A product type non-parametric estimator of the conditional density by quantile transform and copula representation.

Olivier P. Faugeras

A product type non-parametric estimator of the conditional density by quantile transform and copula representation.

Olivier P. Faugeras

L.S.T.A, Université Paris 6
175, rue du Chevaleret, 75013 Paris, France
e-mail: olivier.faugeras@gmail.com

Abstract: We present a new non-parametric estimator of the conditional density of the kernel type. It is based on an efficient transformation of the data by quantile transform. By use of the copula representation, it turns out to have a remarkable product form. We study its asymptotic properties and compare its bias and variance to competitors based on nonparametric regression.

AMS 2000 subject classifications: 62G007, 62M20, 62M10.
Keywords and phrases: conditional density, copula, nonparametric estimation, quantile transform, regression.

1. Introduction

1.1. Motivations

For predicting the response Y of a real valued input variable X at a given location x from an independent identically distributed sample $((X_i, Y_i); i = 1, \ldots, n)$, it is of great interest of estimating not only the conditional mean or *regression function* $E(Y|X = x)$, but the full *conditional density* $f(y|x)$. Indeed, estimating the conditional density is much more informative, allowing not only to recalculate the (predicted) conditional expected value $E(Y|X)$ and conditional standard deviation from the density, but also to provide the general shape of the conditional density. This is especially important for multi-modal densities, which often arise from nonlinear phenomenas, where the expected value might be nowhere near a mode. Therefore, considering the expected value as the best predictor, (which is the case from a mathematical standpoint for a decision based on the choice, yet arbitrary, of the $\mathbb{L}_2$ norm) is questionable. Moreover, for situations in which confidence intervals are preferred to point estimates, the estimated conditional density is an object of obvious interest.

A natural approach to estimate the conditional density $f(y|x)$ of Y given $X = x$ would be to exploit the identity

$$f(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} \tag{1}$$

where f_{XY} and f_X denote the joint density of (X, Y) and X, respectively. By introducing Parzen-Rosenblatt kernel estimators of these densities, namely

$$\hat{f}_{n,XY}(x,y) := \frac{1}{n}\sum_{i=1}^{n} K'_{h'}(X_i - x)K_h(Y_i - y)$$
$$\hat{f}_{n,X}(x) := \frac{1}{n}\sum_{i=1}^{n} K'_{h'}(X_i - x)$$

where K and K' are kernels with their associated sequence of bandwidth h and h' going to zero as $n \to \infty$, one can construct the quotient

$$\hat{f}^1_n(y|x) := \frac{\hat{f}_{n,XY}(x,y)}{\hat{f}_{n,X}(x)}$$

and obtain an estimator of the conditional density.

As pointed out by numerous authors, see e.g. Fan and Yao [2005] chapter 6, this approach is equivalent to the one arising from considering this conditional density estimation problem in a regression framework. Indeed, let $F(y|x)$ be the cumulative conditional distribution function of Y given $X = x$. It stems from the fact that

$$E\left(1_{|Y-y|\leq h}|X = x\right) = F(y + h|x) - F(Y - h|x) \approx 2h.f(y|x)$$

as $h \to 0$, that, if one replace the expectation in the above expression by its empirical counterpart, one can apply the usual local averaging methods and perform a regression estimation on the synthetic data $((1/2h)1_{|Y_i-y|\leq h};\ i = 1, \ldots, n)$. By a Bochner type theorem, one can even replace the transformed data by its smoothed version

$$Y'_i := K_h(Y_i - y) := \frac{1}{h}K\left(\frac{Y_i - y}{h}\right).$$

In particular, the popular Nadaraya-Watson regression estimator

$$\hat{f}^2_n(y|x) := \frac{\sum_{i=1}^n Y'_i.K'_{h'}(X_i - x)}{\sum_{i=1}^n K'_{h'}(X_i - x)}$$

reduces itself to the same estimator of the conditional density of the kernel type as before

$$\hat{f}^2_n(y|x) := \frac{\sum_{i=1}^n K_h(Y_i - y).K'_{h'}(X_i - x)}{\sum_{i=1}^n K'_{h'}(X_i - x)} = \hat{f}^1_n(y|x).$$

However, these equivalent approaches suffer from several drawbacks: first, by its form as a quotient of two estimators, the probabilistic behavior of the Nadaraya-Watson estimator (or its local polynomial counterpart) is tricky to study. It is usually dealt with by a centering at expectation for both numerator

and denominator and a linearisation of the inverse, see e.g. Ferraty and Vieu [2006], Fan and Yao [2005], or Bosq [1998] for details. Second, at a conceptual level, one could argue that implementing regression estimation techniques in this setting is, in a sense, unnatural: estimating a density, even if it a is conditional one, should resort to density estimation techniques only. Finally, practical implementations of these estimators can lead to numerical instability when the denominator is close to zero.

To remedy these problems, we propose an estimator which builds on the idea of using synthetic data, i.e. a representation of the data more adapted to the problem than the original one. By transforming the data, the estimator turns out to have a remarkable *product* form. Its study then reveals to be particularly simple: it reduces to the ones already done on nonparametric density estimation.

The rest of the paper is organized as follows: in the rest of this section, a brief overview of the literature is sketched. In section 2 we introduce the quantile transform and the copula representation which leads to the definition of our estimator (section 3). In section 4, the main asymptotic results about our conditional density estimator are established and compared in section 5 to those of other competitors. Proofs are mainly based on a series of preliminary lemmas which are given in the appendix. For sake of simplicity and clarity of exposition, we have limited ourselves to unidimensional input variables X. However, all results can be easily extended to the multivariate case.

1.2. Overview of the literature

Nonparametric conditional density estimation was not much investigated since it was first studied by Rosenblatt [1969]. Recent years have witnessed a renewed interested, starting with Hyndman, Bashtannyk and Grunwald [1996], who improved Rosenblatt's kernel based estimators. See also the book of Ferraty and Vieu [2006] for an extension for functional data.

Taking advantage of the regression formulation, Fan, Yao and Tong [1996] proposed a conditional density estimator which generalises the kernel one by use of the local polynomial techniques. In particular, it allows to tackle with the bias issues of the kernel smoothing. However, and unlike the former, it is no longer guaranteed to have positive value nor to integrate to 1 with respect to y. With these issues in mind, Hyndman and Yao [1998] built on local polynomial techniques and suggested two improved methods, the first one based on locally fitting a log-linear model and the second one on constrained local polynomial modeling. An overview can be found in Fan and Yao [2005] (chapter 6 and 10). Very recently, Györfi and Kohler [2007] studied a partitioning type estimate and studied its properties in total variation norm.

2. The quantile transform and copula function

2.1. The quantile transform

The idea of transforming the data is not new. It has been used to improve the range of applicability and performance of classical estimation techniques, e.g. to deal with skewed data, heavy tails, or restrictions on the support (see e.g. Devroye and Lugosi [2001] chapter 14 and the references therein, and also Van der Vaart [1998] chapter 3.2 for the related topic of variance stabilizing transformations). In order to make inference on Y from X, a natural question which then arises is, what is the "best" transformation, if this question has a sense. As one can note from the above references, the "best" transformation is very linked to the distribution of the underlying data. We will see below that the natural candidate is the quantile transform.

The quantile transform is a well-known probabilistic trick which dates back at least to Skorohod [1956] and the so-called Skorohod Representation Theorem. It is used to reduce proofs for arbitrary real valued random variables X to ones for random variables U uniformly distributed on the interval $[0,1]$, e.g. in empirical process theory. Moreover, it is at the core of some invariance properties in statistics and probability theory: for example, one can show by this device, that the law of the Kolmogorov-Smirnov statistic

$$\sup_{x\in R} |F_n(x) - F(x)|$$

is independent of F. (See e.g. Shorack and Wellner [1986], chapter 1).

