Beta kernels and transformed kernels applications to copulas and quantiles

Arthur Charpentier

Université Rennes 1

arthur.charpentier@univ-rennes1.fr

http ://freakonometrics.blog.free.fr/

UNIVERSITE DE RENNES 1

Statistical seminar, Université Laval, Québec, April 2011

Agenda

- General introduction and kernel based estimation

Copula density estimation

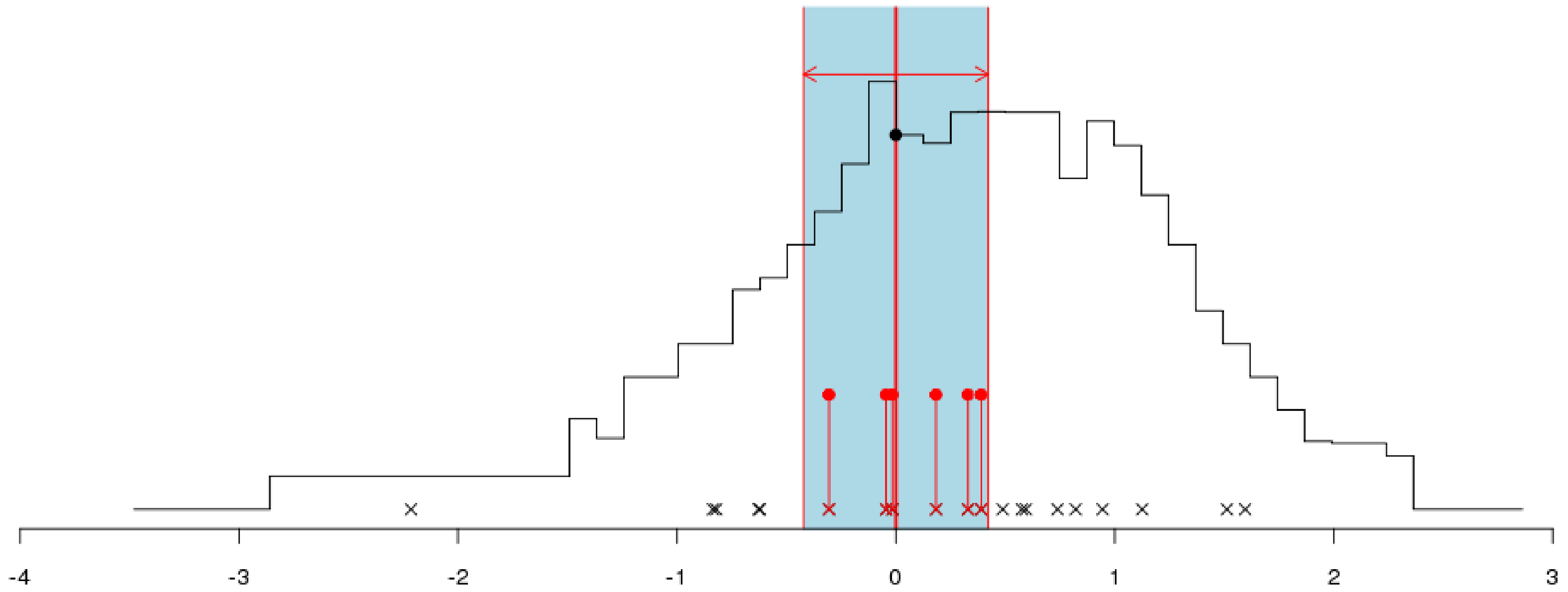
- Kernel based estimation and bias
- Beta kernel estimation
- Transforming observations

Quantile estimation

- Parametric estimation
- Semiparametric estimation, extreme value theory
- Nonparametric estimation
- Transforming observations

Moving histogram to estimate a density

A natural way to estimate a density at x from an i.i.d. sample $\{X_1, \cdots, X_n\}$ is to count (and then normalized) how many observations are in a neighborhood of x, e.g. $|x - X_i| \leq h$,

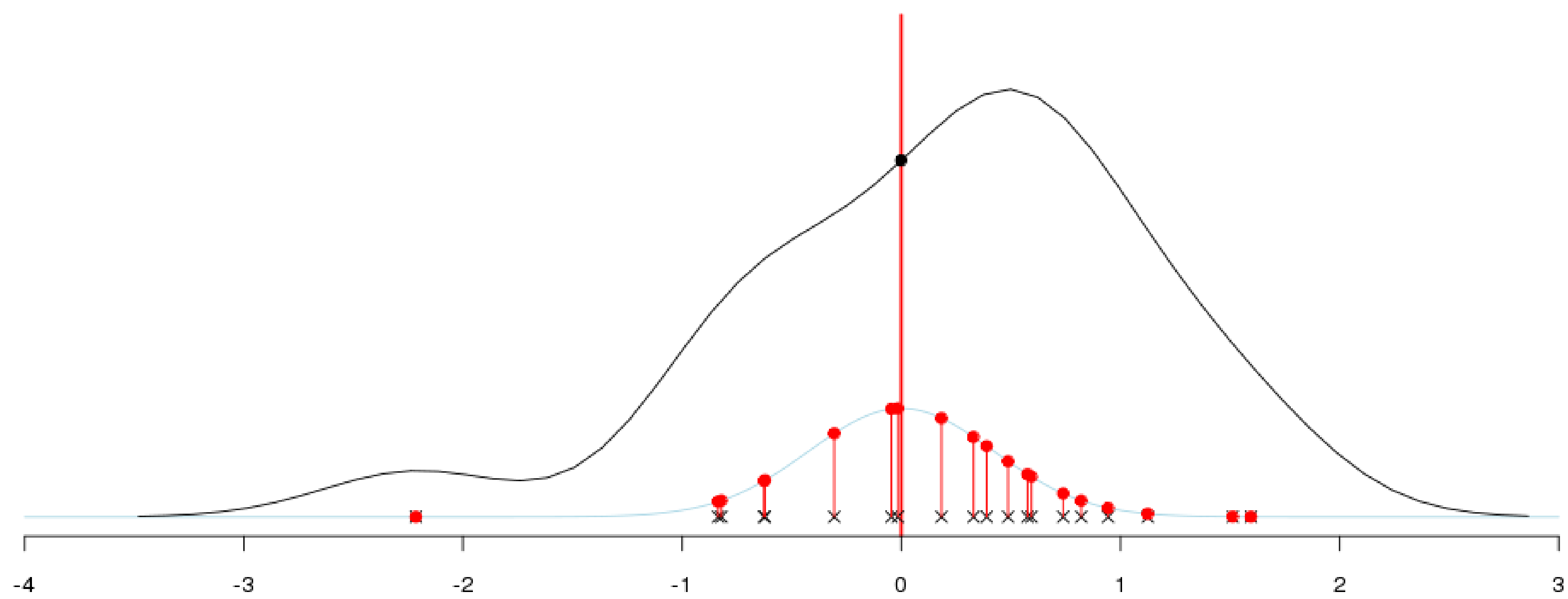


Kernel based estimation

Instead of a step function $\mathbf{1}(|x - X_i| \leq h)$ consider so smoother functions,

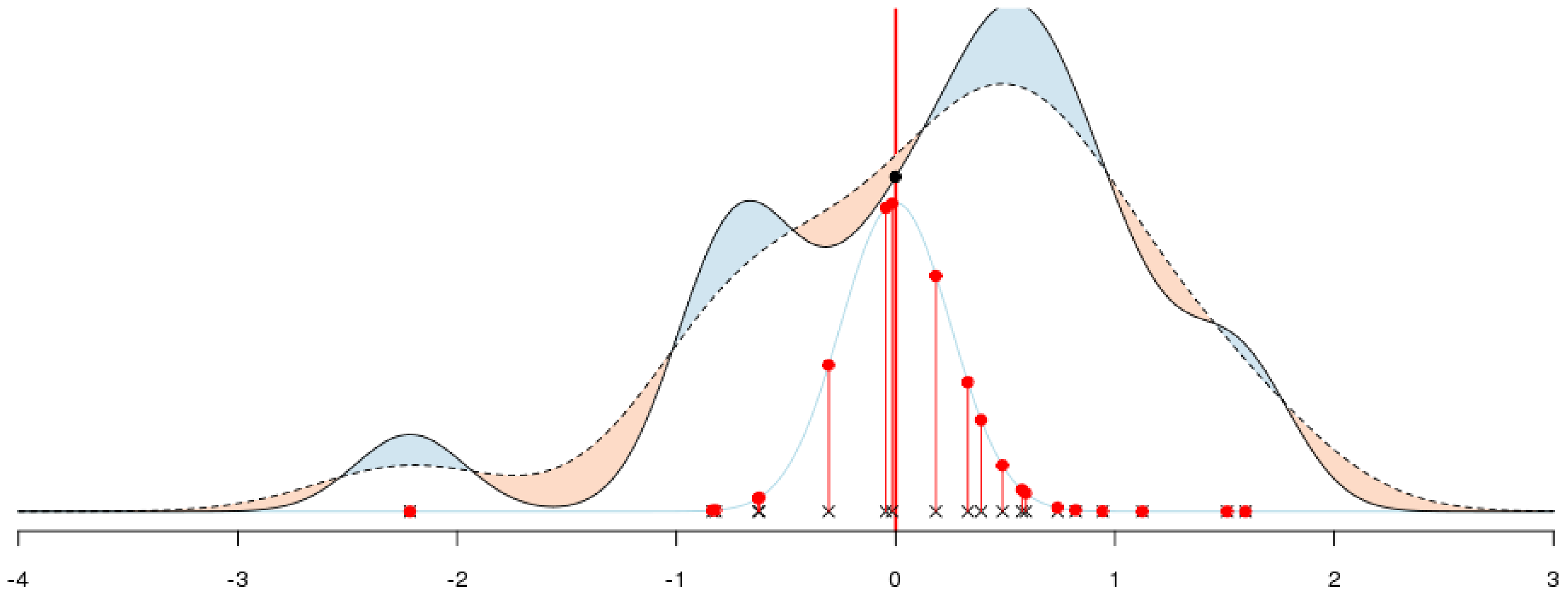
$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where $K(\cdot)$ is a kernel i.e. a non-negative real-valued integrable function such that $\displaystyle\int_{-\infty}^{+\infty} K(u)\,du = 1$ so that $\widehat{f}_h(\cdot)$ is a density , and $K(\cdot)$ is symmetric, i.e. $K(-u) = K(u)$.


-4
-3
-2
-1
0
1
2
3

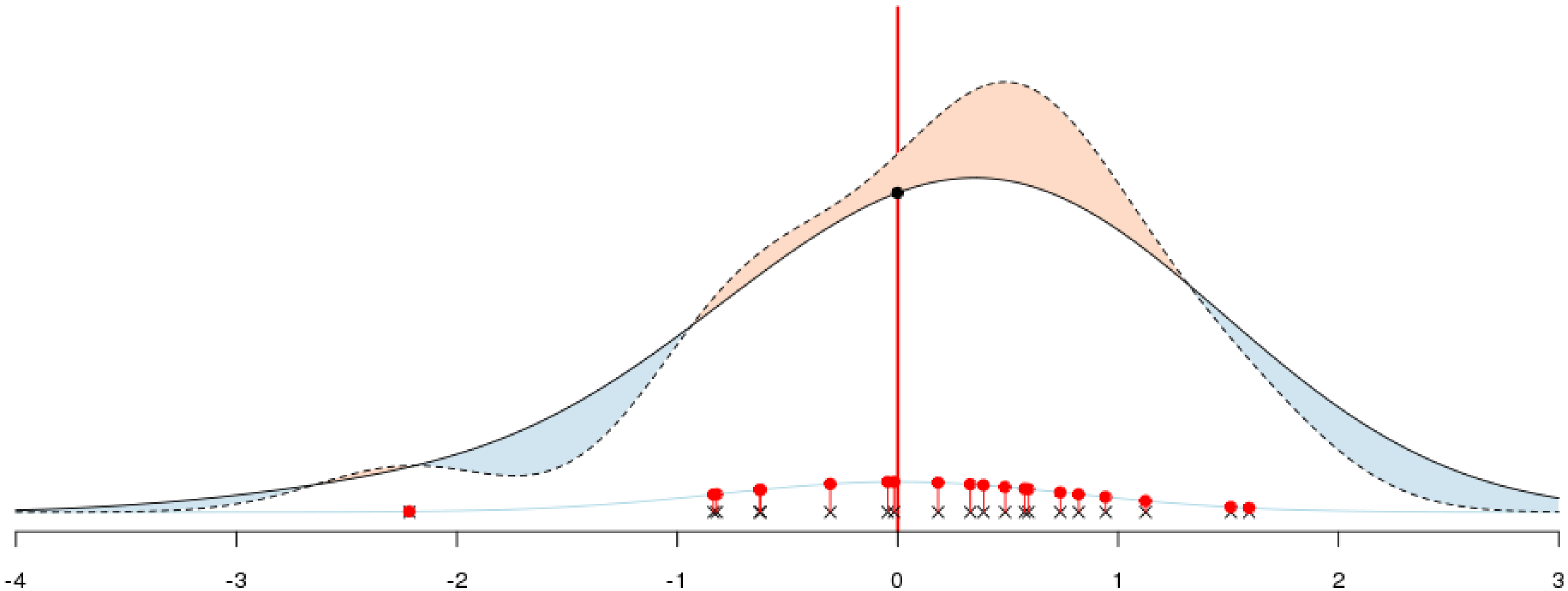
Standard kernels are

- uniform (rectangular) $K(u) = \frac{1}{2}\,\mathbf{1}_{\{|u|\leq 1\}}$
- triangular $K(u) = (1 - |u|)\,\mathbf{1}_{\{|u|\leq 1\}}$
- Epanechnikov $K(u) = \frac{3}{4}(1 - u^2)\,\mathbf{1}_{\{|u|\leq 1\}}$
- Gaussian $K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$



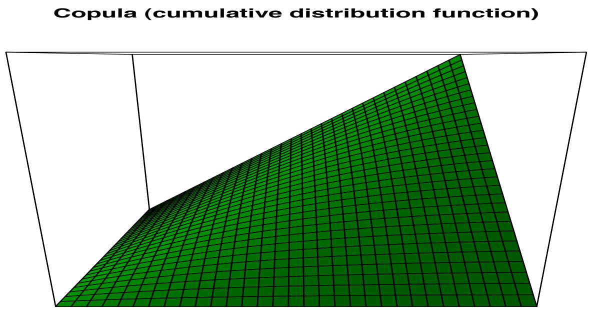
Standard kernels are

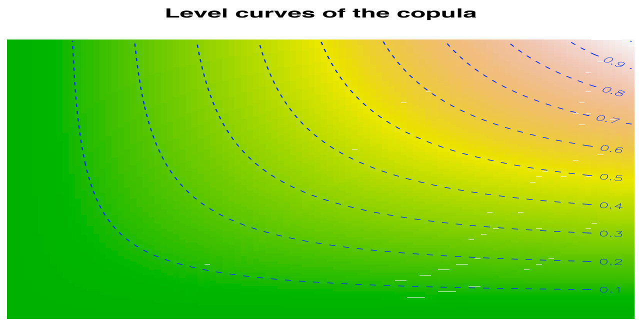
- uniform (rectangular) $K(u) = \dfrac{1}{2}\,\mathbf{1}_{\{|u|\leq 1\}}$
- triangular $K(u) = (1 - |u|)\,\mathbf{1}_{\{|u|\leq 1\}}$
- Epanechnikov $K(u) = \dfrac{3}{4}(1 - u^2)\,\mathbf{1}_{\{|u|\leq 1\}}$
- Gaussian $K(u) = \dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{1}{2}u^2\right)$

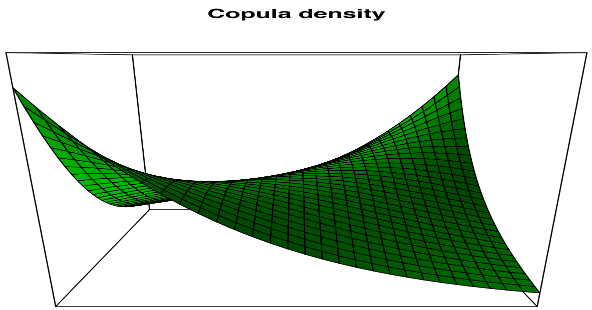


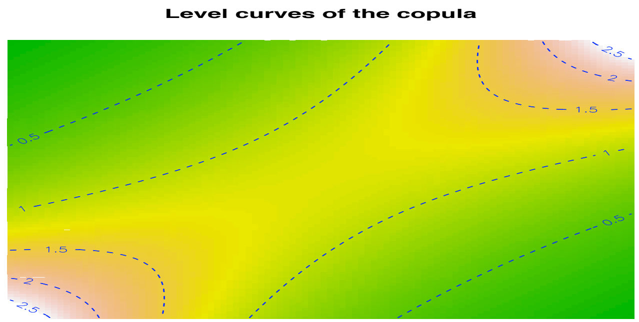
word about copulas

A 2-dimensional *copula* is a 2-dimensional cumulative distribution function restricted to $[0,1]^2$ with standard uniform margins.









If C is twice differentiable, let c denote the density of the copula.

Sklar theorem : Let C be a copula, and F_X and F_Y two marginal distributions, then $F(x,y) = C(F_X(x), F_Y(y))$ is a bivariate distribution function, with $F \in \mathcal{F}(F_X, F_Y)$.

Conversely, if $F \in \mathcal{F}(F_X, F_Y)$, there exists C such that $F(x,y) = C(F_X(x), F_Y(y)$. Further, if F_X and F_Y are continuous, then C is unique, and given by

$$C(u,v) = F(F_X^{-1}(u), F_Y^{-1}(v)) \text{ for all } (u,v) \in [0,1] \times [0,1]$$

We will then define the copula of F, or the copula of (X,Y).

Motivation

Example Loss-ALAE : consider the following dataset, were the X_i's are loss amount (paid to the insured) and the Y_i's are allocated expenses. Denote by R_i and S_i the respective ranks of X_i and Y_i. Set $U_i = R_i/n = \hat{F}_X(X_i)$ and $V_i = S_i/n = \hat{F}_Y(Y_i)$.

Figure 1 shows the log-log scatterplot $(\log X_i, \log Y_i)$, and the associate copula based scatterplot (U_i, V_i).

Figure 2 is simply an histogram of the (U_i, V_i), which is a nonparametric estimation of the copula density.

Note that the histogram suggests strong dependence in upper tails (the interesting part in an insurance/reinsurance context).

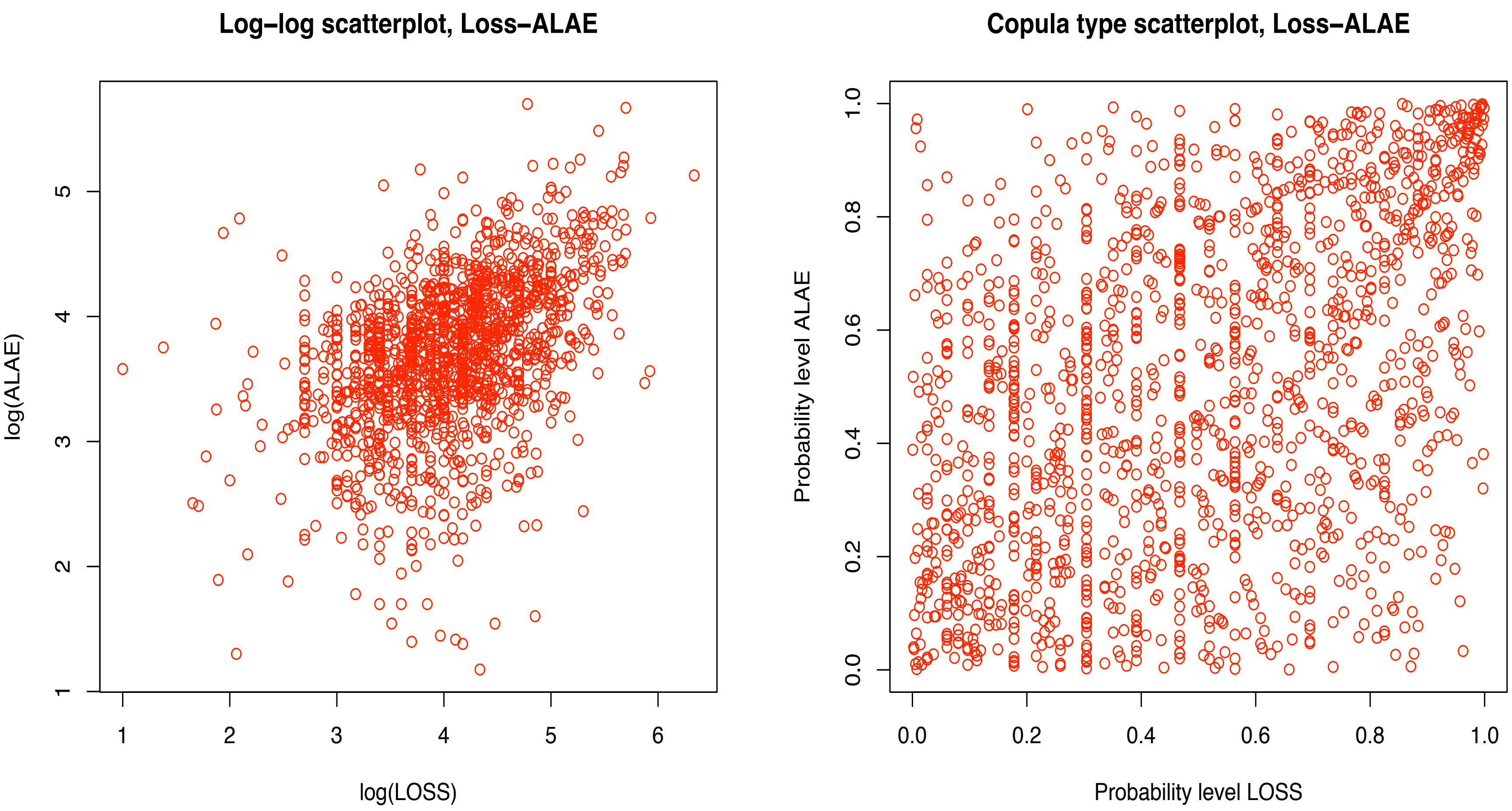


Figure 1 – Loss-ALAE, scatterplots (log-log and copula type).

