Prediction via the quantile-copula conditional density estimator

Olivier P. Faugeras
Toulouse School of Economics
(University Toulouse 1 Capitole - Gremaq)
Manufacture des Tabacs, 21 Allée de Brienne
bureau MF319
31000 Toulouse, France

September 19, 2011

Abstract

To make a prediction of a response variable from an explanatory one which takes into account features such as multimodality, a nonparametric approach based on an estimate of the conditional density is advocated and considered. In particular, we build point and interval predictors based on the quantile-copula estimator of the conditional density by Faugeras [8]. The consistency of these predictors is proved through a uniform consistency result of the conditional density estimator. Eventually, the practical implementation of these predictors is discussed. A simulation on a real data set illustrates the proposed methods.

Key Words: Nonparametric prediction, Modal regression, Level-set, Conditional density estimation, Quantile transform, Copulas.

1 Introduction

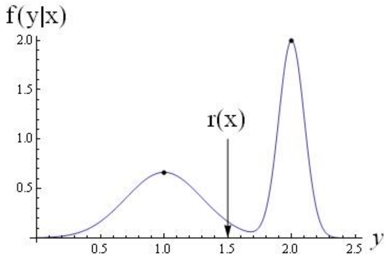
Let X, Y be a couple of real-valued random variables. To what extent one can predict the value of the response variable Y from the explanatory one X? Classical decision theory *à la* Wald [20] recommends to consider a distance or loss function $L : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$ in order to measure the performance of the prediction, and to minimize the corresponding risk or expected loss

$$R(X, Y) = EL(X, Y). \tag{1}$$

It is standard practice among statisticians to use the squared loss, $L(x, y) = (x - y)^2$, so that the risk is minimized by setting as the Bayes (or probabilistic) point predictor the conditional mean $E[Y|X]$. From a statistical standpoint, the problem thus reduces to estimating the regression function $m(x) = E\left[Y|X = x\right]$ by $\hat{m}(x)$, from a training sample (X_i, Y_i), $i = 1, \ldots, n$, and to set as a statistical predictor $\hat{m}(x_0)$ as the best prediction of Y given a newly observed value of $X = x_0$.

Yet, consider the toy model depicted in figure 1, where is plotted the conditional density $f(y|x)$ at the location x where ones wants to make a prediction. The conditional density is assumed here to be a mixture with equal weights of two Normal distribution, each one centered at $y = 1$ and $y = 2$, respectively. In that case, the best prediction corresponding to the conditional mean would be given by $r(x) = 1.5$. However, the practitioner could argue 1.5 is a bad prediction, as he rather observes about half of the outcomes some values of Y concentrated around 1, and the other half values around 2. This toy example is to stress upon the subjectivity of a decision based solely on the estimation of the regression function, as is the focus in many statistical studies: the predictive distribution here is not well summarized by the "average" value to appear, viz. the mean, but better by the "most likely" values to appear, viz. the modes.

Figure 1: A toy example of predictive density



More generally, one can consider the statistician should first estimate the full conditional distribution to fully quantify the input of X on Y and then, once the general shape of the conditional density is given, to build some sensible point predictors and predictive sets. This is especially relevant if the predictive distribution is multi-modal or skewed, which often arises in applications with non-Gaussian or non-linear phenomena.

To that purpose, we propose a methodology to design point and interval predictors based on conditional density estimation, by building upon the quantile-copula conditional density estimator proposed by the same author in [8]. This estimator was shown to be particularly interesting when one wants to make a prediction on y for values of x far from the observed data.

The rest of the article is organized as follows: the quantile-copula estimator of the conditional density is briefly presented in section 2, together with some uniform asymptotic convergence theorems which extend those of [8]. From this conditional density estimator, point and set predictors corresponding to the conditional mode and level sets are defined in sections 3 and 4 respectively, together with a study of their asymptotic consistency and discussions regarding their implementation. An illustration on a real data set is conducted in section 5. Some proofs and auxiliary results are deferred to the appendix 6.

2 The quantile-copula conditional density estimator

2.1 Definition of the Quantile-copula estimator

Nonparametric estimators of the conditional density $f(y|x)$ are either built upon estimators of joint and marginal densities or are based on nonparametric regression on synthetic data, see [8] for an account and references. The Quantile-copula estimator of [8] is based on the idea of transforming the data X and Y by their respective marginal distributions F and G, and the representation of the joint c.d.f. F_{XY} of (X, Y) by means of the copula function C as

$$F_{XY}(x, y) = C(F(X), G(Y)), \tag{2}$$

where F and G are the c.d.f. of X, and Y respectively, see e.g. [13, 15] for some background on copulas. Differentiating formula (2), the conditional density writes

$$f(y|x) = g(y)c(F(x), G(y)) \tag{3}$$

where g is the density of Y and c is the copula density of (X, Y), viz. the density of the vector $(F(X), G(Y))$ with uniform marginals. From (3), a nonparametric estimator can be built as

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

where $\hat{g}$ is the kernel estimate of g, viz.

$$\hat{g}(y) := \frac{1}{nh_n}\sum_{i=1}^{n} K_0\left(\frac{y - Y_i}{h_n}\right),$$

F_n and G_n are the empirical distribution function of F and G, and $\hat{c}_n$ is a kernel estimator of c based on the approximate data $(F_n(X_1), G_n(Y_1)), \ldots, (F_n(X_n), G_n(Y_n))$, viz.

$$\hat{c}_n(u,v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K\left(\frac{u - F_n(X_i)}{a_n}\right) K\left(\frac{v - G_n(Y_i)}{a_n}\right). \tag{4}$$

We refer to [8] for pointwise consistency results and discussions on advantages of the product shape of the estimator compared to ratio-shaped competitors.

2.2 Uniform consistency results of the conditional density estimator

In order to obtain the consistency of the statistical point and interval predictors of sections 3 and 4, uniform consistency results of the conditional density estimator on a compact set are required. Beforehand, we present the notations and assumptions used throughout the paper.

We note the ith moment of a (multivariate) kernel K as $m_i(K) := \int u^i K(u)du$ and the $\mathbb{L}_p$ norm of a function h by $||h||_p := \int h^p$. Let $\simeq$ stands for the order of equivalence of the bandwidths, i.e. $h_n \simeq u_n$ means that $h_n = c_n u_n$ with $c_n \to c > 0$. The support of the densities function f and g are noted by $\text{supp}(f) = \overline{\{x \in \mathbb{R}; f(x) > 0\}}$ and $\text{supp}(g) = \overline{\{y \in \mathbb{R}; g(y) > 0\}}$, where $\overline{A}$ stands for the closure of a set A.