First things first, let's recall the definition of the generalised inverse of an increasing function:

Definition 2.1. *For a non-decreasing function $F : \mathbb{R} \to [0,1]$, its generalised inverse Q is defined as*

$$Q(t) := \inf \{x : F(x) \geqslant t\}.$$

If F is continuous, then Q is uniquely defined. We then have that $x \leq Q(t)$ if and only if $F(x) \leq t$.

The quantile transform is based on the following well-known theorem:

Theorem 2.2. *For any real valued random variable X with cumulative distribution function F and quantile function Q, the following properties hold :*

(i) Whenever F is continuous, the random variable $U = F(X)$ is uniformly distributed on $(0,1)$;
(ii) Conversely, when F is arbitrary, if U is a uniformly distributed random variable on $(0,1)$, one has the distributional identity: $X \stackrel{d}{=} Q(U)$.

Proof. See e.g. Shorack and Wellner [1986] chap. 1. □

As a consequence, given a sample $(X_1, \ldots, X_n)$ of random variables with common c.d.f. F sitting on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, one can always enlarge

this probability space to carry a sequence $(U_1, \ldots, U_n)$ of uniform $(0,1)$ random variables such that $U_i = F(X_i)$, that is to say to construct a pseudo-sample with a *prescribed uniform* marginal distribution.

2.2. The copula representation

Formally, a copula is a bi-(or multi)variate distribution function whose marginal distribution functions are uniform on the interval $[0,1]$. Indeed, solving a problem formulated by Fréchet [51], Sklar [59] proved the following fundamental result:

Theorem 2.3. *For any bivariate cumulative distribution function $F_{X,Y}$ on $\mathbb{R}^2$, if the marginal distribution functions F_X, F_Y are continuous, then there exists some function $C : [0,1]^2 \to [0,1]$, called the dependence or copula function, such as*

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)) \ , \ -\infty \leq x, y \leq +\infty \tag{2}$$

This representation is unique with respect to (F_X, F_Y). The copula function C is itself a cumulative distribution function on $[0,1]^2$ with uniform marginals.

This theorem gives a representation of the bivariate c.d.f. as a function of each univariate c.d.f. In other words, the copula function captures the dependence structure among the components X and Y of the vector (X,Y), irrespectively of the marginal distribution F_X and F_Y. Simply put, it allows to deal with the randomness of the dependence structure and the randomness of the marginals *separately*.

From now on, we assume that the copula function $C(u,v)$ has a density $c(u,v)$ with respect to the Lebesgue measure on $[0,1]^2$ and that F_X and F_Y are differentiable with densities f_X and f_Y. Formula (2) enables us to derive explicit formulas of the following quantities:

- the joint density,

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} = f_X(x) f_Y(y) c(F_X(x), F_Y(y))$$

  where $c(u,v) := \frac{\partial^2 C(u,v)}{\partial u \partial v}$ is the above mentioned copula density;
- the conditional density,

$$f_{Y|X}(x,y) = \frac{f_{XY}(x,y)}{f_X(x)} = f_Y(y) c(F_X(x), F_Y(y)) \tag{3}$$

and symmetrically for the laws of $X|Y$. For more details regarding copulas and their properties, one can consult for example the book of Joe [1997].

As we argued in our introduction, Probability does not mix well with Algebra: the law of a random variable transformed by an algebraic operation, and especially the inverse, is not easy to tackle with. Formula (3) is thus of considerable importance since it has turned the conditional density formula (1) of

the ratio type into a product one. This formula is the backbone of our article where the product form will be especially relevant from a statistical point of view, when we will turn to the estimation issue, which we now deal with in the next section.

3. Presentation of the estimator of the conditional density

From now on, we simplify notations and note f and F the density and c.d.f. of X, and g and G those of Y. We can rewrite the previously stated formula (3) as

$$f_{Y|X}(x,y) = g(y)c(F(x),G(y)).$$

A natural plug-in approach to build an estimator of the conditional density is to use estimators of each of the following quantities:

1. the marginal density g of Y,
2. the c.d.f $F(x)$ and $G(y)$ of X and Y respectively,
3. the second crossed derivative $c(u,v)$ of the copula function $C(u,v)$.

To this purpose, we propose to use:

1. a Parzen-Rosenblatt kernel type non parametric estimator of the marginal density g of Y,
$$\hat{g}_n(y) := \frac{1}{n}\sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)$$
2. the empirical distribution functions $F_n(x)$ and $G_n(y)$ for $F(x)$ and $G(y)$ respectively,
$$F_n(x) := \frac{1}{n}\sum_{j=1}^{n} 1_{X_j \leq x} \tag{4}$$
$$G_n(y) := \frac{1}{n}\sum_{j=1}^{n} 1_{Y_j \leq y}. \tag{5}$$

Concerning the copula density $c(u,v)$, one can note that $c(u,v)$ is the joint density of the transformed variables $(U,V) = (F(X),G(Y))$. Therefore, $c(u,v)$ can be estimated by the bivariate Parzen-Rosenblatt kernel type non parametric density (pseudo) estimator,

$$c_n(u,v) := \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{(u,v)-(U_i,V_i)}{h'_n}\right)$$

where K is a bivariate kernel and h'_n its associated bandwidth. For simplicity, we restrict ourselves to product kernels, and the pseudo estimator c_n writes itself

$$c_n(u,v) := \frac{1}{n}\sum_{i=1}^{n} K_1\left(\frac{u-U_i}{a_n}\right) K_2\left(\frac{v-V_i}{b_n}\right).$$

Nonetheless, since F and G are unknown, the random variables (U_i, V_i) are not observable, i.e. c_n is not a true statistic. Therefore, we approximate the pseudo-sample $(U_i, V_i), i = 1, \ldots, n$ by its empirical counterpart $(F_n(X_i), G_n(Y_i)), i = 1, \ldots, n$ where F_n and G_n are the empirical distribution functions (4) and (5) respectively. We therefore obtain a genuine estimator of $c(u, v)$

$$\hat{c}_n(u, v) := \frac{1}{n} \sum_{i=1}^{n} K_1 \left(\frac{u - F_n(X_i)}{a_n} \right) K_2 \left(\frac{v - G_n(Y_i)}{b_n} \right).$$

Eventually, the conditional density estimator writes itself

$$\hat{f}_n(y|x) := \left[\frac{1}{n} \sum_{i=1}^{n} K_0 \left(\frac{y - Y_i}{h_n} \right) \right] . \left[\frac{1}{n} \sum_{i=1}^{n} K_1 \left(\frac{F_n(x) - F_n(X_i)}{a_n} \right) K_2 \left(\frac{F_n(y) - G_n(Y_i)}{b_n} \right) \right]$$

or, under a more compact form,

$$\hat{f}_n(y|x) := \hat{g}_n(y) \hat{c}_n(F_n(x), G_n(y)). \tag{6}$$

To our knowledge, the estimator studied in this paper has never been proposed in the literature. However, some connections can be made to the ones proposed by Gasser and Muller [1979] in the context of regression estimation, which tackles the issue of having a random denominator. Indeed, their estimator of the regression function, which result from an improvement of the one initially proposed by Priestley and Chao [1972], can be considered as a convolution type estimator, which first transform the design to a uniform (random) one. These estimators are shown below:

$$m_n^{GM(1)} = \frac{1}{h_n} \sum_{i=1}^{n-1} \left\{ \int_{X_{i,n}}^{X_{i+1,n}} K \left(\frac{x - u}{h_n} \right) du \right\} Y_{[i]}$$

$$m_n^{GM(2)} = \frac{1}{h_n} \sum_{i=1}^{n} (X_{i+1,n} - X_{i,n}) K \left(\frac{x - X_{i,n}}{h_n} \right) Y_{[i]}$$

where $X_{i,n}$ denotes the ith order statistic of the sample $(X_1, \ldots, X_n)$ and $Y_{[i]}$ its corresponding Y value.

