

Article

Projection Pursuit Through ϕ -Divergence Minimisation

Jacques Touboul

Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, 175 rue du Chevaleret, 75013 Paris, France; E-Mail: jack_touboul@hotmail.com

Received: 8 April 2010; in revised form: 27 May 2010 / Accepted: 31 May 2010 /

Published: 14 June 2010

Abstract: In his 1985 article (“Projection pursuit”), Huber demonstrates the interest of his method to estimate a density from a data set in a simple given case. He considers the factorization of density through a Gaussian component and some residual density. Huber’s work is based on maximizing Kullback–Leibler divergence. 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.

Keywords: projection pursuit; minimum ϕ -divergence; elliptical distribution; goodness-of-fit; copula; regression

Classification: 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 transformed through a Gaussianization. Through a recursive approach, this process is iterated to find another structure in the remaining data, until no further structure can be evidenced in the data left at the end.

Friedman [1] and Huber [2] 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 Kullback–Leibler divergence.

For a very long time, the two methodologies exposed by each of the above authors were thought to be equivalent but Zhu [3] 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, *i.e.*, in case of density estimation. In the present article, we will therefore only focus on Huber’s study while taking into account the 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 Kullback–Leibler divergence. 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, \dots) of vectors of \mathbb{R}_*^d and a sequence of densities $f^{(i)}$.

Remark 1.1. Huber stops his algorithm when the Kullback–Leibler divergence 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.*, $f = g \prod_{i=1}^j \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$ since by induction $f^{(j)} = f \prod_{i=1}^j \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$. 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 $f = g \prod_{i=1}^d \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$.

He can also stop his algorithm when the Kullback–Leibler divergence equals zero without the condition $j \leq d$ is met. Therefore, since by induction we have $f^{(j)} = f \prod_{i=1}^j \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$ with $f^{(0)} = f$, we obtain

$g = f \prod_{i=1}^j \frac{g_{a_i}}{f_{a_i}^{(i-1)}}$. Consequently, we derive a representation of f as $f = g \prod_{i=1}^j \frac{f_{a_i}^{(i-1)}}{g_{a_i}}$.

Finally, he obtains $K(f^{(0)}, g) \geq K(f^{(1)}, g) \geq \dots \geq 0$ 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, \dots) 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 Kullback–Leibler divergence 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 Kullback–Leibler divergence equals zero, since $K(f, g^{(j)}) = 0$ implies $g^{(j)} = f$. Therefore, since by induction we have $g^{(j)} = g \prod_{i=1}^j \frac{f_{a_i}}{g_{a_i^{(i-1)}}}$ with $g^{(0)} = g$, we then obtain $f = g \prod_{i=1}^j \frac{f_{a_i}}{g_{a_i^{(i-1)}}}$.

Second, he gets $K(f, g^{(0)}) \geq K(f, g^{(1)}) \geq \dots \geq 0$ with $g^{(0)} = g$.

1.3. Proposal

Let us first introduce the concept of ϕ –divergence.

Let φ be a strictly convex function defined by $\varphi : \overline{\mathbb{R}^+} \rightarrow \overline{\mathbb{R}^+}$, and such that $\varphi(1) = 0$. We define a 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

$$D_\phi(Q, P) = \int \varphi\left(\frac{dQ}{dP}\right) dP$$

or $D_\phi(q, p) = \int \varphi\left(\frac{q(x)}{p(x)}\right) p(x) dx$, if P and Q present p and q as density respectively.

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 Appendix A.1 page 1604.

Now, let us introduce our algorithm.

We start with performing the $D_\phi(g, f) = 0$ 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

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

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, \dots) 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 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 representation of f from g : To obtain an approximation of f , we stop our algorithm when the divergence equals zero, i.e., $D_\phi(g^{(j)}, f) = 0$ 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 $g^{(j)} = g \prod_{i=1}^j \frac{f_{a_i}}{g_{a_i}^{(i-1)}}$ with $g^{(0)} = g$, we then obtain $f = g \prod_{i=1}^j \frac{f_{a_i}}{g_{a_i}^{(i-1)}}$.

Second, we get $D_\phi(g^{(0)}, f) \geq D_\phi(g^{(1)}, f) \geq \dots \geq 0$ 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 [4], Toma [5] as well as Huber [6].

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 $D_\phi(g \frac{f_3}{g_3}, f) = D_\phi(n \cdot f_3, f) = D_\phi(f, f) = 0$ as $f_3 = h$, i.e., the function $a \mapsto D_\phi(g \frac{f_a}{g_a}, f)$ reaches zero for $e_3 = (0, 0, 1)'$ —where f_3 and g_3 are the third marginal densities of f and g respectively.

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 $D_\phi(g^{(1)}, f) = 0$ 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

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

To recapitulate our method, if $D_\phi(g, f) = 0$, we derive f from the relationship $f = g$; should a sequence $(a_i)_{i=1, \dots, j}$, $j < d$, of vectors in \mathbb{R}_*^d defining $g^{(j)}$ and such that $D_\phi(g^{(j)}, f) = 0$ exist, then $f(\cdot/a_i^\top x, 1 \leq i \leq j) = g(\cdot/a_i^\top x, 1 \leq i \leq j)$, i.e., f coincides with g on the complement of the vector subspace generated by the family $\{a_i\}_{i=1, \dots, j}$ —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, \dots, 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 regression analysis. Finally, we will present simulations and an application to real datasets.

2. The Algorithm

2.1. The model

As explained by Friedman [1] and Diaconis [7], 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 ϕ -divergences.

Elliptical laws

The interest of this class lies in the fact that conditional densities with elliptical distributions are also elliptical—see Cambanis [8], Landsman [9]. 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 :

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

- with Σ , being a $d \times d$ positive-definite matrix and with μ , being a d -column vector,
- with ξ_d , being referred as the “density generator”,
- with c_d , being a normalisation constant, such that $c_d = \frac{\Gamma(d/2)}{(2\pi)^{d/2}} \left(\int_0^\infty x^{d/2-1} \xi_d(x) dx \right)^{-1}$,
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 an $m \times d$ 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 multivariate elliptical distribution is elliptic, i.e.,

$$X = (X_1, X_2, \dots, X_d) \sim E_d(\mu, \Sigma, \xi_d) \Rightarrow X_i \sim E_1(\mu_i, \sigma_i^2, \xi_1), f_{X_i}(x) = \frac{c_1}{\sigma_i} \xi_1\left(\frac{1}{2}\left(\frac{x - \mu_i}{\sigma_i}\right)^2\right), 1 \leq i \leq d.$$

2/ Corollary 5 of Cambanis [8] 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 [9] shows that multivariate Gaussian distributions derive from $\xi_d(x) = e^{-x}$. He also shows that if $X = (X_1, \dots, X_d)$ 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 a multiple of 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.

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 non-null linearly independent vectors a_j , with $1 \leq j \leq d$, of \mathbb{R}^d , such that

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

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 $X = (X_1, \dots, X_d)$ 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}, \dots, X_d) is independent from (X_1, \dots, X_j) . Hence, the density of (X_{j+1}, \dots, X_d) given (X_1, \dots, X_j) is n .

Let us assume that $D_\phi(g^{(j)}, f) = 0$, for some $j \leq d$. We then get $\frac{f(x)}{f_{a_1} f_{a_2} \dots f_{a_j}} = \frac{g(x)}{g_{a_1}^{(1-1)} g_{a_2}^{(2-1)} \dots g_{a_j}^{(j-1)}}$, since, by induction, we have $g^{(j)}(x) = g(x) \frac{f_{a_1}}{g_{a_1}^{(1-1)}} \frac{f_{a_2}}{g_{a_2}^{(2-1)}} \dots \frac{f_{a_j}}{g_{a_j}^{(j-1)}}$.

Consequently, the fact that conditional densities with elliptical distributions are also elliptical enables us to infer that

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

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

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 D.1—page 1607—implies that

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

which is equivalent to having $D_\phi(g^{(j)}, f) = 0$ —since by induction $g^{(j)} = g \frac{f_{a_1}}{g_{a_1}^{(1-1)}} \frac{f_{a_2}}{g_{a_2}^{(2-1)}} \dots \frac{f_{a_j}}{g_{a_j}^{(j-1)}}$.

