$ with $g^{(k)}(./x_{k+1}) = \overline{g}^{(k)}(./x_{k+1}) * E_{d-1}(0, \sigma^2 I_{d-1}, \xi_{d-1})$ and with $g_{k+1}^{(k)}(x_{k+1}) = \overline{g}_{k+1}^{(k)}(x_{k+1}) * E_1(0, \sigma^2, \xi_1)$. Finally, replacing $g_{k+1}^{(k)}$ with $f_{k+1} = \overline{f}_{k+1} * E_1(0, \sigma^2, \xi_1)$, we conclude this induction with $g^{(k+1)} = g^{(k)}(./x_{k+1}) f_{k+1}(x_{k+1})$.

Now, let us consider ψ (rep. $\overline{\psi}$, $\psi^{(k)}$, $\overline{\psi}^{(k)}$) the characteristic function of f (resp. $\overline{f}$, $g^{(k)}$, $\overline{g}^{(k)}$). We then have $\psi(s) = \overline{\psi}(s)\Psi(\frac{1}{2}\sigma^2|s|^2)$ and $\psi^{(k)}(s) = \overline{\psi}^{(k)}(s)\Psi(\frac{1}{2}\sigma^2|s|^2)$. Hence, ψ and $\psi^{(k)}$ are less or equal to $\Psi(\frac{1}{2}\sigma^2|s|^2)$ which is integrable by hypothesis, i.e. ψ and $\psi^{(k)}$ are absolutely integrable. We then obtain $g^{(k)}(x) = (2\pi)^{-d} \int \psi^{(k)}(s) e^{-is^\top x} ds$ and $f(x) = (2\pi)^{-d} \int \psi(s) e^{-is^\top x} ds$.

Moreover, since the sequence $(\psi^{(k)})$ uniformly converges and since ψ and $\psi^{(k)}$ are less or equal to $\Psi(\frac{1}{2}\sigma^2|s|^2)$, then the dominated convergence theorem implies that $\lim_k |f(x) - g^{(k)}(x)| \leq (2\pi)^{-d} \int \lim_k |\psi(s) - \psi^{(k)}(s)| ds = 0$ a.s. i.e. $\lim_k sup_x |f(x) - g^{(k)}(x)| = 0$ a.s. Finally, since, by hypothesis, $(2\pi)^{-d} \int |\psi(s) - \psi^{(k)}(s)| ds \leq 2(2\pi)^{-d} \int \Psi(\frac{1}{2}\sigma^2|s|^2) ds < \infty$, then the above limit and the dominated convergence theorem imply that $\lim_k \int |f(x) - g^{(k)}(x)| dx = 0$. □

Proof of corollary 3.1. Through the dominated convergence theorem and through theorem 3.4, we get the result using a reductio ad absurdum. □

Proof of lemma F.11.

lemme F.11. *Let consider the sequence* (a_i) *defined in (2.3) page 6.*
We then have $\lim_n \lim_k K(\check{g}_n^{(k)} \frac{f_{a_k,n}}{[\check{g}^{(k)}]_{a_k,n}}, f_n) = 0$ *a.s.*

Proof. Trough the relationship (2.3) and through remark D.1 page 22 as well as the additive relation of proposition D.1, we can say that $0 \leq .. \leq K(g^{(\infty)}, f) \leq .. \leq K(g^{(k)}, f) \leq .. \leq K(g, f)$, where $g^{(\infty)} = \lim_k g^{(k)}$ which is a density by construction. And through proposition C.2, we obtain that $K(g^{(\infty)}, f) = 0$, i.e.

$0 = K(g^{(\infty)}, f) \leq \ldots \leq K(g^{(k)}, f) \leq \ldots \leq K(g, f)$, (*).

Moreover, let $(g_n^{(k)})_k$ be the sequence of densities such that $g_n^{(k)}$ is the kernel estimate of $g^{(k)}$. Since we derive from remark F.1 page 23 an integrable upper bound of $g_n^{(k)}$, for all k, which is greater than f - see also the definition of φ in the proof of theorem 3.4 -, then the dominated convergence theorem implies that, for any k, $\lim_n K(g_n^{(k)}, f_n) = K(g^{(k)}, f)$, i.e., from a certain given rank n_0, we have $0 \leq .. \leq K(g_n^{(\infty)}, f_n) \leq .. \leq K(g_n^{(k)}, f_n) \leq .. \leq K(g_n, f_n)$, (**).