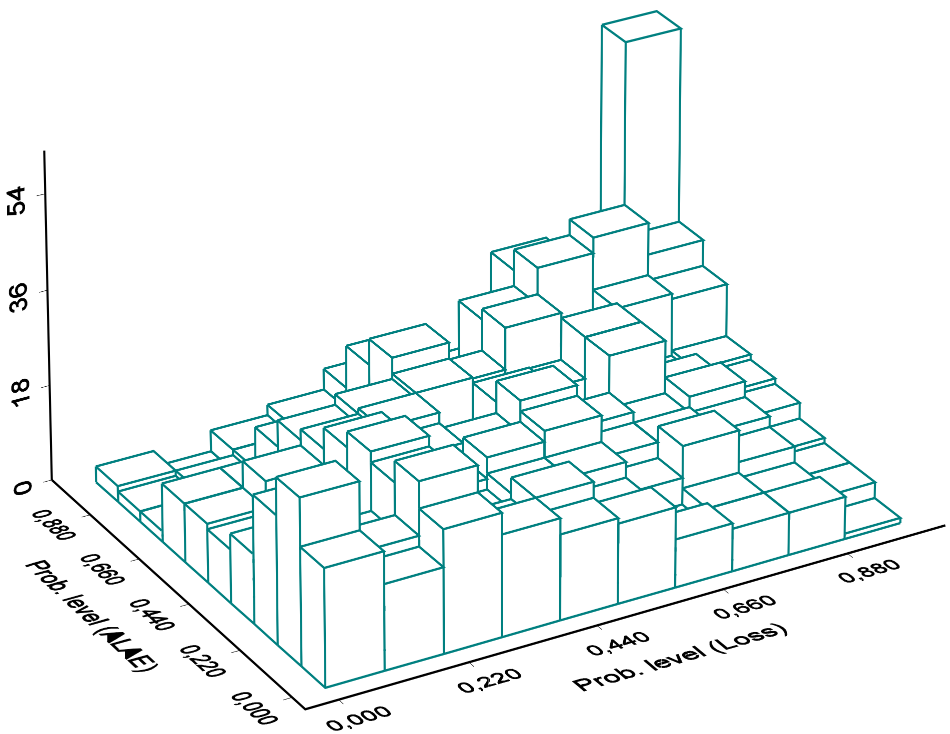


Figure 2 – Loss-ALAE, histogram of copula type transformation.

Why nonparametrics, instead of parametrics ?

In parametric estimation, assume the the copula density c_θ belongs to some given family $\mathcal{C} = \{c_\theta, \theta \in \Theta\}$. The tail behavior will crucially depend on the tail behavior of the copulas in $\mathcal{C}$

Example : Table below show the probability that both X and Y exceed high thresholds ($X > F_X^{-1}(p)$ and $Y > F_Y^{-1}(p)$), for usual copula families, where parameter θ is obtained using maximum likelihood techniques.

p	Clayton	Frank	Gaussian	Gumbel	Clayton*	max/min
90%	1.93500%	2.73715%	4.73767%	4.82614%	5.66878%	2.93
95%	0.51020%	0.78464%	1.99195%	2.30085%	2.78677%	5.46
99%	0.02134%	0.03566%	0.27337%	0.44246%	0.55102%	25.82
99.9%	0.00021%	0.00037%	0.01653%	0.04385%	0.05499%	261.85

Probability of exceedances, for given parametric copulas, $\tau = 0.5$.

Figure 3 shows the graphical evolution of $p \mapsto \mathbb{P}\left(X > F_X^{-1}(p), Y > F_Y^{-1}(p)\right)$. If the original model is an multivariate student vector (X, Y), the associated probability is the upper line. If either marginal distributions are misfitted (e.g. Gaussian assumption), or the dependence structure is mispecified (e.g. Gaussian assumption), probabilities are always underestimated.

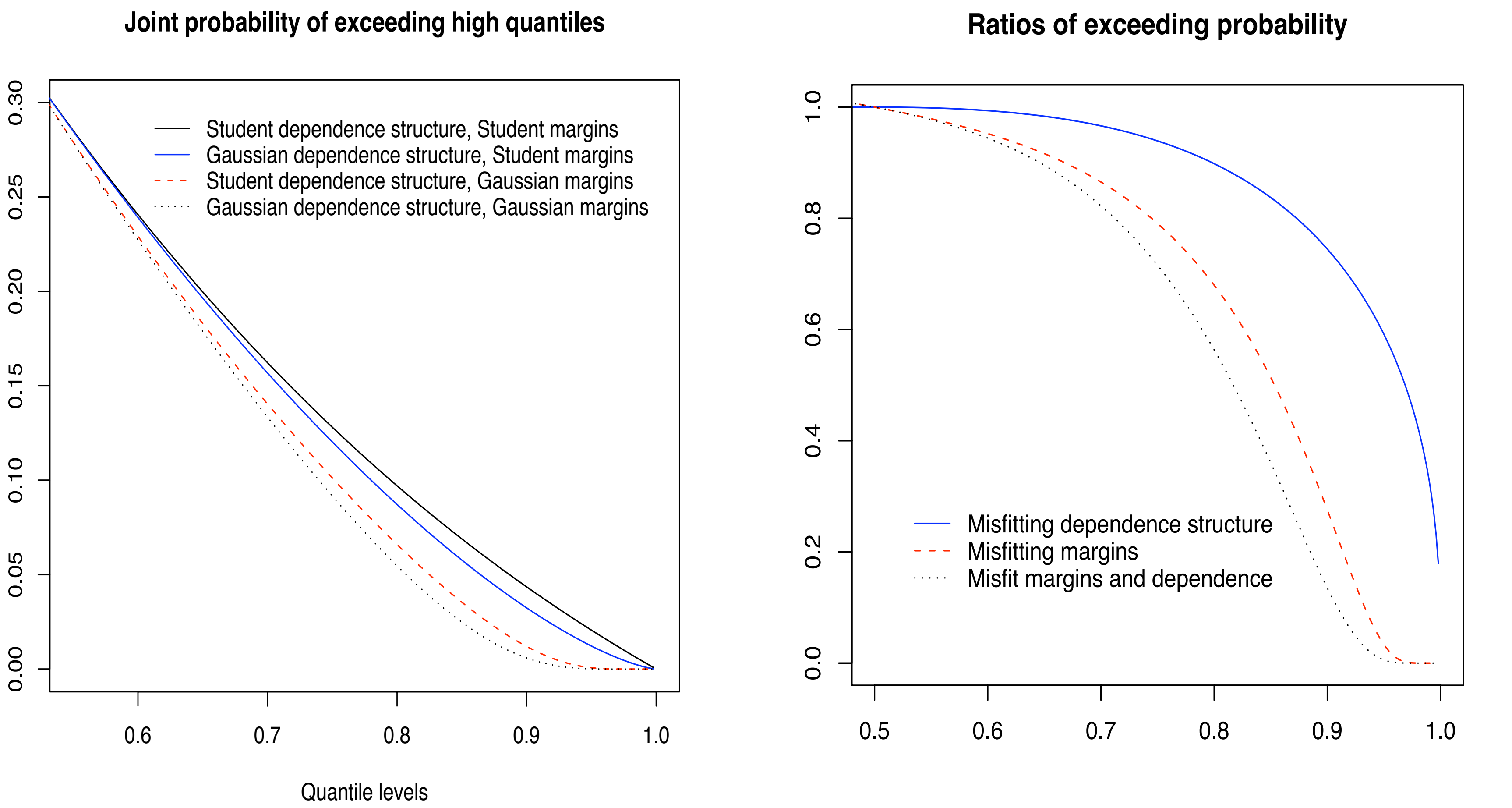


Figure 3 – $p \mapsto \mathbb{P}\left(X > F_X^{-1}(p), Y > F_Y^{-1}(p)\right)$ when (X, Y) is a Student t random vector, and when either margins or the dependence structure is mispectified.

Kernel estimation for bounded support density

Consider a kernel based estimation of density f,

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where K is a kernel function, given a n sample $X_1, ..., X_n$ of positive random variable ($X_i \in [0, \infty[$). Let K denote a symmetric kernel, then

$$\mathbb{E}(\widehat{f}(0)) = \frac{1}{2} f(0) + O(h)$$

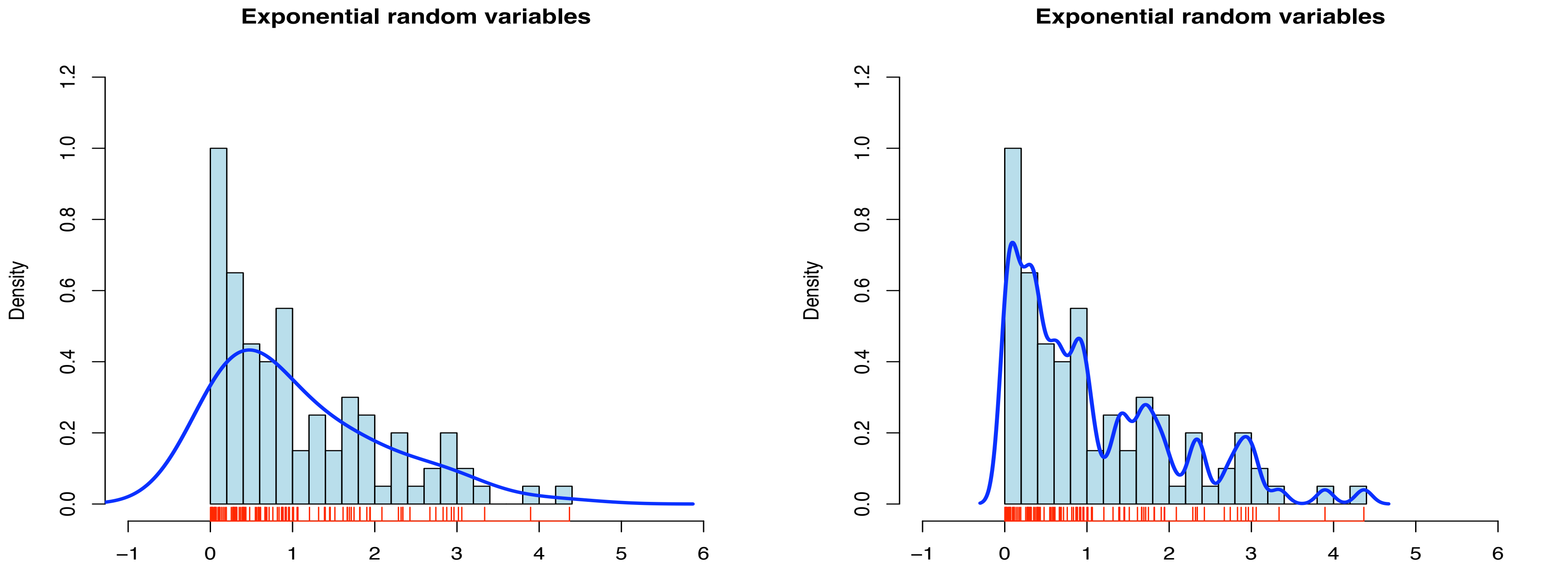


Figure 4 – Density estimation of an exponential density, 100 points.

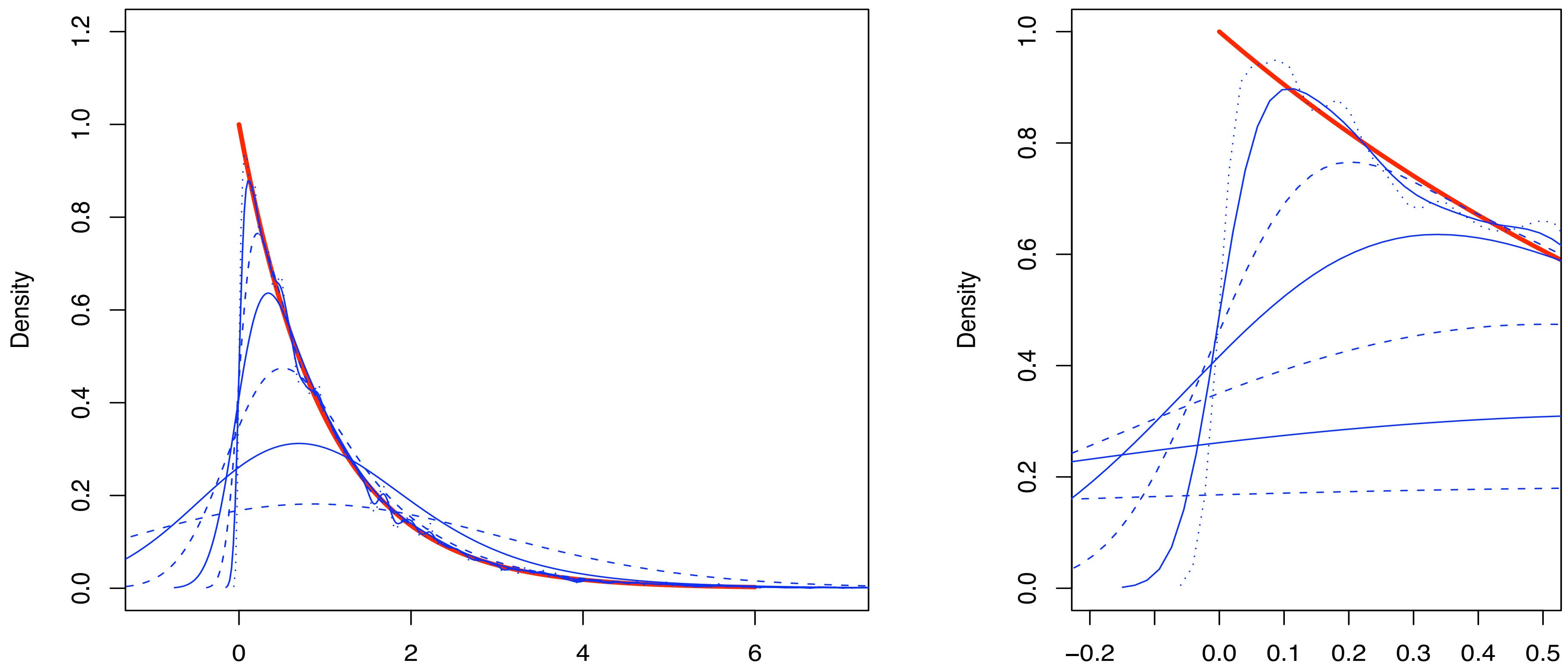


Figure 5 – Density estimation of an exponential density, 10,000 points.

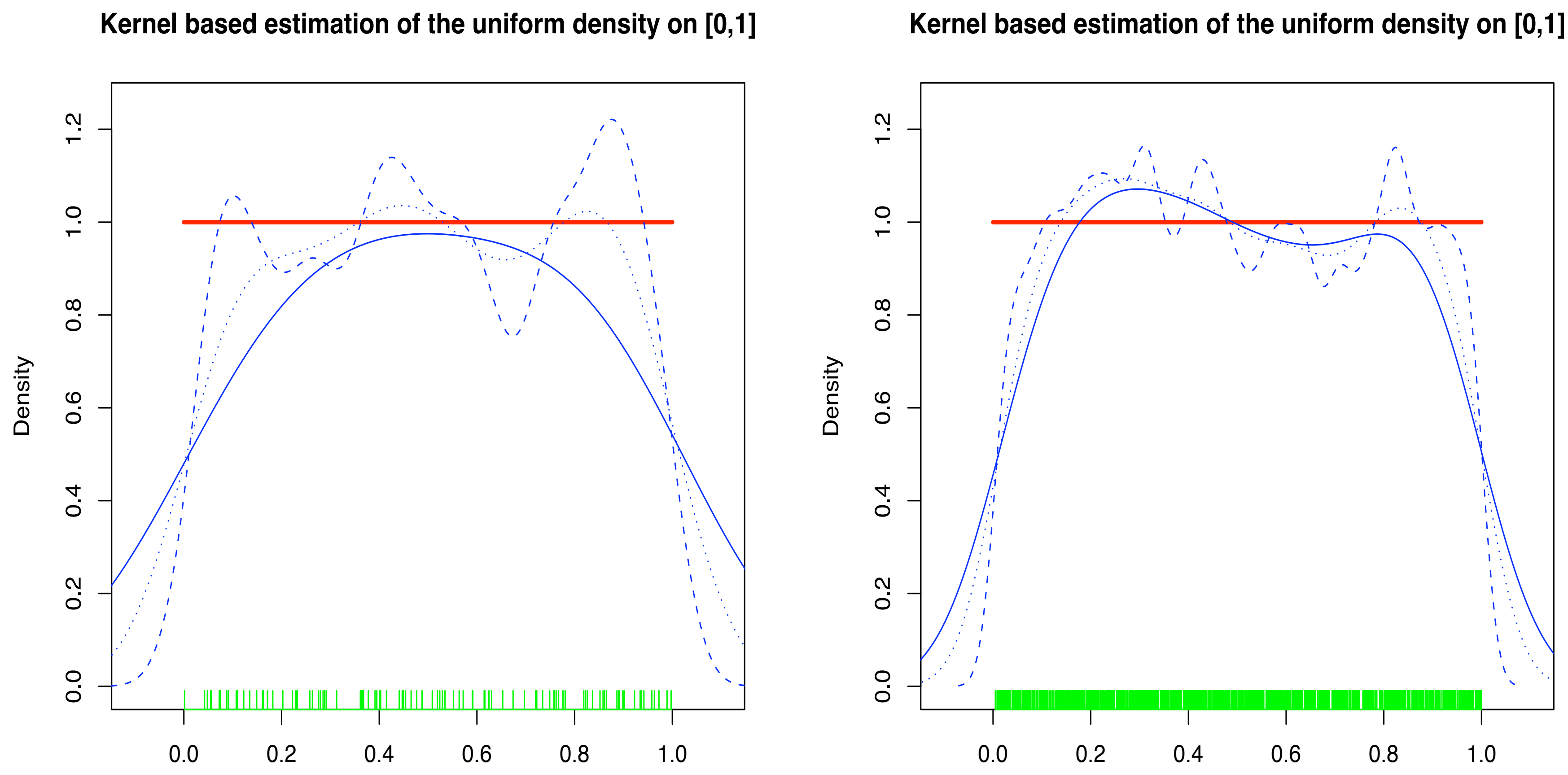


Figure 6 – Density estimation of an uniform density on $[0, 1]$.

How to get a proper estimation on the border

Several techniques have been introduce to get a better estimation on the border,

- boundary kernel (Müller (1991))
- mirror image modification (Deheuvels & Hominal (1989), Schuster (1985))
- transformed kernel (Devroye & Györfi (1981), Wand, Marron & Ruppert (1991))

In the particular case of densities on $[0, 1]$,

- Beta kernel (Brown & Chen (1999), Chen (1999, 2000)),
- average of histograms inspired by Dersko (1998).

Local kernel density estimators

The idea is that the bandwidth $h(x)$ can be different for each point x at which $f(x)$ is estimated. Hence,

$$\widehat{f}(x,h(x)) = \frac{1}{nh(x)}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h(x)}\right),$$

(see e.g. LOFTSGAARDEN & QUESENBERRY (1965)).

Variable kernel density estimators

The idea is that the bandwidth h can be replaced by n values $\alpha(X_i)$. Hence,

$$\widehat{f}(x,\alpha) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\alpha(X_i)} K\left(\frac{x-X_i}{\alpha(X_i)}\right),$$

(see e.g. ABRAMSON (1982)).

The transformed Kernel estimate

The idea was developed in Devroye & Györfi (1981) for univariate densities.

Consider a transformation $T : \mathbb{R} \rightarrow [0,1]$ strictly increasing, continuously differentiable, one-to-one, with a continuously differentiable inverse.

Set $Y = T(X)$. Then Y has density

$$f_Y(y) = f_X(T^{-1}(y)) \cdot (T^{-1})'(y).$$

If f_Y is estimated by $\widehat{f}_Y$, then f_X is estimated by

$$\widehat{f}_X(x) = \widehat{f}_Y(T(x)) \cdot T'(x).$$

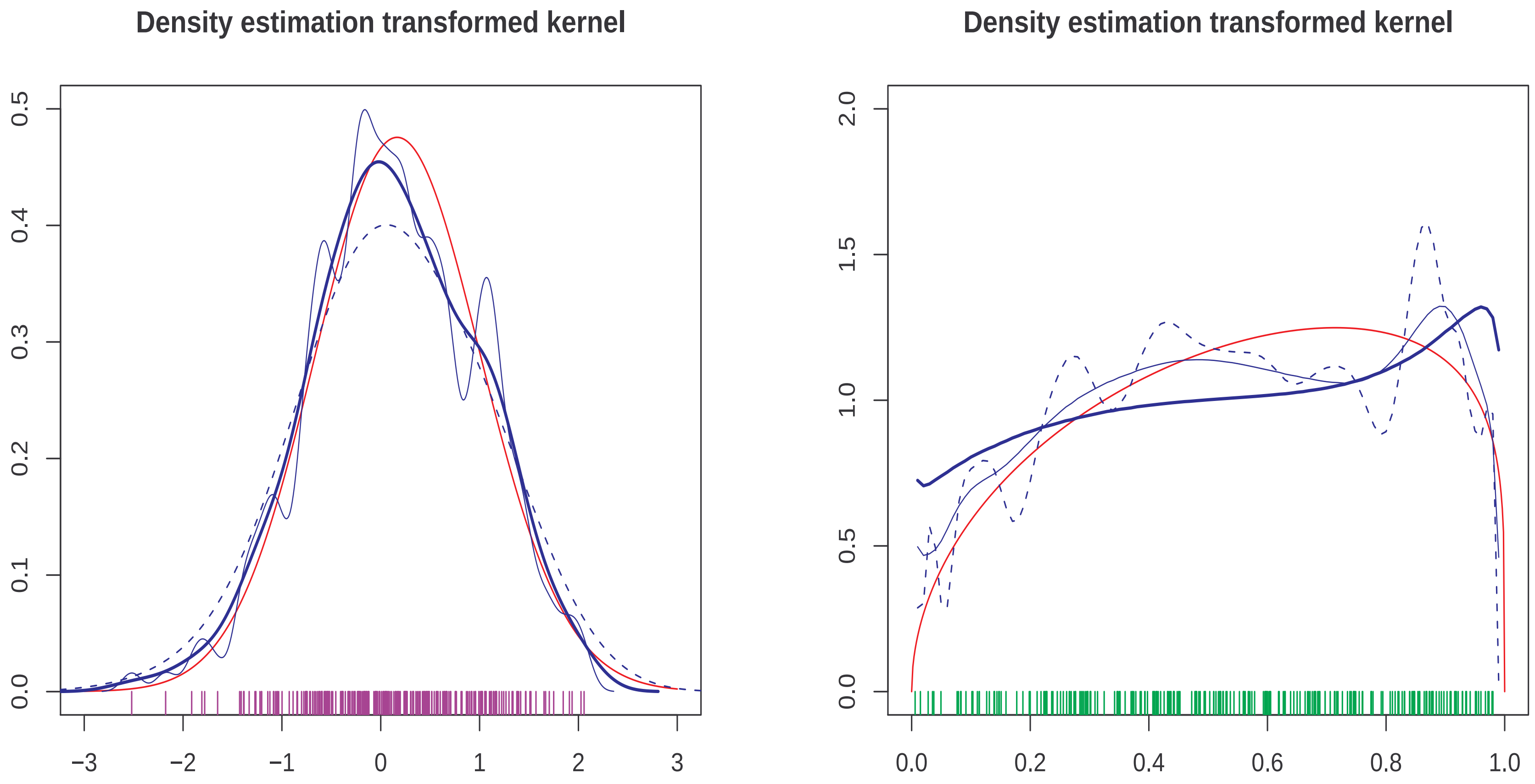


Figure 7 – The transform kernel principle (with a Φ^{-1}-transformation).