To state our results, we will have to make some regularity assumptions on the kernels and the densities which, although far from being minimal, are somehow customary in kernel density estimation (see section 6.1 for discussions and details). Set x be a fixed point in the interior of $\text{supp}(f)$.

Assumption A

(i) the c.d.f. F of X and G of Y are strictly increasing and differentiable;

(ii) the densities g and c are twice continuously differentiable with bounded second derivatives on their support;

(iii) the density g (respectively c) is uniformly continuous and non-vanishing almost everywhere on a compact set $J := [a,b]$ (respectively $D \subset (0,1)^2$), included in the interior of $\text{supp}(g)$ (respectively $\text{supp}(c)$).

Moreover, we assume that the kernels K_0 and K satisfy the following:

Assumption B

(i) K and K_0 are of bounded support and of bounded variation;

(ii) $0 \leq K \leq C$ and $0 \leq K_0 \leq C$ for some constant C;

(iii) K and K_0 are second order kernels: $m_0(K) = \int K = 1$, $m_1(K) = \int xK(x)dx = 0$ and $m_2(K) = \int x^2K(x)dx < +\infty$, and the same for K_0;

(iv) K it is twice differentiable with bounded second partial derivatives.

We have the following uniform consistency result:

Theorem 2.1. *Let the regularity conditions* ***A*** *(i)-(iii) and* ***B*** *(i)-(iv) be satisfied. If $h_n \simeq (\ln n/n)^{1/5}$ and $a_n \simeq (\ln n/n)^{1/6}$, then, for x in the interior of supp(f) and $[a,b]$ included in the interior of supp(g),*

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

and

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

Proof. As in [8], the main ingredient of the proof follows from the decomposition:

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

where c_n is in (4), the analogue of $\hat{c}_n$ but based on the pseudo-data $(F(X_1), G(Y_1)), \ldots, (F(X_n), G(Y_n))$ instead of the approximate ones. We proceed one step further in the decomposition of each terms, by centering at fixed locations,

$$\begin{aligned}D_1 = &[\hat{g}_n(y) - g(y)]\,[\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y))] \\ &+ [\hat{g}_n(y) - g(y)]\,[\hat{c}_n(F(x), G(y)) - c_n(F(x), G(y))] \\ &+ [\hat{g}_n(y) - g(y)]\,[c_n(F(x), G(y)) - c(F(x), G(y))] \\ &+ [\hat{g}_n(y) - g(y)]\,[c(F(x), G(y))]\end{aligned}$$

$$\begin{aligned}D_2 = &\,g(y)\,[\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y))] \\ &+ g(y)\,[\hat{c}_n(F(x), G(y)) - c_n(F(x), G(y))] \\ &+ g(y)\,[c_n(F(x), G(y)) - c(F(x), G(y))]\end{aligned}$$

Taking the supremum norm for $y \in [a,b]$ and applying propositions 6.2 and 6.1 of section 6, together with the uniform consistency results of the kernel density estimators of theorems 6.1 and 6.2 applied to g_n and c_n, also recalled in section 6, yields the claimed result. □

3 Point prediction by the conditional mode

3.1 Construction of the modal predictor

Depending on the loss function L considered in equation (1) to measure the performance of the prediction, several Bayes (or probabilistic) point predictors are obtainable. The focus in this article is on the $0-1$ loss, viz. $L(x,y) = \mathbb{I}_{x \neq y}$, which leads the "most likely value" or conditional mode,

$$\theta(x) := \arg\sup_{y} f(y|x).$$

For additional references regarding the use of the mode for prediction purposes, refer e.g. to Scott [17], which advocates its use to account for situations as those discussed in the introduction. The interest in the conditional mode lies also that its estimator can be directly obtained from an estimator of the conditional density, following the approaches of [16, 14, 6, 7, 9] among others.

Indeed, define the corresponding statistical point predictor by its empirical counterpart in a plug-in setting as follows: Set S a compact subset of $\mathbb{R}$. In order to assure the existence of the desired object, we assume that $f(y|x)$ is such that:

(R) There exists an $\eta > 0$, an unique $y_0 \in S$ such that $f(.|x)$ is strictly increasing on $(y_0 - \eta, y_0)$, and strictly decreasing on $(y_0, y_0 + \eta)$.

Under this assumption, the local maximizing problem of $f(y|x)$ on $S' = (y_0 - \eta, y_0 + \eta)$ has a unique solution, which is exactly y_0. Therefore, the conditional mode is uniquely defined on this interval:

Definition 3.1. *Under assumption (R), the statistical modal predictor is defined as*

$$\hat{\theta}_n(x) := \arg\sup_{y \in S'} \hat{f}_n(y|x).$$

3.2 Asymptotic properties of the conditional mode predictor

From the uniform consistency result of the quantile-copula density estimator (2), we have the following consistency result of the conditional mode predictor:

Proposition 3.1. *if $f(.|x)$ follows assumption (R), and the conditions for uniform consistency of the conditional density on a compact set of theorem 2.1, then,*

$$\hat{\theta}_n(x) \overset{a.s.}{\to} \theta(x).$$

Proof. Let k, n be integers. By assumption (R), $f(.|x)$ is continuous and strictly increasing on $(\theta(x)-\eta, \theta(x))$. therefore, the inverse function $f^{-1}(.|x)$ exists and is continuous. Thus, by continuity of the latter at the point $f(\theta(x)|x)$, for any $\varepsilon > 0$,

$$\exists \delta_1(\varepsilon) > 0, \forall y \in (\theta(x)-\eta, \theta(x)), |f(y|x)-f(\theta(x)|x)| \leq \delta_1(\varepsilon) \Rightarrow |y-\theta(x)| \leq \varepsilon.$$

Similarly,

$$\exists \delta_2(\varepsilon) > 0, \forall y \in (\theta(x), \theta(x)+\eta), |f(y|x)-f(\theta(x)|x)| \leq \delta_2(\varepsilon) \Rightarrow |y-\theta(x)| \leq \varepsilon,$$

so that,

$$\exists \delta(\varepsilon) > 0, \forall y \in (\theta(x)-\eta, \theta(x)+\eta), |f(y|x)-f(\theta(x)|x)| \leq \delta(\varepsilon) \Rightarrow |y-\theta(x)| \leq \varepsilon.$$