Consequently, through lemma F.12 page 28, there exists a k such that

$0 \leq .. \leq K(\Psi_{n,k}^{(\infty)}, f_n) \leq .. \leq K(g_n^{(\infty)}, f_n) \leq .. \leq K(\Psi_{n,k-1}^{(\infty)}, f_n) \leq .. \leq K(g_n, f_n)$, (***)

where $\Psi_{n,k}^{(\infty)}$ is a density such that $\Psi_{n,k}^{(\infty)} = \lim_k g_n^{(k)}$.
Finally, through the dominated convergence theorem and taking the limit as n in (***) we get
$0 = K(g^{(\infty)}, f) = \lim_n K(g_n^{(\infty)}, f_n) \geq \lim_n K(\Psi_{n,k}^{(\infty)}, f_n) \geq 0$.
The dominated convergence theorem enables us to conclude:
$0 = \lim_n K(\Psi_{n,k}^{(\infty)}, f_n) = \lim_n \lim_k K(g_n^{(k)}, f_n)$. □

Proof of lemma F.12.

lemme F.12. *With the notation of the proof of lemma F.11, we have*
$0 \leq .. \leq K(\Psi_{n,k}^{(\infty)}, f_n) \leq .. \leq K(g_n^{(\infty)}, f_n) \leq .. \leq K(\Psi_{n,k-1}^{(\infty)}, f_n) \leq .. \leq K(g_n, f_n)$, *(***)*

Proof. First, as explained in section D, we have $K(f^{(k)}, g) - K(f^{(k+1)}, g) = K(f_{a_{k+1}}^{(k)}, g_{a_{k+1}})$. Moreover, through remark D.1 page 22, we also derive that $K(f^{(k)}, g) = K(g^{(k)}, f)$. Then, $K(f_{a_{k+1}}^{(k)}, g_{a_{k+1}})$ is the decreasing step of the relative entropies in (*) and leading to $0 = K(g^{(\infty)}, f)$. Similarly, the very construction of (**), implies that $K(f_{a_{k+1},n}^{(k)}, g_{a_{k+1},n})$ is the decreasing step of the relative entropies in (**) and leading to $K(g_n^{(\infty)}, f_n)$.
Second, through the conclusion of the section D and lemma 14.2 of Huber's article, we obtain that $K(f_{a_{k+1},n}^{(k)}, g_{a_{k+1},n})$ converges - in decreasing and in k - towards a positive function of n - that we will call ξ_n.
Third, the convergence of $(g^{(k)})_k$ - see proposition C.2 - implies that, for any given n, the sequence $(K(g_n^{(k)}, f_n))_k$ is not finite. Then, through relationship (∗∗), there exists a k such that $0 < K(g_n^{(k-1)}, f_n) - K(g_n^{(\infty)}, f_n) < \xi_n$.
Thus, since $Q \mapsto K(Q, P)$ is l.s.c. - see property A.2 page 18 - relationship (**) implies (***). □

Proof of theorem 3.1. First, by the very definition of the kernel estimator $\breve{g}_n^{(0)} = g_n$ converges towards g. Moreover, the continuity of $a \mapsto f_{a,n}$ and $a \mapsto g_{a,n}$ and proposition 3.3 imply that $\breve{g}_n^{(1)} = \breve{g}_n^{(0)} \frac{f_{a,n}}{\breve{g}_{a,n}^{(0)}}$ converges towards $g^{(1)}$. Finally, since, for any k, $\breve{g}_n^{(k)} = \breve{g}_n^{(k-1)} \frac{f_{\breve{a}_k,n}}{\breve{g}_{\breve{a}_k,n}^{(k-1)}}$, we conclude by an immediat induction. □

Proof of theorem C.2.

relationship (C.1). Let us consider $\Psi_j = \{\frac{f_{\breve{a}_j}(\breve{a}_j{}^\top x)}{[\breve{g}^{(j-1)}]_{\breve{a}_j}(\breve{a}_j{}^\top x)} - \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}\}$. Since f and g are bounded, it is easy to prove that from a certain rank, we get, for any x given in $\mathbb{R}^d$
$|\Psi_j| \leq max(\frac{1}{[\breve{g}^{(j-1)}]_{\breve{a}_j}(\breve{a}_j{}^\top x)}, \frac{1}{[g^{(j-1)}]_{a_j}(a_j^\top x)}) |f_{\breve{a}_j}(\breve{a}_j{}^\top x) - f_{a_j}(a_j^\top x)|$.