The end of our algorithm implies that f coincides with g on the complement of the vector subspace generated by the family $\{a_i\}_{i=1, \dots, j}$. Therefore, the nullity of the ϕ -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, $D_\phi(g^{(j)}, f) = 0$ is equivalent to*

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

More generally, the above proposition leads us to defining the co-support of f as the vector space generated from vectors a_1, \dots, 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, \dots, a_j which solves the problem $D_\phi(g^{(j)}, f) = 0$ 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, \dots, 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 D.2*

2.2. Stochastic outline of our algorithm

Let X_1, X_2, \dots, X_m (resp. Y_1, Y_2, \dots, Y_m) be a sequence of m independent random vectors with same density f (resp. g). As customary in nonparametric ϕ -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, \dots, X_n and Y_1, Y_2, \dots, Y_n —extracted respectively from X_1, X_2, \dots, X_m and Y_1, Y_2, \dots, Y_m —since the estimates are bounded below by some positive deterministic sequence θ_m —see Appendix B.

Let \mathbb{P}_n be the empirical measure of the subsample X_1, X_2, \dots, 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, \dots, X_n (resp. $a^\top X_1, a^\top X_2, \dots, a^\top X_n$).

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 $a_k = \operatorname{arg\,min}_{a \in \mathbb{R}_*^d} D_\phi(g^{(k-1)} \frac{f_a}{g_a^{(k-1)}}, f)$ (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 ϕ -divergence between f_n and $g \frac{f_{a,n}}{g_a}$ and which estimates a_1 :

Proposition B.1 page 1606 and Lemma D.3 page 1607 enable us to minimize the ϕ -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 1589 shows us that this vector tends to a_1 .

Finally, we define the density $\check{g}_m^{(1)}$ as $\check{g}_m^{(1)} = g \frac{f_{\check{a}_1, m}}{g_{\check{a}_1}^{(1)}}$ 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, \dots, X_n (resp. $Y_1^{(1)}, Y_2^{(1)}, \dots, Y_n^{(1)}$) extracted from X_1, X_2, \dots, X_m (resp. $Y_1^{(1)}, Y_2^{(1)}, \dots, Y_m^{(1)}$)—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 Appendix B.

Let \mathbb{P}_n be the empirical measure of the subsample X_1, X_2, \dots, 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, \dots, X_n (resp. $Y_1^{(1)}, Y_2^{(1)}, \dots, Y_n^{(1)}$ and $a^\top X_1, a^\top X_2, \dots, a^\top X_n$ as well as $a^\top Y_1^{(1)}, a^\top Y_2^{(1)}, \dots, a^\top Y_n^{(1)}$). 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 ϕ -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 1606 and Lemma D.3 page 1607 enable us to minimize the ϕ -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 1589 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 $(\check{a}_1, \check{a}_2, \dots)$ of vectors in \mathbb{R}_*^d estimating the co-vectors 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 Appendix C.

In this section, to be more legible we replace g with $g^{(k-1)}$. Let

$$\Theta = \mathbb{R}^d, \Theta^{D_\phi} = \{b \in \Theta \mid \int \varphi^*(\varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)})) d\mathbf{P} < \infty\}$$

$$M(b, a, x) = \int \varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)}) g(x) \frac{f_a(a^\top x)}{g_a(a^\top x)} dx - \varphi^*(\varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)}))$$

$$\mathbb{P}_n M(b, a) = \int M(b, a, x) d\mathbb{P}_n, \mathbf{P} M(b, a) = \int M(b, a, x) d\mathbf{P}$$

where \mathbf{P} is the probability measure presenting f as density.

Similarly as in chapter V of Van der Vaart [10], let us define :

(H1) : For all $\varepsilon > 0$, there is $\eta > 0$, such that for all $c \in \Theta^{D_\phi}$ 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 $(n \geq n_0 \Rightarrow \sup_{a \in \Theta} \sup_{c \in \{\Theta^{D_\phi}\}^c} \mathbb{P}_n M(c, a) < Z)$

(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 $I_{a_k} = \frac{\partial^2}{\partial a^2} D_\phi(g \frac{f_{a_k}}{g_{a_k}}, f)$, and $x \rightarrow \rho(b, a, x) = \varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)}) \frac{g(x) f_a(a^\top x)}{g_a(a^\top x)}$, putting:

(H5) : The function φ is \mathcal{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 (λ -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 \mathcal{C}^3 in a neighbourhood V_k of (a_k, a_k) for all x ; and the partial derivatives of $(b, a) \mapsto M(b, a)$ are all dominated in V_k by a \mathbf{P} -integrable function $H(x)$.

(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.

(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 and such that $K(g, f) \geq \int |f(x) - g(x)| dx$.

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 $D_\phi(gr, f)$ in r :

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

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

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.

Following Broniatowski [11], let us introduce the estimate of $D_\phi(g_{g_a}^{f_{a,n}}, f_n)$, through

$$\check{D}_\phi(g_{g_a}^{f_{a,n}}, f_n) = \int M(a, a, x) d\mathbb{P}_n(x)$$

Proposition 3.2. Let \check{a} be such that $\check{a} := \arg \inf_{a \in \mathbb{R}^d} \check{D}_\phi(g_{g_a}^{f_{a,n}}, f_n)$.

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., $\check{g}_n^{(k)}(x) = g(x) \prod_{j=1}^k \frac{f_{\check{a}_j, n}(\check{a}_j^\top x)}{[\check{g}^{(j-1)}]_{\check{a}_j, n}(\check{a}_j^\top x)}$.

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 $\check{g}_n^{(k)} \rightarrow_n g^{(k)}$ 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 different cases, namely for any given x , with the L^1 distance and with the Kullback–Leibler divergence:

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 $|K(\check{g}_n^{(k)}, f) - K(g^{(k)}, f)| = O_{\mathbf{P}}(n^{-\frac{2}{2+d}})$.

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

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

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

where $\mathcal{A} = \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) (\mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k))$, $\mathcal{C} = \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k)$ and

$$\mathcal{B} = \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k).$$

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 $\check{g}_n^{(k)}$ 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 $D_\phi(g_n^{(k-1)} \frac{f_{a_k,n}}{g_{a_k,n}^{(k-1)}}, f_{a_k,n})$ converges towards zero as k and then as n go to infinity:

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

3.3.2. Testing of the criteria

In this paragraph, through a test of our criteria, namely $a \mapsto D_\phi(\check{g}_n^{(k)} \frac{f_{a,n}}{[\check{g}^{(k)}]_{a,n}}, f_n)$, 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

$$\sqrt{n}(\text{Var}_{\mathbb{P}}(M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)))^{-1/2}(\mathbb{P}_n M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n) - \mathbb{P}_n M(a_k, a_k)) \xrightarrow{\mathcal{L}aw} \mathcal{N}(0, I),$$

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. Thus, in the case when $D_\phi(g^{(k-1)} \frac{f_{a_k}}{g_{a_k}^{(k-1)}}, f) = 0$, we obtain

Corollary 3.2. We have $\sqrt{n}(\text{Var}_{\mathbb{P}}(M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n)))^{-1/2} \mathbb{P}_n M(\check{c}_n(\check{\gamma}_n), \check{\gamma}_n) \xrightarrow{\mathcal{L}aw} \mathcal{N}(0, I)$.

Hence, we propose the test of the null hypothesis

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

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

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

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 arising from a realization of the sequences (X_1, \dots, X_n) and (Y_1, \dots, 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 $D_\phi(g^{(2)}, f) = 0$ 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 D.4 page 1608—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.

Let us assume that our algorithm implies that $D_\phi(g^{(d)}, f) = 0$.

Hence, we have, for any $x \in \mathbb{R}^d$, $g(x) \prod_{k=1}^d \frac{f_{a_k}(a_k^\top x)}{[g^{(k-1)}]_{a_k}(a_k^\top x)} = f(x)$, i.e., $\frac{g(x)}{\prod_{k=1}^d g_{a_k}(a_k^\top x)} = \frac{f(x)}{\prod_{k=1}^d f_{a_k}(a_k^\top x)}$, since Lemma D.4 page 1608 implies that $g_{a_k}^{(k-1)} = g_{a_k}$ if $k \leq d$.

Moreover, the family $(a_i)_{i=1\dots d}$ is a basis of \mathbb{R}^d —see Lemma D.5 page 1608. Hence, putting $A = (a_1, \dots, a_d)$ 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 $\frac{\partial^d}{\partial y_1 \dots \partial y_d} \tilde{C}_f = \frac{\partial^d}{\partial y_1 \dots \partial y_d} \tilde{C}_g$.