The Beta Kernel estimate

The Beta-kernel based estimator of a density with support $[0,1]$ at point x, is obtained using beta kernels, which yields

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(X_i, \frac{u}{b}+1, \frac{1-u}{b}+1\right)$$

where $K(\cdot,\alpha,\beta)$ denotes the density of the Beta distribution with parameters α and β,

$$K(x,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbf{1}_{\{x\in[0,1]\}}.$$

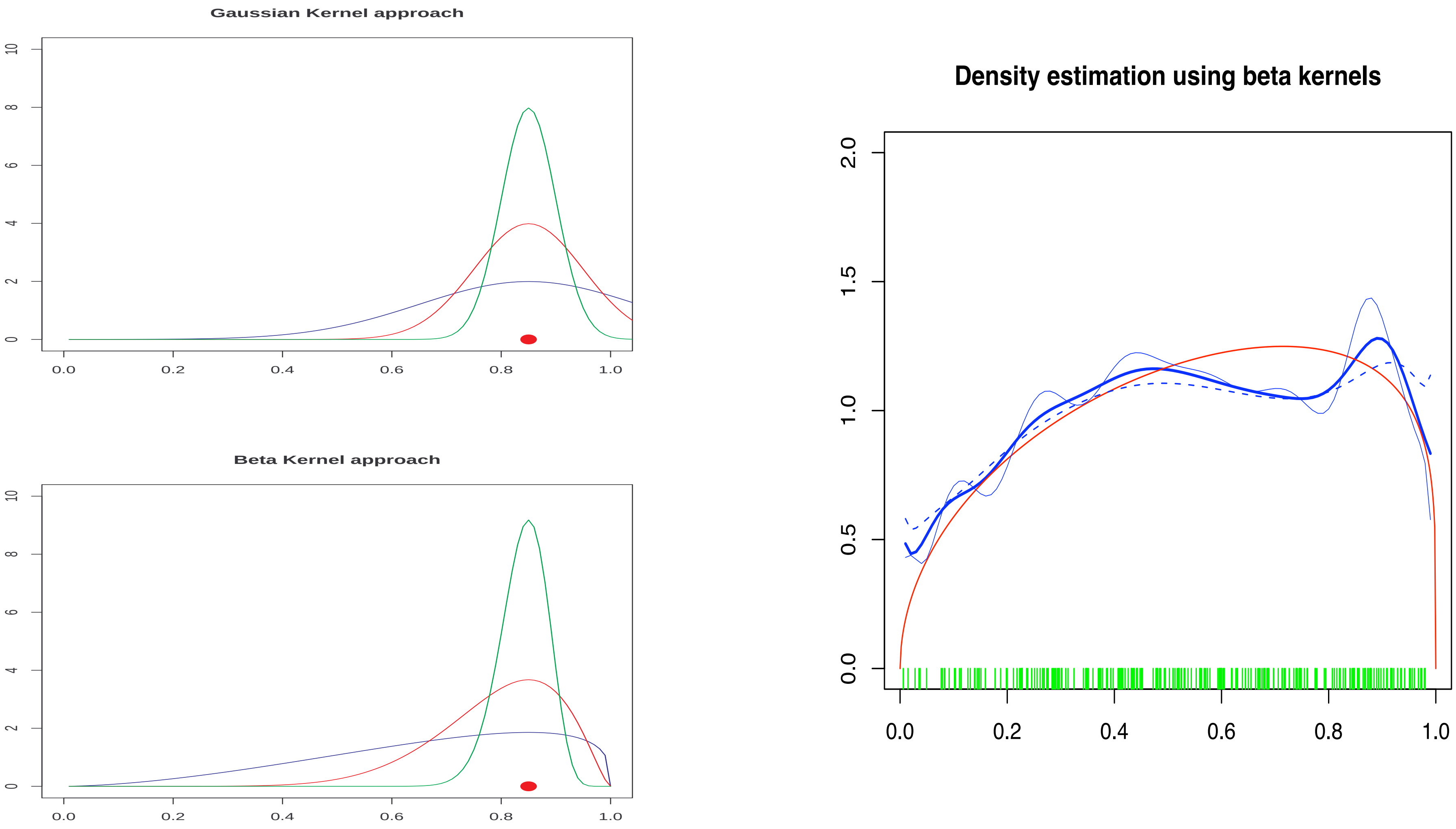


Figure 8 – The beta-kernel estimate.

Copula density estimation : the boundary problem

Let $(U_1, V_1), ..., (U_n, V_n)$ denote a sample with support $[0,1]^2$, and with density $c(u,v)$, which is assumed to be twice continuously differentiable on $(0,1)^2$.

If K denotes a symmetric kernel, with support $[-1,+1]$, then for all $(u,v) \in [0,1] \times [0,1]$, in any corners (e.g. $(0,0)$)

$$\mathbb{E}(\widehat{c}(0,0,h)) = \frac{1}{4} \cdot c(u,v) - \frac{1}{2}[c_1(0,0) + c_2(0,0)] \int_0^1 \omega K(\omega) d\omega \cdot h + o(h).$$

on the interior of the borders (e.g. $u = 0$ and $v \in (0,1)$),

$$\mathbb{E}(\widehat{c}(0,v,h)) = \frac{1}{2} \cdot c(u,v) - [c_1(0,v)] \int_0^1 \omega K(\omega) d\omega \cdot h + o(h).$$

and in the interior ($(u,v) \in (0,1) \times (0,1)$),

$$\mathbb{E}(\widehat{c}(u,v,h)) = c(u,v) + \frac{1}{2}[c_{1,1}(u,v) + c_{2,2}(u,v)] \int_{-1}^1 \omega^2 K(\omega) d\omega \cdot h^2 + o(h^2).$$

On borders, there is a multiplicative bias and the order of the bias is $O(h)$ (while it is $O(h^2)$ in the interior).

If K denotes a symmetric kernel, with support $[-1,+1]$, then for all $(u,v)\in[0,1]\times[0,1]$, in any corners (e.g. $(0,0)$)

$$Var(\widehat{c}(0,0,h)) = c(0,0)\left(\int_0^1 K(\omega)^2 d\omega\right)^2 \cdot \frac{1}{nh^2} + o\left(\frac{1}{nh^2}\right).$$

on the interior of the borders (e.g. $u=0$ and $v\in(0,1)$),

$$Var(\widehat{c}(0,v,h)) = c(0,v)\left(\int_{-1}^1 K(\omega)^2 d\omega\right)\left(\int_0^1 K(\omega)^2 d\omega\right) \cdot \frac{1}{nh^2} + o\left(\frac{1}{nh^2}\right).$$

and in the interior ($(u,v)\in(0,1)\times(0,1)$),

$$Var(\widehat{c}(u,v,h)) = c(u,v)\left(\int_{-1}^1 K(\omega)^2 d\omega\right)^2 d \cdot \frac{1}{nh^2} + o\left(\frac{1}{nh^2}\right).$$

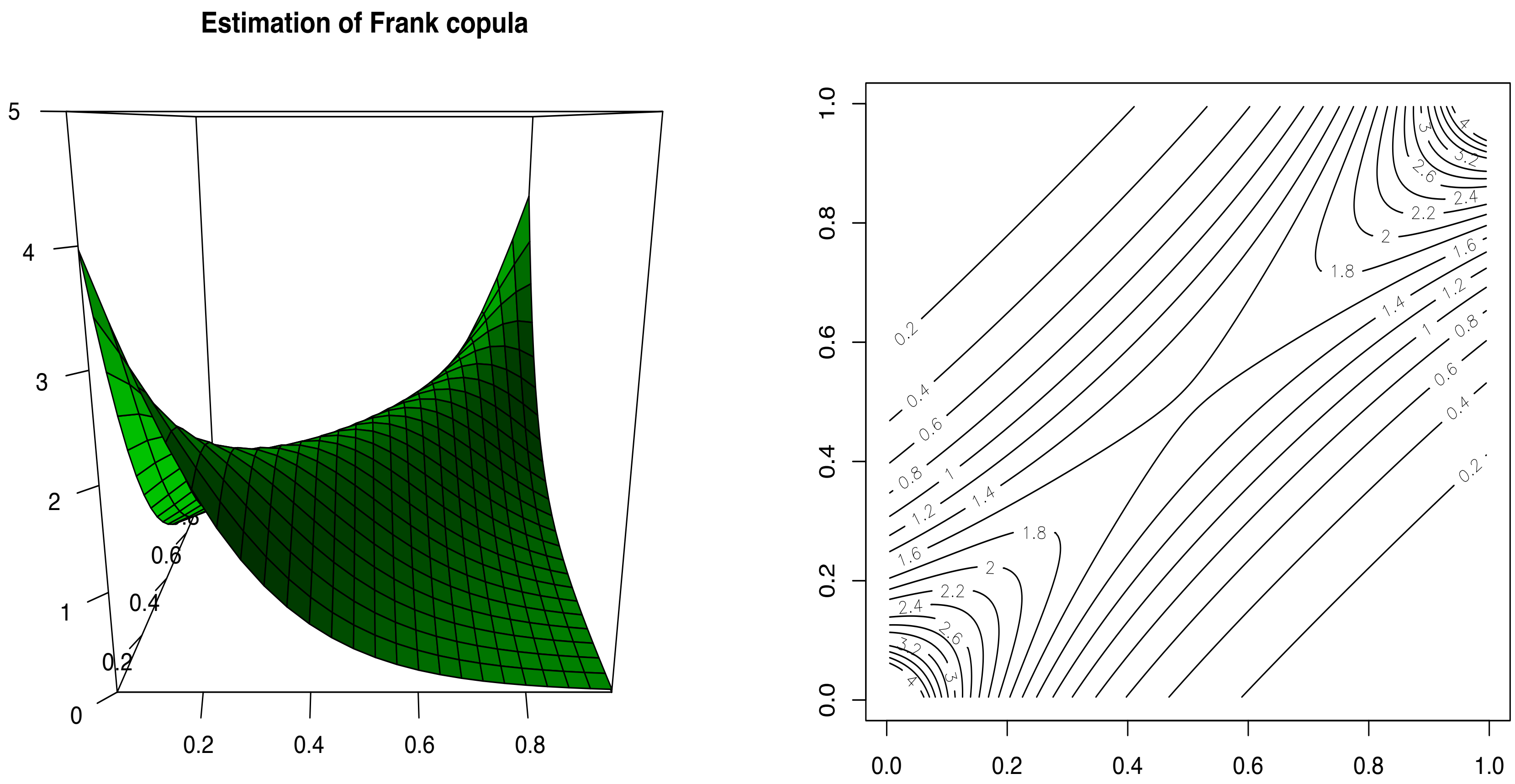


Figure 9 – Theoretical density of Frank copula.

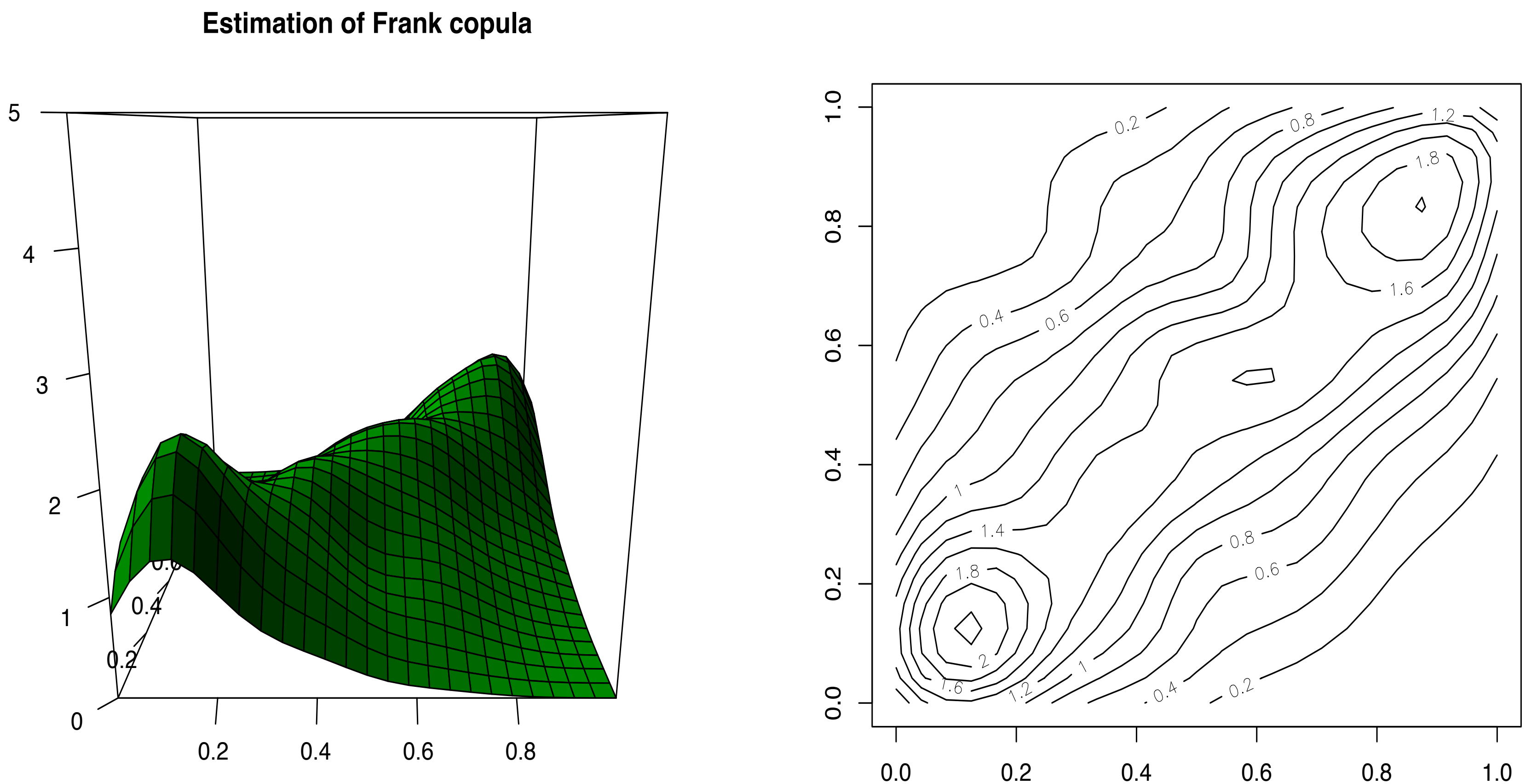


Figure 10 – Estimated density of Frank copula, using standard Gaussian (independent) kernels, $h = h^*$.

Transformed kernel technique

Consider the kernel estimator of the density of the $(X_i, Y_i) = (G^{-1}(U_i), G^{-1}(V_i))$'s, where G is a strictly increasing distribution function, with a differentiable density.

Since density f is continuous, twice differentiable, and bounded above, for all $(x, y) \in \mathbb{R}^2$,

$$\widehat{f}(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right),$$

satisfies

$$\mathbb{E}(\widehat{f}(x,y)) = f(x,y) + O(h^2),$$

as long as $\int \omega K(\omega) = 0$. And the variance is

$$Var(\widehat{f}(x,y)) = \frac{f(x,y)}{nh^2} \left(\int K(\omega)^2 d\omega\right)^2 + o\left(\frac{1}{nh^2}\right),$$

and asymptotic normality can be obtained,

$$\sqrt{nh^2}\left(\widehat{f}(x,y) - \mathbb{E}(\widehat{f}(x,y))\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, f(x,y)\left(\int K(\omega)^2 d\omega\right)^2).$$

Since

$$f(x,y)=g(x)g(y)c[G(x),G(y)]. \tag{1}$$

can be inverted in

$$c(u,v)=\frac{f(G^{-1}(u),G^{-1}(v))}{g(G^{-1}(u))g(G^{-1}(v))}, \qquad (u,v)\in[0,1]\times[0,1], \tag{2}$$

one gets, substituting $\widehat{f}$ in (2)

$$\widehat{c}(u,v)=\frac{1}{nh\cdot g(G^{-1}(u))\cdot g(G^{-1}(v))}\sum_{i=1}^{n}K\left(\frac{G^{-1}(u)-G^{-1}(U_i)}{h},\frac{G^{-1}(v)-G^{-1}(V_i)}{h}\right), \tag{3}$$

Therefore,

$$\mathbb{E}(\widehat{c}(u,v,h))=c(u,v)+\frac{o(h)}{g(G^{-1}(u))g(G^{-1}(v))}.$$

Similarly,

$$\begin{aligned} Var(\widehat{c}(u,v,h)) &= \frac{1}{g(G^{-1}(u))g(G^{-1}(v))}\left[\frac{c(u,v)}{nh^2}\left(\int K(\omega)^2 d\omega\right)^2\right] \\ &+ \frac{1}{g(G^{-1}(u))^2 g(G^{-1}(v))^2}\, o\left(\frac{1}{nh^2}\right). \end{aligned}$$

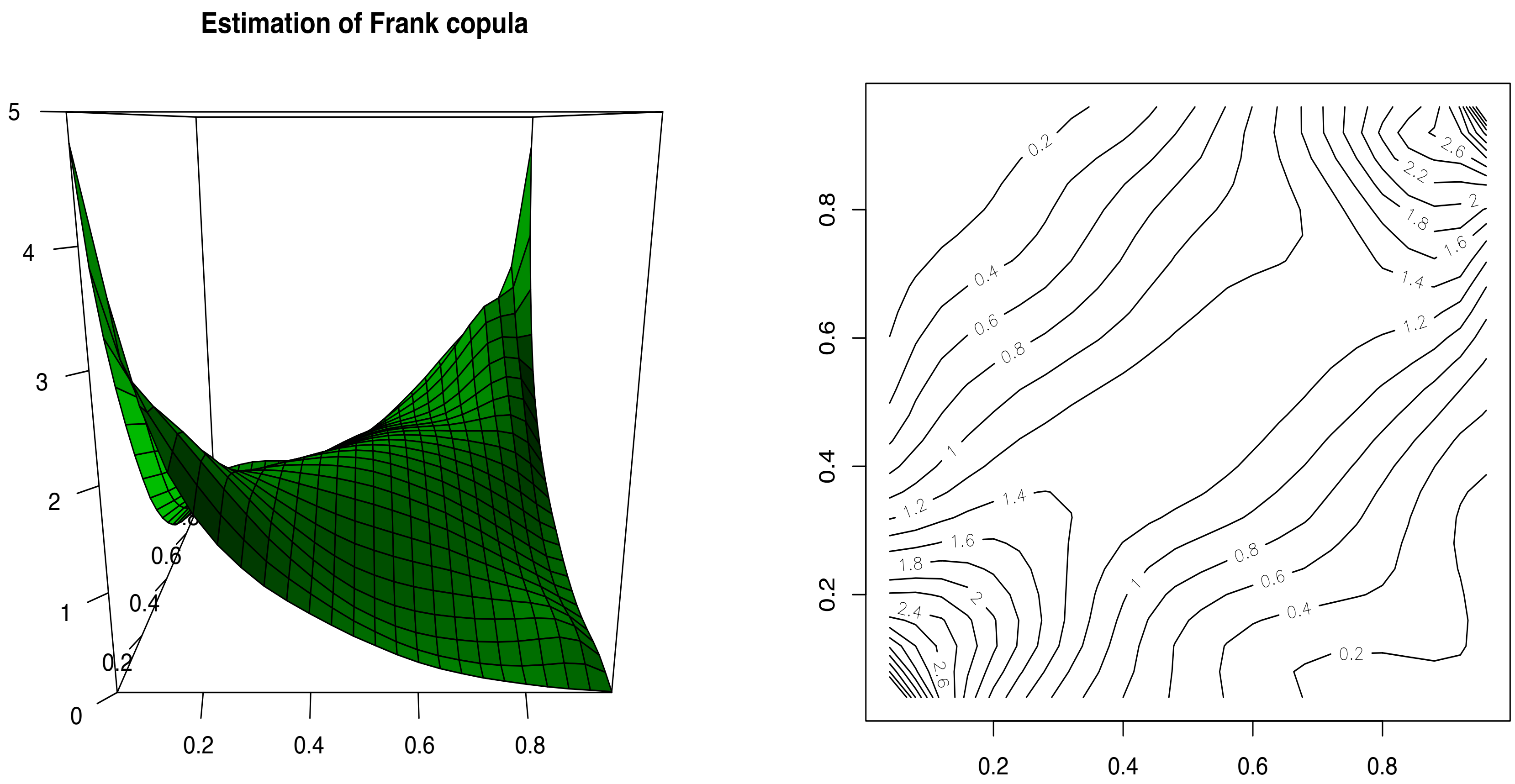


Figure 11 – Estimated density of Frank copula, using a Gaussian kernel, after a Gaussian normalization.