Remark **F.2.** *First, based on what we stated earlier, for any given x and from a certain rank, there is a constant R>0 independent from n, such that*
$max(\frac{1}{[\breve{g}^{(j-1)}]_{\breve{a}_j}(\breve{a}_j{}^\top x)}, \frac{1}{[g^{(j-1)}]_{a_j}(a_j^\top x)}) \leq R = R(x) = O(1)$.
Second, since $\breve{a}_k$ is an M−estimator of a_k, its convergence rate is $O_{\mathbf{P}}(n^{-1/2})$.

Thus using simple functions, we infer an upper and lower bound for $f_{\breve{a}_j}$ and for f_{a_j}. We therefore reach the following conclusion:

$$|\Psi_j| \leq O_{\mathbf{P}}(n^{-1/2}). \tag{F.3}$$

We finally obtain
$|\Pi_{j=1}^k \frac{f_{\breve{a}_j}(\breve{a}_j{}^\top x)}{[\breve{g}^{(j-1)}]_{\breve{a}_j}(\breve{a}_j{}^\top x)} - \Pi_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}| = \Pi_{j=1}^k \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)} |\Pi_{j=1}^k \frac{f_{\breve{a}_j}(\breve{a}_j{}^\top x)}{[\breve{g}^{(j-1)}]_{\breve{a}_j}(\breve{a}_j{}^\top x)} \frac{[g^{(j-1)}]_{a_j}(a_j^\top x)}{f_{a_j}(a_j^\top x)} - 1|$.
Based on relationship (F.3), the expression $\frac{f_{\breve{a}_j}(\breve{a}_j{}^\top x)}{[\breve{g}^{(j-1)}]_{\breve{a}_j}(\breve{a}_j{}^\top x)} \frac{[g^{(j-1)}]_{a_j}(a_j^\top x)}{f_{a_j}(a_j^\top x)}$ tends towards 1 at a rate

of $O_{\mathbf{P}}(n^{-1/2})$ 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 $O_{\mathbf{P}}(n^{-1/2})$. Thus from a certain rank, we get

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

In conclusion, we obtain $|\check{g}^{(k)}(x) - g^{(k)}(x)| = g(x)|\Pi_{j=1}^{k} \frac{f_{\check{a}_j}(\check{a}_j^\top x)}{[\check{g}^{(j-1)}]_{\check{a}_j}(\check{a}_j^\top x)} - \Pi_{j=1}^{k} \frac{f_{a_j}(a_j^\top x)}{[g^{(j-1)}]_{a_j}(a_j^\top x)}| \leq O_{\mathbf{P}}(n^{-1/2})$.

relationship (C.2). The relationship C.1 of theorem C.2 implies that $|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1| = O_{\mathbf{P}}(n^{-1/2})$ because, for any given x, $g^{(k)}(x)|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1| = |\check{g}^{(k)}(x) - g^{(k)}(x)|$. Consequently, there exists a smooth function C of $\mathbb{R}^d$ in $\mathbb{R}^+$ such that

$\lim_{n\to\infty} n^{-1/2} C(x) = 0$ and $|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1| \leq n^{-1/2} C(x)$, for any x.

We then have $\int |\check{g}^{(k)}(x) - g^{(k)}(x)| dx = \int g^{(k)}(x)|\frac{\check{g}^{(k)}(x)}{g^{(k)}(x)} - 1| dx \leq \int g^{(k)}(x) C(x) n^{-1/2} dx$.

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

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

This implies that $\sup_{x\in\mathbb{R}^d} g^{(k)}(x) C(x) < \infty$ *a.s.*, i.e. $\sup_{x\in\mathbb{R}^d} C(x) < \infty$ *a.s.* since $g^{(k)}$ has been assumed to be positive and bounded - see remark F.1.

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 $\int |\check{g}^{(k)}(x) - g^{(k)}(x)| dx \leq \sup C . n^{-1/2} = O_{\mathbf{P}}(n^{-1/2})$. □

relationship (C.3). We have

$K(\check{g}^{(k)}, f) - K(g^{(k)}, f) = \int f(\varphi(\frac{\check{g}^{(k)}}{f}) - \varphi(\frac{g^{(k)}}{f})) dx \leq \int f\, S |\frac{\check{g}^{(k)}}{f} - \frac{g^{(k)}}{f}| dx = S \int |\check{g}^{(k)} - g^{(k)}| dx$

with the line before last being derived from theorem A.1 page 18 and where $\varphi : x \mapsto x ln(x) - x + 1$ is a convex function and where $S > 0$. We get the same expression as the one found in our Proof of Relationship (C.2) section, we then obtain $K(\check{g}^{(k)}, f) - K(g^{(k)}, f) \leq O_{\mathbf{P}}(n^{-1/2})$. Similarly, we get $K(g^{(k)}, f) - K(\check{g}^{(k)}, f) \leq O_{\mathbf{P}}(n^{-1/2})$. We can therefore conclude. □

Proof of lemma F.13.

lemme F.13. *We keep the notations introduced in Appendix B. It holds* $n = O(m^{\frac{1}{2}})$.