Finally, we perform a statistical test of the null hypothesis $(H_0) : \frac{\partial^d}{\partial y_1 \dots \partial y_d} \tilde{C}_f = \frac{\partial^d}{\partial y_1 \dots \partial y_d} \tilde{C}_g$ versus the alternative $(H_1) : \frac{\partial^d}{\partial y_1 \dots \partial y_d} \tilde{C}_f \neq \frac{\partial^d}{\partial y_1 \dots \partial y_d} \tilde{C}_g$. Since, under (H_0) , we have $D_\phi(g^{(d)}, f) = 0$, then, as explained in Section 3.3, 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 data. 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 $d > 1$,

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

did indeed exist, with $j < d$, with (a_1, \dots, 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 $\sigma^2 \cdot I_d$, $\sigma^2 > 0$, 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 [9]. Let f be a density on \mathbb{R}^d which can be deconvoluted with ϕ , i.e., $f = \bar{f} * \phi = \int_{\mathbb{R}^d} \bar{f}(x) \phi(t - x) dt$, where \bar{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.,

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

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 \bar{f} :

Theorem 3.7. *It holds $\lim_n \lim_k \check{g}_n^{(k)} = \bar{f} * \phi$ 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, \dots, X_d) (resp. (Y_1, \dots, Y_d)) as a vector with density f (resp. g —see Section 2.1).

Remark 3.2. *In this paragraph, we will work in the L^2 space. Then, we will first only consider the ϕ -divergences which are greater than or equal to the L^2 distance—see Vajda [12]. Note also that the co-vectors of f can be obtained in the L^2 space—see Lemma D.3 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 3.3. *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 [14] implies that $E(Y_1/Y_2) = E(Y_1) + \frac{\text{Cov}(Y_1, Y_2)}{\text{Var}(Y_2)}(Y_2 - E(Y_2))$ and then $X_1 = E(Y_1) + \frac{\text{Cov}(Y_1, Y_2)}{\text{Var}(Y_2)}(Y_2 - E(Y_2)) + \varepsilon$.

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 D.6 implies the existence of an orthogonal and free family $(b_i)_{i=j+1, \dots, d}$ of \mathbb{R}_*^d such that $\mathbb{R}^d = \text{Vect}\{a_i\} \oplus \text{Vect}\{b_k\}$ and such that

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

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

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

Corollary 3.4. *If g is a Gaussian density with same mean and variance as f , and if $\text{Cov}(X_i, X_j) = 0$ for any $i \neq j$, then, the regression of $b_k^\top X$, $k = 1, \dots, d$, on $(a_1^\top X, \dots, a_j^\top X)$ 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 $E(b_k^\top X / a_1^\top X, \dots, a_j^\top X)$.*

4. Simulations

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

In each example, our program will follow our algorithm and will aim at creating a sequence of densities $(g^{(j)})$, $j = 1, \dots, k$, $k < d$, such that $g^{(0)} = g$, $g^{(j)} = g^{(j-1)} f_{a_j} / [g^{(j-1)}]_{a_j}$ and $D_\phi(g^{(k)}, f) = 0$, with D_ϕ being a divergence and $a_j = \arg \inf_b D_\phi(g^{(j-1)} f_b / [g^{(j-1)}]_b, f)$, for all $j = 1, \dots, k$. Moreover, in the second example, we will study the robustness of our method with two outliers. In the third and the fourth example, defining (X_0, X_1) as a vector with f as density, we will study the regression of X_1 on X_0 . And finally, in the fifth 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

$$f(x) = \text{Gaussian}(x_1 + x_2) \cdot \text{Gaussian}(x_0 + x_2) \cdot \text{Gumbel}(x_0 + x_1)$$

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 with a density—that we will name g —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:

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

Then, Corollary 3.3 enables us to estimate a_1 by the following $0.9(=\alpha)$ level confidence ellipsoid

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

And, we obtain

Table 1. Simulation 1: Numerical results of the optimisation.

Our Algorithm	
	minimum : 0.0201741
Projection Study 0 :	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 $x = (2, 0, \dots, 0)$ and 98 values of a random variable X with a density f defined by

$$f(x) = \text{Gumbel}(x_0) \cdot \text{Normal}(x_1, \dots, x_9)$$

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 $a_1 = (1, 0, \dots, 0)$. To get this result, we perform the following test

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

We estimate a_1 by the following $0.9(=\alpha)$ level confidence ellipsoid

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

And, we obtain

Table 2. Simulation 2: Numerical results of the optimisation.

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 : $H_0 : a_1 \in \mathcal{E}_1 : \text{True}$
$H(\text{Est. of } g^{(1)}, g^{(1)})$	3.042174

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

Simulation 4.3 (With the Cressie-Read divergence (D_{ϕ})).

We are in dimension $2(=d)$, and we consider a sample of $50(=n)$ values of a random variable

$X = (X_0, X_1)$ with a density law f defined by

$$f(x) = \text{Gumbel}(x_0) \cdot \text{Normal}(x_1)$$

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 with a density—that we will name g —presenting the 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

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

Then, Corollary 3.3 enables us to estimate a_1 by the following $0.9(=\alpha)$ level confidence ellipsoid

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

And, we obtain

Table 3. Simulation 3: Numerical results of the optimisation.

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

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

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

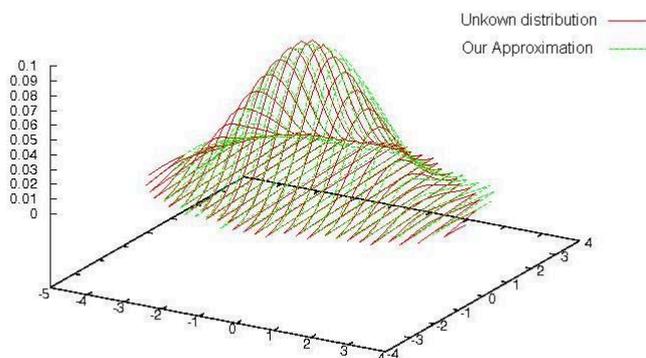
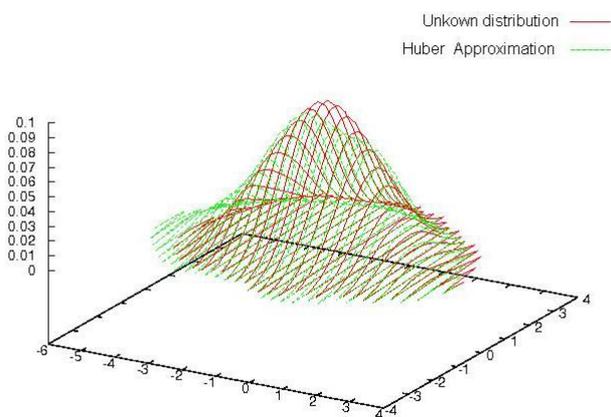


Figure 2. Graph of the distribution to estimate (red) and of Huber’s estimate (green).



At present, keeping the notations of this simulation, let us study the regression of X_1 on X_0 .

Our algorithm leads us to infer that the density of X_1 given X_0 is the same as the density of Y_1 given Y_0 . Moreover, Property A.1 implies that the co-factors of f are the same for any divergence.

Consequently, applying Theorem 3.8 implies that $X_1 = E(Y_1/Y_0) + \varepsilon$, where ε is a centered random variable orthogonal to $E(X_1/X_0)$.

Thus, since g is a Gaussian density, Remark 3.3 implies that

$$X_1 = E(Y_1) + \frac{Cov(Y_1, Y_0)}{Var(Y_0)}(Y_0 - E(Y_0)) + \varepsilon$$

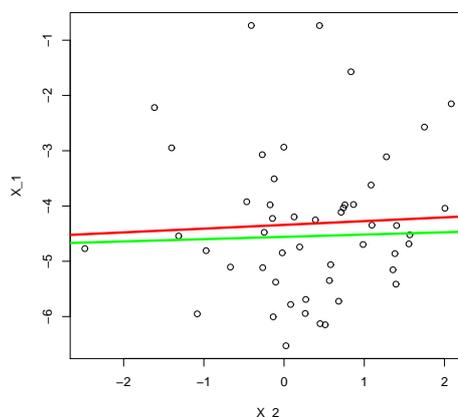
Now, using the least squares method, we estimate α_1 and α_2 such that $X_1 = \alpha_1 + \alpha_2 \cdot X_0 + \varepsilon$.