4. Asymptotic results

For stating our results, we will have to make some regularity assumptions on the densities f, c and the kernels K_0, K_1, K_2. One will note that these assumptions are far from being minimal but are somehow customary in nonparametric density estimation (See section 6.2 for details). These assumptions are presented below: We note the ith moment of a generic kernel (possibly multivariate) K as $m_i(K) := \int u^i K(u) du$, and the $\mathbb{L}_p$ norm of a function h by $||h||_p := \int h^p$.

Assumptions on the Kernel:

- **Assumption (K-0)**
  (i) K is of bounded support and of bounded variation;
  (ii) $K \geq 0$ and is bounded above by a constant C;
  (iii) K is a first order kernel: $m_0(K) = 1$, $m_1(K) = 0$ and $m_2(K) < +\infty$.
- **Assumption (K-1)** K satisfies a Lipschitz condition. For a bivariate K of the product type, one can write it as follows: There exists constants C_1 and C_2, such that for every $(u, v) \in [0, 1]^2 \times [0, 1]^2$,

$$|K_1(u_1)K_2(u_2) - K_1(v_1)K_2(v_2)| \leq C_1 |u_1 - v_1| + C_2 |u_2 - v_2|$$

- **Assumption (K-2)** K is twice differentiable with bounded second partial derivatives.

Regularity assumptions on the density:

- **Assumptions (f-0)**: The density is twice differentiable with bounded second derivative on its support.
- **Assumptions (f'-0)**: In addition to the previous assumption, the density is bounded and non-vanishing on an interval $[a, b]$.

In the remainder of this paper, we will always suppose that g and c satisfy assumptions **(f-0)**, and the kernels K_0 and K assumptions **(K-0)**.

4.1. Weak consistency of the estimator

We have the following weak consistency theorem:

Theorem 4.1. *If the bivariate kernel K satisfy the Lipschitz condition **(K-1)**, and if $h_n = O(n^{-1/5})$, $a_n = b_n = O(n^{-1/6})$, then*

$$\hat{f}_n(y|x) = f(y|x) + O_p(n^{-1/3}).$$

Proof. We have the following decomposition,

$$\begin{aligned}
\hat{f}_n(y|x) - f(y|x) &= \hat{g}_n(y)\hat{c}_n(F_n(x), G_n(y)) - g(y)c(F(x), G(y)) \\
&= (\hat{g}_n(y) - g(y))\, \hat{c}_n(F_n(x), G_n(y)) \\
&+ g(y)\, (\hat{c}_n(F_n(x), G_n(y)) - c_n(F(x), G(y))) \\
&+ g(y)\, (c_n(F(x), G(y)) - c(F(x), G(y))) \\
&:= D_1 + D_2 + D_3.
\end{aligned}$$

Since $\|K\|_\infty < C$, $\|\hat{c}_n\|_\infty < C$, which in turn entails

$$|D_1| \leq C\, |\hat{g}_n(y) - g(y)|\,.$$

By invoking classical consistency results for the kernel unidimensional density estimator of section 6 (lemma 6.2), we get

$$|D_1| = O_P\left(\frac{1}{\sqrt{nh_n}}\right) + O(h_n^2) = O_p(n^{-2/5})$$

for an optimal choice of $h_n = O(n^{-1/5})$. By invoking again consistency results for the kernel bidimensional density estimator (lemma 6.3), we also get

$$D_3 = O_P(n^{-1/3})$$

for a choice of $a_n = b_n = O(n^{-1/6})$. Therefore, fot this choice of a_n and b_n, lemma (6.4), entails

$$D_2 = O_P\left(\frac{1}{\sqrt{n}a_n}\right) = O_P(n^{-1/3}).$$

We may then conclude. □

We also have weak consistency results uniformly on sets:

Corollary 4.2. *In addition, if g and c also satisfies assumption **(f'-0)**, then $h_n = O((\ln n/n)^{1/5})$, $a_n = b_n = O((\ln n/n)^{1/6})$ entails*

$$\sup_{y\in[a,b]} |\hat{f}_n(y|x) = f(y|x)| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/3}\right).$$

Proof. Use the same decomposition as before and majorize in uniform norm. Then use the results in uniform norm of section 6 (lemma 6.2 and 6.3). □

Remark 1. *Our estimator is optimal in the sense that it reaches the minimax rate of convergence.*

4.2. Almost sure convergence (strong consistency)

Theorem 4.3. *If the bivariate kernel K satisfy the Lipschitz condition **(K-1)**, then, for $h_n = O((\ln n/n)^{1/5})$, $a_n = b_n = O((\ln n/n)^{1/6})$, we have*

$$\hat{f}_n(y|x) = f(y|x) + O_{a.s.}((\ln n/n)^{1/3}).$$

Proof. It follows the same lines as the preceding theorem, but uses the a.s. consistency results of the kernel density estimators in lemmas 6.2 and 6.3 and corollary 6.5. It is therefore omitted. □

Corollary 4.4. *With the same hypothesis, we have that*

$$\sup_{y\in R}\left|\hat{f}_n(y|x) - f(y|x)\right| = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{1/3}\right).$$

Proof. Omitted for the same reasons. □

4.3. Convergence in distribution

Theorem 4.5. *If the kernel K satisfy the assumption* ***(K-2)*** *of lemma (6.6), then $h_n = O(n^{-1/5})$, $a_n = b_n = O(n^{-1/6})$ entails*

$$n^{1/3}\left(\hat{f}_n(y|x) - f(y|x)\right) \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, g(y)f(y|x)||K||_2^2\right).$$

Proof. Use the decomposition,

$$\begin{aligned}
\hat{f}_n(y|x) - f(y|x) &= \hat{g}_n(y)\hat{c}_n(F_n(x), G_n(y)) - g(y)c(F(x), G(y)) \\
&= (\hat{g}_n(y) - g(y))\,\hat{c}_n(F_n(x), G_n(y)) \\
&+ g(y)\,(\hat{c}_n(F_n(x), G_n(y)) - c_n(F(x), G(y))) \\
&+ g(y)\,(c_n(F(x), G(y)) - c(F(x), G(y))) \\
&:= D_1 + D_2 + D_3.
\end{aligned}$$

Since, $\hat{g}_n(y) = g(y) + O_P(n^{-2/5})$ and $|\hat{c}_n| \leq C$ for a bounded bivariate kernel, we have $n^{1/3}|D_1| = O_P(n^{-1/15}) = o_P(1)$. By the second approximation lemma (6.6), $|D_2| = O_P(n^{-1/2})$ and $n^{1/3}|D_2| = O_P(n^{-1/6}) = o_P(1)$. The last term, is asymptotically normal at rate $n^{1/3}$ by the result of section 6, lemma 6.3:

$$n^{1/3}\left[c_n(u,v) - c(u,v)\right] \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, c(u,v)\,\|K\|_2^2\right).$$

That is to say,

$$n^{1/3}g(y)\left[c_n(F(x), G(y)) - c(F(x), G(y))\right] \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, g^2(y)c(F(x), G(y))\,\|K\|_2^2\right).$$

An application of Slutsky's lemma yields the desired result. □

For a vector $(y_1, \ldots, y_d)$, one can get a multidimensional version of the convergence in distribution (fidi convergence):

Theorem 4.6. *With the same assumptions,*

$$n^{1/3}\left(\left(\frac{\hat{f}_n(y_i|x) - f(y_i|x)}{\sqrt{g(y_i)f(y_i|x)}\,\|K\|_2}\right), i = 1, ..., m\right) \overset{d}{\rightsquigarrow} N^{(m)}$$

where $N^{(m)}$ is the standard m-variate normal distribution.