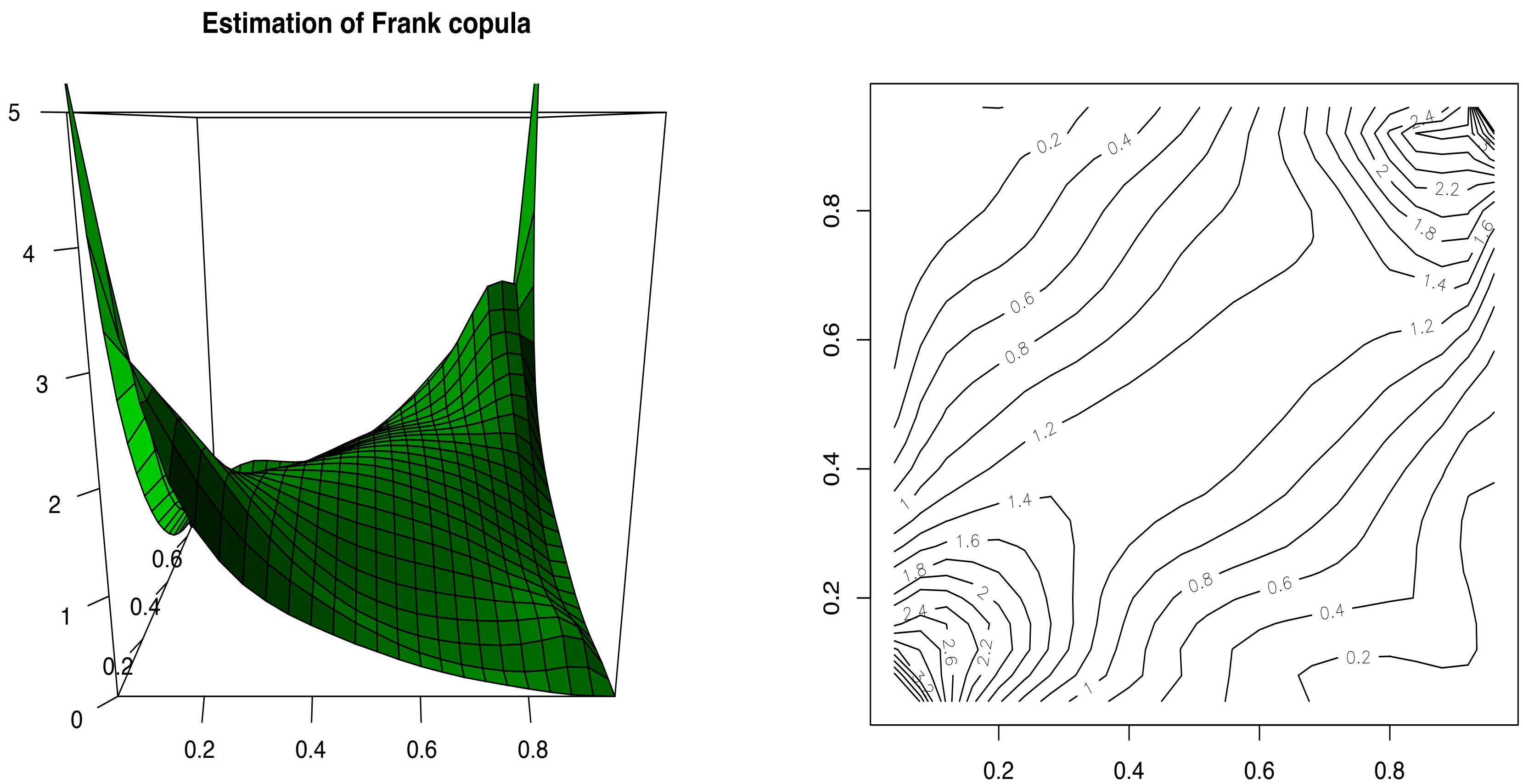
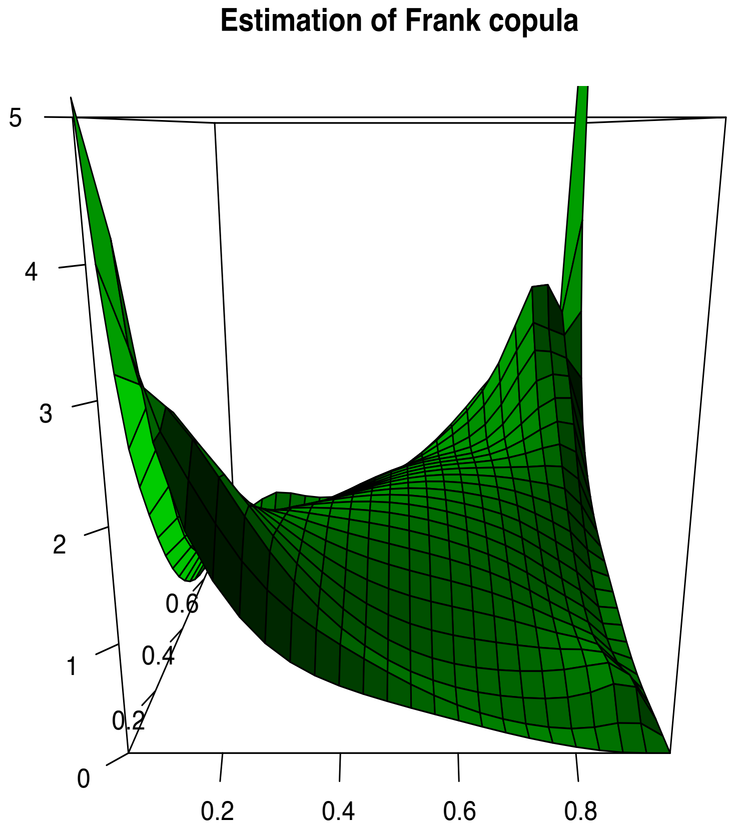


Figure 12 – Estimated density of Frank copula, using a Gaussian kernel, after a Student normalization, with 5 degrees of freedom.



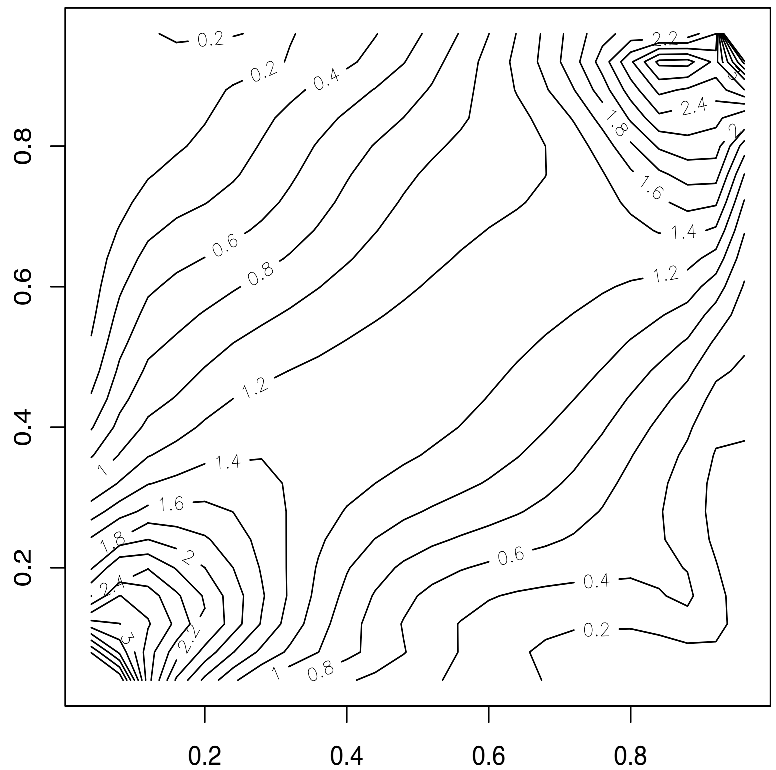


Figure 13 – Estimated density of Frank copula, using a Gaussian kernel, after a Student normalization, with 3 degrees of freedom.

Bivariate Beta kernels

The Beta-kernel based estimator of the copula density at point (u, v), is obtained using product beta kernels, which yields

$$\widehat{c}(u,v) = \frac{1}{n}\sum_{i=1}^{n} K\left(X_i, \frac{u}{b}+1, \frac{1-u}{b}+1\right) \cdot K\left(Y_i, \frac{v}{b}+1, \frac{1-v}{b}+1\right),$$

where $K(\cdot, \alpha, \beta)$ denotes the density of the Beta distribution with parameters α and β.

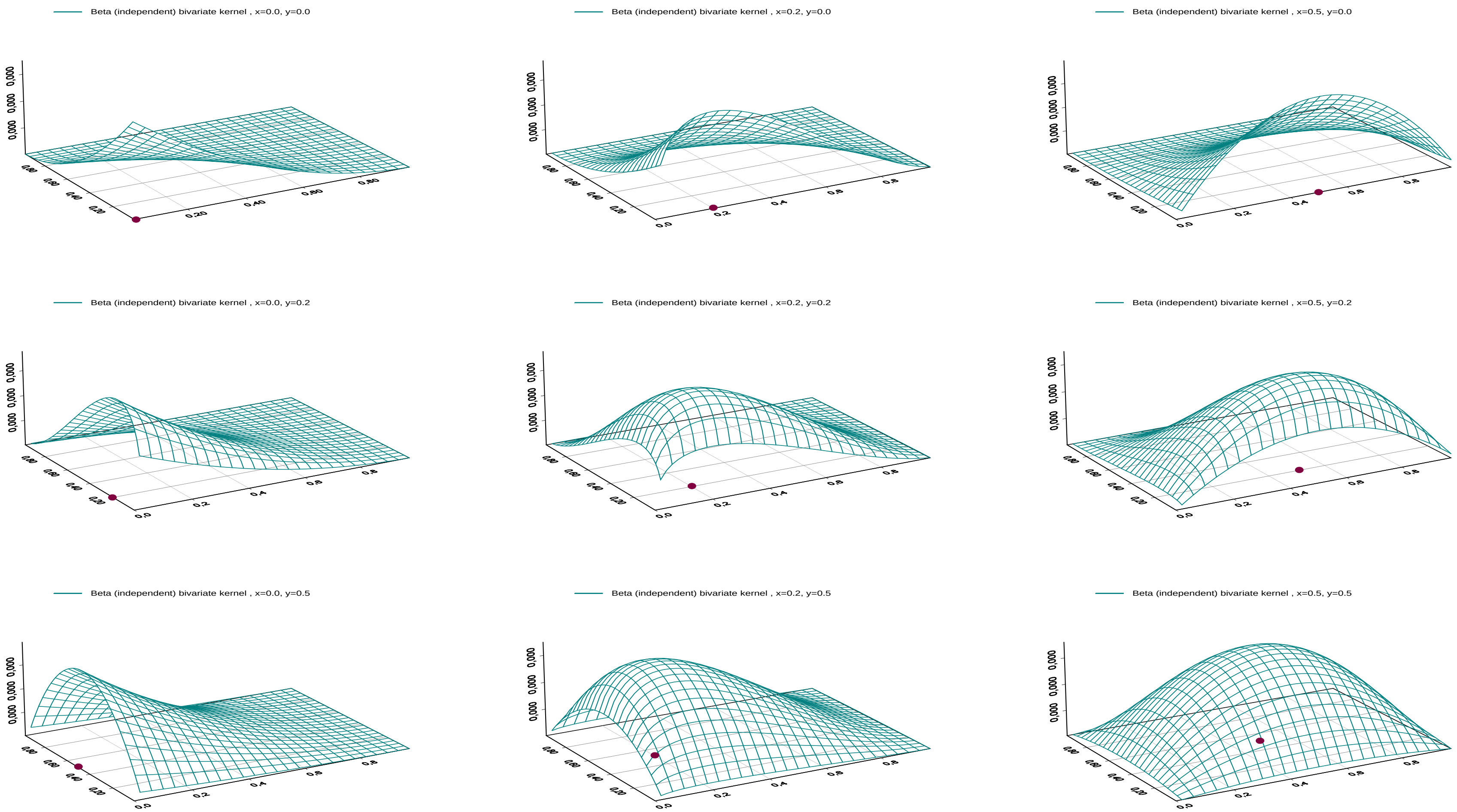


Figure 14 – Shape of bivariate Beta kernels $K(\cdot, x/b+1, (1-x)/b+1) \times K(\cdot, y/b+1, (1-y)/b+1)$ for $b = 0.2$.

Assume that the copula density c is twice differentiable on $[0,1]\times[0,1]$. Let $(u,v)\in[0,1]\times[0,1]$. The bias of $\widehat{c}(u,v)$ is of order b, i.e.

$$\mathbb{E}(\widehat{c}(u,v)) = c(u,v) + \mathcal{Q}(u,v)\cdot b + o(b),$$

where the bias $\mathcal{Q}(u,v)$ is

$$\mathcal{Q}(u,v) = (1-2u)c_1(u,v) + (1-2v)c_2(u,v) + \frac{1}{2}\left[u(1-u)c_{1,1}(u,v) + v(1-v)c_{2,2}(u,v)\right].$$

The bias here is $O(b)$ (everywhere) while it is $O(h^2)$ using standard kernels.

Assume that the copula density c is twice differentiable on $[0,1]\times[0,1]$. Let $(u,v)\in[0,1]\times[0,1]$. The variance of $\widehat{c}(u,v)$ is in corners, e.g. $(0,0)$,

$$Var(\widehat{c}(0,0)) = \frac{1}{nb^2}[c(0,0)+o(n^{-1})],$$

in the interior of borders, e.g. $u=0$ and $v\in(0,1)$

$$Var(\widehat{c}(0,v)) = \frac{1}{2nb^{3/2}\sqrt{\pi v(1-v)}}[c(0,v)+o(n^{-1})],$$

and in the interior,$(u, v) \in (0,1) \times (0,1)$

$$Var(\widehat{c}(u,v)) = \frac{1}{4nb\pi\sqrt{v(1-v)u(1-u)}}[c(u,v) + o(n^{-1})].$$

Remark From those properties, an (asymptotically) optimal bandwidth b can be deduced, maximizing asymptotic mean squared error,

$$b^* \equiv \left(\frac{1}{16\pi n \mathcal{Q}(u,v)^2} \cdot \frac{1}{\sqrt{v(1-v)u(1-u)}} \right)^{1/3}.$$

Note (see Charpentier, Fermanian & Scaillet (2005)) that all those results can be obtained in dimension $d \geq 2$.

Example For $n = 100$ simulated data, from Frank copula, the optimal bandwidth is $b \sim 0.05$.

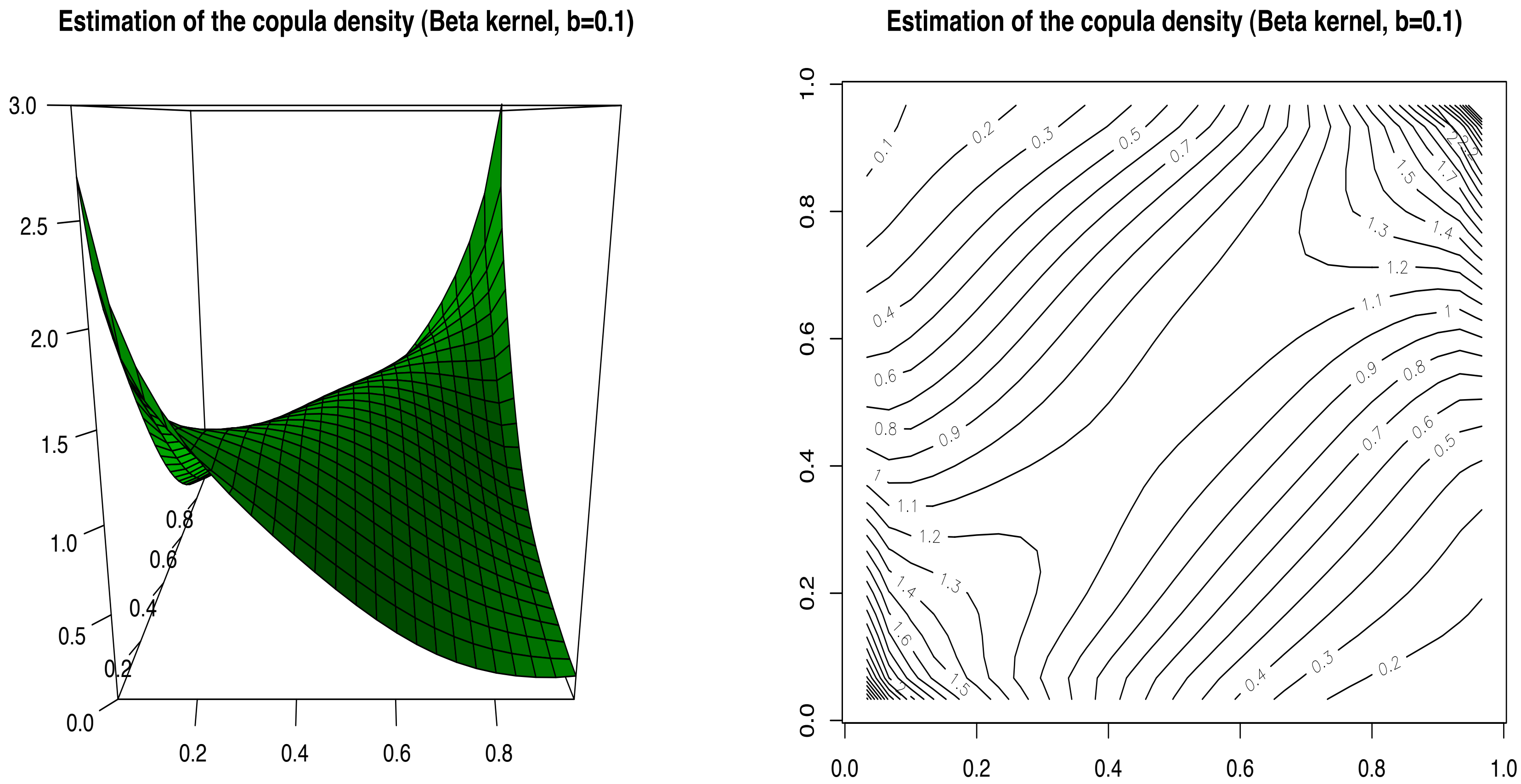


Figure 15 – Estimated density of Frank copula, Beta kernels, $b = 0.1$

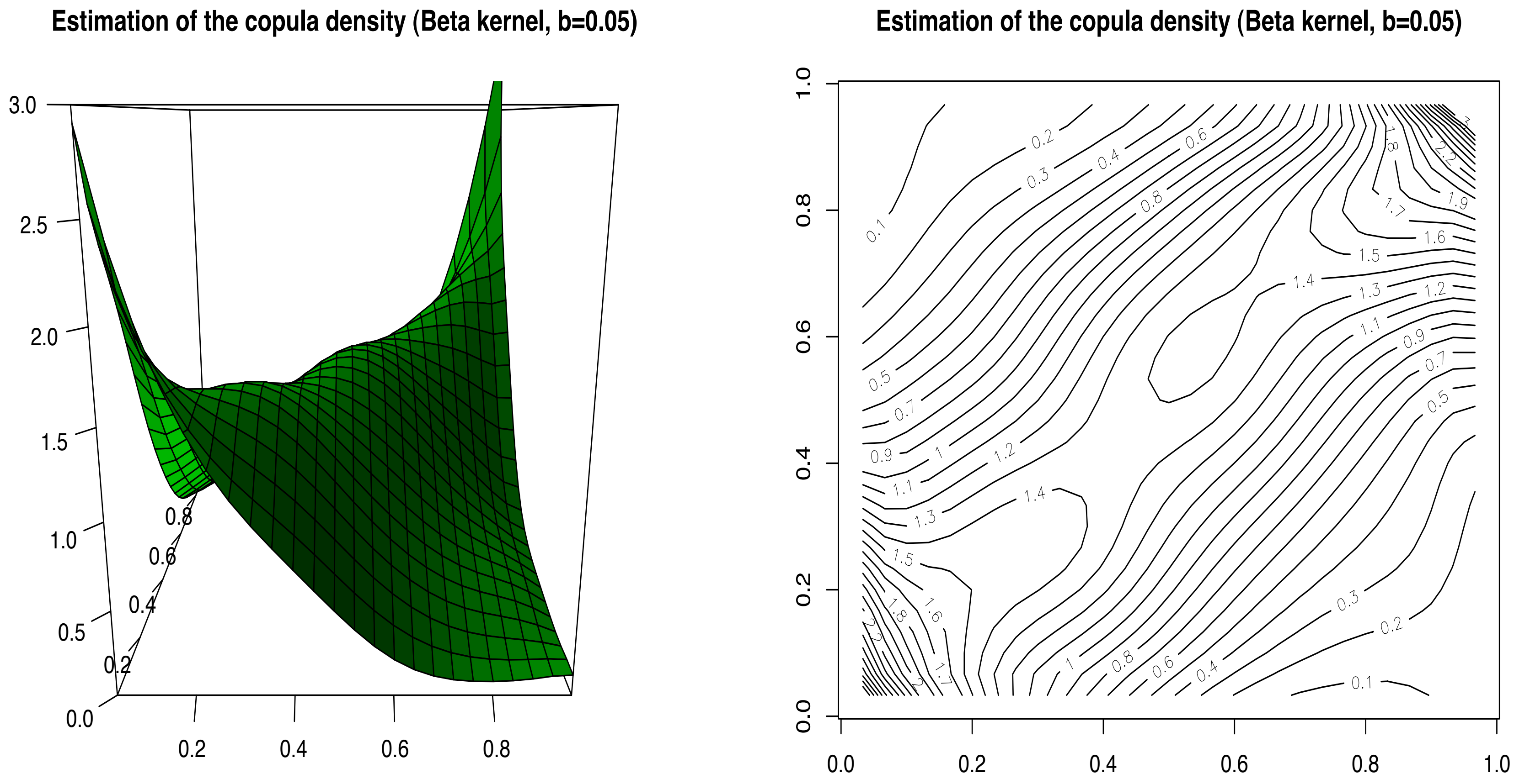


Figure 16 – Estimated density of Frank copula, Beta kernels, $b = 0.05$

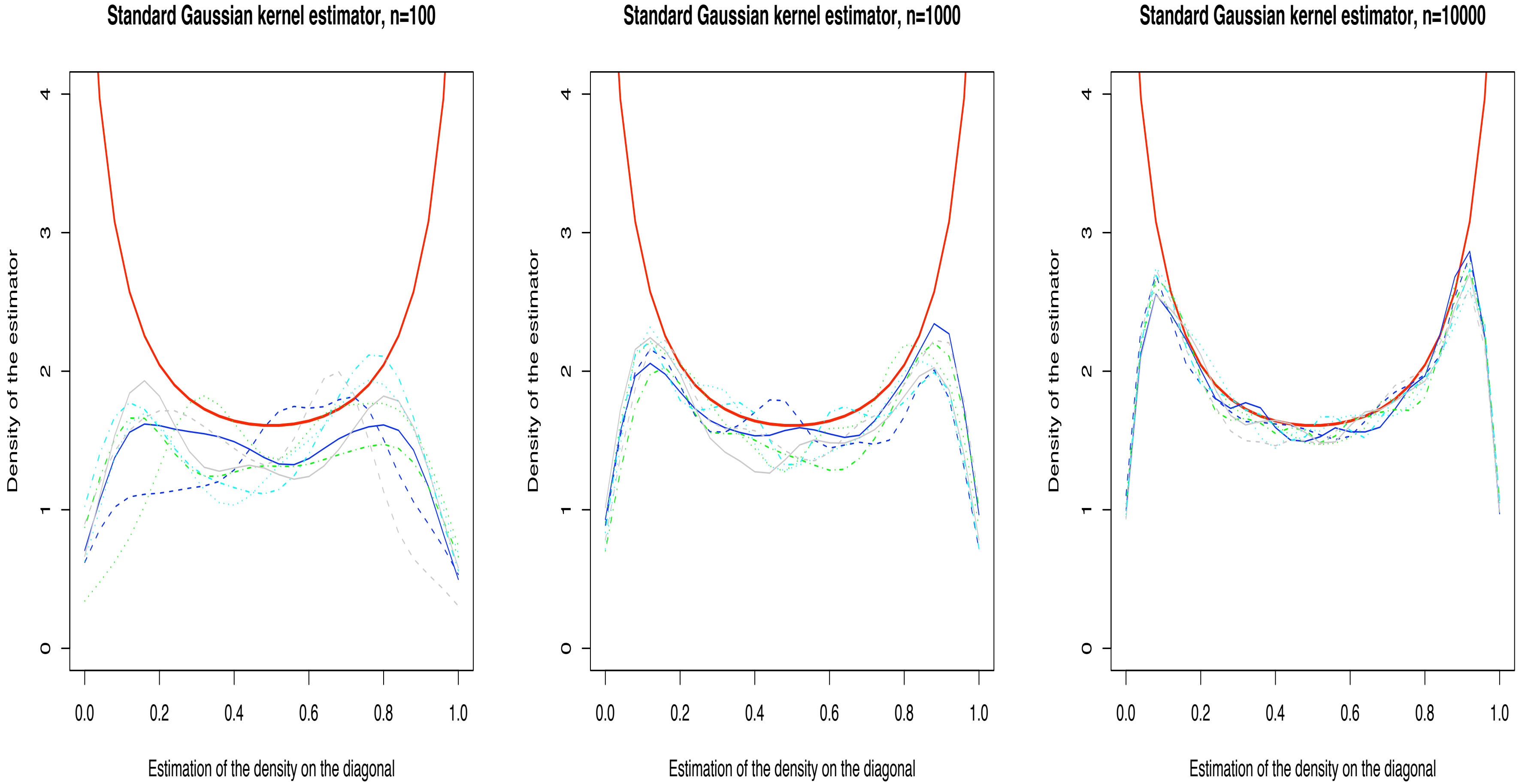


Figure 17 – Density estimation on the diagonal, standard kernel.