Thus, the following table presents the results of our regression and of the least squares method if we assume that ε is Gaussian.

Table 4. Simulation 3: Numerical results of the regression.

Our Regression	$E(Y_1)$	-4.545483
	$Cov(Y_1, Y_0)$	0.0380534
	$Var(Y_0)$	0.9190052
	$E(Y_0)$	0.3103752
	correlation (Y_1, Y_0)	0.02158213
Least squares method	α_1	-4.34159227
	Std Error of α_1	0.19870
	α_2	0.06803317
	Std Error of α_2	0.21154
	correlation (X_1, X_0)	0.04888484

Figure 3. Graph of the regression of X_1 on X_0 based on the least squares method (red) and based on our theory (green).



Simulation 4.4 (With the Cressie-Read divergence (D_ϕ)).

We are in dimension 2(=d), and we consider a sample of 500(=n) values of a random variable $X = (X_0, X_1)$ with a density law f defined by

$$f(x) = \text{Gumbel}(x_1 - x_0) \cdot \text{Normal}(x_1 + x_0)$$

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 with a density—that we will name g —presenting the 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

$H_0 : a_1 = (1, -1)$ versus $(H_1) : a_1 \neq (1, -1)$. Then, Corollary 3.3 enables us to estimate a_1 by the following $0.9(=\alpha)$ level confidence ellipsoid

$$\mathcal{E}_1 = \{b \in \mathbb{R}^2; (\text{Var}_{\mathbf{P}}(M(b, b)))^{(-1/2)} \mathbb{P}_n M(b, b) \leq q_{\alpha}^{N(0,1)} / \sqrt{n} \simeq 0, 2533 / \sqrt{500} = 0.01132792\}$$

And, we obtain

Table 5. Simulation 4: Numerical results of the optimisation.

Our Algorithm	
Projection Study 0 :	minimum : 0.010920
	at point : (1.09,-0.9701)
	P-Value : 0.889400
Test :	$H_0 : a_1 \in \mathcal{E}_1 : \text{True}$
D_{ϕ} (Kernel Estimation of $g^{(1)}, g^{(1)}$)	5.25077

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

At present, keeping the notations of this simulation, let us study the regression of $X_1 + X_0$ on $X_1 - X_0$. Our algorithm leads us to infer that the density of $X_1 + X_0$ given $X_1 - X_0$ is the same as the density of $Y_1 + Y_0$ given $Y_1 - Y_0$. Moreover, Property A.1 implies that the co-factors of f are the same for any divergence. Consequently, putting $U = X_1 + X_0, V = X_1 - X_0, U' = Y_1 + Y_0$ and $V' = Y_1 - Y_0$, and since $\{(1, 1)', (1, -1)'\}$ is an orthogonal basis, we can therefore infer from Theorem 3.8 that $U = E(U'/V') + \varepsilon$, where ε is a centered random variable orthogonal to $E(U/V)$.

Thus, since g is a Gaussian density, Remark 3.3 implies that

$$U = E(U') + \frac{\text{Cov}(U', V')}{\text{Var}(V')} (V' - E(V')) + \varepsilon$$

In other words, we apply the same reasoning as the one used in the regression studies in Simulation 4.3 to (U, V) instead of (X_1, X_0) . This is possible since $\{(1, 1)', (1, -1)'\}$ is an orthogonal basis of \mathbb{R}^2 , i.e., we implement a change in basis from the canonical basis of \mathbb{R}^2 to $\{(1, 1)', (1, -1)'\}$.

Thus, in the canonical basis $U = E(U'/V') + \varepsilon$ becomes $X_1 + X_0 = E(Y_1 + Y_0/Y_1 - Y_0) + \varepsilon$, i.e., we obtain that

$$X_1 + X_0 = E(Y_1 + Y_0) + \frac{\text{Cov}(Y_1+Y_0, Y_1-Y_0)}{\text{Var}(Y_1-Y_0)} (Y_1 - Y_0 - E(Y_1 - Y_0)) + \varepsilon$$

where ε is a centered random variable orthogonal to $E(X_1 + X_0/X_1 - X_0)$.

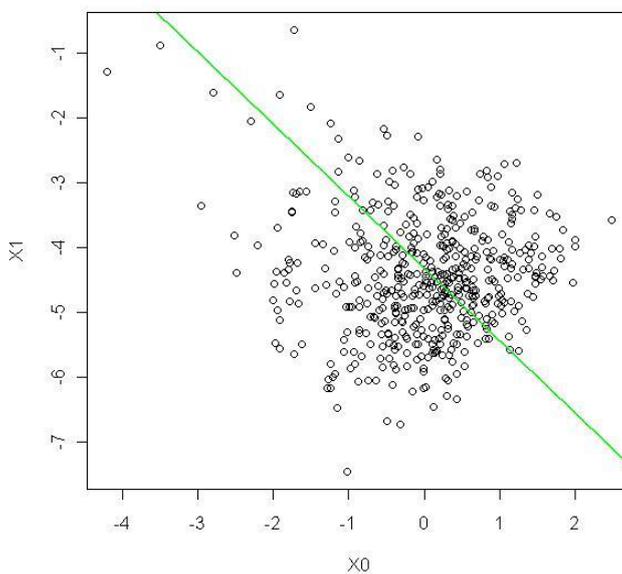
The following table presents the results of our regression.

We simulate 10 times the regression and we obtain a and b such that $X_1 = a + bX_0 + \varepsilon :$

Table 6. Simulation 4: Numerical results of the regression.

Simulation	a	Std Error of a	b	Std Error of b
1	-4.83739	0.11149	-0.95861	0.04677
2	-4.56895	0.09989	-0.88577	0.04225
3	-4.4926	0.1057	-1.2085	0.0452
4	-4.70619	0.10350	-1.04549	0.04235
5	-4.40331	0.10248	-1.00890	0.0438
6	-4.61757	0.09813	-1.20890	0.04649
7	-4.40572	0.09172	-1.16085	0.04091
8	-4.39581	0.10174	-1.38696	0.04487
9	-4.42780	0.10018	-0.93672	0.04066
10	-4.55394	0.09923	-0.98065	0.04382

Figure 4. Graph of the regression of X_1 on X_0 based on our theory (green).



Simulation 4.5 (With the Kullback-Leibler divergence K).

We are in dimension 2(=d), and we use the Kullback–Leibler divergence 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 with a density—that we will name g —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

$$\mathcal{E}_2 = \{b \in \mathbb{R}^2; (Var_{\mathbf{P}}(M(b, b)))^{(-1/2)} \mathbb{P}_n M(b, b) \leq q_{\alpha}^{N(0,1)} / \sqrt{n} \simeq 0, 2533 / 7.0710678 = 0.0358220\}$$

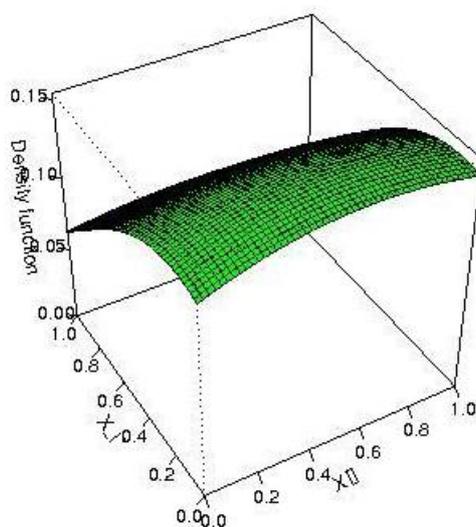
And, we obtain

Table 7. Simulation 5: Numerical results of the optimisation.

Our Algorithm	
Projection Study number 0 :	minimum : 0.445199 at point : (1.0142,0.0026) P-Value : 0.94579
Test :	$H_1 : a_1 \notin \mathcal{E}_1 : \text{True}$
Projection Study number 1 :	minimum : 0.0263 at point : (0.0084,0.9006) P-Value : 0.97101
Test :	$H_0 : a_2 \in \mathcal{E}_2 : \text{True}$
$K(\text{Kernel Estimation of } g^{(2)}, g^{(2)})$	4.0680

Therefore, we can conclude that H_0 is verified.