Proof. Omitted. It follows the lines of e.g. Bosq [1998], theorem 2.3. □

4.4. Asymptotic Bias

Theorem 4.7. *With assumptions* ***(K-2)*** *of lemma 6.5, and the choice of the bandwidth $h_n = O(n^{-1/5})$, $a_n = b_n = O(n^{-1/6})$, we have*

$$E_0 := E(\hat{f}_n(y|x)) - f(y|x) = n^{-1/3}\frac{g(y)}{2}\nabla^2 c(F(x), G(y), K) + o(n^{-1/3})$$

with

$$\nabla^2 c(a,b,K) = \sum_{1\leq i,j\leq 2} \frac{\partial^2 c(a,b)}{\partial u_i \partial u_j} \int_{\mathbb{R}^2} u_i u_j K(u) du.$$

Proof. We omit x and y. We still have the additive decomposition,

$$\begin{aligned}\hat{f}_n(y|x) - f(y|x) &= (\hat{g}_n - g)\hat{c}_n + g\,(\hat{c}_n - c_n) + g\,(c_n - c)\\ &:= D_1 + D_2 + D_3.\end{aligned}$$

The last term is, up to a multiplicative factor, the bias of the kernel density estimator c_n, in dimension 2. Therefore, lemma 6.3 yields

$$E(D_3) = g(y)E(c_n - c) = \frac{g(y)a_n^2}{2} \sum_{1\leqslant i,j\leqslant 2} \frac{\partial^2 c(u,v)}{\partial u_i \partial u_j} \int u_i u_j K(u) du + o(a_n^2).$$

We will show below that the other terms are negligible compared to D_3: We go further in the decomposition of D_1

$$\begin{aligned}(\hat{g}_n - g)\hat{c}_n &= (\hat{g}_n - g)(\hat{c}_n - c_n) + (\hat{g}_n - g)(c_n - c) + (\hat{g}_n - g)c\\ &:= D_{11} + D_{12} + D_{13}.\end{aligned}$$

By lemma 6.2 the bias of $\hat{g}_n(y)$ is

$$E(\hat{g}_n(y)) - g(y) = \frac{m_2(K_0)}{2} g''(y) h_n^2 + o(h_n^2).$$

By Cauchy-Schwarz inequality, we can bound the product terms as D_{12} and D_{11} as follows

$$E(D_{12}) = E(\hat{g}_n - g)(c_n - c) \leq \left(E(\hat{g}_n - g)^2\right)^{1/2} \left(E(c_n - c)^2\right)^{1/2}$$

and

$$E(D_{11}) = E(\hat{g}_n - g)(\hat{c}_n - c_n) \leq \left(E(\hat{g}_n - g)^2\right)^{1/2} \left(E(\hat{c}_n - c_n)^2\right)^{1/2}.$$

By lemma 6.3 $|\hat{c}_n - c_n| \xrightarrow{P} 0$ and $|\hat{c}_n - c_n|$ is trivially asymptotically uniformly integrable since the kernels are bounded. Therefore, $E(\hat{c}_n - c_n)^2 = o(1)$ and the term $E(D_{11}) = E(\hat{g}_n - g)(\hat{c}_n - c_n) = o(h_n^2)$ is asymptotically negligible. The term $(E(c_n - c)^2)^{1/2}$ is the root of the MSE of c_n, and is of order a_n^2 while $(E(\hat{g}_n - g)^2)^{1/2}$ is the root of the MSE of $\hat{g}_n$, and is of order h_n^2. Therefore $E(D_{12}) = O(h_n^2 a_n^2)$ is also negligible. In turn, $E(D_1) = O(h_n^2)$ is negligible compared to $E(D_3)$.

For the last term D_2, first note that $||\hat{c}_n - c_n||_\infty$ is bounded uniformly in n. By Fatou's reversed lemma,

$$\limsup E||\hat{c}_n - c_n||_\infty \leq E \limsup ||\hat{c}_n - c_n||_\infty. \tag{7}$$

Now, a careful analysis of lemma 6.5 shows that $||\hat{c}_n - c_n||_\infty$ is bounded above by terms such as

$$C||F - F_n||_\infty A$$

with C a constant, A a random variable depending of x, y, and such as $A \to E(A) < +\infty$ a.s.. The law of iterated logarithm (lemma 6.1) entails

$$\limsup ||F_n - F||_\infty = \frac{1}{2}\sqrt{\frac{\ln_2 n}{n}} \quad \text{a.s..}$$

Since $\limsup A = E(A)$ a.s.,

$$\limsup ||\hat{c}_n - c_n||_\infty \leq C'\sqrt{\frac{\ln_2 n}{n}} \quad \text{a.s.}$$

where C' is a constant. In turn, together with (7),

$$\limsup E(||\hat{c}_n - c_n||_\infty) \leq C'\sqrt{\frac{\ln_2 n}{n}},$$

yieding $E(||\hat{c}_n - c_n||_\infty) = o(n^{-1/3})$. Therefore $E(D_2) = o(n^{-1/3})$ is negligible compared to $E(D_3)$. □

4.5. Asymptotic Variance and Mean Square Error

The asymptotic variance has already been derived in theorem 4.5:

$$V_0 := Var(\hat{f}(y|x)) = n^{-2/3} g(y) f(y|x) ||K||_2^2 + o(n^{-2/3})$$

Together with the computation of the asymptotic bias of the preceding theorem, we get the asymptotic mean squared error as a corollary:

Corollary 4.8. *with the previous assumptions, the Asymptotic Mean Squared Error (AMSE) is*

$$AMSE = n^{-2/3} g(y) \left(\frac{g(y) \left(\nabla^2 c(F(x), G(y), K) \right)^2}{4} + f(y|x) ||K||_2^2 \right) + o(n^{-2/3})$$

which can also be written as

$$AMSE = n^{-\frac{2}{3}} f^2(y) \left(\frac{\left(\nabla^2 c(F(x), G(y), K) \right)^2}{4} + c(F(x), G(y)) ||K||_2^2 \right) + o(n^{-\frac{2}{3}}).$$

5. Comparison with other estimators

5.1. Presentation of alternative estimators

For convenience, we recall below the definition of other estimators of the conditional density encountered in the literature and summarize their bias and variance properties. We will note the bias of the ith estimator $\hat{f}_n^i(y|x)$ by E_i and its variance by V_i.

1. **Double kernel estimator**: as defined in the introduction section of our paper by the following ratio,

$$\hat{f}_n^{(1)}(y|x) := \frac{\frac{1}{n}\sum_{i=1}^{n} K'_{h_1}(X_i - x)K_{h_2}(Y_i - y)}{\frac{1}{n}\sum_{i=1}^{n} K'_{h_1}(X_i - x)}.$$

   - Bias:

$$E_1 = \frac{h_1^2 m_2(K)}{2}\left(2\frac{f'(x)}{f(x)}\frac{\partial f(y|x)}{\partial x} + \frac{\partial^2 f(y|x)}{\partial x^2} + \left(\frac{h_2}{h_1}\right)^2 \frac{\partial^2 f(y|x)}{\partial y^2}\right) + o\left(h_1^2 + h_2^2\right)$$

   - Variance:

$$V_1 = \frac{\|K\|_2^2 f(y|x)}{nh_1h_2 f(x)}\left(\|K\|_2^2 - h_2 f(y|x)\right) + o\left(\frac{1}{nh_1h_2}\right)$$