By construction, $\hat{\theta}_k(x) \in (\theta(x) - \eta, \theta(x) + \eta)$, so that,

$$\exists \delta(\varepsilon) > 0, |f(\hat{\theta}_k(x)|x) - f(\theta(x)|x)| \leq \delta(\varepsilon) \Rightarrow |\hat{\theta}_k(x) - \theta(x)| \leq \varepsilon$$

and finally,

$$\exists \delta(\varepsilon) > 0, P(\sup_{k \geq n} |\hat{\theta}_k(x) - \theta(x)| > \varepsilon) \leq P(\sup_{k \geq n} |f(\hat{\theta}_k(x)|x) - f(\theta(x)|x)| > \delta(\varepsilon)). \tag{5}$$

On the other hand, it comes from the triangle inequality that

$$\begin{aligned} \left| f(\theta(x)|x) - f(\hat{\theta}_k(x)|x) \right| &\leq \left| \hat{f}_k(\theta(x)|x) - f(\theta(x)|x) \right| + \left| \hat{f}_k(\hat{\theta}_k(x)|x) - f(\hat{\theta}_k(x)|x) \right| \\ &\leq 2 \sup_{y \in (\theta(x)-\eta, \theta(x)+\eta)} \left| \hat{f}_k(y|x) - f(y|x) \right| \end{aligned}$$

and uniform almost sure convergence of the conditional mode estimator on a compact set of theorem 2.1 entails that

$$\forall \delta > 0, \lim_{n \to \infty} P\left(\sup_{k \geq n} \sup_{y \in (\theta(x)-\eta, \theta(x)+\eta)} \left| \hat{f}_k(y|x) - f(y|x) \right| > \delta \right) = 0,$$

thus $\hat{\theta}_n(x) \overset{a.s.}{\to} \theta(x)$ by equation (5). □

3.3 A remark on the practical implementation of the conditional mode predictor

Set $\mathcal{S}_{Y|X} = \{y : f(y|x) > 0\}$ the support of the conditional density. In practice, the search of the conditional mode can be difficult and time-consuming to implement. Indeed, as the conditional mode estimator is defined as the maximizer of $\hat{f}(y|x)$, i.e. $\hat{\theta}(x) = \arg\sup_{y \in \mathcal{S}_{Y|X}} \hat{f}(y|x)$, one has *a priori* to compute the estimator of the conditional density on a large number of y values in $\mathcal{S}_{Y|X}$ to find the largest value of the estimated conditional density.

Therefore, we would like to mention a method to ease the computation of the conditional mode predictor, proposed in the papers by Abraham, Biau, Cadre [1, 2]. An alternative is to maximize the estimator on the Y data $D_n := \{y_1, \ldots, y_n\}$, i.e. to set $\tilde{\theta}(x) = \arg\max_{y \in D_n} \hat{f}(y|x)$. The maximisation is thus performed on a set of finite cardinality, and can be quickly implemented. According to the asymptotics developed in these papers, one has that $\tilde{\theta}(x) - \hat{\theta}(x) \overset{a.s.}{\to} 0$ as $n \to \infty$, under suitable regularity conditions.

4 Prediction by intervals

4.1 Predictive intervals and level sets

Similarly to the well-known case in estimation, where point estimates can be replaced by confidence intervals, one may wish to summarise the predictive probability distribution by defining a region of the sample space covering a specified probability, i.e. to define a set $\mathcal{C}_\alpha(x)$ such that

$$P(Y \in \mathcal{C}_\alpha(x)|X = x) = \alpha.$$

In the present context, there are numerous ways to construct such predictive intervals or sets covering a specified conditional probability. Among proposals, one may cite the interval symmetric around the mean, the interval symmetric around the median, the interval between the $\frac{1-\alpha}{2}$ and $\frac{1+\alpha}{2}$ quantiles, the interval of shortest length, the interval that minimizes the probability of covering a given family of sets. et cetera... Hyndman [12] provides a detailed discussion of the issues involved in defining such a probability region in the unconditional case. In the conditional case we are interested in, note the coverage region depends on a fixed, chosen x.

To make use of the conditional density and its estimator, it is natural to advocate for an approach based on the level sets of the conditional density, also called the Highest Density Region (HDR) by Hyndman [12], which allows to incorporate the features mentioned in the introductory example such as multimodality.

Definition 4.1. *The level set (probabilistic) predictor is the set $\mathcal{C}_\alpha$ consisting of points y,*

$$\mathcal{C}_\alpha := \{y : f(y|x) \geq f_\alpha\} \tag{6}$$

where f_α is the largest constant such as the prediction set has coverage probability α,

$$P(Y \in C_\alpha(x)|X = x) \geq \alpha. \tag{7}$$

In case of multimodality, $\mathcal{C}_\alpha(x)$ takes the form of an union of possibly disjoint intervals, say $\mathcal{C}_\alpha(x) = \bigcup I_\alpha(x)$, where each $I_\alpha(x) := [y_\alpha, y^\alpha]$, with $y_\alpha \leq y^\alpha$. Each extremity of these subintervals is such that $f(y_\alpha|x) = f(y^\alpha|x) = f_\alpha$. Note, as shown in [11], that this approach also allows to give an informative and convenient graphical display of the predicted regions by drawing confidence bands corresponding to, e.g. 50 % and 99% coverage probability.

A plug-in strategy to define the corresponding statistical predictor is discussed in the next two subsections.

4.2 Determination of the level by a density quantile approach

In order to determine the corresponding interval statistical predictor, a first step is to determine, for a given coverage probability α, the corresponding cut-off level f_α of equation (7). To that purpose, we assume x is fixed and follow the approach proposed by Hyndman [12]. For Y with conditional density $f(y|x)$, define the random variable $Z = f(Y|x)$. Then,

$$Y \in \mathcal{C}_\alpha \Leftrightarrow f(Y|x) \geq f_\alpha \Leftrightarrow Z \geq f_\alpha.$$

Therefore, $P(Y \in \mathcal{C}_\alpha) = \alpha \Leftrightarrow P(Z \geq f_\alpha) = \alpha$. So f_α is the $1 - \alpha$ quantile of Z. It thus can be estimated by the sample quantile from a set of i.i.d. observations $Z_1, \ldots, Z_n$ from the distribution of $Z = f(Y|X = x)$. As $f(y|x)$ is unknown, it has to be estimated by $\hat{f}(y|x)$. Therefore, the following two practical approaches to determine the level of the level-set can be proposed:

1. A Bootstrap technique for estimating f_α is to generate a i.i.d. pseudo-sample $(\hat{Y}_1, \ldots, \hat{Y}_N)$ from the estimated distribution $\hat{f}(y|x)$ of $f(y|x)$.