Proof. Let N be the random variable such that

$N = \Sigma_{j=1}^{m} \mathbf{1}_{\{f_m(X_j)\geq\theta_m,\ g(Y_j)\geq\theta_m\}}$. Since the events $\{f_m(X_j) \geq \theta_m\}$ and $\{g(Y_j) \geq \theta_m\}$ are independent from one another and since $\{g(Y_j) \geq \theta_m\} \subset \{g_m(Y_j) \geq -y_m + \theta_m\}$, we can say that

$n = m.\mathbf{P}(f_m(X_j) \geq \theta_m,\ g(Y_j) \geq \theta_m) \leq m.\mathbf{P}(f_m(X_j) \geq \theta_m).\mathbf{P}(g_m(Y_j) \geq -y_m + \theta_m)$.

Consequently, let us study $\mathbf{P}(f_m(X_i) \geq \theta_m)$. Let $(\xi_i)_{i=1\ldots m}$ be the sequence such that, for any i and any x in $\mathbb{R}^d$, $\xi_i(x) = \Pi_{l=1}^{d} \frac{1}{(2\pi)^{1/2} h_l} e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} - \int \Pi_{l=1}^{d} \frac{1}{(2\pi)^{1/2} h_l} e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x) dx$. Hence, for any given j and conditionally to $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_m$, the variables $(\xi_i(X_j))_{i=1\ldots m}^{i\neq j}$ are i.i.d. and centered, have same second moment, and are such that

$|\xi_i(X_j)| \leq \Pi_{l=1}^{d} \frac{1}{(2\pi)^{1/2} h_l} + \Pi_{l=1}^{d} \frac{1}{(2\pi)^{1/2} h_l} \int |f(x)| dx = 2.(2\pi)^{-d/2} \Pi_{l=1}^{d} h_l^{-1}$ since $\sup_x e^{-\frac{1}{2}x^2} \leq 1$.

Moreover, noting that $f_m(x) = \frac{1}{m}\Sigma_{i=1}^{m} \xi_i(x) + (2\pi)^{-d/2} \frac{1}{m} \Sigma_{i=1}^{m} \Pi_{l=1}^{d} h_l^{-1} \int e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x) dx$,

we have $f_m(X_j) \geq \theta_m \Leftrightarrow \frac{1}{m}\Sigma_{i=1}^{m} \xi_i(X_j) + (2\pi)^{-d/2} \frac{1}{m} \Sigma_{i=1}^{m} \Pi_{l=1}^{d} h_l^{-1} \int e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x) dx \geq \theta_m$

$\Leftrightarrow \frac{1}{m-1} \Sigma_{\substack{i=1 \\ i\neq j}}^{m} \xi_i(X_j) \geq (\theta_m - (2\pi)^{-d/2} \frac{1}{m} \Sigma_{i=1}^{m} \Pi_{l=1}^{d} h_l^{-1} \int e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x) dx - \frac{1}{m} \xi_j(X_j)) \frac{m}{m-1}$

with $\xi_j(X_j) = 0$. Then, defining t (resp. ε) as $t = 2.(2\pi)^{-d/2} \Pi_{l=1}^{d} h_l^{-1}$ (resp.

$\varepsilon = (\theta_m - (2\pi)^{-d/2} \Pi_{l=1}^{d} h_l^{-1} \frac{1}{m} \Sigma_{i=1}^{m} \Pi_{l=1}^{d} \int e^{-\frac{1}{2}(\frac{x_l - X_{il}}{h_l})^2} f(x) dx) \frac{m}{m-1}$), the Bennet's inequality -Devroye (1985) page 160- implies that $\mathbf{P}(\frac{1}{m-1} \Sigma_{\substack{i=1 \\ i\neq j}}^{m} \xi_i(X_j) \geq \varepsilon / X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_m) \leq 2.exp(-\frac{(m-1)\varepsilon^2}{4t^2})$.