2. **Local polynomial estimator**: Set

$$R(\theta, x, y) := \sum_{i=1}^{n}\left(K_{h_2}(Y_i - y) - \sum\nolimits_{j=0}^{r}\theta_j(X_i - x)^j\right)^2 K'_{h_1}(X_i - x),$$

   then the local polynomial estimator is defined as

$$\hat{f}_n^{(2)}(y|x) := \hat{\theta}_0,$$

   where $\hat{\theta}_{xy} := (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_r)$ is the value of θ which minimizes $R(\theta, x, y)$. This local polynomial estimator, although it has a superior bias than the kernel one, is no longer restricted to be non-negative and does not integrate to 1, except in the special case $r = 0$. From results of Fan, Yao and Tong [1996], we get for the local linear estimator (see Fan and Yao p256),

   - Bias:

$$E_2 = \frac{h_1^2 m_2(K')}{2}\frac{\partial^2 f(y|x)}{\partial x^2} + \frac{h_2^2 m_2(K)}{2}\frac{\partial^2 f(y|x)}{\partial y^2} + o(h_1^2 + h_2^2)$$

   - Variance:

$$V_2 = \frac{||K||_2^2||K'||_2^2 f(y|x)}{nh_1h_2 f(x)} + o\left(\frac{1}{nh_1h_2}\right)$$

3. **Local parametric estimator**: Set

$$R_1(\theta, x, y) := \sum_{i=1}^{n}\left(K_{h_2}(Y_i - y) - A(X_i - x, \theta)\right)^2 K'_{h_1}(X_i - x)$$

where

$$A(x,\theta) = l\left(\sum_{j=0}^{r} \theta_j (X_i - x)^j\right)$$

and $l(.)$ is a monotonic function mapping $\mathbb{R} \mapsto \mathbb{R}^+$, e.g. $l(u) = \exp(u)$. Then

$$\hat{f}_n^{(3)}(y|x) := A(0,\hat{\theta}) = l(\hat{\theta}_0).$$

- Bias:

$$E_3 = h_1^2 \eta(K') \left(\frac{\partial^2 f(y|x)}{\partial x^2} - \frac{\partial^2 A(0,\theta_{xy})}{\partial x^2}\right) + \frac{h_2^2 m_2(K)}{2} \frac{\partial^2 f(y|x)}{\partial y^2} + o(h_1^2 + h_2^2)$$

- Variance:

$$V_3 = \frac{\tau(K,K')^2 f(y|x)}{n h_1 h_2 f(x)} + o\left(\frac{1}{n h_1 h_2}\right)$$

where η and τ are kernel dependent constants.

4. **Constrained local polynomial estimator**: A simple device to force the local polynomial estimator to be positive is to set $\theta_0 = \exp(\alpha)$ in the definition of R_0 to be minimized. The constrained local polynomial estimator $\hat{f}_n^4(y|x)$ is then defined analogously as the local polynomial estimator $\hat{f}_n^2(y|x)$. We have:
   - Bias:

$$E_4 := h_1^2 \frac{m_2(K')}{2} \frac{\partial^2 f(y|x)}{\partial x^2} + h_2^2 \frac{m_2(K)}{2} \frac{\partial^2 f(y|x)}{\partial y^2} + o(h_1^2 + h_2^2)$$

   - Variance:

$$V_4 = \frac{\|K\|_2^2 f(y|x)}{n h_1 h_2 f(x)} + o\left(\frac{1}{n h_1 h_2}\right)$$

5.2. Asymptotic Bias and Variance comparison

All estimators have (hopefully) the same order in their asymptotic bias and variance terms. The main difference lies in the constant terms which depend on unknown densities.

Bias: Contrary to all the alternative estimators whose bias involve derivatives of the full conditional density, one can note that our estimator's one only involves the density of Y and the derivatives of the copula density. To make things more explicit, the terms involved, e.g. in the local polynomial estimator, write themselves as the sum of the derivatives of the conditional densiy, that is to say

$$\approx \frac{\partial^2 f(y|x)}{\partial x^2} + \frac{\partial^2 f(y|x)}{\partial y^2} = f'(x)g(y)\frac{\partial c(F(x),G(y))}{\partial u} + f^2(x)g(y)\frac{\partial^2 c(F(x),G(y))}{\partial u^2} + 2g'(y)g(y)\frac{\partial c(F(x),G(y))}{\partial v} + g^3(y)\frac{\partial^2 c(F(x),G(y))}{\partial v^2}$$

whereas our $(g(y)/2)\nabla^2 c(F(x), G(y), K)$ term writes itself, modulo the constants involved by the kernel, as

$$\approx g(y)\left(\frac{\partial^2 c(F(x),G(y))}{\partial u^2}+\frac{\partial^2 c(F(x),G(y))}{\partial v^2}+2\frac{\partial^2 c(F(x),G(y))}{\partial u\partial v}\right).$$

It then becomes clear that we have a simpler expression, which does not involve the density and its derivative of X nor the derivative of the Y density, as is the case for the competitors.

Variance: The variance of our estimator involves a product of the density $g(y)$ of Y by the conditional density $f(y|x)$,

$$g(y)f(y|x) = g^2(y)c(F(x), G(y)$$

whereas competitors involve the ratio of $f(y|x)$ by the density $f(x)$ of X

$$\frac{f(y|x)}{f(x)} = \frac{g(y)}{f(x)}c(F(x), G(y)).$$

It is a remarkable feature of the estimator we propose, that its variance does not involve directly $f(x)$, as is the case for the competitors, but only its contribution to Y, through the copula density. This reflects the ability announced in the introduction of the copula representation to have effectively separated the randomness pertaining to Y alone, from the dependence structure of (X, Y). Moreover, our estimator also does not suffer from the unstable nature of competitors who, due to their intrinsic ratio structure, get an explosive variance for small value of the density $f(x)$, making conditional estimation difficult, e.g. in the tail of the distribution of X.

6. Appendix : auxiliary results

In this section, we gather some preliminary results which we will need as basic tools for the demonstrations of section 4. In subsection 6.1, we recall classical results about the convergence of the Kolmogorov-Sminorv statistic. Next, we make a brief overview of kernel density estimation and apply these results to the estimators $\hat{g}_n$ (section 6.2) and c_n (section 6.3). Eventually, we need two approximation lemmas of $\hat{c}_n$ by c_n to prove the consistency and asymptotic normality of our estimator, in sections 6.4 and 6.5 respectively.

6.1. Approximation of the pseudo-variables $F(X_i)$ by their estimates $F_n(X_i)$

Let us note $\|F\|_\infty$, the infinite (also called uniform) norm

$$\|F\|_\infty := \sup_{x\in\mathbb{R}} |F(x)|.$$

Let $(X_i, i = 1, \ldots, n)$, be an i.i.d. sample of the random variable X with common c.d.f. F. The Kolmogorov-Smirnov (K-S) statistic is defined as $D_n := \|F_n - F\|_\infty$. We have already seen that it is invariant w.r.t to the c.d.f. F. The famous Glivenko-Cantelli theorem asserts its convergence to zero in probability: $\|F_n - F\|_\infty = O_P(1)$. Later Kolmogorov and Smirnov derived a central limit theorem for a continuous F

$$\sqrt{n} \sup_x |F_n(x) - F(x)| \overset{d}{\rightsquigarrow} \mu$$

yielding $\|F_n - F\|_\infty = O_P(1/\sqrt{n})$. Chung [1949] derived the optimal a.s. rate for i.i.d. observations:

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\ln\ln n}}. \|F_n - F\|_\infty = \frac{1}{2} \text{ a.s.}$$

which entails

$$\|F_n - F\|_\infty = O_{a.s.}\left(\sqrt{\frac{\ln\ln n}{n}}\right).$$

Remark 2. *This kind of theorems can be considerably generalized and rederived from functional central limit theorems of the Donsker type and invariance principles. They allow to give upper bounds for the suprema of empirical processes indexed by sets (the sets in our case would be $]-\infty, x]$) or functions.*

Let's collect these results in an approximation lemma:

Lemma 6.1. *For an i.i.d. sample from a continuous c.d.f. F,*

$$\|F_n - F\|_\infty = O_{a.s.}\left(\sqrt{\frac{\ln\ln n}{n}}\right) \tag{8}$$

$$\|F_n - F\|_\infty = O_P\left(\frac{1}{\sqrt{n}}\right). \tag{9}$$

As said earlier, although the random variables $(U_i) = (F(X_i))$ are not observable, since F is unknown, one can naturally approximate them by the statistics $F_n(X_i)$. The lemmas above gives the speed of this approximation : since

$$|F(X_i) - F_n(X_i)| \leq \sup_{x\in R} |F(x) - F_n(x)| = \|F_n - F\|_\infty \quad \text{a.s.},$$

we have that, for every $(i \leq n) \in \mathbb{N}^2$,

$$|F(X_i) - F_n(X_i)| = O_{a.s.}\left(\sqrt{\frac{\ln\ln n}{n}}\right) \tag{10}$$

$$|F(X_i) - F_n(X_i)| = O_P\left(\frac{1}{\sqrt{n}}\right) \tag{11}$$

with the suitable previous assumptions.