Figure 5. Graph of the estimate of $(x_0, x_1) \mapsto c_{\rho}(F_{Gumbel}(x_0), F_{Exponential}(x_1))$.



Application to real datasets

Let us now apply our theory to real datasets.

Let us for instance study the moves in the stock prices of Nokia and Sanofi from January 11, 2010 to May 10, 2010. We thus gather 84(=n) data from these stock prices—see data below.

Let us also consider X_1 (resp. X_2) the random variable defining the stock price of Nokia (resp. Sanofi). We will assume—as it is commonly done in mathematical finance—that the stock market abides by the classical hypotheses of the Black–Scholes model—see [13].

Consequently, X_1 and X_2 each present a log-normal distribution as probability distribution. Let f be the density of vector $(\ln(X_1), X_2)$, let us now apply our algorithm to f with the Kullback–Leibler divergence as ϕ -divergence. Let us generate then a Gaussian random variable Y with a density—that we will name g —presenting same mean and variance as f .

We first assume that there exists a vector a such that $D_\phi(g \frac{f_a}{g_a}, f) = 0$.

In order to verify this hypothesis, our reasoning will be the same as in Simulation 4.1. Indeed, we assume that this vector is a co-factor of f . Consequently, Corollary 3.3 enables us to estimate a by the following $0.9(=\alpha)$ level confidence ellipsoid

$$\mathcal{E}_1 = \{b \in \mathbb{R}^2; (Var_{\mathbf{P}}(M(b, b)))^{(-1/2)} \mathbb{P}_n M(b, b) \leq q_\alpha^{N(0,1)} / \sqrt{n} \simeq 0,2533 / \sqrt{84} = 0.02763730\}$$

And, we obtain

Table 8. Numerical results of the optimisation.

Our Algorithm	
Projection Study 0 :	minimum : 0.017345
	at point : (0.027,3.18)
	P-Value : 0.890210
Test :	$H_0 : a_1 \in \mathcal{E}_1 : \text{True}$
K(Kernel Estimation of $g^{(1)}, g^{(1)}$)	2.7704005

Therefore, we conclude that $f = g^{(1)}$, i.e., our hypothesis is confirmed.

Consequently, as explained in Simulations 4.3 and 4.4, we can say that

$$\log(X_1) = 0.027.X_2 + 3.18 + \varepsilon$$

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

Finally, using the least squares method, we estimate α_1 and α_2 such that $\log(X_1) = \alpha_1 + \alpha_2.X_2 + \varepsilon$. Thus, the following table presents the results of the least squares method if we assume that ε is Gaussian:

Table 9. Numerical results of the regression.

Simulation	α_1	Std Error of α_1	α_2	Std Error of α_2
1	3.153694	0.230380	0.026578	0.004236

Figure 6. Graph of the regression of log of Nokia on Sanofi based on the least squares method (red) and based on our theory (green).

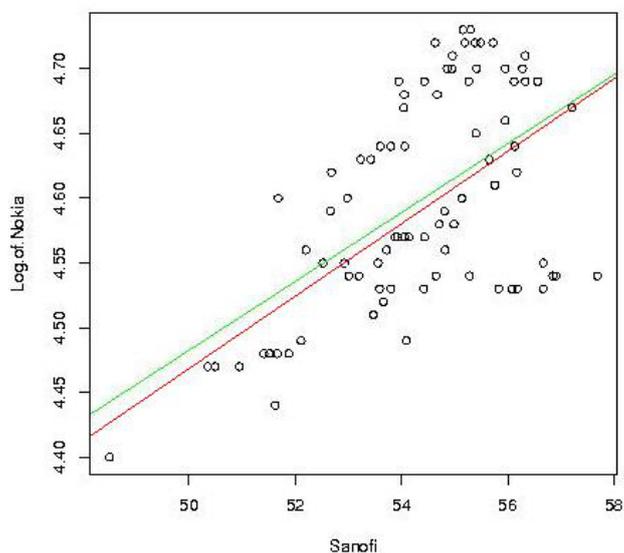


Table 10. Stock prices of Nokia and Sanofi.

Date	Nokia	Log-of-Nokia	Sanofi	Date	Nokia	Log-of-Nokia	Sanofi
10/05/10	84.75	4.44	51.62	07/05/10	81.85	4.4	48.5
06/05/10	87.3	4.47	50.35	05/05/10	87.75	4.47	50.95
04/05/10	87.25	4.47	50.49	03/05/10	87.85	4.48	51.51
30/04/10	87.8	4.48	51.66	29/04/10	87.85	4.48	51.41
28/04/10	87.85	4.48	51.88	27/04/10	89	4.49	52.11
26/04/10	89.2	4.49	54.09	23/04/10	90.7	4.51	53.47
22/04/10	92.75	4.53	53.59	21/04/10	108.4	4.69	53.95
20/04/10	108.9	4.69	54.43	19/04/10	108.3	4.68	54.05
16/04/10	106.8	4.67	54.04	15/04/10	109.9	4.7	54.95
14/04/10	109.8	4.7	54.86	13/04/10	108.3	4.68	54.67
12/04/10	109.1	4.69	55.27	09/04/10	110.1	4.7	55.41
08/04/10	110.7	4.71	54.96	07/04/10	113.2	4.73	55.3
06/04/10	112.4	4.72	54.64	01/04/10	113.3	4.73	55.16
31/03/10	112.4	4.72	55.19	30/03/10	112.5	4.72	55.39
29/03/10	111.8	4.72	55.49	26/03/10	112.5	4.72	55.72
25/03/10	111.4	4.71	56.33	24/03/10	110.2	4.7	55.95
23/03/10	109.1	4.69	56.12	22/03/10	109.2	4.69	56.33
19/03/10	108.5	4.69	56.57	18/03/10	108.4	4.69	56.56
17/03/10	109.9	4.7	56.28	16/03/10	107	4.67	57.21

Table 11. Stock prices of Nokia and Sanofi.

Date	Nokia	Log-of-Nokia	Sanofi	Date	Nokia	Log-of-Nokia	Sanofi
15/03/10	105.3	4.66	55.95	12/03/10	105	4.65	55.4
11/03/10	103	4.63	55.65	10/03/10	104	4.64	56.13
09/03/10	101.5	4.62	56.17	08/03/10	100.7	4.61	55.75
05/03/10	100.2	4.61	55.76	04/03/10	98.7	4.59	54.81
03/03/10	99.8	4.6	55.14	02/03/10	97.25	4.58	54.99
01/03/10	95.85	4.56	54.82	26/02/10	95.85	4.56	53.72
25/02/10	94.55	4.55	52.92	24/02/10	96.3	4.57	53.92
23/02/10	96.2	4.57	54.05	22/02/10	96.7	4.57	54.14
19/02/10	97.3	4.58	54.71	18/02/10	96.6	4.57	54.43
17/02/10	96.1	4.57	53.88	16/02/10	94.95	4.55	53.56
15/02/10	93.65	4.54	53.2	12/02/10	93.55	4.54	53.01
11/02/10	94.6	4.55	52.52	10/02/10	95.55	4.56	52.2
09/02/10	98.4	4.59	52.66	08/02/10	99.2	4.6	52.98
05/02/10	99.8	4.6	51.68	04/02/10	102.6	4.63	53.42
03/02/10	103.9	4.64	54.06	02/02/10	103.8	4.64	53.8
01/02/10	102.4	4.63	53.23	29/01/10	103.6	4.64	53.6
28/01/10	101.8	4.62	52.68	27/01/10	92.55	4.53	53.8
26/01/10	92.7	4.53	54.42	25/01/10	91.9	4.52	53.66
22/01/10	94.1	4.54	54.65	21/01/10	93.7	4.54	55.28
20/01/10	92.75	4.53	56.67	19/01/10	93.6	4.54	57.69
18/01/10	94.55	4.55	56.67	15/01/10	93.55	4.54	56.85
14/01/10	93.7	4.54	56.91	13/01/10	92.5	4.53	56.18
12/01/10	92.35	4.53	55.83	11/01/10	93	4.53	56.08

5. 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. $500^{-\frac{4}{4+d}}$ and $100^{-\frac{4}{4+d}}$) for the AMISE of Simulations 4.1, 4.2 and 4.3 (resp. Simulations 4.4 and 4.5). This choice leads us to simulate 50 (resp. 500 and 100) random variables—see Scott [15] page 151—none of which have been discarded to obtain the truncated sample. This has also been the case in our application to real datasets.