Finally, since the X_i are i.i.d. and since $\int(\int \Pi_{l=1}^d e^{-\frac{1}{2}(\frac{x_l-y_l}{h_l})^2} \ f(x)dx)f(y)dy < 1$, then the law of large numbers implies that $\frac{1}{m}\Sigma_{i=1}^m \int \Pi_{l=1}^d e^{-\frac{1}{2}(\frac{x_l-X_{il}}{h_l})^2} \ f(x)dx \to_m \int\int \Pi_{l=1}^d e^{-\frac{1}{2}(\frac{x_l-y_l}{h_l})^2} f(x)f(y)dxdy$ a.s. Consequently, since $0 < \nu < \frac{1}{4+d}$ - see remark B.1 - and since $e^{-x} \leq x^{-\frac{1}{2}}$ when $x > 0$, we obtain, after calculation, that, from a certain rank, $exp(-\frac{(m-1)\varepsilon^2}{4t^2}) = O(m^{-\frac{1}{4}})$, i.e., from a certain rank, $\mathbf{P}(f_m(Y_j) \geq \theta_m) = O(m^{-\frac{1}{4}})$. Similarly, we infer $\mathbf{P}(g(Y_j) \geq \theta_m) = O(m^{-\frac{1}{4}})$. In conclusion, we can say that $n = m.\mathbf{P}(f_m(X_j) \geq \theta_m).\mathbf{P}(g_m(Y_j) \geq \theta_m) = O(m^{\frac{1}{2}})$. Similarly, we derive the same result as above for any step of our method. □

Proof of theorem 3.2. First, from lemma F.10, we derive that, for any x,
$\sup_{a\in\mathbb{R}_*^d} |f_{a,n}(a^\top x) - f_a(a^\top x)| = O_{\mathbf{P}}(n^{-\frac{2}{4+d}})$. Then, let us consider $\Psi_j = \frac{f_{\check{a}_j,n}(\check{a}_j{}^\top x)}{\check{g}_{\check{a}_j,n}^{(j-1)}(\check{a}_j{}^\top x)} - \frac{f_{a_j}(a_j^\top x)}{g_{a_j}^{(j-1)}(a_j^\top x)}$, we have
$\Psi_j = \frac{1}{\check{g}_{\check{a}_j,n}^{(j-1)}(\check{a}_j{}^\top x) g_{a_j}^{(j-1)}(a_j^\top x)} ((f_{\check{a}_j,n}(\check{a}_j{}^\top x) - f_{a_j}(a_j^\top x))g_{a_j}^{(j-1)}(a_j^\top x) + f_{a_j}(a_j^\top x)(g_{a_j}^{(j-1)}(a_j^\top x) - \check{g}_{\check{a}_j,n}^{(j-1)}(\check{a}_j{}^\top x)))$,
i.e. $|\Psi_j| = O_{\mathbf{P}}(n^{-\frac{1}{2}\mathbf{1}_{d=1}-\frac{2}{4+d}\mathbf{1}_{d>1}})$ since $f_{a_j}(a_j^\top x) = O(1)$ and $g_{a_j}^{(j-1)}(a_j^\top x) = O(1)$. We can therefore conclude similarly as in theorem C.2. □

Proof of theorem 3.3. We get ththeorem through theorem C.3 and proposition B.1. □