6.2. Convergence of the kernel density estimator $\hat{g}_n$

We recall below some classical results about the convergence of the Parzen-Rosenblatt kernel non-parametric estimator $\hat{f}_n$ of a d-variate density. Since its inception by Rosenblatt [1956] and Parzen [1962] , it has been studied by a great deal of authors. See e.g. Scott [1992], Prakasa Rao [1983], Nadaraya [1989] for details. See also Bosq [1998] chapter 2.

It is well known that the bias of the kernel density estimator depends on the degree of smoothness of the underlying density, measured by its number of derivatives or its Lipschitz order. In order to get the convergence of the bias to zero, it suffices to assume that the density is continuous (See Parzen [1962]). To get further information on the rate of convergence of the estimator, it is necessary to make further assumptions. Moreover, for kernel functions with unbounded support, the rate of convergence also depends on the tail behaviour of the kernel (See Stute [1982]). Therefore, for clarity of exposition and simplicity of notations, we will make the customary assumptions that the density is twice differentiable and that the kernel is of bounded support. We then have the following results:

- Bias: if assumptions **(f-0)** and **(K-0)** are verified, then for a x in the interior of the support of f, with $h_n \to 0$ and $nh_n^d \to \infty$:

$$E\hat{f}_n(x) = f(x) + \frac{h_n^2}{2} \int_{\mathbb{R}^d} \sum_{1 \le i,j \le d} \frac{\partial^2 f(x)}{\partial x_i \partial x_i} z_i z_j K(z) dz + o(h_n^2).$$

- Variance: With the same assumptions,

$$Var\left[\hat{f}_n(x)\right] = \frac{f(x)}{nh_n^d} \|K_0\|_2^2 + o\left(\frac{1}{nh_n^d}\right).$$

- Pointwise asymptotic normality: under the previous conditions,

$$\sqrt{nh_n^d}\left(\hat{f}_n(x) - E\hat{f}_n(x)\right) \overset{d}{\rightsquigarrow} \mathcal{N}(0, f(x)\|K_0\|_2^2).$$

For a choice of the bandwidth as $h_n = O(n^{-1/(d+4)})$, which realizes the optimal trade-off between the bias and variance, one gets the following rates for the convergence

- in probability:
$$\left|\hat{f}_n(x) - f(x)\right| = O_p(n^{-2/(d+4)})$$
which is the optimal speed of convergence in the minimax sense in the class of density functions with bounded second derivatives, according to Stone [1980].
- in law:
$$n^{2/(d+4)}\left[\hat{f}_n(x) - f(x)\right] \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, f(x)\|K_0\|_2^2\right).$$

One can refine these results by a chaining argument to get uniform rate of convergence on a compact set (see Bickel and Rosenblatt [1973]): for f bounded and non-vanishing on $[a,b]$,

$$\sup_{x\in[a,b]}\left|\hat{f}_n(x)-E\hat{f}_n(x)\right| = O_p\left[\left(\frac{\ln n}{nh_n}\right)^{1/2}\right].$$

Therefore, for the choice of the bandwidth $h_n = O((\ln n/n)^{1/d+4})$ which realizes the optimal trade-off between the bias and variance, one gets the following result in probability:

$$\sup_{x\in[a,b]}\left|\hat{f}_n(x)-f(x)\right| = O_p\left[\left(\frac{\ln n}{n}\right)^{2/(d+4)}\right]$$

which is the optimal speed in the minimax sense in the class of density functions with bounded second derivatives, according to Hasminskii [1978].

For almost sure results, we have (see e.g. Stute [1982], Bosq [1998] chapter 2 and Fan and Yao [2005] chapter 5), under similar hypothesis and $h_n = O((\ln n/n)^{1/(d+4)})$, that

- pointwisely, for a fixed value of x in the interior of the support of f, $\hat{f}_n(x)-f(x) = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{2/(d+4)}\right)$
- on a compact set,

$$\sup_{x\in[a,b]}\left|\hat{f}_n(x)-f(x)\right| = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{2/(d+4)}\right).$$

Applied to our case ($d=1$), we can summarize these results for further reference in the following lemma for the estimator $\hat{g}_n$ of the density g of Y:

Lemma 6.2. *If the kernel K_0 and the density g of Y satisfy assumption **(K-0)** and **(f-0)** respectively, then for a point y in the interior of the support of g, and a bandwidth chosen such as $h_n = O(n^{-1/5})$, we have*

$$|\hat{g}_n(y)-g(y)| = O_p(n^{-2/5})$$

$$n^{2/5}\left[\hat{g}_n(y)-g(y)\right] \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, g(y)\left\|K_0\right\|_2^2\right).$$

With the same assumptions, but for a bandwidth choice of $h_n = O((\ln n/n)^{1/5})$,

$$\hat{g}_n(y)-g(y) = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{2/5}\right).$$

*If, in addition, g satisfies assumption **(f'-0)**, then, for a choice of a bandwidth such as $h_n = O((\ln n/n)^{1/5})$,*

$$\sup_{y\in[a,b]}|\hat{g}_n(y)-g(y)| = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{2/5}\right).$$

6.3. Convergence of $c_n(u,v)$

Once one convinces oneself that $c_n(u,v)$ is simply the kernel density estimator of the bivariate density $c(u,v)$ of the pseudo-variables (U,V), one directly draws its convergence properties by applying the results of the preceding subsection with $d=2$:

Lemma 6.3. *If the bivariate kernel $K = K_1K_2$ and the bivariate density c satisfy assumptions **(K-0)** and **(f-0)** respectively, then, for a choice of $a_n = b_n = O(n^{-1/6})$, for every $(u,v) \in [0,1]^2$, we have:*

- *Pointwise consistency: $c_n(u,v) - c(u,v) = O_P(n^{-1/3})$;*
- *Bias: with obvious notations,*

$$Ec_n(u,v) = c(u,v) + \frac{a_n^2}{2} \sum_{1 \leqslant i,j \leqslant 2} \frac{\partial^2 c(u,v)}{\partial u_i \partial u_j} \int u_i u_j K(u) du + o(a_n^2);$$

- *Asymptotic normality:*

$$n^{1/3} \left[c_n(u,v) - c(u,v) \right] \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, c(u,v) \left\| K \right\|_2^2\right).$$

For almost sure results, we have, with the previous assumptions and for a choice of $a_n = b_n = O((\frac{\ln n}{n})^{1/6})$,

- *pointwisely, for fixed values of u and v,*

$$c_n(u,v) - c(u,v) = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{1/3}\right),$$

- *on a compact set, if c satisfy assumption **(f'-0)**,*

$$\sup_{(u,v)\in[0,1]^2} |c_n(u,v) - c(u,v)| = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{1/3}\right).$$