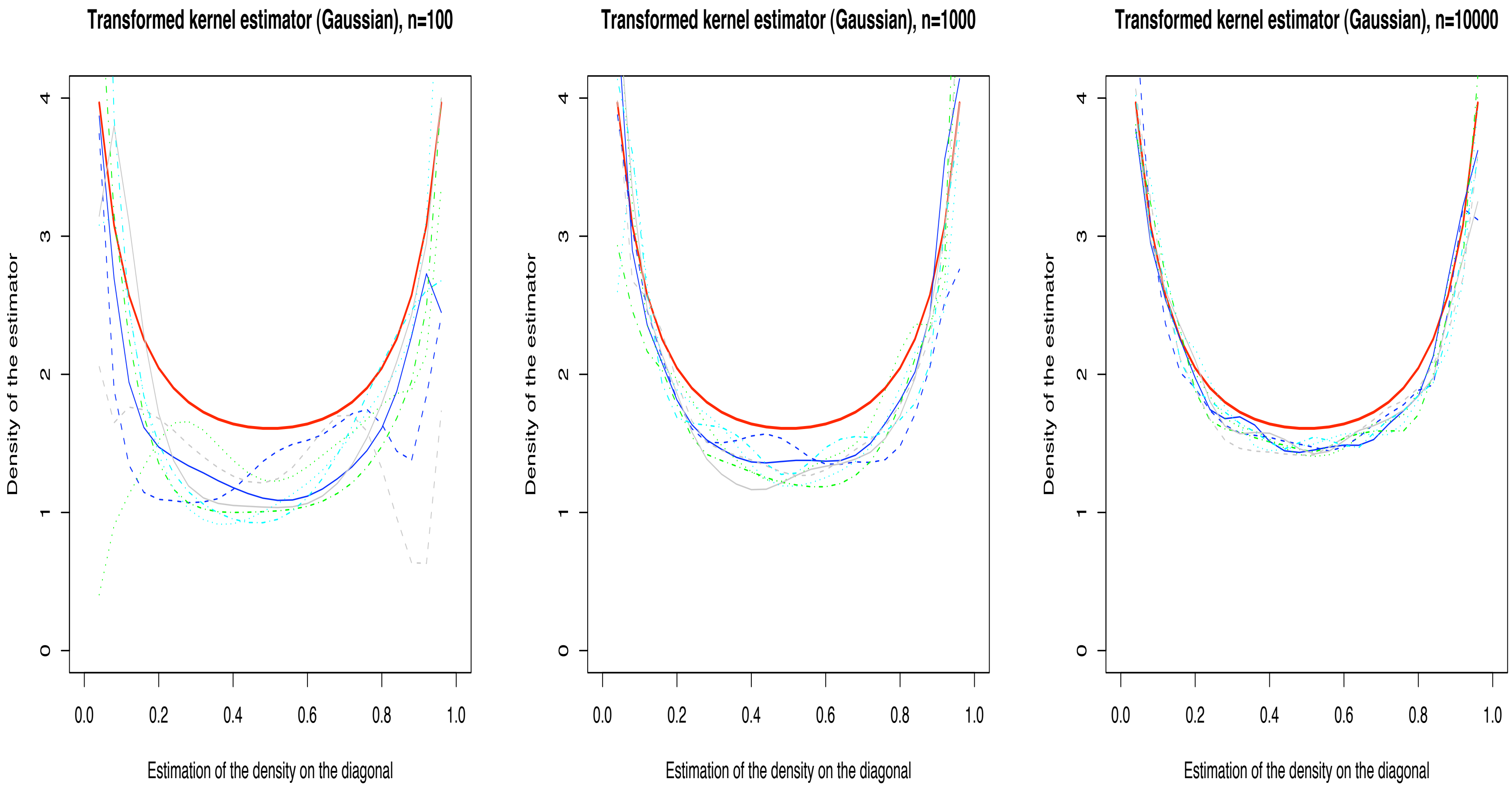


Figure 18 – Density estimation on the diagonal, transformed kernel.

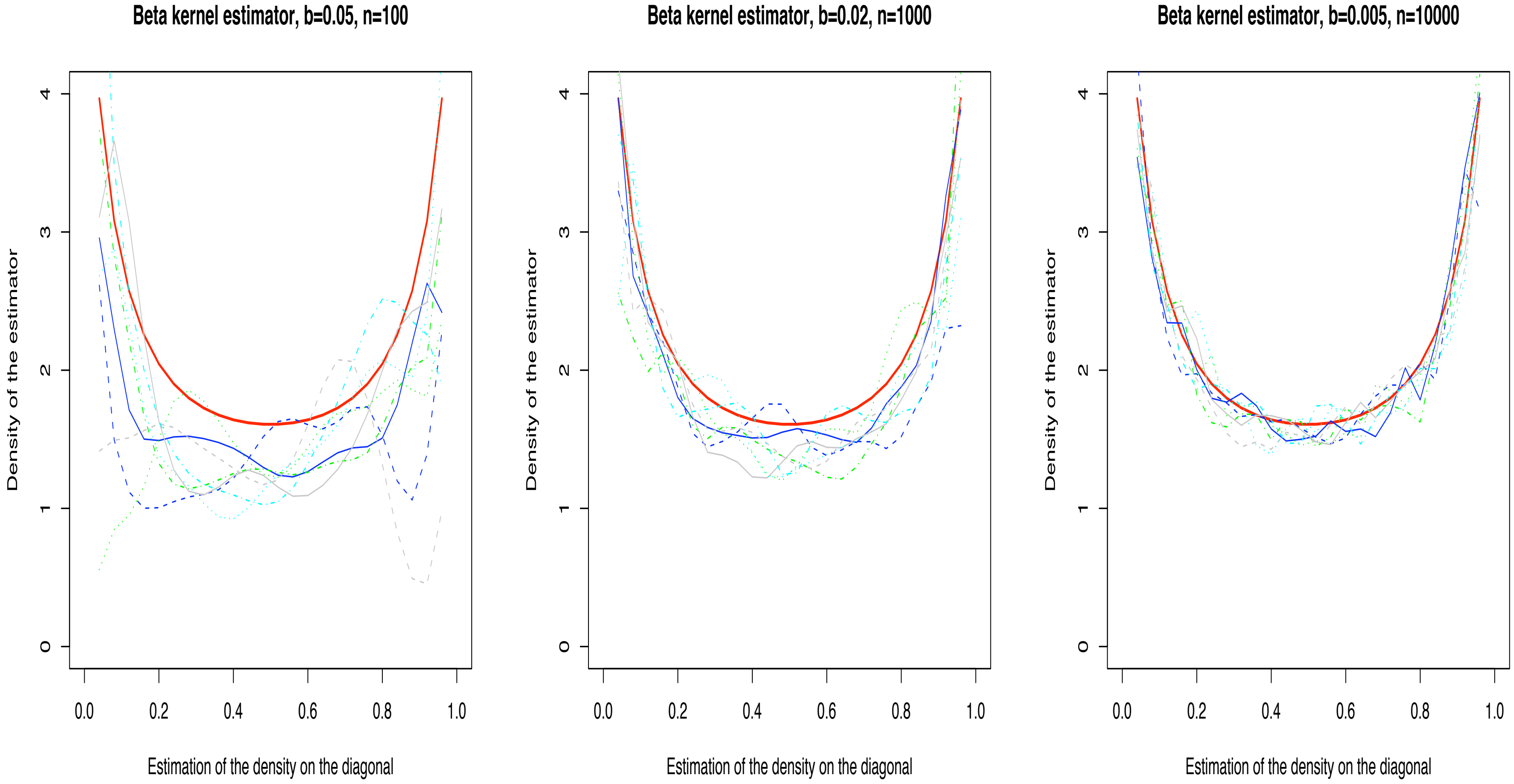


Figure 19 – Density estimation on the diagonal, Beta kernel.

Copula density estimation

Gijbels & Mielniczuk (1990) : given an i.i.d. sample, a natural estimate for the normed density is obtained using the transformed sample $(\widehat{F}_X(X_1), \widehat{F}_Y(Y_1)), ..., (\widehat{F}_X(X_n), \widehat{F}_Y(Y_n))$, where $\widehat{F}_X$ and $\widehat{F}_Y$ are the empirical distribution function of the marginal distribution. The copula density can be constructed as some density estimate based on this sample (Behnen, Husková & Neuhaus (1985) investigated the kernel method).

The natural kernel type estimator $\widehat{c}$ of c is

$$c(u,v) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{u-\widehat{F}_X(X_i)}{h}, \frac{v-\widehat{F}_Y(Y_i)}{h}\right), (u,v) \in [0,1].$$

"*this estimator is not consistent in the points on the boundary of the unit square*."

Copula density estimation and pseudo-observations

Example : in linear regression, residuals are pseudo observations.

$$\varepsilon_i = H(X_i, Y_i) = Y_i - \alpha - \beta X_i$$

$$\widehat{\varepsilon}_i = \widehat{H}_n(X_i, Y_i) = Y_i - \widehat{\alpha}_n - \widehat{\beta}_n X_i$$

Example : when dealing with copulas, ranks U_i, V_i yield pseudo-observations.

$$(U_i, V_i) = H(X_i, Y_i) = (F_X(X_i), F_Y(Y_i))$$

$$(\widehat{U}_i, \widehat{V}_i) = \widehat{H}_n(X_i, Y_i) = (\widehat{F}_X(X_i), \widehat{F}_Y(Y_i))$$

(see Genest & Rivest (1993)).

More formally, let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ denote a series of observations of $\boldsymbol{X}$ $(\in \boldsymbol{\mathcal{X}})$, stationary and ergodic.

Let $H : \boldsymbol{\mathcal{X}} \to \mathbb{R}^d$ and set $\boldsymbol{\varepsilon}_i = H(\boldsymbol{X}_i)$ (non-observable).

If H is estimated by $\widehat{H}_n$ then $\widehat{\boldsymbol{\varepsilon}}_i = \widehat{H}_n(\boldsymbol{X}_i)$ are called pseudo-observations.

Let $\widehat{K}_n$ denote the empirical distribution function of those pseudo-observations,

$$\widehat{K}_n(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\widehat{\boldsymbol{\varepsilon}}_i \leq \boldsymbol{t}) \text{ where } \boldsymbol{t} \in \mathbb{R}^d.$$

Further, if K denotes the distribution function of $\boldsymbol{\varepsilon} = H(\boldsymbol{X})$, then define the empirical process based on pseudo-observations,

$$\mathbb{K}_n(t) = \sqrt{n}\left(\widehat{K}_n(\boldsymbol{t}) - K(\boldsymbol{t})\right)$$

As proved in Ghoudi & Rémillard (1998, 2004), this empirical process converges weakly.

Figure **??** shows scatterplots when margins are known (i.e. $(F_X(X_i), F_Y(Y_i))$'s), and when margins are estimated (i.e. $(\hat{F}_X(X_i), \hat{F}_Y(Y_i)$'s). Note that the pseudo sample is more "*uniform*", in the sense of a lower discrepancy (as in Quasi Monte Carlo techniques, see e.g. Niederreiter (1992)).

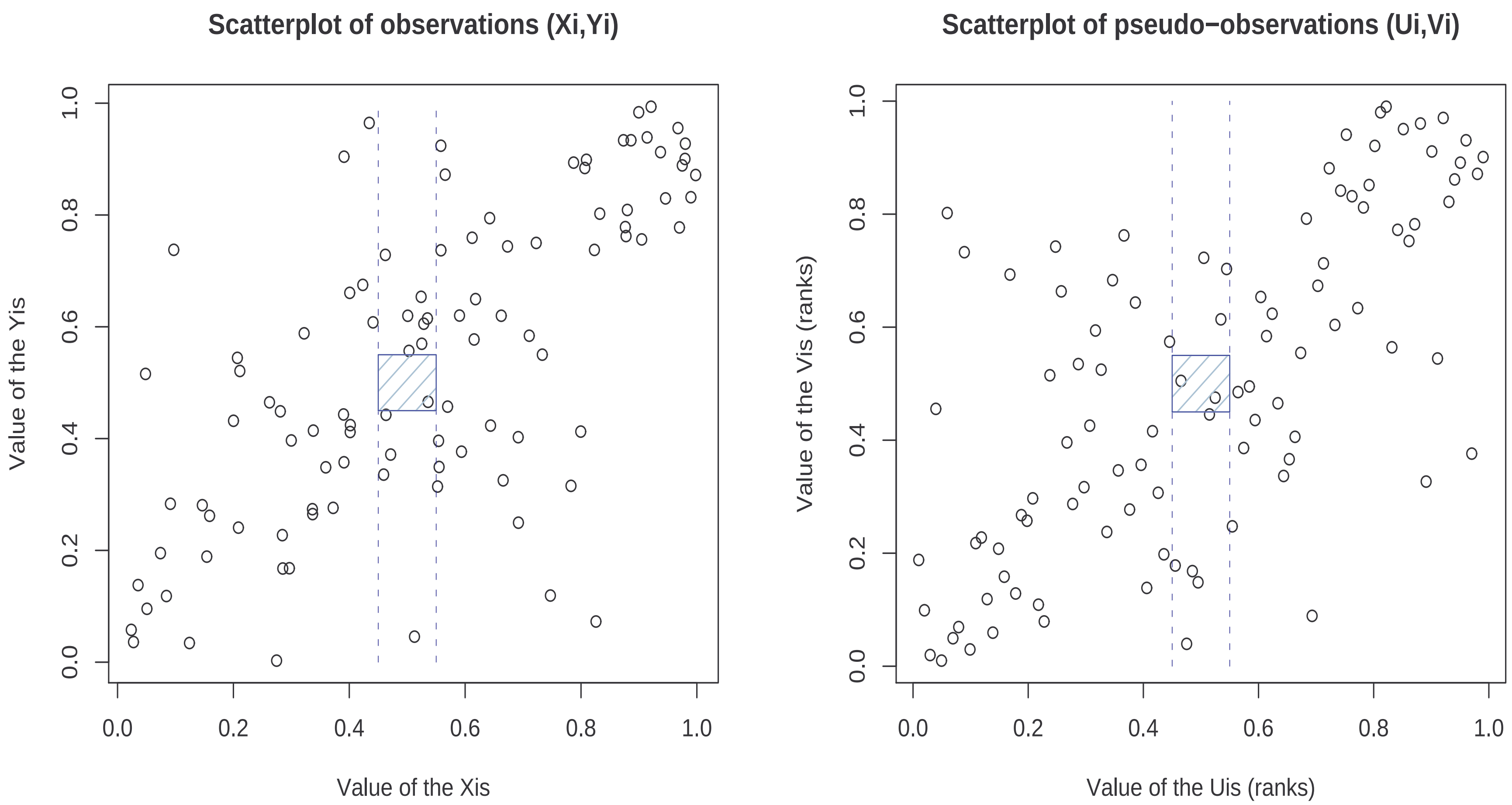


Figure 20 – Observations and pseudo-observation, 500 simulated observations from Frank copula (X_i, Y_i) and the associate pseudo-sample $(\hat{F}_X(X_i), \hat{F}_Y(Y_i))$.

Because samples are more "*uniform*" using ranks and pseudo-observations, the variance of the estimator of the density, at some given point $(u, v) \in (0, 1) \times (0, 1)$ is usually smaller. For instance, Figure 21 shows the impact of considering pseudo observations, i.e. substituting $\hat{F}_X$ and $\hat{F}_Y$ to unknown marginal distributions F_X and F_Y. The dotted line shows the density of $\hat{c}(u, v)$ from a $n = 100$ sample (U_i, V_i) (from Frank copula), and the straight line shows the density of $\hat{c}(u, v)$ from the sample $(\hat{F}_U(U_i), \hat{F}_V(V_i))$ (i.e. ranks of observations).

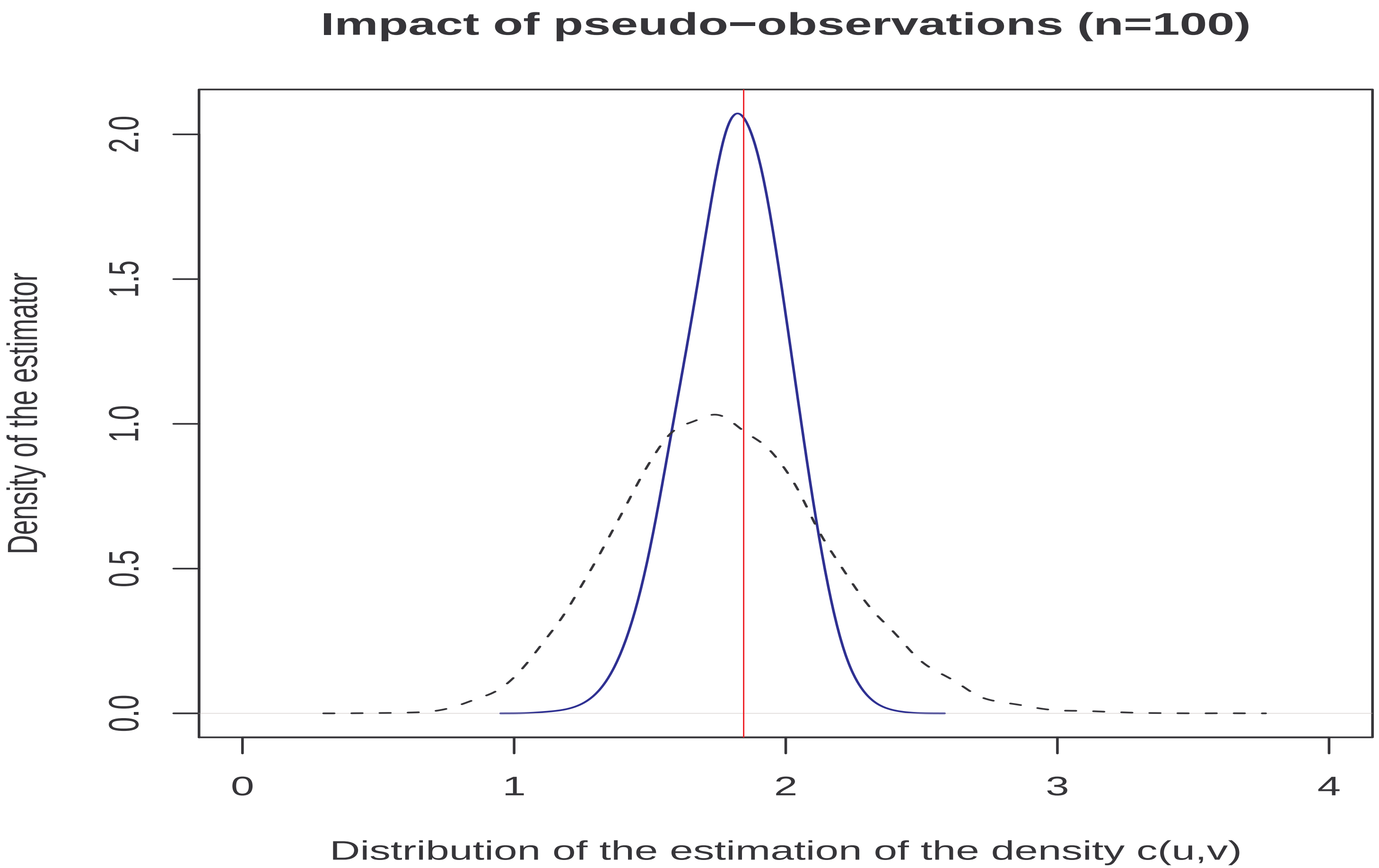


Figure 21 – The impact of estimating from pseudo-observations.