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 Kullback–Leibler divergence—our method requires a considerably shorter computation time and also in the superior robustness of our method.

6. Conclusions

Projection Pursuit is useful in evidencing characteristic structures as well as one-dimensional projections and their associated distributions in multivariate data. Huber [2] shows us how to achieve it through maximization of the Kullback–Leibler divergence.

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.

References

1. Friedman, J.H.; Stuetzle, W.; Schroeder, A. Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **1984**, *79*, 599–608.
2. Huber, P.J. Projection pursuit. *Ann. Statist.* **1985**, *13*, 435–525; With discussion.
3. Zhu, M. On the forward and backward algorithms of projection pursuit. *Ann. Statist.* **2004**, *32*, 233–244.
4. Yohai, V.J. Optimal robust estimates using the Kullback–Leibler divergence. *Stat. Probab. Lett.* **2008**, *78*, 1811–1816.
5. Toma, A. Optimal robust M-estimators using divergences. *Stat. Probab. Lett.* **2009**, *79*, 1–5.
6. Huber, P.J. *Robust Statistics*; Wiley: Hoboken, NJ, USA, 1981; republished in paperback, 2004.
7. Diaconis, P.; Freedman, D. Asymptotics of graphical projection pursuit. *Ann. Statist.* **1984**, *12*, 793–815.
8. Cambanis, S.; Huang, S.; Simons, G. On the theory of elliptically contoured distributions. *J. Multivariate Anal.* **1981**, *11*, 368–385.
9. Landsman, Z.M.; Valdez, E.A. Tail conditional expectations for elliptical distributions. *N. Am. Actuar. J.* **2003**, *7*, 55–71.
10. Van der Vaart, A.W. *Asymptotic Statistics*; In *Cambridge Series in Statistical and Probabilistic Mathematics*; Cambridge University Press: Cambridge, MA, USA, 1998; Volume 3.
11. Broniatowski, M.; Keziou, A. Parametric estimation and tests through divergences and the duality technique. *J. Multivariate Anal.* **2009**, *100*, 16–36.
12. Vajda, I. χ^α -divergence and generalized Fisher's information. Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes; Czech Technical University in Prague: Prague, Czech, 1971, dedicated to the memory of Antonín Spacek; Academia: Prague, Czech; pp. 873–886.
13. Black, F.; Scholes, M.S. The pricing of options and corporate liabilities. *J. Polit. Econ.* **1973**, *3*, 637–654.
14. Saporta, G. *Probabilités, Analyse des données et Statistique*; Technip: Paris, France, 2006.

15. Scott, D.W. *Multivariate Density Estimation. Theory, Practice, and Visualization*; John Wiley and Sons: New York, NY, USA, 1992.
16. Cressie, N.; Read, T.R.C. Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc.* **1984**, *Ser. B* *46*, 440–464.
17. Csiszár, I. On topology properties of f -divergences. *Studia Sci. Math. Hungar.* **1967**, *2*, 329–339.
18. Liese, F.; Vajda, I. *Convex Statistical Distances*; In *Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]*; B.G. Teubner Verlagsgesellschaft: Leipzig, Germany, 1987; Volume 95.
19. Pardo, L. Statistical inference based on divergence measures. In *Statistics: Textbooks and Monographs*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2006; Volume 185.
20. Zografos, K.; Ferentinos, K.; Papaioannou, T. φ -divergence statistics: sampling properties and multinomial goodness of fit and divergence tests. *Comm. Statist. Theory Methods* **1990**, *19*, 1785–1802.
21. Azé, D. *Eléments d'analyse convexe et variationnelle*; Ellipse: Minneapolis, MN, USA, 1997.
22. Touboul, J. Projection pursuit through ϕ -divergence minimisation. *arXiv:0912.2883*, 2009.
23. Bosq, D.; Lecoutre J.-P. *Livre—Théorie De L'Estimation Fonctionnelle*; Economica: Hoboken, NJ, USA, 1999.

Appendix

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 $\varphi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, 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

$$D_\phi(Q, P) = \int \varphi\left(\frac{dQ}{dP}\right) dP \quad (\text{A.1})$$

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 [16], Csiszár [17] and the books of Liese [18], Pardo [19] and Zografos [20]). They are defined by a specific φ . Indeed,

- with the Kullback–Leibler divergence, we associate $\varphi(x) = x \ln(x) - x + 1$
- 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 $\varphi(x) = |x - 1|$.

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

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

Property A.2. The divergence function $Q \mapsto D_\phi(Q, P)$ is 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 Liese [18]). If $T : (X, A) \rightarrow (Y, B)$ is measurable and if $D_\phi(P, Q) < \infty$, then $D_\phi(P, Q) \geq D_\phi(PT^{-1}, QT^{-1})$, with equality being reached when T is surjective for (P, Q) .

Theorem A.1 (theorem III.4 of Azé [21]). Let $f : I \rightarrow \mathbb{R}$ be a convex function. Then f is a Lipschitz function in all compact intervals $[a, b] \subset \text{int}\{I\}$. In particular, f is continuous on $\text{int}\{I\}$.

A.2. Miscellaneous

In the present section, all demonstrations can be found in Touboul [22].

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

Lemma A.2. For all $c > 0$, we have $\Gamma_c \subset \overline{B_{L^1}(f, c)}$, where $B_{L^1}(f, c) = \{p \in L^1; \|f - p\|_1 \leq c\}$.

Lemma A.3. G is closed in L^1 for the topology of the uniform convergence.

Lemma A.4. Let consider the sequence (a_i) defined in (2.3) page 1587.

We then have $\lim_n \lim_k K(\check{g}_n^{(k)} \frac{f_{a_{k,n}}}{[g^{(k)}]_{a_{k,n}}}, f_n) = 0$ a.s.

In the case where f is known and keeping the notations introduced in Section 3.1, we have

Proposition A.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.

Theorem A.2. Assuming (H0) to (H3) hold, for any $k = 1, \dots, d$ and any $x \in \mathbb{R}^d$, we have $|\check{g}^{(k)}(x) - g^{(k)}(x)| = O_{\mathbf{P}}(n^{-1/2})$ and $\int |\check{g}^{(k)}(x) - g^{(k)}(x)| dx = O_{\mathbf{P}}(n^{-1/2})$ as well as $|K(\check{g}^{(k)}, f) - K(g^{(k)}, f)| = O_{\mathbf{P}}(n^{-1/2})$.

Theorem A.3. Assuming that (H1) to (H3), (H6) and (H8) hold. Then,

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

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

B. Study of the sample

Let X_1, X_2, \dots, X_m be a sequence of independent random vectors with same density f . Let Y_1, Y_2, \dots, 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 [23])—with $L(u) = \ln(u \vee e)$:

$$(\mathcal{H}yp): h_m \searrow_m 0, mh_m \nearrow_m \infty, mh_m/L(h_m^{-1}) \rightarrow_m \infty \text{ and } L(h_m^{-1})/LLm \rightarrow_m \infty.$$

Let us consider

$$B_1(n, a) = \frac{1}{n} \sum_{i=1}^n \varphi' \left\{ \frac{f_{a,n}(a^\top Y_i) g_n(Y_i)}{g_{a,n}(a^\top Y_i) f_n(Y_i)} \right\} \text{ and } B_2(n, a) = \frac{1}{n} \sum_{i=1}^n \varphi^* \left\{ \varphi' \left\{ \frac{f_{a,n}(a^\top X_i) g_n(X_i)}{g_{a,n}(a^\top X_i) f_n(X_i)} \right\} \right\}.$$

Our goal is to estimate the minimum of $D_\phi(g \frac{f_a}{g_a}, f)$. To do this, it is necessary for us to truncate our samples:

Let us consider now a positive sequence θ_m such that $\theta_m \rightarrow 0$, $y_m/\theta_n^2 \rightarrow 0$, 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 D.7— $y_m^{(1)}/\theta_m^2 \rightarrow 0$, where $y_m^{(1)}$ is defined by $|\varphi(\frac{g_m(x) f_{b,m}(b^\top x)}{f_m(x) g_{b,m}(b^\top x)}) - \varphi(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)})| \leq y_m^{(1)}$, for all b in \mathbb{R}_*^d and all x in \mathbb{R}^d , and finally $\frac{y_m^{(2)}}{\theta_m^2} \rightarrow 0$, where $y_m^{(2)}$ is defined by $|\varphi'(\frac{g_m(x) f_{b,m}(b^\top x)}{f_m(x) g_{b,m}(b^\top x)}) - \varphi'(\frac{g(x) f_b(b^\top x)}{f(x) g_b(b^\top x)})| \leq y_m^{(2)}$, 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, \dots, X_n and Y_1, Y_2, \dots, 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 [11] and in Section 3.1, it holds*

$$\lim_{n \rightarrow \infty} \sup_{a \in \mathbb{R}_*^d} |(B_1(n, a) - B_2(n, a)) - D_\phi(g_{\frac{f_a}{g_a}}, f)| = 0.$$

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

C. Hypotheses’ discussion

C.1. Discussion of (H2).