6.4. A first approximation lemma of $\hat{c}_n$ by c_n

In order to prove the consistency of the estimator, we need to prove the approximation lemma of this section. To this end, we make the assumption that the bivariate kernel $K = K_1K_2$ verify the Lipschitz hypothesis **(K-1)**, i.e., there exist two constants C_1 and C_2 such that for every $(u,v) \in [0,1]^2 \times [0,1]^2$,

$$|K_1(u_1)K_2(u_2) - K_1(v_1)K_2(v_2)| \leq C_1 |u_1 - v_1| + C_2 |u_2 - v_2|.$$

The following lemma gives an approximation rate of the copula density estimator $\hat{c}_n(F_n(x), G_n(y))$

$$\hat{c}_n(F_n(x), G_n(y)) = \frac{1}{n} \sum_{i=1}^{n} K_1\left(\frac{F_n(x) - F_n(X_i)}{a_n}\right) K_2\left(\frac{G_n(y) - G_n(Y_i)}{b_n}\right)$$

by its analogue $c_n(F(x),F(y))$ in the space of the pseudo-variables $(U,V) := (F(X),G(Y))$:

$$c_n(F(x),G(y)) = \frac{1}{n}\sum_{i=1}^{n} K_1\left(\frac{F(x)-F(X_i)}{a_n}\right) K_2\left(\frac{G(y)-G(Y_i)}{b_n}\right).$$

Lemma 6.4. *If the kernel $K(u,v) = K_0(u)K_1(v)$ follows hypothesis **(K-1)**, then*

$$\sup_{(x,y)\in\mathbb{R}^2} |\hat{c}_n(F_n(x),G_n(y)) - c_n(F(x),G(y))| = O_P\left(\frac{1}{\sqrt{n}\inf(a_n,b_n)}\right)$$

Proof. For every $(x,y)\in\mathbb{R}^2$, we have a.s.

$$\begin{aligned}
&|\hat{c}_n(F_n(x),G_n(y)) - c_n(F(x),G(y))| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left| K_1\left(\frac{F_n(x)-F_n(X_i)}{a_n}\right) K_2\left(\frac{G_n(y)-G_n(Y_i)}{b_n}\right)\right. \\
&\left. -K_1\left(\frac{F(x)-F(X_i)}{a_n}\right) K_2\left(\frac{G(y)-G(Y_i)}{b_n}\right)\right| \\
&\leq \frac{C_1}{na_n}\sum_{i=1}^{n} |F_n(x)-F(x)+F(X_i)-F_n(X_i)| \\
&+ \frac{C_2}{nb_n}\sum_{i=1}^{n} |G_n(y)-G(y)+G(Y_i)-G_n(Y_i)|.
\end{aligned}$$

Yet, $|F(X_i)-F_n(X_i)| \leq \sup_{x\in R}|F(x)-F_n(x)| := \|F_n-F\|_\infty$, and the same for $G-G_n$. Consequently, by using the approximation result of section 6.1 (lemma 6.1),

$$\begin{aligned}
|\hat{c}_n(F_n(x),G_n(y)) - c_n(F(x),G(y))| &\leq \frac{2C_1}{a_n}\|F_n-F\|_\infty + \frac{2C_2}{b_n}\|G_n-G\|_\infty \\
&= O_P\left(\frac{1}{\sqrt{n}a_n} + \frac{1}{\sqrt{n}b_n}\right) \\
&= O_P\left(\frac{1}{\sqrt{n}\inf(a_n,b_n)}\right)
\end{aligned}$$

which had to be proved. □

Remark 3. *In particular, for a choice of $a_n = b_n = O(n^{-1/6})$, one gets the approximation rate $n^{-1/3}$.*

Corollary 6.5. *With the same hypotheses,*

$$\sup_{(x,y)\in\mathbb{R}^2} |\hat{c}_n(F_n(x),G_n(y)) - c_n(F(x),G(y))| = O_{a.s.}\left(\sqrt{\frac{2\ln\ln n}{n}}\frac{1}{\inf(a_n,b_n)}\right).$$

Proof. It follows the same lines of the previous demonstration but uses the a.s. bounds instead of the in probability ones of lemma (6.1). □

6.5. A second approximation lemma

In order to prove the asymptotic normality of the estimator, we need to prove the approximation lemma of this section.

For simplicity, we use the same bandwidths for the bivariate kernel: $a_n = b_n$. Moreover, set

$$K(a,b) := K_1\left(\frac{a}{a_n}\right) K_2\left(\frac{b}{a_n}\right)$$

and let's introduce the following notation:

$$c_n(u,U,v,V) := \frac{1}{n}\sum_{i=1}^{n} K_1\left(\frac{u-U_i}{a_n}\right) K_2\left(\frac{v-V_i}{a_n}\right)$$

to stress the fact that the copula density estimator is calculated from from the sample paths of (U,V). We will make the slightly stronger assumption **(K-2)** on the bivariate kernel $K = K_1K_2$, i.e. that K is twice differentiable with bounded second partial derivatives.

We are going to show the following approximation lemma:

Lemma 6.6. *For every fixed (x,y), if K satisfies assumption **(K-2)**, then*

$$c_n(F_n(x),F_n(X),G_n(y),G_n(Y)) - c_n(F(x),F(X),G(y),G(Y)) = O_P(1/\sqrt{n}).$$

Proof. Set

$$\begin{aligned}\Delta_n(x,y) &:= c_n(F_n(x),F_n(X),G_n(y),G_n(Y)) - c_n(F(x),F(X),G(y),G(Y))\\ &= \frac{1}{n}\sum_{i=1}^{n}\left[K\left(F_n(x)-F_n(X_i),G_n(y)-G_n(Y_i)\right)\right.\\ &\left.-K\left(F(x)-F(X_i),G(y)-G(Y_i)\right)\right]\end{aligned}$$

and introduce the following random variables

$$\begin{aligned}Z_{i,n}(x) &:= F_n(x) - F_n(X_i) - F(x) + F(X_i)\\ Z'_{i,n}(y) &:= G_n(y) - G_n(Y_i) - G(y) + G(Y_i).\end{aligned}$$

For all $i \in \mathbb{N}$, $|F(X_i) - F_n(X_i)| \leq \|F_n - F\|_\infty$ a.s. . We thus have uniformly in i the a.s following bound

$$\|Z_{i,n}\|_\infty \leq 2\|F_n - F\|_\infty \tag{12}$$

and similarly for $Z'_{i,n}(y)$. Since K is twice countinously differentiable, its Taylor expansion writes itself for $a > 0$ and $b > 0$

$$\begin{aligned}K(a,b) &= a\frac{\partial K(0,0)}{\partial a} + b\frac{\partial K(0,0)}{\partial b}\\ &+ \frac{a^2}{2}\frac{\partial^2 K(a_0,b_0)}{\partial a^2} + \frac{b^2}{2}\frac{\partial^2 K(a_0,b_0)}{\partial b^2} + ab\frac{\partial^2 K(a_0,b_0)}{\partial a \partial b}\end{aligned}$$

for an $a_0 \in [0,a]$ and $b_0 \in [0,b]$. Applied to our case, it gives

$$\begin{aligned}\Delta_n(x,y) = &\frac{1}{n}\sum_{i=1}^n (Z_{i,n}(x))\frac{\partial K(F(x)-F(X_i),G(y)-G(Y_i))}{\partial a}\\
&+\frac{1}{n}\sum_{i=1}^n (Z'_{i,n}(y))\frac{\partial K(F(x)-F(X_i),G(y)-G(Y_i))}{\partial b}\\
&+\frac{1}{2n}\sum_{i=1}^n Z^2_{i,n}(x)\frac{\partial^2 K(a_{i,n}(x),b_{i,n}(y))}{\partial a^2}\\
&+\frac{1}{2n}\sum_{i=1}^n Z'^2_{i,n}(y)\frac{\partial^2 K(a_{i,n}(x),b_{i,n}(y))}{\partial b^2}\\
&+\frac{1}{n}\sum_{i=1}^n Z_{i,n}(x)Z'_{i,n}(y)\frac{\partial^2 K(a_{i,n}(x),b_{i,n}(y))}{\partial a\partial b}\end{aligned}$$