Proof of theorem C.3. First of all, let us remark that hypotheses $(H1)$ to $(H3)$ imply that $\check{\gamma}_n$ and $\check{c}_n(a_k)$ converge towards a_k in probability.
Hypothesis $(H4)$ enables us to derive under the integrable sign after calculation,
$\mathbf{P}\frac{\partial}{\partial b}M(a_k,a_k) = \mathbf{P}\frac{\partial}{\partial a}M(a_k,a_k) = 0$,
$\mathbf{P}\frac{\partial^2}{\partial a_i\partial b_j}M(a_k,a_k) = \mathbf{P}\frac{\partial^2}{\partial b_j\partial a_i}M(a_k,a_k) = \int \varphi"(\frac{gf_{a_k}}{fg_{a_k}})\frac{\partial}{\partial a_i}\frac{gf_{a_k}}{fg_{a_k}}\frac{\partial}{\partial b_j}\frac{gf_{a_k}}{fg_{a_k}} \ f\,dx$,
$\mathbf{P}\frac{\partial^2}{\partial a_i\partial a_j}M(a_k,a_k) = \int \varphi'(\frac{gf_{a_k}}{fg_{a_k}})\frac{\partial^2}{\partial a_i\partial a_j}\frac{gf_{a_k}}{fg_{a_k}} \ f\,dx$,
$\mathbf{P}\frac{\partial^2}{\partial b_i\partial b_j}M(a_k,a_k) = -\int \varphi"(\frac{gf_{a_k}}{fg_{a_k}})\frac{\partial}{\partial b_i}\frac{gf_{a_k}}{fg_{a_k}}\frac{\partial}{\partial b_j}\frac{gf_{a_k}}{fg_{a_k}} \ f\,dx$,
and consequently $\mathbf{P}\frac{\partial^2}{\partial b_i\partial b_j}M(a_k,a_k) = -\mathbf{P}\frac{\partial^2}{\partial a_i\partial b_j}M(a_k,a_k) = -\mathbf{P}\frac{\partial^2}{\partial b_j\partial a_i}M(a_k,a_k)$, which implies,
$\frac{\partial^2}{\partial a_i\partial a_j}K(g\frac{f_{a_k}}{g_{a_k}},f) = \mathbf{P}\frac{\partial^2}{\partial a_i\partial a_j}M(a_k,a_k) - \mathbf{P}\frac{\partial^2}{\partial b_i\partial b_j}M(a_k,a_k)$,
$= \mathbf{P}\frac{\partial^2}{\partial a_i\partial a_j}M(a_k,a_k) + \mathbf{P}\frac{\partial^2}{\partial a_i\partial b_j}M(a_k,a_k), = \mathbf{P}\frac{\partial^2}{\partial a_i\partial a_j}M(a_k,a_k) + \mathbf{P}\frac{\partial^2}{\partial b_j\partial a_i}M(a_k,a_k)$.
The very definition of the estimators $\check{\gamma}_n$ and $\check{c}_n(a_k)$, implies that $\begin{cases} \mathbb{P}_n\frac{\partial}{\partial b}M(b,a) = 0 \\ \mathbb{P}_n\frac{\partial}{\partial a}M(b(a),a) = 0 \end{cases}$
ie $\begin{cases} \mathbb{P}_n\frac{\partial}{\partial b}M(\check{c}_n(a_k),\check{\gamma}_n) = 0 \\ \mathbb{P}_n\frac{\partial}{\partial a}M(\check{c}_n(a_k),\check{\gamma}_n) + \mathbb{P}_n\frac{\partial}{\partial b}M(\check{c}_n(a_k),\check{\gamma}_n)\frac{\partial}{\partial a}\check{c}_n(a_k) = 0, \end{cases}$ i.e. $\begin{cases} \mathbb{P}_n\frac{\partial}{\partial b}M(\check{c}_n(a_k),\check{\gamma}_n) = 0 \ (E0) \\ \mathbb{P}_n\frac{\partial}{\partial a}M(\check{c}_n(a_k),\check{\gamma}_n) = 0 \ (E1) \end{cases}$.
Under $(H5)$ and $(H6)$, and using a Taylor development of the $(E0)$ (resp. $(E1)$) equation, we infer there exists $(\overline{c}_n,\overline{\gamma}_n)$ (resp. $(\tilde{c}_n,\tilde{\gamma}_n)$) on the interval $[(\check{c}_n(a_k),\check{\gamma}_n),(a_k,a_k)]$ such that
$-\mathbb{P}_n\frac{\partial}{\partial b}M(a_k,a_k) = [(\mathbf{P}\frac{\partial^2}{\partial b\partial b}M(a_k,a_k))^\top + o_{\mathbf{P}}(1), (\mathbf{P}\frac{\partial^2}{\partial a\partial b}M(a_k,a_k))^\top + o_{\mathbf{P}}(1)]a_n$.
(resp. $-\mathbb{P}_n\frac{\partial}{\partial a}M(a_k,a_k) = [(\mathbf{P}\frac{\partial^2}{\partial b\partial a}M(a_k,a_k))^\top + o_{\mathbf{P}}(1), (\mathbf{P}\frac{\partial^2}{\partial a^2}M(a_k,a_k))^\top + o_{\mathbf{P}}(1)]a_n$)
with $a_n = ((\check{c}_n(a_k) - a_k)^\top, (\check{\gamma}_n - a_k)^\top)$. Thus we get

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

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

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

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. □

Proof of proposition C.2. Let us consider ψ (resp. $\psi^{(k)}$) the characteristic function of f (resp. $g^{(k-1)}$). Let also consider the sequence (a_i) defined in (2.3) page 6.
We have $|\psi(t) - \psi^{(k)}(t)| \leq \int |f(x) - g^{(k)}(x)|dx \leq K(g^{(k)}, f)$. As explained in section 14 of Huber's article and through remark D.1 page 22 as well as through the additive relation of proposition D.1, we can say that $\lim_k K(g^{(k-1)} \frac{f_{a_k}}{[g^{(k-1)}]_{a_k}}, f) = 0$. Consequently, we get $\lim_k g^{(k)} = f$.