Then, $(\hat{Z}_1, \dots, \hat{Z}_N) := (\hat{f}(\hat{Y}_1|x), \dots, \hat{f}(\hat{Y}_N|x))$ will be a i.i.d. pseudo-sample from the distribution of Z. The level f_α is estimated by the sample quantile of the Z_i as

$$\hat{f}_\alpha := \hat{Z}_{j_\alpha,N},$$

with $j_\alpha = \lfloor (1-\alpha)N \rfloor$ and where $\hat{Z}_{j,N}$ denotes the jth order statistic of the sample $\hat{Z}_1, \dots, \hat{Z}_N$.

2. Alternatively, a more direct approach, especially if n is large, is to use the same set of observations $(Y_1, \dots, Y_n)$, and to calculate the quantile from the synthetic sample $\tilde{Z} = (\tilde{Z}_1, \dots, \tilde{Z}_n) := (\hat{f}_n(Y_1|x), \dots, \hat{f}_n(Y_n|x))$. The estimated value is defined analogously by

$$\hat{f}_\alpha := \tilde{Z}_{j\alpha,n}.$$

4.3 Calculation of predictive intervals

A natural plug-in estimate of the predictive set $\mathcal{C}_\alpha(x)$ defined by equation (6), would be to set

$$\mathcal{C}_{\alpha,n}(x) := \{y : \hat{f}_n(y|x) \geq \hat{f}_\alpha\},$$

where $\hat{f}_\alpha$ is the above mentioned estimate of the level f_α. Practically, recall that $\mathcal{C}_\alpha(x)$ is made up of the different subintervals $I_\alpha(x) = [y_\alpha, y^\alpha]$. The corresponding statistical interval estimate $\hat{I}_\alpha(x) = [\hat{y}_\alpha, \hat{y}^\alpha]$ with $\hat{y}_\alpha \leq \hat{y}^\alpha$ is then obtained by solving for y the equation $\hat{f}_n(y|x) = \hat{f}_\alpha$, viz.

$$\hat{y}_\alpha = \hat{f}_n^{-1}(\hat{f}_\alpha|x) \text{ and } \hat{y}^\alpha = \hat{f}_n^{-1}(\hat{f}_\alpha|x).$$

In the following, we assume the existence of these inverses, that is to say we consider that the level is reasonably chosen.

Convergence of the estimated predictive intervals is then obtained from the uniform convergence of the conditional density estimator, as shown in the next proposition.

Proposition 4.1. *Assume $\hat{f}_\alpha \overset{a.s.}{\to} f_\alpha$. Then $\hat{y}_\alpha \overset{a.s}{\to} y_\alpha$ and $\hat{y}^\alpha \overset{a.s}{\to} y^\alpha$, thus $\lambda(\mathcal{C}_{\alpha,n} \Delta \mathcal{C}_\alpha) \overset{a.s.}{\to} 0$.*

Proof. We do the proof only for $\hat{y}_\alpha$, the proof for $\hat{y}^\alpha$ being similar. Introduce the estimate y^*_α of y_α, had we known the true value f_α, i.e.

$$\hat{f}_n(y^*_\alpha|x) = f_\alpha.$$

Then,

$$P(|\hat{y}_\alpha - y_\alpha| > \epsilon) \leq P(|\hat{y}_\alpha - y^*_\alpha| > \epsilon/2) + P(|y^*_\alpha - y_\alpha| > \epsilon/2).$$

Since $\hat{f}_n^{-1}(.|x)$ is continuous at y_α, for every $\epsilon > 0$, there exists a $\delta_\epsilon > 0$, such that $|\hat{f}_n(y|x) - \hat{f}_n(y_\alpha|x)| \leq \delta_\epsilon/2$ implies $|y - y_\alpha| \leq \epsilon$. In particular, for $y = y^*_\alpha$, there exists a δ_ϵ such that

$$\begin{aligned} P(|y^*_\alpha - y_\alpha| > \epsilon/2) &\leq P(|\hat{f}_n(y^*_\alpha|x) - \hat{f}_n(y_\alpha|x)| > \delta_\epsilon) \\ &\leq P(|f_\alpha - \hat{f}_n(y_\alpha|x)| > \delta_\epsilon) \\ &\leq P(|f(y_\alpha|x) - \hat{f}_n(y_\alpha|x)| > \delta_\epsilon) \end{aligned}$$

and almost sure convergence of the conditional density estimator yields almost sure convergence of the y^*_α to y_α. Similarly, by continuity of $\hat{f}_n^{-1}(.|x)$ at y^*_α, there exists $\delta'_\epsilon > 0$, such that

$$P(|\hat{y}_\alpha - y^*_\alpha| > \epsilon/2) \leq P(|\hat{f}_n(\hat{y}_\alpha|x) - \hat{f}_n(y^*_\alpha|x)| > \delta'_\epsilon)$$

and almost sure convergence of $\hat{f}_\alpha \overset{a.s.}{\to} f_\alpha$ means that $|\hat{f}_n(\hat{y}_\alpha|x) - \hat{f}_n(y^*_\alpha|x)| \overset{a.s.}{\to} 0$, yielding $\hat{y}_\alpha - y_\alpha \overset{a.s.}{\to} 0$. □