Roots of '*transformed kernel*'

CHAPTER 9

The Transformed Kernel Estimate

The *transformed kernel estimate* (Devroye et al., 1983) is based upon a transformation T: $R^1 \to [0,1]$ which is strictly monotonically increasing, continuously differentiable, one-to-one and onto, and which has a continuously differentiable inverse. The transformed data sequence is $Y_1, \ldots, Y_n$, where $Y_i = T(X_i)$. Note that Y_1 has density

$$g(x) = f(T^{-1}(x))T^{-1\prime}(x).$$

Now, g is estimated by g_n from $Y_1, \ldots, Y_n$, and f is estimated by

$$f_n(x) = g_n(T(x))T'(x). \qquad (2)$$

The key observation is that if g_n is a density on $[0,1]$, the f_n is a density on R^1, and furthermore,

$$\int |f_n - f| = \int |g_n - g|.$$

2. CHOOSING A TRANSFORMATION

Choosing a transformation is not a sinecure. In a vast number of applications, one suspects that f belongs to a certain family of densities (usually a parametric family), or at least is close to a given member of this family. If the family is a parametrized by θ, with distribution function F_θ, the natural approach is to estimate θ by $\tilde{\theta}$ in a *robust* manner, and use $F_{\tilde{\theta}}$ in the expression of the optimal transformation T. Throughout we use the same h, that is, the optimal h for the isosceles triangular density on $[0,1]$.

3. ESTIMATION OF DENSITIES WITH LARGE TAILS

There are two factors that determine the efficiency of the kernel estimate: discontinuities or sharp oscillations, and large tails. The former factor, captured for smooth densities by $\int|f''|$, is infinite for densities with simple discontinuities such as the uniform density on $[0,1]$. The latter factor, measured by $\int\sqrt{f}$, is infinite for densities with a large tail such as the Cauchy density. We have seen that when one or both of these factors is infinite, we must have $n^{2/5}E(J_n)\to\infty$ for the standard kernel estimate, regardless of the choice of h as a function of n.

An isolated bump in *any* density estimate is associated with one of the data points $X_1, \ldots, X_n$: X_i defines an isolated bump if there exists an interval $[a, b]$ with the property that $X_i \in [a, b]$, no other point X_j belongs to $[a, b]$, $\int_a^b f_n > 0$, and $f_n = 0$ on $[a - \varepsilon, a) \cup (b, b + \varepsilon]$ for some $\varepsilon > 0$. Assume, for example, that we are using the kernel estimate with Epanechnikov's kernel. Then X_i defines an isolated bump if and only if $[X_i - 2h, X_i + 2h]$ contains no data point except X_i. Thus, in the graph of f_n, $[X_i - h, X_i + h]$ appears as a separate hill, and it would seem that the data point "X_i" is wasted. Note also that the number of isolated bumps is invariant under strictly monotone transformations such as the ones considered in this chapter.

The total number of isolated bumps, B_n, bounds from below the number of hills in a graph. For example, when we are estimating a unimodal density, we would like the number of separate hills to be 1 and $B_n = 0$. As we will show in this section, this is usually not the case. For example, for the normal density with optimal h, $E(B_n)$ increases at least as $n^{1/5}/\sqrt{\log n}$, and the situation gets worse for longer-tailed densities. We will also show that for the triangular density, $E(B_n) = o(1)$.

THEOREM 2 (Densities with a Regularly Varying Tail). *Let f be strictly monotonically decreasing on $[0, \infty)$ with uniquely defined inverse, and let f be 0 on $(-\infty, 0)$ for the sake of convenience. Assume further that f is regularly varying at ∞ with exponent $r < -1$, that is,*

$$\lim_{x \to \infty} \frac{f(tx)}{f(x)} = t^r, \quad \text{all } t > 0.$$

If $h \to 0$, $nh \to \infty$, then

$$E(B_n) \geq \frac{L(n)}{(nh)^{1/r} h}$$

for some slowly varying function L (i.e., a regularly varying function with exponent 0).

Consider the transformed kernel estimate with Epanechnikov kernel K, and smoothing factor $h = \frac{1}{2}(5/6\pi n)^{1/5}$ (which is optimal for the triangular density on [0, 1]). We will not worry for the time being about transformations T_n: $R^1 \to [0,1]$ and the corresponding normalizations, because, as we have seen, this is an asymptotically negligible detail. We have the following densities:

- f: density of $X_1, \ldots, X_n$ (the data).
- g: density of $Y_i = T_n(X_i)$, given $X_1, \ldots, X_n$.
- g^*: density of $T(X_1)$, where T is some given transformation.
- g_n: transformed kernel estimate based upon $Y_1, \ldots, Y_n$.
- g_n^*: transformed kernel estimate based upon $Z_i = T(X_i)$, $1 \le i \le n$.

Consistency 251

For variable transformations T, we must worry about the consistency of the resulting estimate.

The transformation $Y_i = T(X_i)$ is usually of the form

$$Y_i = T_n(X_i; X_1, \ldots, X_n),$$

Using a parametric approach

If $F_X \in \mathcal{F} = \{F_\theta, \theta \in \Theta\}$ (assumed to be continuous), $q_X(\alpha) = F_\theta^{-1}(\alpha)$, and thus, a natural estimator is

$$\widehat{q}_X(\alpha) = F_{\widehat{\theta}}^{-1}(\alpha), \tag{4}$$

where $\widehat{\theta}$ is an estimator of θ (maximum likelihood, moments estimator...).

Using the Gaussian distribution

A natural idea (that can be found in classical financial models) is to assume Gaussian distributions : if $X \sim \mathcal{N}(\mu, \sigma)$, then the α-quantile is simply

$$q(\alpha) = \mu + \Phi^{-1}(\alpha)\sigma,$$

where $\Phi^{-1}(\alpha)$ is obtained in statistical tables (or any statistical software), e.g. $u = -1.64$ if $\alpha = 90\%$, or $u = -1.96$ if $\alpha = 95\%$.

Definition1

Given a n sample $\{X_1, \cdots, X_n\}$, the (Gaussian) parametric estimation of the α-quantile is

$$\widehat{q}_n(\alpha) = \widehat{\mu} + \Phi^{-1}(\alpha)\widehat{\sigma},$$

Using a parametric models

Actually, is the Gaussian model does not fit very well, it is still possible to use Gaussian approximation

If the variance is finite, $(X - \mathbb{E}(X))/\sigma$ might be closer to the Gaussian distribution, and thus, consider the so-called Cornish-Fisher approximation, i.e.

$$Q(X, \alpha) \sim \mathbb{E}(X) + z_\alpha \sqrt{V(X)}, \tag{5}$$

where

$$\widehat{z}_\alpha = \Phi^{-1}(\alpha) + \frac{\zeta_1}{6}[\Phi^{-1}(\alpha)^2 - 1] + \frac{\zeta_2}{24}[\Phi^{-1}(\alpha)^3 - 3\Phi^{-1}(\alpha)] - \frac{\zeta_1^2}{36}[2\Phi^{-1}(\alpha)^3 - 5\Phi^{-1}(\alpha)],$$

where ζ_1 is the skewness of X, and ζ_2 is the excess kurtosis, i.e. i.e.

$$\zeta_1 = \frac{\mathbb{E}([X - \mathbb{E}(X)]^3)}{\mathbb{E}([X - \mathbb{E}(X)]^2)^{3/2}} \text{ and } \zeta_2 = \frac{\mathbb{E}([X - \mathbb{E}(X)]^4)}{\mathbb{E}([X - \mathbb{E}(X)]^2)^2} - 3. \tag{6}$$

Using a parametric models

Definition2

Given a n sample $\{X_1, \cdots, X_n\}$, the Cornish-Fisher estimation of the α-quantile is

$$\widehat{q}_n(\alpha) = \widehat{\mu} + \widehat{z}_\alpha \widehat{\sigma}, \text{ where } \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2},$$

and

$$z_\alpha = \Phi^{-1}(\alpha) + \frac{\widehat{\zeta_1}}{6}[\Phi^{-1}(\alpha)^2 - 1] + \frac{\widehat{\zeta_2}}{24}[\Phi^{-1}(\alpha)^3 - 3\Phi^{-1}(\alpha)] - \frac{\widehat{\zeta_1^2}}{36}[2\Phi^{-1}(\alpha)^3 - 5\Phi^{-1}(\alpha)], \quad (7)$$

where $\widehat{\zeta_1}$ is the natural estimator for the skewness of X, and $\widehat{\zeta_2}$ is the natural estimator of the excess kurtosis, i.e. $\widehat{\zeta_1} = \dfrac{\sqrt{n(n-1)}}{n-2} \dfrac{\sqrt{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^3}{\left(\sum_{i=1}^{n}(X_i - \widehat{\mu})^2\right)^{3/2}}$ and

$$\widehat{\zeta_2} = \frac{n-1}{(n-2)(n-3)}\left((n+1)\widehat{\zeta_2'} + 6\right) \text{ where } \widehat{\zeta_2'} = \frac{n\sum_{i=1}^{n}(X_i - \widehat{\mu})^4}{\left(\sum_{i=1}^{n}(X_i - \widehat{\mu})^2\right)^2} - 3.$$

Parametrics estimator and error model

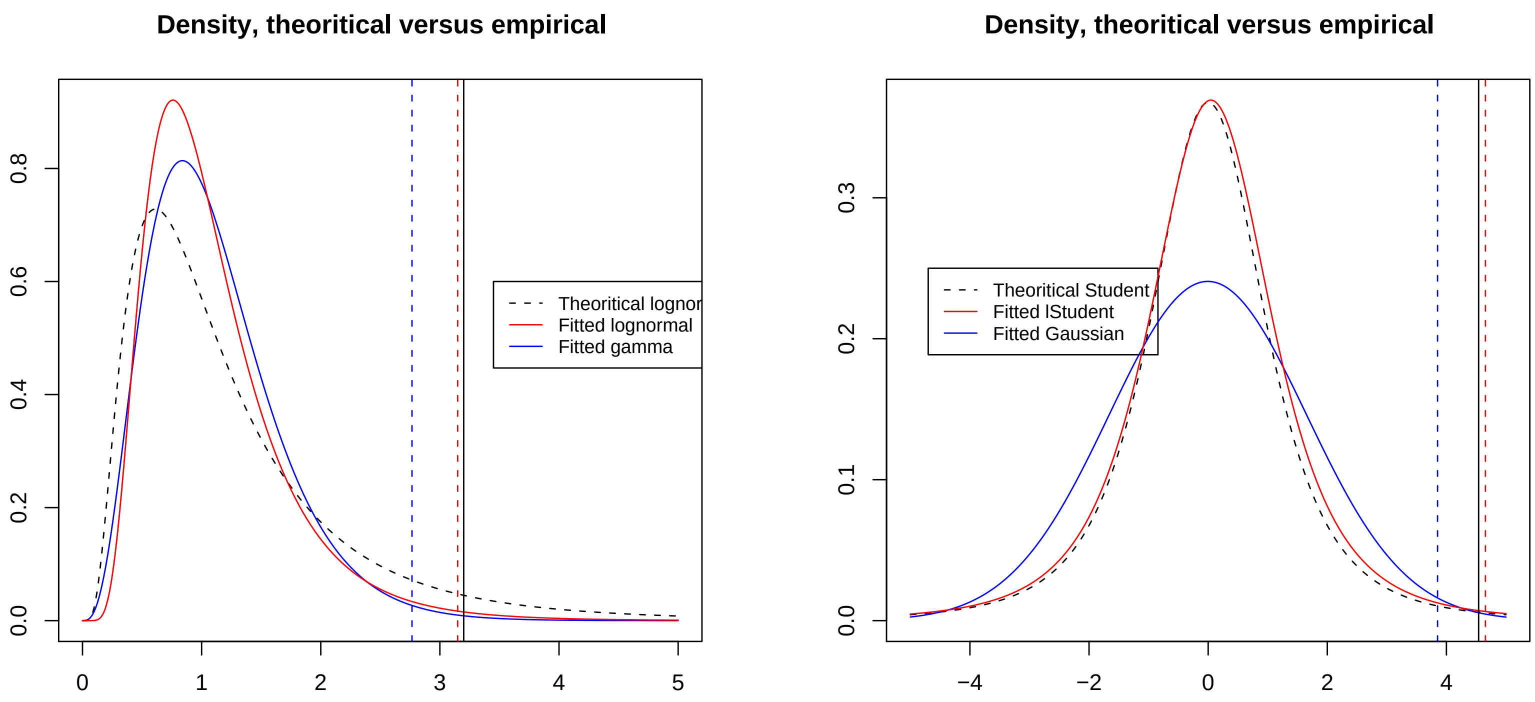


Figure 22 – Estimation of Value-at-Risk, model error.

Using a semiparametric models

Given a n-sample $\{Y_1, \ldots, Y_n\}$, let $Y_{1:n} \leq Y_{2:n} \leq \ldots \leq Y_{n:n}$ denotes the associated order statistics.

If u large enough, $Y - u$ given $Y > u$ has a Generalized Pareto distribution with parameters ξ and β (Pickands-Balkema-de Haan theorem).

If $u = Y_{n-k:n}$ for k large enough, and if $\xi_>0$, denote by $\widehat{\beta}_k$ and $\widehat{\xi}_k$ maximum likelihood estimators of the Genralized Pareto distribution of sample $\{Y_{n-k+1:n} - Y_{n-k:n}, ..., Y_{n:n} - Y_{n-k:n}\}$,

$$\widehat{Q}(Y, \alpha) = Y_{n-k:n} + \frac{\widehat{\beta}_k}{\widehat{\xi}_k}\left(\left(\frac{n}{k}(1-\alpha)\right)^{-\widehat{\xi}_k} - 1\right), \qquad (8)$$

An alternative is to use Hill's estimator if $\xi > 0$,

$$\widehat{Q}(Y, \alpha) = Y_{n-k:n}\left(\frac{n}{k}(1-\alpha)\right)^{-\widehat{\xi}_k}, \widehat{\xi}_k = \frac{1}{k}\sum_{i=1}^{k} \log Y_{n+1-i:n} - \log Y_{n-k:n}. \qquad (9)$$

On nonparametric estimation for quantiles

For continuous distribution $q(\alpha) = F_X^{-1}(\alpha)$, thus, a natural idea would be to consider $\widehat{q}(\alpha) = \widehat{F}_X^{-1}(\alpha)$, for some nonparametric estimation of F_X.

Definition3

The empirical cumulative distribution function F_n, based on sample $\{X_1, \ldots, X_n\}$ is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x).$$

Definition4

The kernel based cumulative distribution function, based on sample $\{X_1, \ldots, X_n\}$ is

$$\widehat{F}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{x} k\left(\frac{X_i - t}{h}\right) dt = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

where $K(x) = \int_{-\infty}^{x} k(t)dt$, k being a *kernel* and h the bandwidth.

Smoothing nonparametric estimators

Two techniques have been considered to smooth estimation of quantiles, either implicit, or explicit.

- consider a linear combinaison of order statistics,

The classical empirical quantile estimate is simply

$$Q_n(p) = F_n^{-1}\left(\frac{i}{n}\right) = X_{i:n} = X_{[np]:n} \text{ where } [\cdot] \text{ denotes the integer part.} \tag{10}$$

The estimator is simple to obtain, but depends only on *one* observation. A natural extention will be to use - at least - two observations, if np is not an integer. The *weighted empirical quantile estimate* is then defined as

$$Q_n(p) = (1-\gamma)\, X_{[np]:n} + \gamma X_{[np]+1:n} \text{ where } \gamma = np - [np].$$

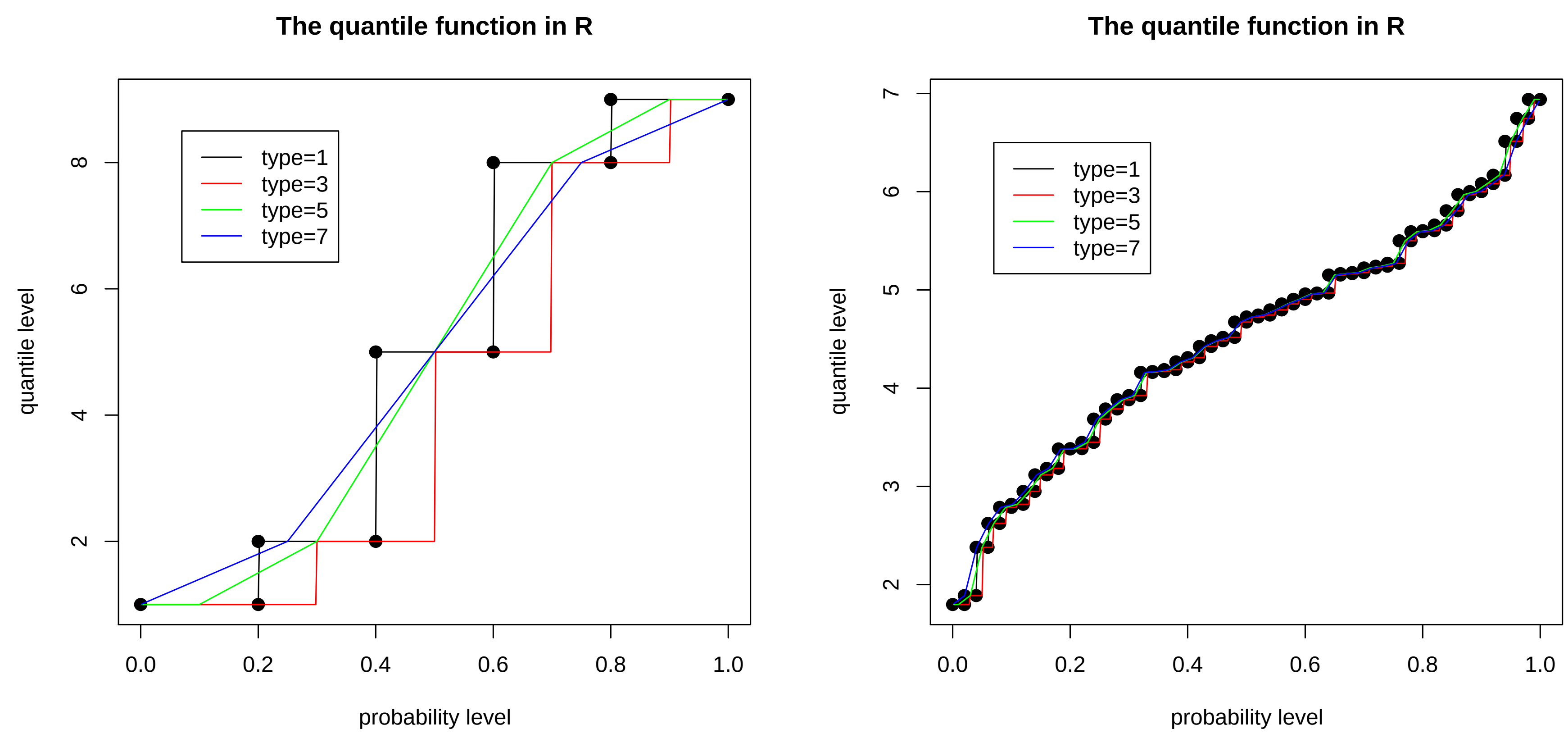


Figure 23 – Several quantile estimators in R.

Smoothing nonparametric estimators

In order to increase efficiency, *L-statistics* can be considered i.e.

$$Q_n(p) = \sum_{i=1}^{n} W_{i,n,p} X_{i:n} = \sum_{i=1}^{n} W_{i,n,p} F_n^{-1}\left(\frac{i}{n}\right) = \int_0^1 F_n^{-1}(t)\, k(p,h,t)\, dt \qquad (11)$$

where F_n is the empirical distribution function of F_X, where k is a kernel and h a bandwidth. This expression can be written equivalently

$$Q_n(p) = \sum_{i=1}^{n} \left[\int_{\frac{(i-1)}{n}}^{\frac{i}{n}} k\left(\frac{t-p}{h}\right) dt \right] X_{(i)} = \sum_{i=1}^{n} \left[\mathbb{K}\left(\frac{\frac{i}{n}-p}{h}\right) - \mathbb{K}\left(\frac{\frac{i-1}{n}-p}{h}\right) \right] X_{(i)} \qquad (12)$$

where again $\mathbb{K}(x) = \displaystyle\int_{-\infty}^{x} k(t)\, dt$. The idea is to give more weight to order statistics $X_{(i)}$ such that i is closed to pn.

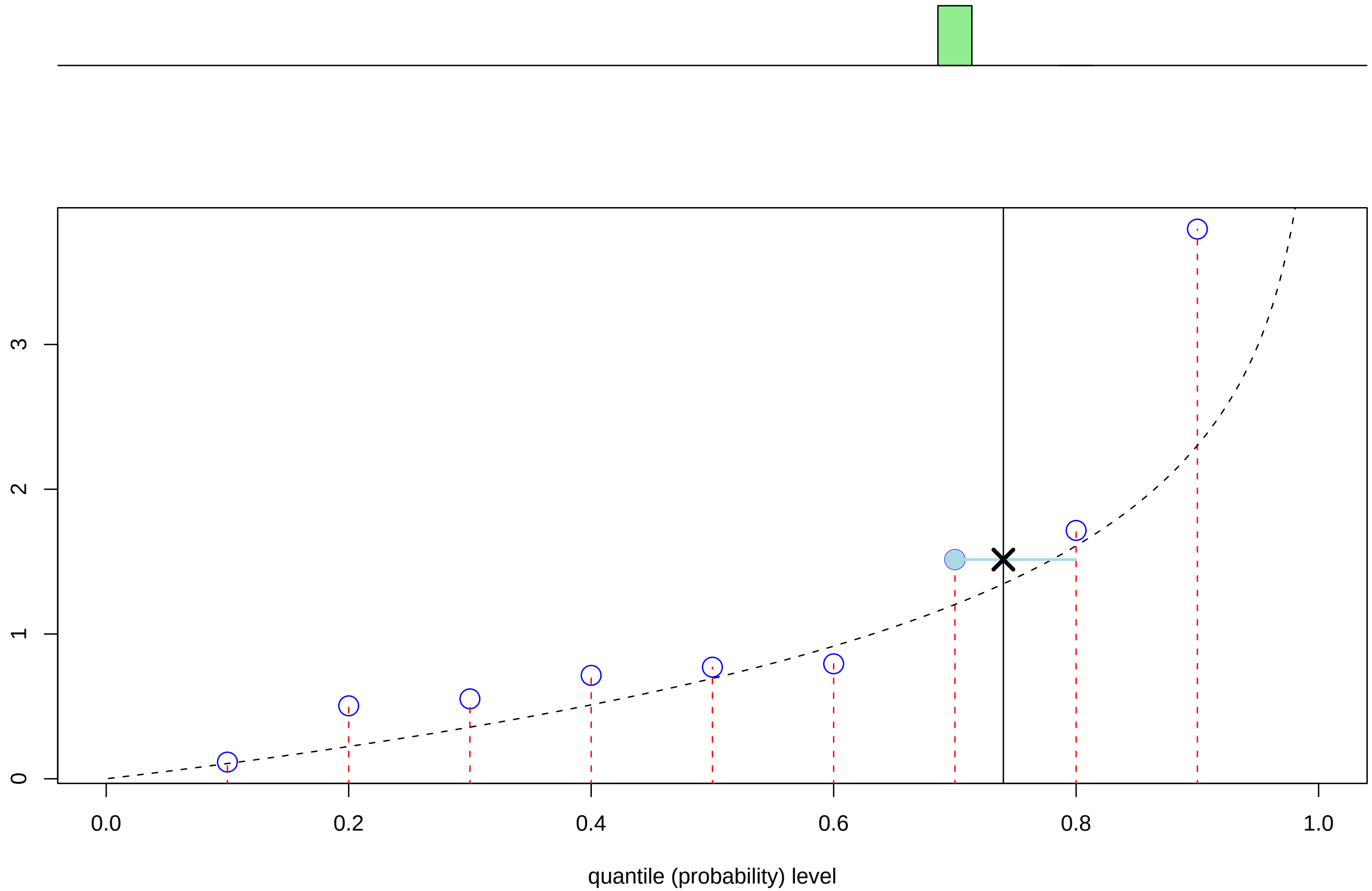


Figure 24 – Quantile estimator as wieghted sum of order statistics.