where $a_{i,n}(x)$ and $b_{i,n}(y)$ are measurable random variables. We have obviously that

$$\begin{aligned}\frac{\partial K(a,b)}{\partial a} &= \frac{1}{a_n}K'_1\left(\frac{a}{a_n}\right)K_2\left(\frac{b}{a_n}\right)\\
\frac{\partial^2 K(a,b)}{\partial a^2} &= \frac{1}{a_n^2}K''_1\left(\frac{a}{a_n}\right)K_2\left(\frac{b}{a_n}\right)\end{aligned}$$

and symmetrically for the other partial derivatives. Therefore for bounded kernels with bounded derivatives of first and second order, there exist a constant C such as

$$\left\|\frac{\partial^2 K(.,.)}{\partial a^2}\right\|_\infty \leq \frac{C}{a_n^2}.$$

Therefore by using (12), we have a.s. the upper bound

$$\begin{aligned}\left|\frac{1}{2n}\sum_{i=1}^n Z^2_{i,n}(x)\frac{\partial^2 K(a_{i,n}(x),b_{i,n}(y))}{\partial a^2}\right| &\leq \frac{1}{2n}\sum_{i=1}^n \|Z_{i,n}\|^2_\infty \frac{C}{a_n^2}\\
&\leq \frac{C}{a_n^2}\|F_n - F\|^2_\infty = O_P\left(\frac{1}{na_n^2}\right)\end{aligned}$$

and similarly for the other second-order terms in the expansion.

For the first order terms, we similarly bounds by using (12) as follows:

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n Z_{i,n}(x)\frac{\partial K(F(x)-F(X_i),G(y)-G(Y_i))}{\partial a}\right|\\
&\leq 2\|F_n - F\|_\infty \frac{1}{n}\sum_{i=1}^n \left|\frac{\partial K(F(x)-F(X_i),G(y)-G(Y_i))}{\partial a}\right|.\end{aligned}$$

It remains to bound in probability,

$$A := \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial K(F(x)-F(X_i),G(y)-G(Y_i))}{\partial a}\right|.$$

We have by stationarity of (X_i, Y_i),

$$\begin{aligned}E(A) &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial K(F(x)-F(X_i),G(y)-G(Y_i))}{\partial a}\right|\right]\\ &= E\left|\frac{\partial K(F(x)-F(X_i),G(y)-G(Y_i))}{\partial a}\right|\\ &= \frac{1}{a_n}E\left|K_1'\left(\frac{F(x)-F(X)}{a_n}\right)K_2\left(\frac{G(y)-G(Y)}{a_n}\right)\right|.\end{aligned}$$

If K_2 is bounded by a constant C,

$$\begin{aligned}E(A) &\le \frac{C}{a_n}E\left|K_1'\left(\frac{F(x)-F(X)}{a_n}\right)\right| = \frac{C}{a_n}E\left|K_1'\left(\frac{u-U}{a_n}\right)\right|\\ &= \frac{C}{a_n}\int_0^1\left|K_1'\left(\frac{u-t}{a_n}\right)\right|dt = C\int_{(u-1)/a_n}^{u/a_n}|K_1'(z)|\,dz\\ &\le C\int_0^1|K_1'(z)|\,dz \le C^2 < \infty\end{aligned}$$

since K_1' is also bounded in absolute value by C, and the bound is uniform in n. Therefore, Markov inequality entails that A is uniformly tight, i.e. bounded in probability

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial K(F(x)-F(X_i),G(y)-G(Y_i))}{\partial a} = O_P(1)$$

and similarly for the other first-order term.

By recollecting all elements, we finally have:

$$\begin{aligned}|\Delta_n(x,y)| &\le O_P\left(\|F_n-F\|_\infty + \|G_n-G\|_\infty\right) + O_P\left(\frac{1}{na_n^2}\right)\\ &\le O_P(1/\sqrt{n}) + O_P\left(\frac{1}{na_n^2}\right)\end{aligned}$$

where the last inequality proceeds from the approximation lemma (6.1). For an $a_n = O(n^{-1/6})$, $1/na_n^2$ is of order $n^{-2/3}$ which is a $o(n^{-1/2})$ and thus yields the claimed result. □

Corollary 6.7. *With the same hypotheses,*

$$|\hat{c}_n(F_n(x), G_n(y)) - c_n(F(x), G(y))| = O_{a.s.}\left(\sqrt{\frac{\ln \ln n}{n}}\right).$$

Proof. It follows the same lines of the previous demonstration but uses the a.s. bounds instead of the in probability ones of lemma (6.1). To bound a.s. the quantity A, one note that by a strong law of large number $A \to E(A)$ a.s., therefore $A = O_{a.s.}(1)$.

□

Acknowledgments

The author would like to thank Prof. Emmanuel Guerre for his thoughtful comments which greatly helped to improve the presentation of this article.

References

[1] D. Bosq. *Nonparametric statistics for stochastic processes*, volume 110 of *Lecture Notes in Statistics*. Springer-Verlag, New York, second edition, 1998. Estimation and prediction.

[2] Kai-Lai Chung. An estimate concerning the Kolmogoroff limit distribution. *Trans. Amer. Math. Soc.*, 67:36–50, 1949.

[3] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[4] Jianqing Fan and Qiwei Yao. *Nonlinear time series*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2005. Nonparametric and parametric methods.

[5] Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.

[6] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

[7] Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon. Sect. A. (3)*, 14:53–77, 1951.

[8] Theo Gasser and Hans-Georg Müller. Kernel estimation of regression functions. In *Smoothing techniques for curve estimation (Proc. Workshop, Heidelberg, 1979)*, volume 757 of *Lecture Notes in Math.*, pages 23–68. Springer, Berlin, 1979.

[9] László Györfi and Michael Kohler. Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory*, 53(5):1872–1879, 2007.

[10] R. Z. Has′minskiĭ. A lower bound for risks of nonparametric density estimates in the uniform metric. *Teor. Veroyatnost. i Primenen.*, 23(4):824–828, 1978.

[11] Rob J. Hyndman, David M. Bashtannyk, and Gary K. Grunwald. Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336, 1996.

[12] Rob J. Hyndman and Qiwei Yao. Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3):259–278, 2002.

[13] Harry Joe. *Multivariate models and dependence concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997.

[14] È. A. Nadaraya. *Nonparametric estimation of probability densities and regression curves*, volume 20 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1989. Translated from the Russian by Samuel Kotz.

[15] Emanuel Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076, 1962.

[16] B. L. S. Prakasa Rao. *Nonparametric functional estimation*. Probability and Mathematical Statistics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1983.

[17] M. B. Priestley and M. T. Chao. Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B*, 34:385–392, 1972.

[18] M. Rosenblatt. Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York, 1969.

[19] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27:832–837, 1956.

[20] David W. Scott. *Multivariate density estimation*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1992. Theory, practice, and visualization, A Wiley-Interscience Publication.

[21] Galen R. Shorack and Jon A. Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986.

[22] M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

[23] A. V. Skorohod. Limit theorems for stochastic processes. *Teor. Veroyatnost. i Primenen.*, 1:289–319, 1956.

[24] Charles J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360, 1980.

[25] Winfried Stute. A law of the logarithm for kernel density estimators. *Ann. Probab.*, 10(2):414–422, 1982.

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