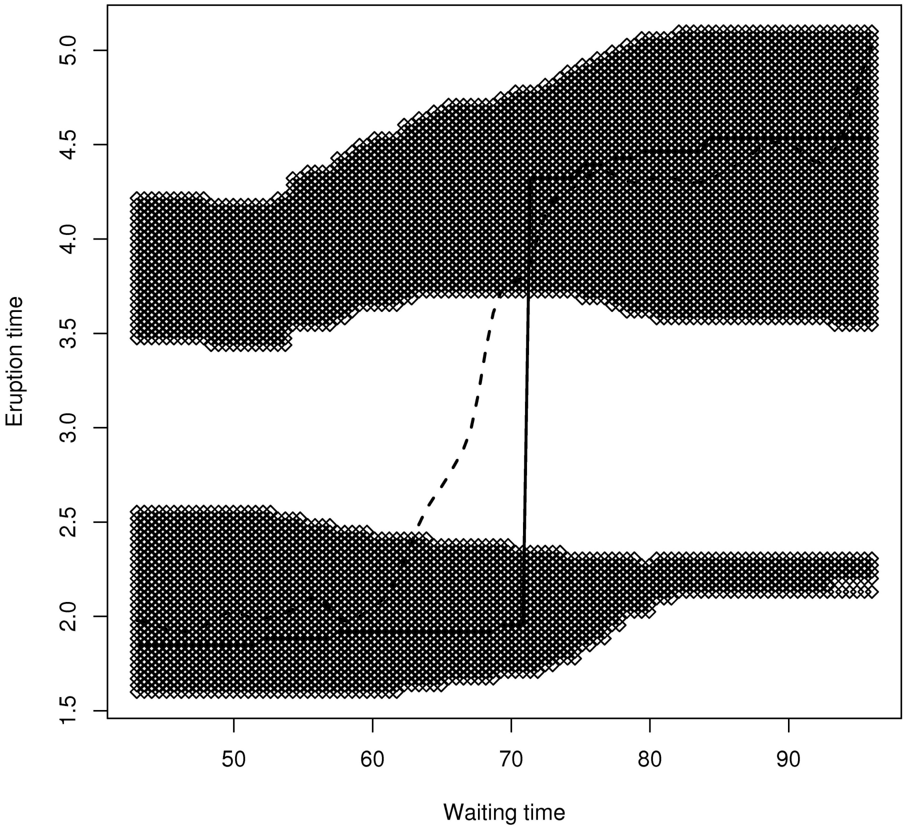
5 An illustration on a real-data set

To complement the asymptotic results obtained, i.e. valid for large samples, the proposed methodology is illustrated in a small sample setting on the Old Faifthful Geyser data set. The data consists in 272 records of the eruption time of the geyser and the waiting time between two successive eruptions. The aim is to predict the eruption time (Y) conditionally on the waiting time (X).

5.1 Small sample implementation

As noted in [8], since the copula density is of compact support $[0,1]^2$, the kernel method of estimation may suffer from boundary bias. Therefore, to alleviate this possible bias issue of the copula density part of the estimator, the quantile-copula density estimator is implemented with the modifications suggested in [8]. In particular, the Beta kernels mentioned herein were used. The bandwidth for the copula density was chosen such that it contains at least a fixed amount (20%) of the data, and for the bandwidth for the Y density by direct plug-in (see [21]). The conditional density was calculated on a regular rectangular grid of 100 values with the edges corresponding to the maximum and minimum values of the data. We computed the conditional mode (black line) as well as 50 percents level sets (shaded area) and the classical Nadaraya-Watson regression estimator (dashed line). The result of the simulation is displayed on figure 5.1.

Figure 2: Prediction of the Eruption duration from the Waiting time for the Old Faithful Geyser data



5.2 Results

The shaded area corresponding to the 50 percent predictive intervals clearly evidences the bimodal and nonlinear nature of the response of the eruption duration conditionally on the waiting time. The point prediction of the eruption duration given by the conditional mode switches between about 2 min., say 1.8 min, and, slightly more than 4 min., say 4.5 min., depending on the waiting time being lower or upper than a threshold value of about 70 min. As advocated in the introduction, such a phenomenon can not be inferred from the regression function. Moreover, the regression proposes as a point prediction a continuum of values between 1.8 min and 4.5 min, whereas the shaded area shows that observing a value in between this continuum (say between 2.5 min and 3.5 min) appears to be very unlikely.

6 Appendix

6.1 Uniform consistency of the kernel density estimators

We recall below for convenience some classical results of convergence of the kernel density estimators uniformly on sets. For additional references, see.g. [17, 21, 4]. In this section only, f denotes a generic density on $\mathbb{R}^d$.

6.1.1 Bias

If f is supposed to be twice differentiable with second partial derivatives uniformly bounded on J, the bias is also uniformly bounded on J: indeed,

$$\sup_{t\in J} |Ef_n(t) - f(t)| = h_n^2/2 \int K(y) y^T \{f''(y)\} y dy + o(h_n^2)$$

where

$$f''(t) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\bigg|_{x=t} \right)$$

is a shorthand for the Hessian of f, and where the $o(.)$ is independent of t. This comes from a so-called uniform Bochner type theorem, see e.g. Bosq and Lecoutre [4].

6.1.2 Uniform convergence in probability

The following theorem is a direct corollary of Bickel and Rosenblatt's [3] convergence result of the norm of the deviation of the kernel density estimator to a double exponential law:

Theorem 6.1 (Bickel and Rosenblatt)**.** *For f bounded and non-vanishing on a compact subset J included in the interior of supp(f), and a bandwidth sequence $h_n \to 0$, such that $nh_n^d \to \infty$, $nh_n^d / \ln n \to \infty$,*

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

Therefore, for the choice of the bandwidth $h_n \simeq (\ln n/n)^{1/d+4}$ which realises the optimal trade-off between the bias and variance, one gets, by combining this result with the one on the bias in section 6.1.1 above, the following result in probability:

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

which is the optimal speed in the minimax sense in the class of density functions with bounded second derivatives, according to Hasminskii [10].

6.1.3 Uniform almost sure convergence

We cite Stute's [18, 19] theorem on the uniform convergence of the kernel density estimator, see also Deheuvels [5], Bosq and Lecoutre [4] :

Theorem 6.2 (Stute)**.** *Let J be a compact subset of $\mathbb{R}^d$, included in the support of f.*

i) If the kernel K is of bounded support, and of finite variation (e.g. if K has bounded partial derivative of order two),

ii) if the density f is uniformly continuous on J, is bounded away from zero and infinity on J: $0 \le m < f|_J < +\infty$,

iii) if the marginal densities f_i of f, $i = 1, \ldots, d$ are bounded away from zero and infinity on J,

iv) if the bandwidth $h_n \to 0$ satisfy $nh_n^d \to +\infty$, $\ln(1/h_n^d) = o(nh_n^d)$, and $\ln(1/h_n^d)/(\ln \ln n) \to +\infty$

then, with probability one,

$$\lim_{n\to\infty} \sup_{t \in J} \sqrt{\frac{nh_n^d}{2 \ln h_n^{-d}}} \left| \frac{f_n(t) - Ef_n(t)}{\sqrt{f(t)}} \right| = \left(\int K^2 \right)^{1/2}.$$

Remark 6.1. *if the last condition on the bandwidth is suppressed, the theorem remains valid with* $\lim$ *replaced with* $\overline{\lim}$. *With the usual choice of bandwidth* $h_n \simeq (\ln n/n)^{1/(d+4)}$ *to deal with the bias, one gets the almost sure uniform convergence of the kernel density estimator at the rate* $(\ln n/n)^{2/(d+4)}$.