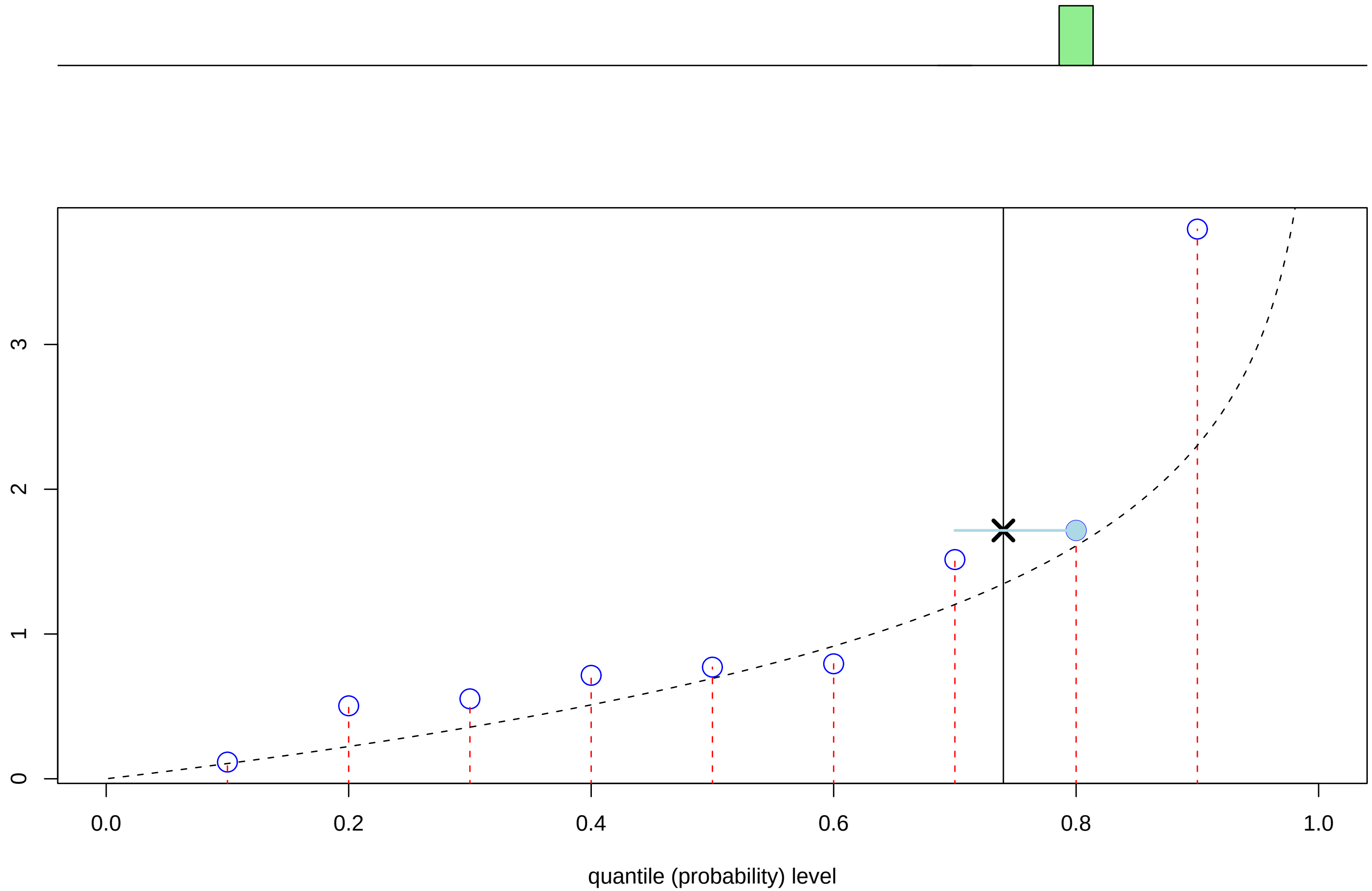


Figure 25 – Quantile estimator as wieghted sum of order statistics.

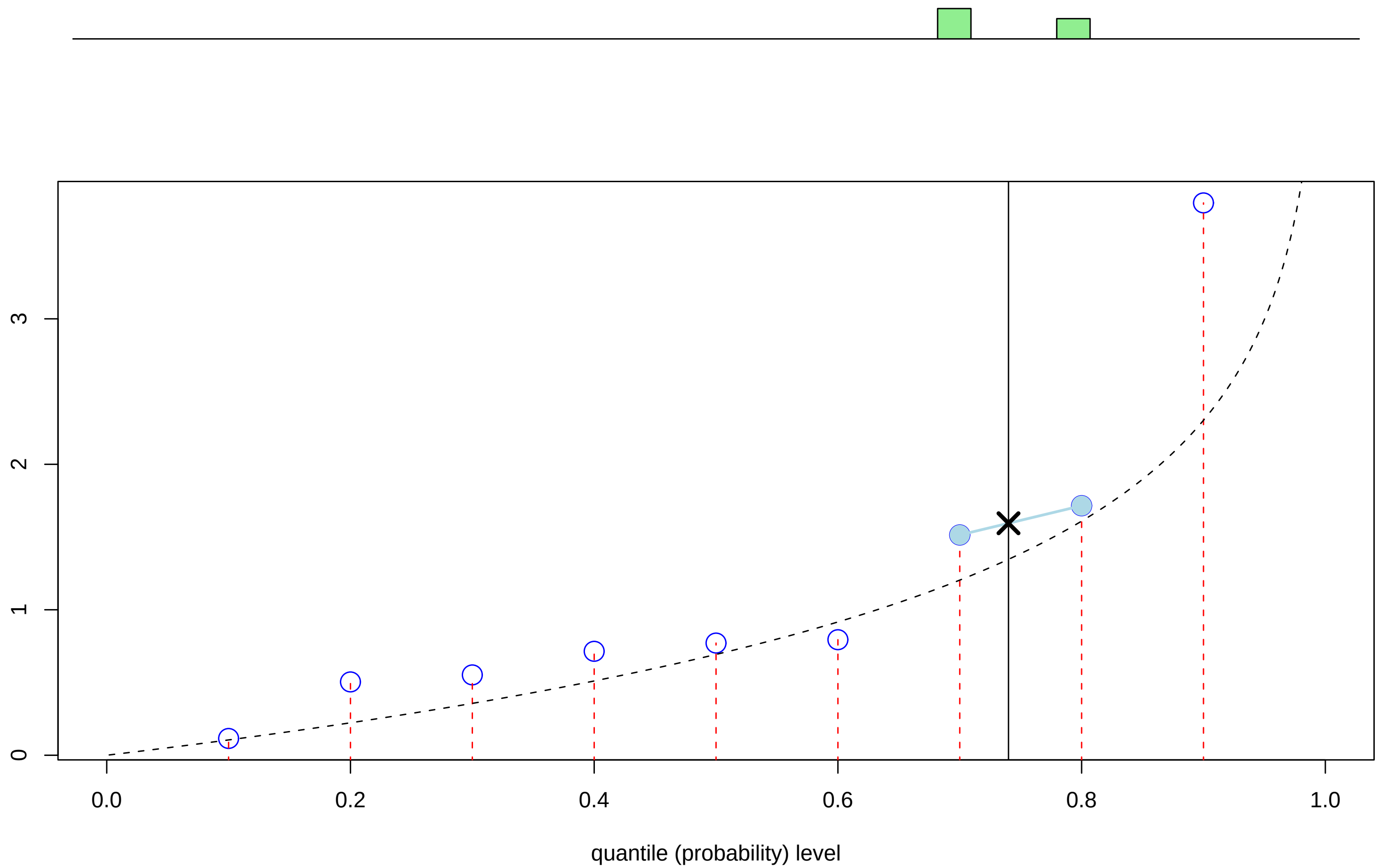


Figure 26 – Quantile estimator as wieghted sum of order statistics.

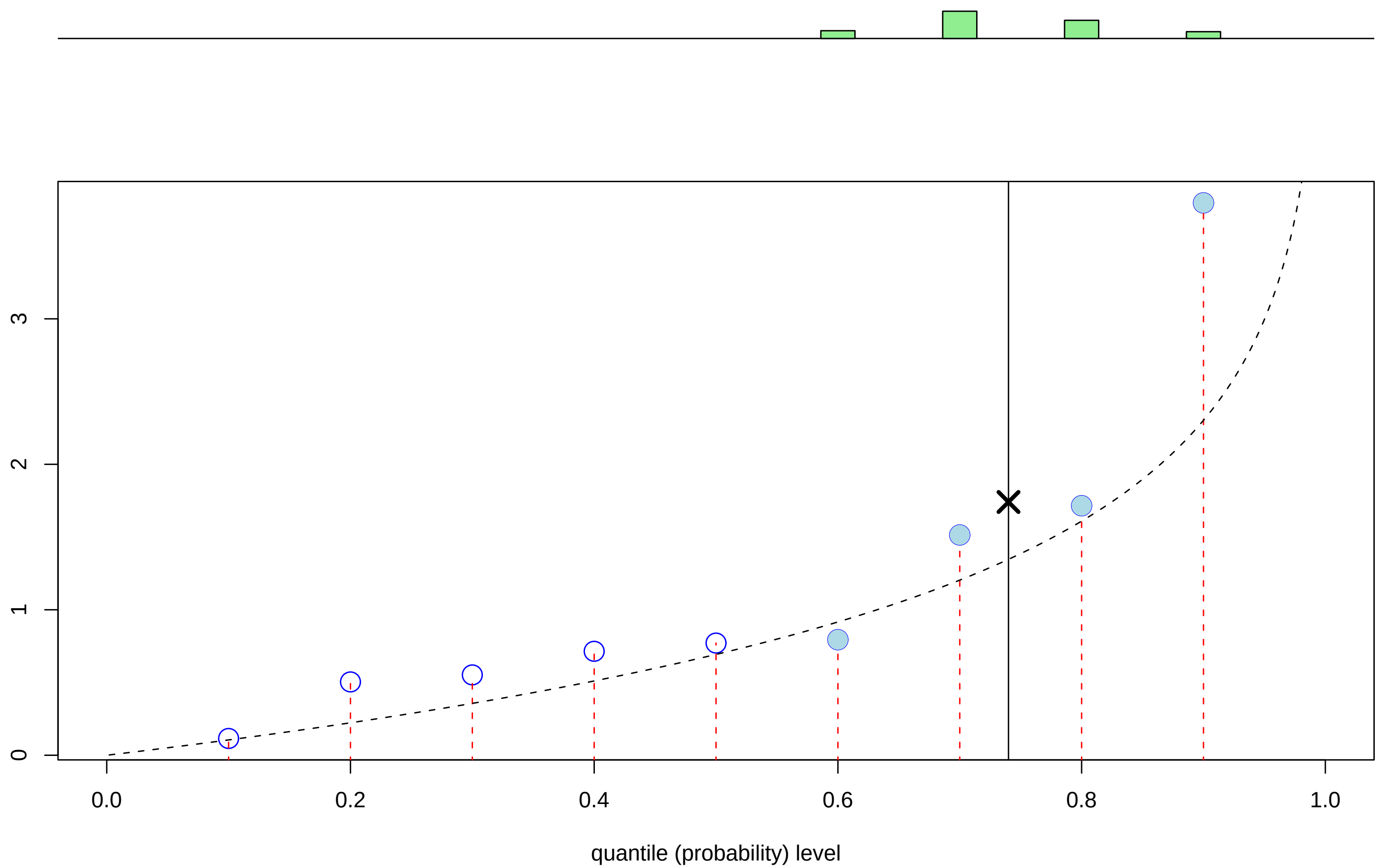


Figure 27 – Quantile estimator as wieghted sum of order statistics.

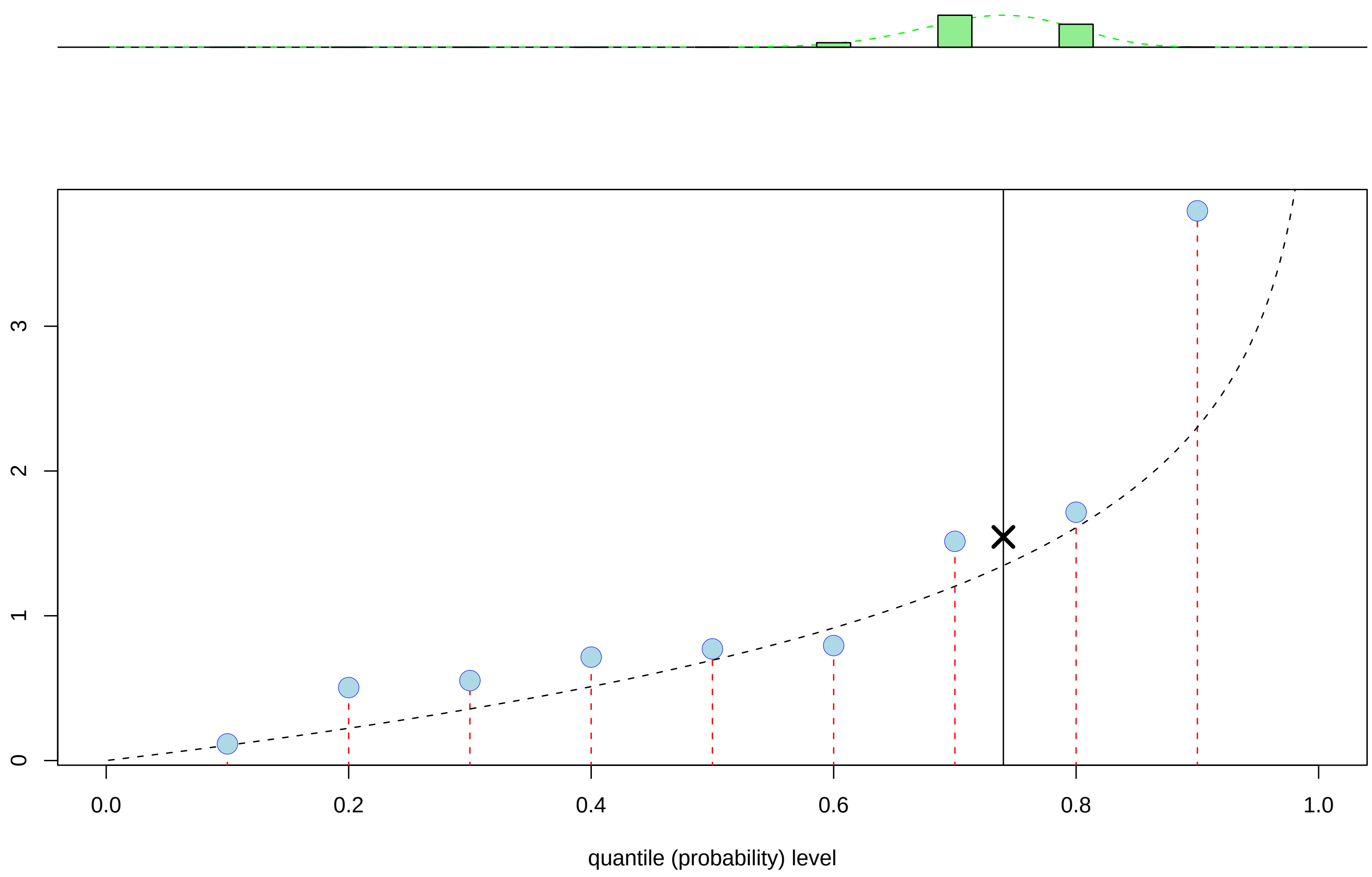


Figure 28 – Quantile estimator as wieghted sum of order statistics.

Smoothing nonparametric estimators

E.g. the so-called Harrell-Davis estimator is defined as

$$Q_n(p) = \sum_{i=1}^{n} \left[\int_{\frac{(i-1)}{n}}^{\frac{i}{n}} \frac{\Gamma(n+1)}{\Gamma((n+1)p)\Gamma((n+1)q)} y^{(n+1)p-1}(1-y)^{(n+1)q-1} \right] X_{i:n},$$

- find a smooth estimator for F_X, and then find (numerically) the inverse,

The α-quantile is defined as the solution of $F_X \circ q_X(\alpha) = \alpha$.

If $\widehat{F}_n$ denotes a continuous estimate of F, then a natural estimate for $q_X(\alpha)$ is $\widehat{q}_n(\alpha)$ such that $\widehat{F}_n \circ \widehat{q}_n(\alpha) = \alpha$, obtained using e.g. Gauss-Newton algorithm.

Improving Beta kernel estimators

Problem : the convergence is not uniform, and there is large second order bias on borders, i.e. 0 and 1.

Chen (1999) proposed a modified Beta 2 kernel estimator, based on

$$k_2(u;b;t) = \begin{cases} k_{\frac{t}{b},\frac{1-t}{b}}(u) & \text{, if } t \in [2b, 1-2b] \\ k_{\rho_b(t),\frac{1-t}{b}}(u) & \text{, if } t \in [0, 2b) \\ k_{\frac{t}{b},\rho_b(1-t)}(u) & \text{, if } t \in (1-2b, 1] \end{cases}$$

where $\rho_b(t) = 2b^2 + 2.5 - \sqrt{4b^4 + 6b^2 + 2.25 - t^2 - \frac{t}{b}}$.

Non-consistency of Beta kernel estimators

Problem : $k(0,\alpha,\beta)=k(1,\alpha,\beta)=0$. So if there are point mass at 0 or 1, the estimator becomes inconsistent, i.e.

$$\begin{aligned}
\widehat{f}_b(x) &= \frac{1}{n}\sum k\left(X_i, 1+\frac{x}{b}, 1+\frac{1-x}{b}\right), x\in[0,1]\\
&= \frac{1}{n}\sum_{X_i\neq 0,1} k\left(X_i, 1+\frac{x}{b}, 1+\frac{1-x}{b}\right), x\in[0,1]\\
&= \frac{n-n_0-n_1}{n}\frac{1}{n-n_0-n_1}\sum_{X_i\neq 0,1} k\left(X_i, 1+\frac{x}{b}, 1+\frac{1-x}{b}\right), x\in[0,1]\\
&\approx (1-\mathbb{P}(X=0)-\mathbb{P}(X=1))\cdot f_0(x), x\in[0,1]
\end{aligned}$$

and therefore $\widehat{F}_b(x)\approx(1-\mathbb{P}(X=0)-\mathbb{P}(X=1))\cdot F_0(x)$, and we may have problem finding a 95% or 99% quantile since the total mass will be lower.

Non-consistency of Beta kernel estimators

Gouriéroux & Monfort (2007) proposed

$$\widehat{f}_b^{(1)}(x) = \frac{\widehat{f}_b(x)}{\int_0^1 \widehat{f}_b(t)dt}, \text{ for all } x \in [0,1].$$

It is called macro-β since the correction is performed globally.

Gouriéroux & Monfort (2007) proposed

$$\widehat{f}_b^{(2)}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{k_\beta(X_i; b; x)}{\int_0^1 k_\beta(X_i; b; t)dt}, \text{ for all } x \in [0,1].$$

It is called micro-β since the correction is performed locally.

Transforming observations ?

In the context of density estimation, Devroye and Györfi (1985) suggested to use a so-called transformed kernel estimate

Given a random variable Y , if H is a strictly increasing function, then the p-quantile of $H(Y)$ is equal to $H(q(Y;p))$.

An idea is to transform initial observations $\{X_1,\cdots,X_n\}$ into a sample $\{Y_1,\cdots,Y_n\}$ where $Y_i=H(X_i)$, and then to use a beta-kernel based estimator, if $H:\mathbb{R}\rightarrow[0,1]$. Then $\widehat{q}_n(X;p)=H^{-1}(\widehat{q}_n(Y;p))$.

In the context of density estimation $\widehat{f}_X(x)=\widehat{f}_Y(H(x))H'(x)$. As mentioned in Devroye and Györfi (1985) (p 245), "*for a transformed histogram histogram estimate, the optimal H gives a uniform $[0,1]$ density and should therefore be equal to $H(x)=F(x)$, for all x*".