Proof of theorem 3.4. We recall that $g_n^{(k)}$ is the kernel estimator of $\breve{g}^{(k)}$. Since the relative entropy is greater than the L^1-distance, we then have

$\lim_n \lim_k K(g_n^{(k)}, f_n) \geq \lim_n \lim_k \int |g_n^{(k)}(x) - f_n(x)|dx$

Moreover, the Fatou's lemma implies that

$\lim_k \int |g_n^{(k)}(x) - f_n(x)|dx \geq \int \lim_k [|g_n^{(k)}(x) - f_n(x)|]dx = \int |[\lim_k g_n^{(k)}(x)] - f_n(x)|dx$

and $\lim_n \int |[\lim_k g_n^{(k)}(x)] - f_n(x)|dx \geq \int \lim_n [|[\lim_k g_n^{(k)}(x)] - f_n(x)|]dx$

$= \int |[\lim_n \lim_k g_n^{(k)}(x)] - \lim_n f_n(x)|dx.$

Trough lemma F.11, we then obtain that $0 = \lim_n \lim_k K(g_n^{(k)}, f_n) \geq \int |[\lim_n \lim_k g_n^{(k)}(x)] - \lim_n f_n(x)|dx \geq 0$, i.e. that $\int |[\lim_n \lim_k g_n^{(k)}(x)] - \lim_n f_n(x)|dx = 0$.
Moreover, for any given k and any given n, the function $g_n^{(k)}$ is a convex combination of multivariate Gaussian distributions. As derived at remark 2.1 of page 5, for all k, the determinant of the covariance of the random vector - with density $g^{(k)}$ - is greater than or equal to the product of a positive constant times the determinant of the covariance of the random vector with density f. The form of the kernel estimate therefore implies that there exists an integrable function φ such that, for any given k and any given n, we have $|g_n^{(k)}| \leq \varphi$.
Finally, the dominated convergence theorem enables us to say that $\lim_n \lim_k g_n^{(k)} = \lim_n f_n = f$, since f_n converges towards f and since $\int |[\lim_n \lim_k g_n^{(k)}(x)] - \lim_n f_n(x)|dx = 0$. □

Proof of theorem C.4. 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):

$\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) = \mathbb{P}_n M(a_k, a_k) + \mathbb{P}_n \frac{\partial}{\partial a} M(a_k, a_k)(\check{\gamma}_n - a_k)^\top + \mathbb{P}_n \frac{\partial}{\partial b} M(a_k, a_k)(\check{c}_n(a_k) - a_k)^\top$

$+\frac{1}{2}\{(\check{\gamma}_n - a_k)^\top \mathbb{P}_n \frac{\partial^2}{\partial a \partial a} M(a_k, a_k)(\check{\gamma}_n - a_k) + (\check{c}_n(a_k) - a_k)^\top \mathbb{P}_n \frac{\partial^2}{\partial b \partial a} M(a_k, a_k)(\check{\gamma}_n - a_k)$

$+(\check{\gamma}_n - a_k)^\top \mathbb{P}_n \frac{\partial^2}{\partial a \partial b} M(a_k, a_k)(\check{c}_n(a_k) - a_k) + (\check{c}_n(a_k) - a_k)^\top \mathbb{P}_n \frac{\partial^2}{\partial b \partial b} M(a_k, a_k)(\check{c}_n(a_k) - a_k)\}$

The lemma below enables us to conclude.

lemme F.14. *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 $\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) = \mathbb{P}_n M(a_k, a_k) + O_{\mathbf{P}}(\frac{1}{n})$,
i.e. $\sqrt{n}(\mathbb{P}_n M(\check{c}_n(a_k), \check{\gamma}_n) - \mathbf{P}M(a_k, a_k)) = \sqrt{n}(\mathbb{P}_n M(a_k, a_k) - \mathbf{P}M(a_k, a_k)) + o_{\mathbf{P}}(1)$.
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 theorem 3.5. Through proposition B.1 and theorem C.4, we derive theorem 3.5.. □

Proof of theorem 3.7. We immediately get the proof from theorem 3.4. □

Proof of theorem 3.8. Since $\Phi(g^{(1)}, f) = 0$, then, through lemma F.9, we deduct that the density of $b_2^\top X / a_1^\top X$, with $a_1 = (0, 1)'$ and $b_2 = (1, 0)'$, is the same as the one of $b_2^\top Y / a_1^\top Y$.
Hence, we derive that $E(X_1/X_2) = E(Y_1/Y_2)$ and also that the regression between X_1 and X_2 is $X_1 = E(Y_1) + \frac{Cov(Y_1, Y_2)}{Var(Y_2)}(Y_2 - E(Y_2)) + \varepsilon$, where ε is a centered random variable such that it is

orthogonal to $E(X_1/X_2)$. □

Proof of theorem 3.9. We infer this proof similarly to the proof of theorem 3.8 section. □

Proof of corollary 3.4. Assuming first that the b_k and the a_i are the canonical basis of $\mathbb{R}^d$. Then, for any $i \neq j$, Y_i is independent from Y_j, i.e. $E(Y_k/Y_1, ..., Y_j) = E(Y_k)$. Consequently, the regression between X_k and $(X_1, ..., X_j)$ is given by $X_k = E(Y_k) + \varepsilon_k$ where ε is a centered random variable such that it is orthogonal to $E(X_k/X_1, ..., X_j)$.