Let us work with the Kullback–Leibler divergence 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^{D_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 .

C.2. Discussion of (H4).

This hypothesis consists in the following assumptions:

- We work with the Kullback–Leibler divergence, (0)
- We have $f(\cdot/a_1^\top x) = g(\cdot/a_1^\top x)$, 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 $A = \{(c, x) \in \mathbb{R}_^d \setminus \{a_1\} \times \mathbb{R}^d; \frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} > \frac{f_c(c^\top x)}{g_c(c^\top x)}, g(x) \frac{f_c(c^\top x)}{g_c(c^\top x)} > f(x)\} = \emptyset$ through a reductio ad absurdum, i.e., if we assume $A \neq \emptyset$.*

Thus, our hypothesis enables us to derive

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

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

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

Thus, our hypothesis enables us to derive

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

We can therefore conclude as above.

Let us now verify (H4):

We have $PM(c, a_1) - PM(c, a) = \int \ln\left(\frac{g(x)f_c(c^\top x)}{g_c(c^\top x)f(x)}\right)\left\{\frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} - \frac{f_c(c^\top x)}{g_c(c^\top x)}\right\}g(x)dx$. 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 $\ln\left(\frac{g(x)f_c(c^\top x)}{g_c(c^\top x)f(x)}\right)$ and $\left\{\frac{f_{a_1}(a_1^\top x)}{g_{a_1}(a_1^\top x)} - \frac{f_c(c^\top x)}{g_c(c^\top x)}\right\}$ 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 $a \neq a_1$.

D. Proofs

Preliminary remark :

Let us note that if $K(g, f) \geq \int |f(x) - g(x)|dx$, a simple reductio ad absurdum enables us to infer that $K(g^{(1)}, f) \geq \int |f(x) - g^{(1)}(x)|dx$. Therefore, through an induction, we immediately obtain that, for any k , $K(g^{(k)}, f) \geq \int |f(x) - g^{(k)}(x)|dx$. Thus, for any k and from a certain rank n , we derive that $K(g_n^{(k)}, f) \geq \int |f(x) - g_n^{(k)}(x)|dx$.

Proof of Lemma D.1.

Lemma D.1. We have $g(\cdot/a_1^\top x, \dots, a_j^\top x) = n(a_{j+1}^\top x, \dots, a_d^\top x) = f(\cdot/a_1^\top x, \dots, a_j^\top x)$.

Putting $A = (a_1, \dots, a_d)$, let us determine f in basis A . Let us first study the function defined by $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^d, x \mapsto (a_1^\top x, \dots, a_d^\top x)$. We can immediately say that ψ is continuous and since A is a basis, its bijectivity is obvious. Moreover, let us study its Jacobian.

$$\text{By definition, it is } J_\psi(x_1, \dots, x_d) = \begin{vmatrix} \frac{\partial \psi_1}{\partial x_1} & \dots & \frac{\partial \psi_1}{\partial x_d} \\ \dots & \dots & \dots \\ \frac{\partial \psi_d}{\partial x_1} & \dots & \frac{\partial \psi_d}{\partial x_d} \end{vmatrix} = \begin{vmatrix} a_{1,1} & \dots & a_{1,d} \\ \dots & \dots & \dots \\ a_{d,1} & \dots & a_{d,d} \end{vmatrix} = |A| \neq 0 \text{ since}$$

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 $\Psi(y) = \tilde{n}(y_{j+1}, \dots, y_d)\tilde{h}(y_1, \dots, y_j)$, 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. And then, if \tilde{g} is the expression of g in basis A , we have $\tilde{g}(\cdot/y_1, \dots, y_j) = \tilde{n}(y_{j+1}, \dots, y_d) = \Psi(\cdot/y_1, \dots, y_j)$, i.e., $g(\cdot/a_1^\top x, \dots, a_j^\top x) = n(a_{j+1}^\top x, \dots, a_d^\top x) = f(\cdot/a_1^\top x, \dots, a_j^\top x)$.

Proof of Lemma D.2.

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

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

Lemma D.3. $\inf_{a \in \mathbb{R}_*^d} D_\phi(g^*, f)$ is reached when the ϕ -divergence is greater than the L^1 distance as well as the L^2 distance.

Indeed, let G be $\{g_{g_a}^{f_a}; a \in \mathbb{R}_*^d\}$ and Γ_c be $\Gamma_c = \{p; K(p, f) \leq c\}$ for all $c > 0$. From Lemmas A.1, A.2 and A.3 (see page 1605), 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 1605) implies that $Q \mapsto D_\phi(Q, P)$ 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 $c = D_\phi(g, f)$, Ω is necessarily not empty because we always have $D_\phi(g_{g_a}^{f_a}, f) \leq D_\phi(g, f)$). Moreover, when the ϕ -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 D.4.

Lemma D.4. 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.

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 $a_i, i = 1, \dots, p$, are the vectors of the canonical basis, since $g^{(p-1)}(x) = g(x) \frac{f_1(x_1)}{g_1(x_1)} \frac{f_2(x_2)}{g_2(x_2)} \dots \frac{f_{p-1}(x_{p-1})}{g_{p-1}(x_{p-1})}$ 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 D.5.

Lemma D.5. If there exists $p, p \leq d$, such that $D_\phi(g^{(p)}, f) = 0$, then the family of $(a_i)_{i=1, \dots, p}$ —derived from the construction of $g^{(p)}$ —is free and orthogonal.

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 $a_1 = (1, 0, \dots, 0)$ and $a_2 = (\alpha, 0, \dots, 0)$, where $\alpha \in \mathbb{R}$, we get $g^{(1)}(x) = g(x_2, \dots, x_d/x_1) f_1(x_1)$ and $f = g^{(2)}(x) = g(x_2, \dots, x_d/x_1) f_1(x_1) \frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}$. Hence $f(x_2, \dots, x_d/x_1) = g(x_2, \dots, x_d/x_1) \frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}$. It consequently implies that $f_{\alpha a_1}(\alpha x_1) = [g^{(1)}]_{\alpha a_1}(\alpha x_1)$ since $1 = \int f(x_2, \dots, x_d/x_1) dx_2 \dots dx_d = \int g(x_2, \dots, x_d/x_1) dx_2 \dots dx_d \frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)} = \frac{f_{\alpha a_1}(\alpha x_1)}{[g^{(1)}]_{\alpha a_1}(\alpha x_1)}$. 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 $\int f(x) dx = 1 \neq +\infty = \int n(a_{j+1}^\top x, \dots, a_d^\top x) h(a_1^\top x, \dots, a_j^\top x) dx$. The use of the same argument as in the proof of Lemma D.2, enables us to infer the orthogonality of $(a_i)_{i=1, \dots, p}$.

Proof of Lemma D.6.

Lemma D.6. If there exists $p, p \leq d$, such that $D_\phi(g^{(p)}, f) = 0$, where $g^{(p)}$ is built from the free and orthogonal family a_1, \dots, a_j , then, there exists a free and orthogonal family $(b_k)_{k=j+1, \dots, d}$ of vectors of \mathbb{R}_*^d , such that $g^{(p)}(x) = g(b_{j+1}^\top x, \dots, b_d^\top x/a_1^\top x, \dots, a_j^\top x) f_{a_1}(a_1^\top x) \dots f_{a_j}(a_j^\top x)$ and such that $\mathbb{R}^d = Vect\{a_i\} \oplus Vect\{b_k\}$.

Through the incomplete basis theorem and similarly as in Lemma D.5, we obtain the result thanks to the Fubini’s theorem.