6.2 Two uniform approximation propositions

The following two propositions are key to prove theorem 2.1. Proposition 6.1 gives the a.s. and in probability rates of approximation of the quantile density estimator $\hat{c}_n$ based on the approximate data $(F_n(X_i), G_n(Y_i))$ from the moving w.r.t n location $(F_n(x), G_n(y))$ to the fixed one $(F(x), G(y))$, and proposition 6.2 the a.s. and in probability rates of approximations of the quantile density estimator $\hat{c}_n$ based on the approximate data by the pseudo estimator c_n based on the pseudo data $(F(X_i), G(Y_i))$.

Proposition 6.1. *Let the regularity assumptions* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be satisfied, then, for a compact set* $D \subset (0,1)^2$, $a_n \to 0$ *and* $na_n^3/\ln n \to \infty$ *entails*

$$\sup_{(x,y)\in D} |\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y))| = O_P\left(\frac{1}{na_n^4} + \frac{\ln n}{n^{1/2}}\right),$$

$$\sup_{(x,y)\in D} |\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y))| = O_{a.s}\left(\frac{\ln\ln n}{na_n^4} + \frac{\ln n(\ln\ln n)^{1/2}}{n^{1/2}}\right).$$

Proof. For convenience, set the norm of a vector as the max norm, $||(x_1, \ldots, x_d)|| = \max_{1\le j\le d} |x_j|$. Set $D = [u_0, u_\infty] \times [v_0, v_\infty] \subset (0,1)^2$ a compact subset where $0 < u_0 \le u_\infty < 1$ and $0 < v_0 \le v_\infty < 1$. Note ^{T} for the transpose of a matrix. Set

$$\Delta_n(x,y) := \hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y)) = \frac{1}{na_n^2}\sum_{i=1}^{n} \Delta_{i,n}(x,y) \qquad (8)$$

with

$$\Delta_{i,n}(x,y) := K\left(\frac{F_n(x) - F_n(X_i)}{a_n}, \frac{G_n(y) - G_n(Y_i)}{a_n}\right) - K\left(\frac{F(x) - F_n(X_i)}{a_n}, \frac{G(y) - G_n(Y_i)}{a_n}\right).$$

For notational simplicity, define

$$Z_n(x,y) := \begin{pmatrix} F_n(x) - F(x) \\ G_n(y) - G(y) \end{pmatrix}, \quad Z_{i,n} := \begin{pmatrix} F(X_i) - F_n(X_i) \\ G(Y_i) - G_n(Y_i) \end{pmatrix}.$$

We first express $\Delta_{i,n}$ at a fixed location by a Taylor expansion, viz.

$$\Delta_{i,n} = Z_n(x,y)^T \frac{\nabla K}{a_n}\left(\frac{F(x)-F_n(X_i)}{a_n}, \frac{G(y)-G_n(Y_i)}{a_n}\right) + \frac{||Z_n||_\infty^2}{a_n^2} R_3 \quad (9)$$

where R_3 is uniformly bounded almost surely by the boundedness assumptions on the second order derivatives of the kernel (assumption **K (iv)**). We then proceed from the data $(F_n(X_i), G_n(Y_i))$ to the pseudo ones $(F(X_i), G(Y_i))$ by another Taylor expansion,

$$\nabla K\left(\frac{F(x)-F_n(X_i)}{a_n}, \frac{G(y)-G_n(Y_i)}{a_n}\right) = \nabla K\left(\frac{F(x)-F(X_i)}{a_n}, \frac{G(y)-G(Y_i)}{a_n}\right) + \frac{Z_{i,n}^T}{a_n} R_2 \quad (10)$$

with $||R_2|| = O_{a.s.}(1)$, again by assumption **K (iv)**. Thus, plugging (9) and (10) in (8),

$$\begin{aligned}\Delta_n(x,y) = {} & \frac{Z_n^T(x,y)}{na_n^3}\sum_{i=1}^n \nabla K\left(\frac{F(x)-F(X_i)}{a_n}, \frac{G(y)-G(Y_i)}{a_n}\right) \\ & + \frac{Z_n^T(x,y)}{na_n^4}\sum_{i=1}^n Z_{i,n}^T R_2 + R_3\frac{||Z_n||_\infty^2}{a_n^4}. \end{aligned} \quad (11)$$

Notice that $|F_n(X_i) - F(X_i)| \leq ||F_n - F||_\infty$ and $|G_n(Y_i) - G(Y_i)| \leq ||G_n - G||_\infty$ a.s. for every $i = 1, \ldots, n$. From Chung-Mogulskii's law of the iterated logarithm,

$$||F_n - F||_\infty = O_{a.s.}\left(\sqrt{\frac{\ln\ln n}{n}}\right), \quad \text{or } = O_P\left(\frac{1}{\sqrt{n}}\right). \quad (12)$$

and similarly for $||G_n - G||$, so that the norm of $Z_{i,n}$ is independent of i and such that

$$||Z_{i,n}|| = O_P(1/\sqrt{n}), \quad \text{or } = O_{a.s.}(\sqrt{\ln\ln n/n}). \quad (13)$$

In the same manner,

$$||Z_n||_\infty = O_P(1/\sqrt{n}), \quad \text{or } = O_{a.s.}(\sqrt{\ln\ln n/n}). \quad (14)$$

Therefore the last two terms in (11) are of order $O_P\left(\frac{1}{na_n^4}\right)$, or $O_{a.s}\left(\frac{\ln\ln n}{na_n^4}\right)$, uniformly for $(x,y) \in D$.

For the first term in (11), by Cauchy-Schwarz inequality,

$$\sup_{(x,y)\in D}\left|\frac{Z_n^T(x,y)}{na_n^3}\sum_{i=1}^n \nabla K\left(\frac{F(x)-F(X_i)}{a_n},\frac{G(y)-G(Y_i)}{a_n}\right)\right|$$
$$\leq ||Z_n||_\infty \sup_{(x,y)\in D}\left\|\frac{1}{na_n^3}\sum_{i=1}^n \nabla K\left(\frac{F(x)-F(X_i)}{a_n},\frac{G(y)-G(Y_i)}{a_n}\right)\right\|$$

From the convergence results of the kernel estimator of the gradient of the density $c(u,v)$ (see Scott [17] and the previous section 6.1), and the assumption **A (ii)** of boundedness of the gradient of the copula density on D, one gets with $na_n^3/\ln n \to \infty$ that

$$\sup_{(x,y)\in D}\left\|\frac{1}{na_n^3}\sum_{i=1}^n \nabla K\left(\frac{F(x)-F(X_i)}{a_n},\frac{G(y)-G(Y_i)}{a_n}\right)\right\| = O_P(\ln n), \quad \text{or } = O_{a.s.}(\ln n)$$