At present, we derive the general case thanks to the methodology used in the proof of lemma F.1 section with the transformation matrix $B = (a_1, ..., a_j, b_{j+1}, ..., b_d)$. □

References

Azé D., *Eléments d'analyse convexe et variationnelle, Ellipse, 1997.*

Bosq D., Lecoutre J.-P. *Livre - Theorie De L'Estimation Fonctionnelle, Economica, 1999.*

Broniatowski M., Keziou A. *Parametric estimation and tests through divergences and the duality technique. J. Multivariate Anal. 100 (2009), no. 1, 16–36.*

Cambanis, Stamatis; Huang, Steel; Simons, Gordon. *On the theory of elliptically contoured distributions. J. Multivariate Anal. 11 (1981), no. 3, 368–385.*

Cressie, Noel; Read, Timothy R. C. *Multinomial goodness-of-fit tests. J. Roy. Statist. Soc. Ser. B 46 (1984), no. 3, 440–464.*

Csiszár, I. *On topology properties of f-divergences. Studia Sci. Math. Hungar. 2 1967 329–339.*

Devroye, Luc; Györfi, László. Distribution free exponential bound for the L_1 error of partitioning-estimates of a regression function. Probability and statistical decision theory, Vol. A (Bad Tatzmannsdorf, 1983), 67–76, Reidel, Dordrecht, 1985

Diaconis, Persi; Freedman, David. *Asymptotics of graphical projection pursuit. Ann. Statist. 12 (1984), no. 3, 793–815.*

Jean Dieudonné, *Calcul infinitésimal. Hermann. 1980*

Friedman, Jerome H.; Stuetzle, Werner; Schroeder, Anne. *Projection pursuit density estimation. J. Amer. Statist. Assoc. 79 (1984), no. 387, 599–608.*

Friedman, Jerome H. *Exploratory projection pursuit. J. Amer. Statist. Assoc. 82 (1987), no. 397, 249–266.*

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.*

Huber Peter J., *Robust Statistics. Wiley, 1981 (republished in paperback, 2004)*

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

Landsman, Zinoviy M.; Valdez, Emiliano A. *Tail conditional expectations for elliptical distributions. N. Am. Actuar. J. 7 (2003), no. 4, 55–71.*

Pardo Leandro, *Statistical inference based on divergence measures, volume 185 of Statistics: Textbooks and Monographs. Chapman & Hall/CRC, Boca Raton, FL, 2006.*

Rockafellar, R. Tyrrell., *Convex analysis. Princeton Mathematical Series, No. 28 Princeton University Press, Princeton, N.J. 1970 xviii+451 pp.*

Saporta Gilbert, *Probabilités, analyse des données et statistique, Technip, 2006.*

Scott, David W., *Multivariate density estimation. Theory, practice, and visualization. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. A Wiley-Interscience Publication. John Wiley and Sons, Inc., New York, 1992. xiv+317 pp. ISBN: 0-471-54770-0.*

Aida Toma *Optimal robust M-estimators using divergences. Statistics and Probability Letters, Volume 79, Issue 1, 1 January 2009, Pages 1-5*

Vajda, Igor, *χ^α-divergence and generalized Fisher's information. Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Tech. Univ. Prague, Prague, 1971; dedicated to the memory of Antonín Spacek), pp. 873–886. Academia, Prague, 1973.*

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

Zhu, Mu. *On the forward and backward algorithms of projection pursuit. Ann. Statist. 32 (2004), no. 1, 233–244.*

Victor J. Yohai *Optimal robust estimates using the Kullback-Leibler divergence. Statistics and Probability Letters, Volume 78, Issue 13, 15 September 2008, Pages 1811-1816.*

Zografos, K. and Ferentinos, K. and Papaioannou, T., *φ-divergence statistics: sampling properties and multinomial goodness of fit and divergence tests, Comm. Statist. Theory Methods,19(5):1785–1802(1990), MR1075502.*