Proof of Lemma D.7.

Lemma D.7. For any continuous density f , we have $y_m = |f_m(x) - f(x)| = O_P(m^{-\frac{2}{4+d}})$.

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 [15], we derive that $b_m(x) = O_{\mathbf{P}}(\sum_{j=1}^d h_j^2)$ 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 3.1.

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, $g_1^*(x_1) = \int g^*(x) dx_2 \dots dx_d = \int h(x_1)g(x) dx_2 \dots dx_d = h(x_1) \int g(x) dx_2 \dots dx_d = h(x_1)g_1(x_1)$.

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

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

$$T : f(\cdot/x_1)f_1(x_1) \mapsto f(\cdot/x_1)f_1(x_1)$$

is measurable.

Second, the above remark implies that

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

Consequently, property A.3 page 1605 infers:

$$D_\phi(g(\cdot/x_1) \frac{g_1(x_1)h(x_1)}{f_1(x_1)}, f(\cdot/x_1)f_1(x_1)) \geq D_\phi(T^{-1}(g(\cdot/x_1) \frac{g_1(x_1)h(x_1)}{f_1(x_1)}), T^{-1}(f(\cdot/x_1)f_1(x_1)))$$

$$= D_\phi(g(\cdot/x_1)f_1(x_1), f(\cdot/x_1)f_1(x_1)), \text{ by the very definition of } T.$$

$$= D_\phi(g \frac{f_1}{g_1}, f), \text{ which completes the proof of this proposition.}$$

Proof of Proposition 3.3. Proposition 3.3 comes immediately from Proposition B.1 page 1606 and Lemma A.1 page 1605.

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

Proof of Theorem 3.2. First, from Lemma D.7, 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

$$\text{have } \Psi_j = \frac{1}{\check{g}_{\check{a}_j,n}^{(j-1)}(\check{a}_j^\top x)g_{a_j}^{(j-1)}(a_j^\top x)}$$

$$((f_{\check{a}_j,n}(\check{a}_j^\top x) - f_{a_j}(a_j^\top x))g_{a_j}^{(j-1)}(a_j^\top x) + f_{a_j}(a_j^\top x)(g_{a_j}^{(j-1)}(a_j^\top x) - \check{g}_{\check{a}_j,n}^{(j-1)}(\check{a}_j^\top x))),$$

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 the proof of Theorem A.2.

Proof of Theorem D.1.

Theorem D.1. In the case where f is known and under the hypotheses assumed in Section 3.1, it holds

$$\sqrt{n} \mathcal{A} \cdot (\check{c}_n(a_k) - a_k) \xrightarrow{\mathcal{L}aw} \mathcal{B} \cdot \mathcal{N}_d(0, \mathbf{P} \|\frac{\partial}{\partial b} M(a_k, a_k)\|^2) + \mathcal{C} \cdot \mathcal{N}_d(0, \mathbf{P} \|\frac{\partial}{\partial a} M(a_k, a_k)\|^2) \text{ and}$$

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

where $\mathcal{A} = \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) (\mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k))$, $\mathcal{C} = \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k)$ and

$$\mathcal{B} = \mathbf{P} \frac{\partial^2}{\partial b \partial b} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k).$$

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,

$$\begin{aligned} \mathbf{P} \frac{\partial}{\partial b} M(a_k, a_k) &= \mathbf{P} \frac{\partial}{\partial a} M(a_k, a_k) = 0, \\ \mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k) &= \mathbf{P} \frac{\partial^2}{\partial b_j \partial a_i} M(a_k, a_k) = \int \varphi'' \left(\frac{g f_{a_k}}{f g_{a_k}} \right) \frac{\partial}{\partial a_i} \frac{g f_{a_k}}{f g_{a_k}} \frac{\partial}{\partial b_j} \frac{g f_{a_k}}{f g_{a_k}} f \, dx, \\ \mathbf{P} \frac{\partial^2}{\partial b_i \partial b_j} M(a_k, a_k) &= - \int \varphi'' \left(\frac{g f_{a_k}}{f g_{a_k}} \right) \frac{\partial}{\partial b_i} \frac{g f_{a_k}}{f g_{a_k}} \frac{\partial}{\partial b_j} \frac{g f_{a_k}}{f g_{a_k}} f \, dx, \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) = \int \varphi' \left(\frac{g f_{a_k}}{f g_{a_k}} \right) \frac{\partial^2}{\partial a_i \partial a_j} \frac{g f_{a_k}}{f g_{a_k}} f \, dx, \\ \text{and consequently } \mathbf{P} \frac{\partial^2}{\partial b_i \partial b_j} M(a_k, a_k) &= - \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) = - \mathbf{P} \frac{\partial^2}{\partial b_j \partial a_i} M(a_k, a_k), \text{ which implies,} \\ \frac{\partial^2}{\partial a_i \partial a_j} K(g \frac{f_{a_k}}{g_{a_k}}, f) &= \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) - \mathbf{P} \frac{\partial^2}{\partial b_i \partial b_j} M(a_k, a_k), \\ &= \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial a_i \partial b_j} M(a_k, a_k) = \mathbf{P} \frac{\partial^2}{\partial a_i \partial a_j} M(a_k, a_k) + \mathbf{P} \frac{\partial^2}{\partial b_j \partial a_i} M(a_k, a_k). \end{aligned}$$

The very definition of the estimators $\check{\gamma}_n$ and $\check{c}_n(a_k)$, implies that $\begin{cases} \mathbb{P}_n \frac{\partial}{\partial b} M(b, a) = 0 \\ \mathbb{P}_n \frac{\partial}{\partial a} M(b(a), a) = 0 \end{cases}$
i.e. $\begin{cases} \mathbb{P}_n \frac{\partial}{\partial b} M(\check{c}_n(a_k), \check{\gamma}_n) = 0 \\ \mathbb{P}_n \frac{\partial}{\partial a} M(\check{c}_n(a_k), \check{\gamma}_n) + \mathbb{P}_n \frac{\partial}{\partial b} M(\check{c}_n(a_k), \check{\gamma}_n) \frac{\partial}{\partial a} \check{c}_n(a_k) = 0, \end{cases}$ *i.e.* $\begin{cases} \mathbb{P}_n \frac{\partial}{\partial b} M(\check{c}_n(a_k), \check{\gamma}_n) = 0 \text{ (E0)} \\ \mathbb{P}_n \frac{\partial}{\partial a} M(\check{c}_n(a_k), \check{\gamma}_n) = 0 \text{ (E1)} \end{cases}$

Under (H5) and (H6), and using a Taylor development of the (E0) (resp. (E1)) equation, we infer there exists $(\bar{c}_n, \bar{\gamma}_n)$ (resp. $(\tilde{c}_n, \tilde{\gamma}_n)$) on the interval $[(\check{c}_n(a_k), \check{\gamma}_n), (a_k, a_k)]$ such that

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

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

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

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

Proof of Theorem 3.3. We derive this theorem through Proposition B.1 and Theorem D.1.

Proof of Theorem 3.4. We recall that $g_n^{(k)}$ is the kernel estimator of $\check{g}^{(k)}$. Since the Kullback–Leibler divergence 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

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

$$\begin{aligned} \text{and } \lim_n \int |[\lim_k g_n^{(k)}(x)] - f_n(x)| dx &\geq \int \lim_n [|[\lim_k g_n^{(k)}(x)] - f_n(x)|] dx \\ &= \int |[\lim_n \lim_k g_n^{(k)}(x)] - \lim_n f_n(x)| dx \end{aligned}$$

Through Lemma A.4, we then obtain that

$$\begin{aligned} 0 = \lim_n \lim_k K(g_n^{(k)}, f_n) &\geq \int |[\lim_n \lim_k g_n^{(k)}(x)] - \lim_n f_n(x)| dx \geq 0, \text{ i.e., that} \\ \int |[\lim_n \lim_k g_n^{(k)}(x)] - \lim_n f_n(x)| dx &= 0. \end{aligned}$$

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 1585, 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 Corollary 3.1. Through the dominated convergence theorem and through Theorem 3.4, we get the result using a reductio ad absurdum.

Proof of Theorem 3.5. Through Proposition B.1 and Theorem A.3, we derive theorem 3.5.

© 2010 by the author; licensee MDPI, Basel, Switzerland. This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution license <http://creativecommons.org/licenses/by/3.0/>.