In turn, with (14), the first term in (11), is an $= O_P(\ln n/n^{-1/2})$ or $O_{a.s.}(\ln n(\ln\ln n/n)^{1/2})$. Recollecting all elements yields the claimed result. □

Proposition 6.2. *Let the regularity assumptions* ***A*** *and* ***B*** *be satisfied, then, for a compact set $D \subset (0,1)^2$, and a bandwidth such as $a_n \simeq \left(\frac{\ln n}{n}\right)^{1/6}$, one has*

$$\sup_{(u,v)\in D}|\hat{c}_n(u,v)-c_n(u,v)| = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{1/3}\right) \text{ or } O_P\left(\left(\frac{\ln n}{n}\right)^{1/3}\right)$$

Proof. We proceed similarly. Write

$$\Delta'(u,v) := \hat{c}_n(u,v) - c_n(u,v) = \frac{1}{na_n^2}\sum_{i=1}^n \Delta_{i,n}(u,v),$$

with

$$\Delta'_{i,n}(u,v) := K\left(\frac{u-F_n(X_i)}{a_n},\frac{v-G_n(Y_i)}{a_n}\right) - K\left(\frac{u-F(X_i)}{a_n},\frac{v-G(Y_i)}{a_n}\right),$$

and define

$$W_{i,n}(u,v) := \nabla K\left(\frac{u-F(X_i)}{a_n},\frac{v-G(Y_i)}{a_n}\right).$$

For every fixed $(u,v)\in[0,1]^2$, since the kernel K is twice differentiable, there exists, by Taylor expansion, random variables $\tilde{U}_{i,n}$ and $\tilde{V}_{i,n}$ such that, almost surely,

$$\begin{aligned}\Delta'(u,v) &= \frac{1}{na_n^3}\sum_{i=1}^{n} Z_{i,n}^T \nabla K\left(\frac{u-F(X_i)}{a_n}, \frac{v-G(Y_i)}{a_n}\right)\\ &+ \frac{1}{2na_n^4}\sum_{i=1}^{n} Z_{i,n}^T \nabla^2 K\left(\frac{u-\tilde{U}_{i,n}}{a_n}, \frac{v-\tilde{V}_{i,n}}{a_n}\right) Z_{i,n}\\ &:= \Delta_1' + \Delta_2',\end{aligned}$$

where ∇K and $\nabla^2 K$ the gradient and the Hessian respectively of the multivariate kernel function K. By centering at expectations, decompose further the first term Δ_1' as,

$$\begin{aligned}\Delta_1' &= \frac{1}{na_n^3}\sum_{i=1}^{n} Z_{i,n}^T W_{i,n}(u,v)\\ &= \frac{1}{na_n^3}\sum_{i=1}^{n} Z_{i,n}^T.\left(W_{i,n}(u,v) - EW_{i,n}(u,v)\right)\\ &+ \frac{1}{na_n^3}\sum_{i=1}^{n} Z_{i,n}^T.EW_{i,n}(u,v)\\ &:= \Delta_{11}' + \Delta_{12}'.\end{aligned}$$

- **Negligibility of Δ_2'**

  By bounding uniformly the Hessian of the kernel, we get that

  $$\sup_{(u,v)\in D} |\Delta_2'(u,v)| \leq \frac{||Z_{i,n}||^2}{a_n^4} R_2,$$

  where $R_2 = O_{a.s.}(1)$ uniformly. With (13), we get eventually that

  $$\sup_{(u,v)\in D} |\Delta_2(u,v)| = O_P(n^{-1}a_n^{-4}),\ \text{or}\ O_{a.s.}((\ln\ln n)/(na_n^4)). \qquad (15)$$

- **Negligibility of Δ_{12}'**

  Notice that $na_n^{-3}\sum_{i=1}^{n} W_{i,n}(u,v)$ is the kernel estimator of the gradient $\nabla c(u,v)$ and that in the expression of the bias of the kernel estimator, equation 6.1.1, the $O(.)$ is uniform in (u,v). Therefore one gets that

  $$\sup_{(u,v)\in D} ||EW_{i,n}(u,v) - a_n^3 \nabla c(u,v)|| = O(a_n^5).$$

Thus,

$$\sup_{(u,v)\in D} |\Delta'_{12}(u,v)| = O_P(1/\sqrt{n}), \text{ or } O_{a.s.}((\ln\ln n/n)^{1/2}). \tag{16}$$

- **Negligibility of Δ'_{11}**

  We use a chaining argument: define a covering of D by M_n^2 compact hypercubes D_k centered in (u_k, v_k),

  $$D_k = \{(u,v) \in D : ||(u,v) - (u_k, v_k)|| \leq 1/M_n\} \ , 1 \leq k \leq M_n^2$$

  with

  $$\overset{\circ}{D}_k \cap \overset{\circ}{D}_{k'} = \emptyset \ , 1 \leq k \neq k' \leq M_n^2.$$

  One can write

  $$\begin{aligned} \sup_{(u,v)\in D} |\Delta'_{11}(u,v)| \leq & \max_{1\leq k\leq M_n^2} \sup_{(u,v)\in D_k} |\Delta'_{11}(u,v) - \Delta'_{11}(u_k,v_k)| \\ & + \max_{1\leq k\leq M_n^2} |\Delta'_{11}(u_k,v_k)| \\ & := (I) + (II). \end{aligned}$$

- **Negligibility of (I)**

  For (I), by boundedness and Lipshitz assumption on the product kernel K, there exists a constant C such that,

  $$||\nabla K(u,v) - \nabla K(u_k,v_k)|| \leq C||(u,v) - (u_k,v_k)||.$$

  Therefore for $(u,v) \in D_k$,

  $$\left\| \nabla K\left(\frac{u - F(X_i)}{a_n}, \frac{v - G(Y_i)}{a_n}\right) - \nabla K\left(\frac{u_k - F(X_i)}{a_n}, \frac{v_k - G(Y_i)}{a_n}\right) \right\| \leq \frac{C}{M_n a_n}$$

  since K is product-shaped. In turn, the same bound is valid by Jensen's inequality for the expectations of the difference, so that

  $$(I) \leq \frac{2C||Z_n||}{M_n a_n^4}. \tag{17}$$

  Setting $M_n = n^{1/2}a_n^{-3} \simeq n/\sqrt{\ln n}$ for $a_n \simeq (\ln n/n)^{1/6}$, one has that $(I) = o_{a.s.}\left(\sqrt{\frac{\ln n}{na_n^2}}\right)$ or $o_P((na_n^2)^{-1/2})$.

- **Negligibility of** (II)

  For the second term, set $A_i(u,v) = W_{i,n}(u,v) - EW_{i,n}(u,v)$, and bound, for each k,

$$\begin{aligned}|\Delta'_{11}(u_k,v_k)| &\leq \frac{||Z_n||}{na_n^3}\sum_{i=1}^{n}||A_i(u_k,v_k)|| \\ &\leq \frac{||Z_n||}{na_n^3}\sum_{i=1}^{n}(||A_i(u_k,v_k)|| - E||A_i(u_k,v_k)|| + E||A_i(u_k,v_k)||) \\ &\leq \frac{||Z_n||}{na_n^3}\sum_{i=1}^{n}\eta_i(u_k,v_k) + \frac{||Z_n||}{na_n^3}\sum_{i=1}^{n}E||A_i(u_k,v_k)||\end{aligned}$$

  where we have set $\eta_i(u_k,v_k) = ||A_i(u_k,v_k)|| - E||A_i(u_k,v_k)||$.

  For the expectation term, as the product kernel is of finite variation, and with the assumption that the gradient of the copula density remains bounded on D, one has that

$$\max_{1\leq k\leq M_n^2} E||A_i(u_k,v_k)|| = O(a_n^3).$$

  In turn,

$$\max_{1\leq k\leq M_n^2}\frac{||Z_n||}{na_n^3}\sum_{i=1}^{n}E||A_i(u_k,v_k)|| = O_P(n^{-1/2}) \text{ , or } O_{a.s.}\left(\left(\frac{\ln\ln n}{n}\right)^{1/2}\right). \tag{18}$$

  It remains to deal with the deviation term

$$\max_{1\leq k\leq M_n^2}\frac{||Z_n||}{na_n^3}\sum_{i=1}^{n}\eta_i(u_k,v_k).$$

  We have

$$P\left(\max_{1\leq k\leq M_n^2}|\sum_{i=1}^{n}\eta_i(u_k,v_k)| > \varepsilon\right) \leq \sum_{k=1}^{M_n^2}P\left(|\sum_{i=1}^{n}\eta_i(u_k,v_k)| > \varepsilon\right)$$

  and apply Hoeffding's inequality to the summand, to get that, for every $\varepsilon > 0$,

$$P\left(|\sum_{i=1}^{n}\eta_i(u_k,v_k)| > \varepsilon\sqrt{n\ln n}\right) \leq \exp\left(-\frac{\varepsilon^2\ln n}{C}\right)$$

for a constant C independent of k, which exists by the boundedness assumption on the gradient of the kernel. Thus,

$$\begin{aligned} P\left(\max_{1\leq k\leq M_n^2}|\sum_{i=1}^n \eta_i(u_k,v_k)| > \varepsilon\sqrt{n\ln n}\right) &\leq M_n^2\exp\left(-\frac{\varepsilon^2\ln n}{C}\right) \\ &\leq \exp\left(\sqrt{2\ln M_n}-\frac{\varepsilon^2\ln n}{C}\right). \end{aligned}$$

For $a_n \simeq (\ln n/n)^{1/6}$ and $M_n = n^{1/2}a_n^{-3} \simeq n/\sqrt{\ln n}$,

$$\exp\left(\sqrt{2\ln M_n}-\frac{\varepsilon^2\ln n}{C}\right) \approx \exp\left(-\frac{\varepsilon^2\ln n}{C}\right) = \frac{1}{n^{\varepsilon^2/C}}$$

which is absolutely summable for an ε large enough. Therefore,

$$\max_{1\leq k\leq M_n^2}|\sum_{i=1}^n \eta_i(u_k,v_k)| = O_{a.co.}\left(\sqrt{n\ln n}\right)$$

and eventually,

$$\frac{||Z_n||}{na_n^3}\max_{1\leq k\leq M_n^2}|\sum_{i=1}^n \eta_i(u_k,v_k)| = O_{a.s.}\left(\frac{\sqrt{\ln n\ln\ln n}}{na_n^3}\right) \tag{19}$$

for the choice $a_n \simeq (\ln n/n)^{1/6}$.

Recollecting elements (15), (16), (17), (18), (19) gives the claimed result for the given choice of a_n.

□

References

[1] Christophe Abraham, Gérard Biau, and Benoît Cadre. Simple estimation of the mode of a multivariate density. *Canad. J. Statist.*, 31(1):23–34, 2003.

[2] Christophe Abraham, Gérard Biau, and Benoît Cadre. On the asymptotic properties of a simple estimate of the mode. *ESAIM Probab. Stat.*, 8:1–11 (electronic), 2004.

[3] P. J. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *Ann. Statist.*, 1:1071–1095, 1973.

[4] Denis Bosq and Jean-Pierre Lecoutre. *Théorie de l'estimation fonctionnelle.* Collection Economie et Statistiques Avancées. Economica, Paris, 1987.

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

[6] William F. Eddy. Optimum kernel estimators of the mode. *Ann. Statist.*, 8(4):870–882, 1980.

[7] William F. Eddy. The asymptotic distributions of kernel estimators of the mode. *Z. Wahrsch. Verw. Gebiete*, 59(3):279–290, 1982.

[8] Olivier P. Faugeras. A quantile-copula approach to conditional density estimation. *Journal of Multivariate Analysis*, 100(9):2083–2099, 2009.

[9] F. Ferraty and P. Vieu. *Nonparametric functional data analysis.* Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

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

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

[12] Rob J. Hyndman. Computing and graphing highest density regions. *American Statistician*, 50:120–126, 1996.

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

[14] V. D. Konakov. The asymptotic normality of the mode of multivariate distributions. *Teor. Verojatnost. i Primenen.*, 18:836–842, 1973.

[15] Roger B. Nelsen. *An introduction to copulas.* Springer Series in Statistics. Springer, New York, second edition, 2006.

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

[17] D. 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.

[18] Winfried Stute. The oscillation behavior of empirical processes. *Ann. Probab.*, 10(1):86–107, 1982.

[19] Winfried Stute. The oscillation behavior of empirical processes: the multivariate case. *Ann. Probab.*, 12(2):361–379, 1984.

[20] Abraham Wald. *Statistical Decision Functions.* John Wiley & Sons Inc., New York, N. Y., 1950.

[21] M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability.* Chapman and Hall Ltd., London, 1995.