Transforming observations ? a monte carlo study

Assume that sample $\{X_1, \cdots, X_n\}$ have been generated from F_{θ_0} (from a familly $\mathcal{F} = (F_\theta, \theta \in \Theta)$. 4 transformations will be considered

- $H = F_{\widehat{\theta}}$ (based on a maximum likelihood procedure)
- $H = F_{\theta_0}$ (theoritical optimal transformation)
- $H = F_\theta$ with $\theta < \theta_0$ (heavier tails)
- $H = F_\theta$ with $\theta > \theta_0$ (lower tails)

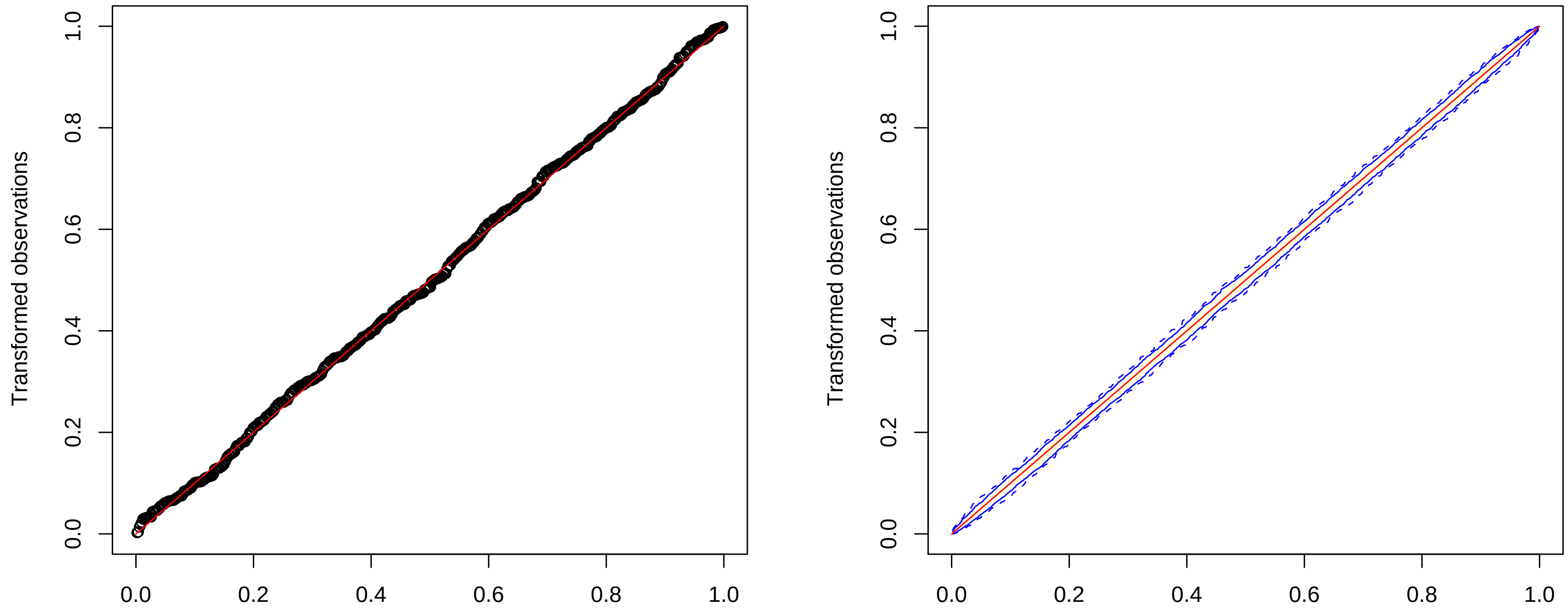


Figure 29 – $\widehat{F}(X_i)$ versus $F_{\hat{\theta}}(X_i)$, i.e. PP plot.

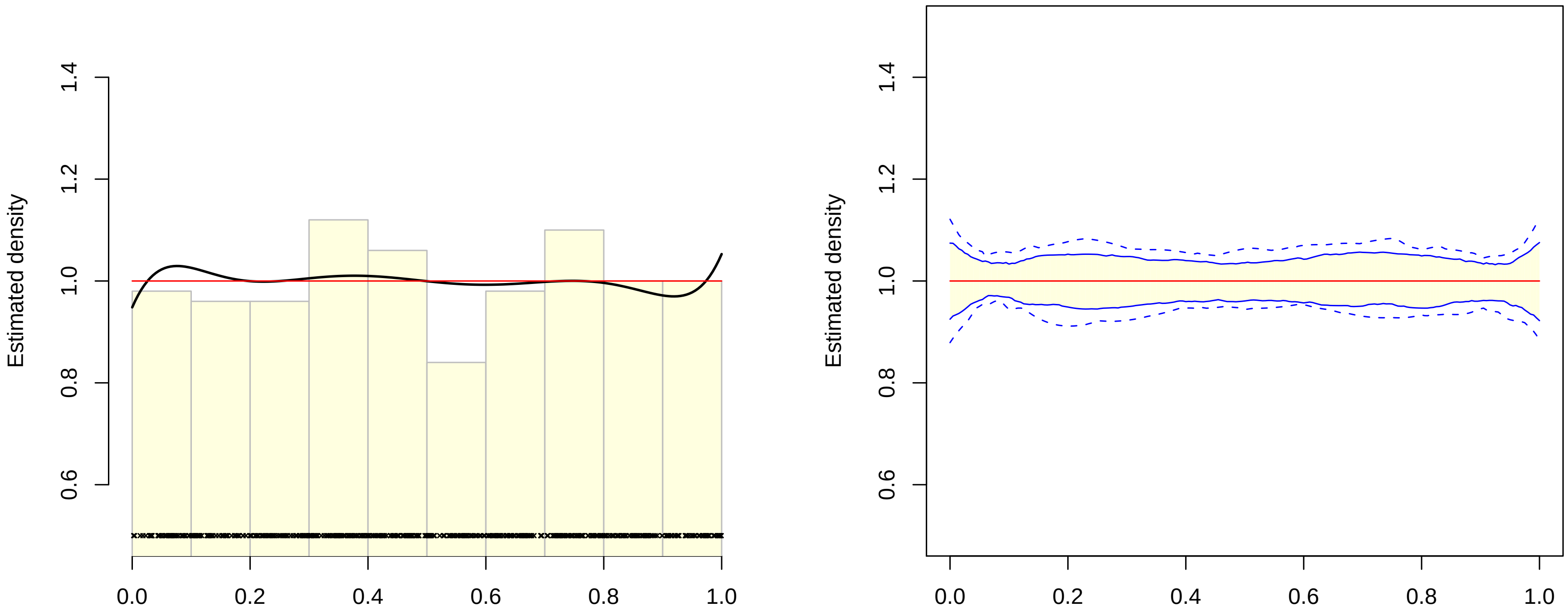


Figure 30 – Nonparametric estimation of the density of the $F_{\hat{\theta}}(X_i)$'s.

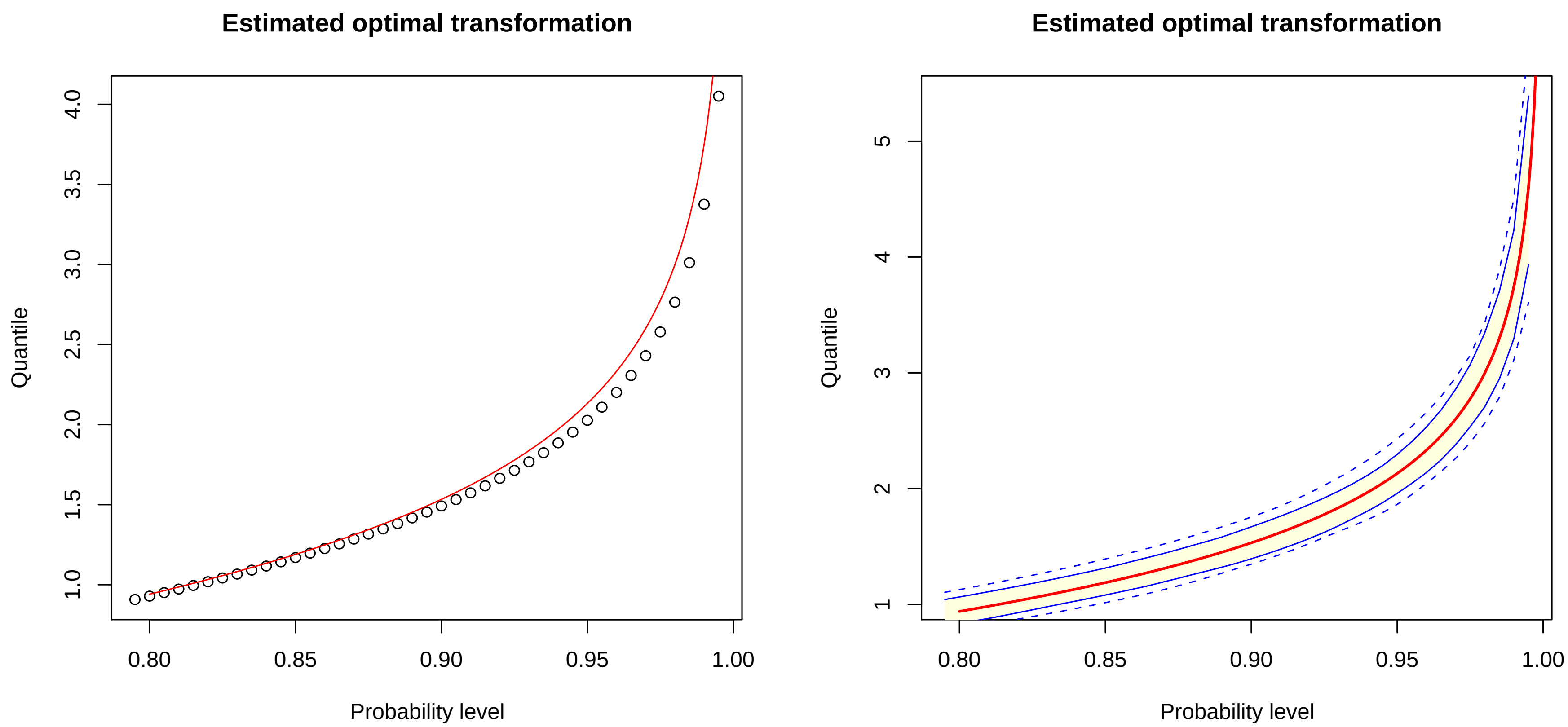


Figure 31 – Nonparametric estimation of the quantile function, $F_{\hat{\theta}}^{-1}(q)$.

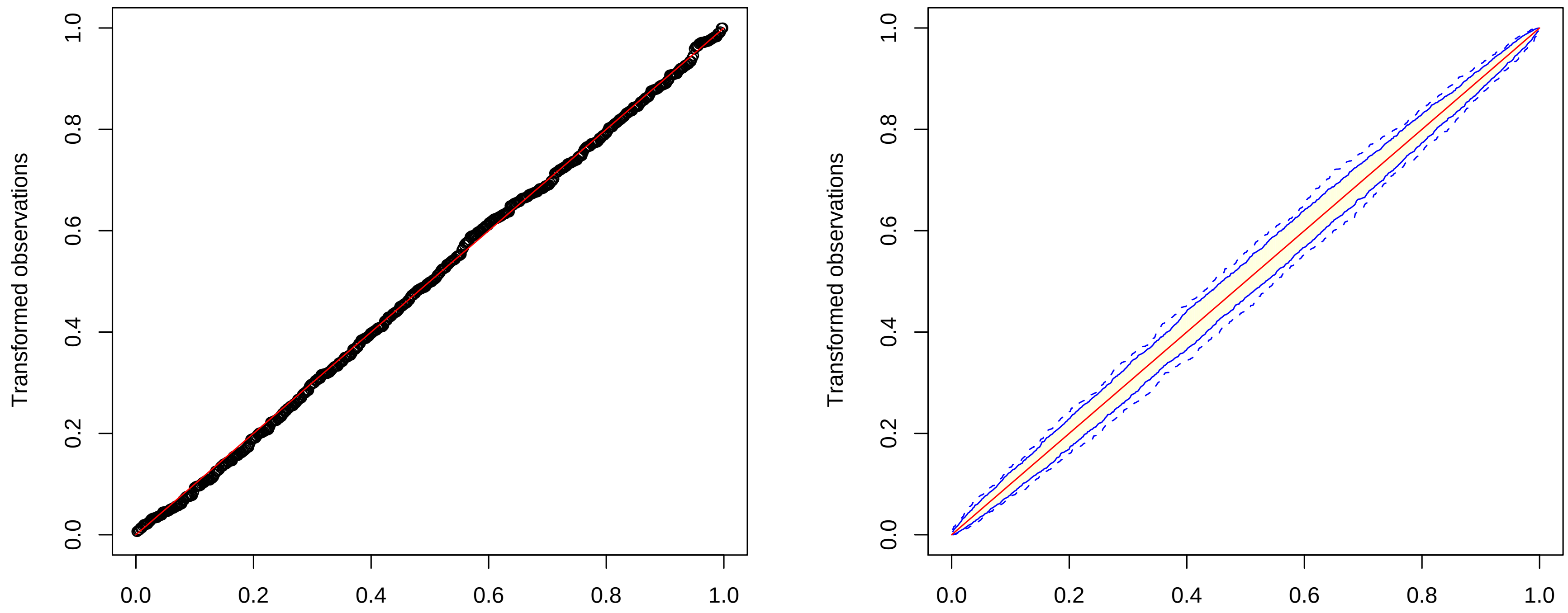


FIGURE 32 – $\widehat{F}(X_i)$ versus $F_{\theta_0}(X_i)$, i.e. PP plot.

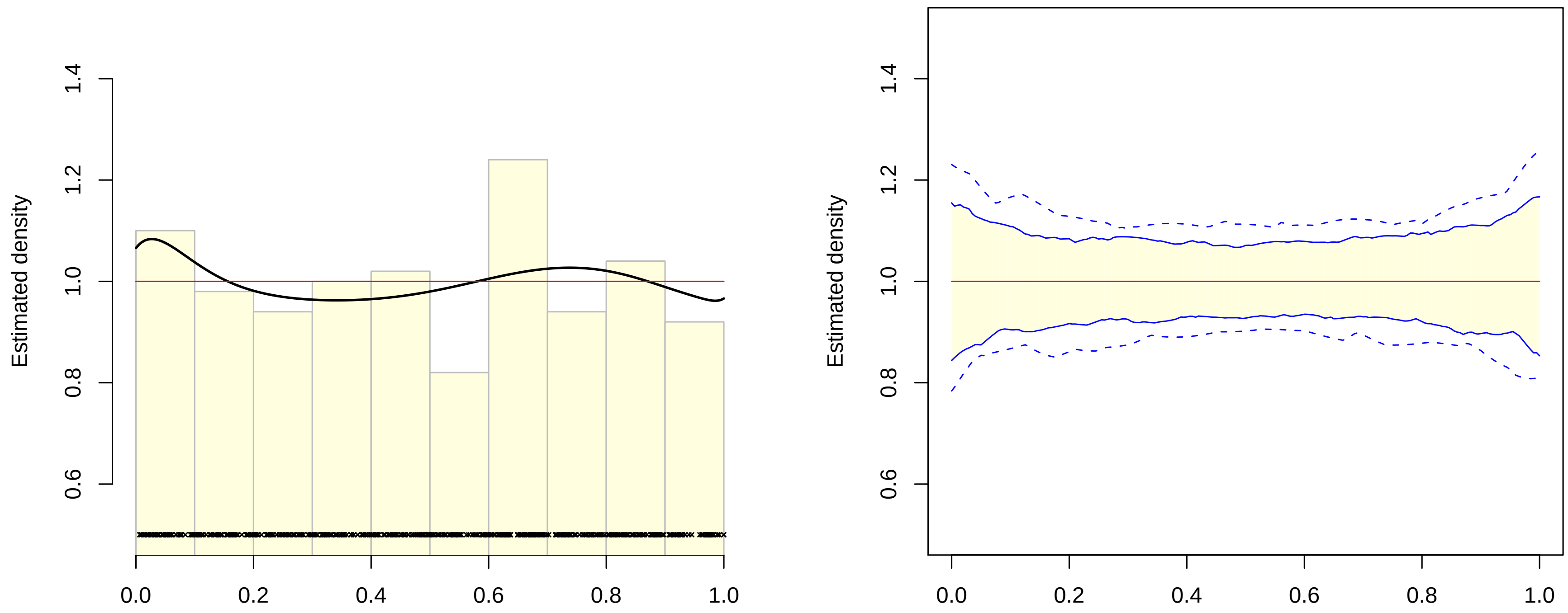


Figure 33 – Nonparametric estimation of the density of the $F_{\theta_0}(X_i)$'s.

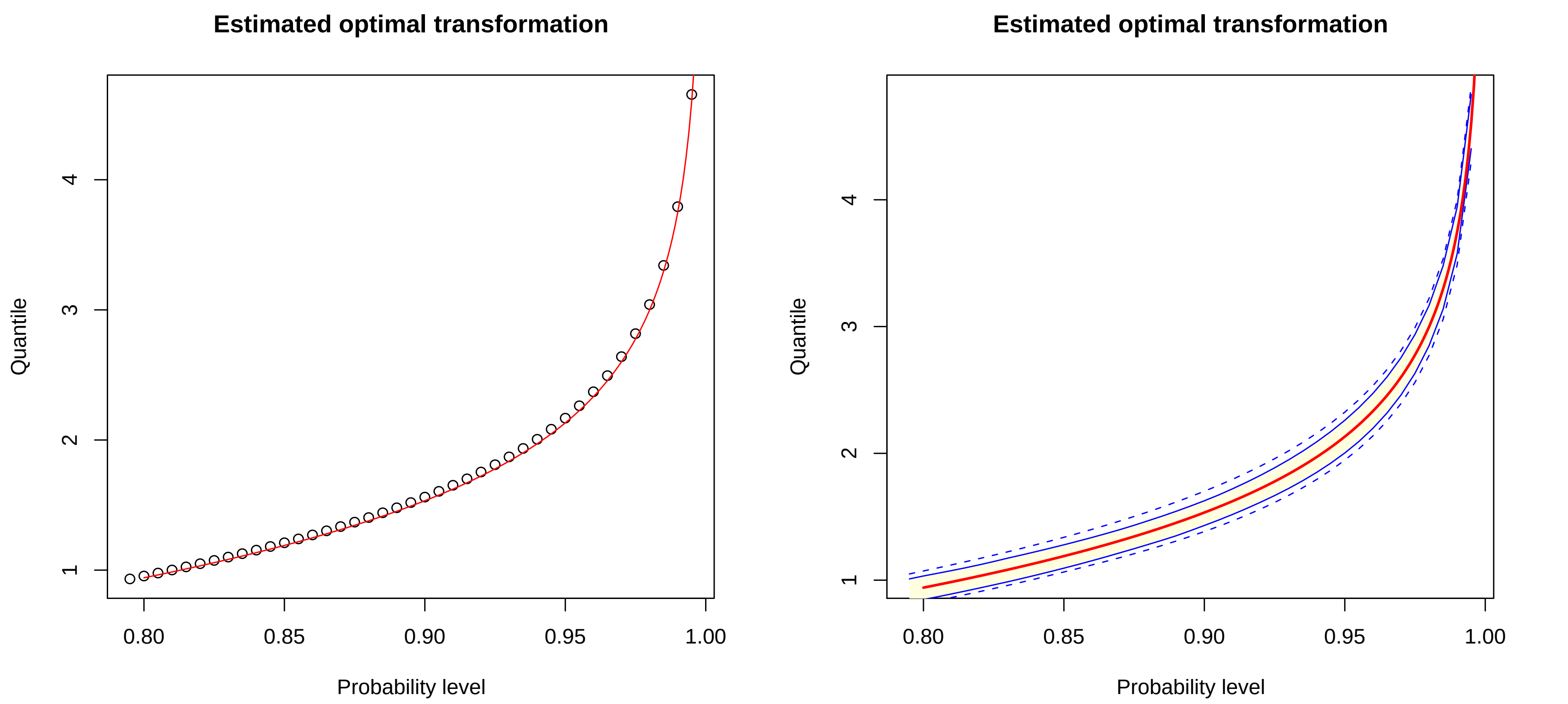


Figure 34 – Nonparametric estimation of the quantile function, $F_{\theta_0}^{-1}(q)$.

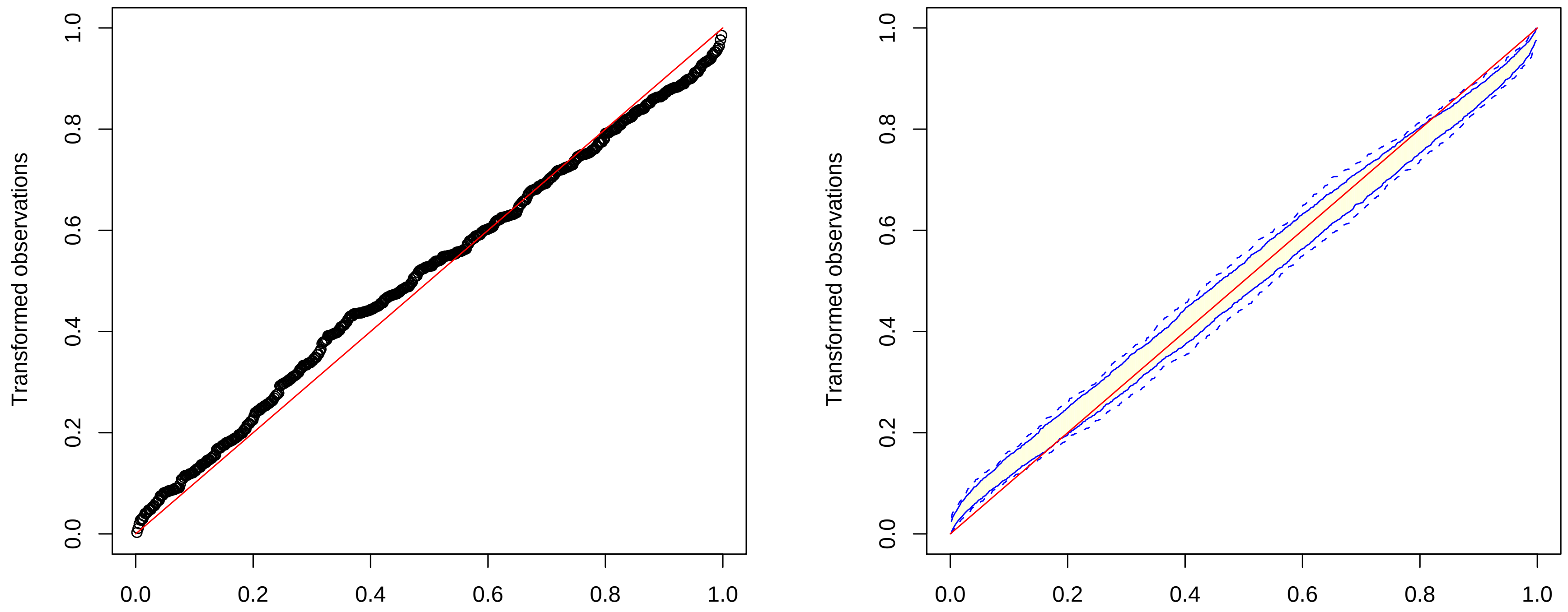


Figure 35 – $\widehat{F}(X_i)$ versus $F_\theta(X_i)$, i.e. PP plot, $\theta < \theta_0$ (heavier tails).

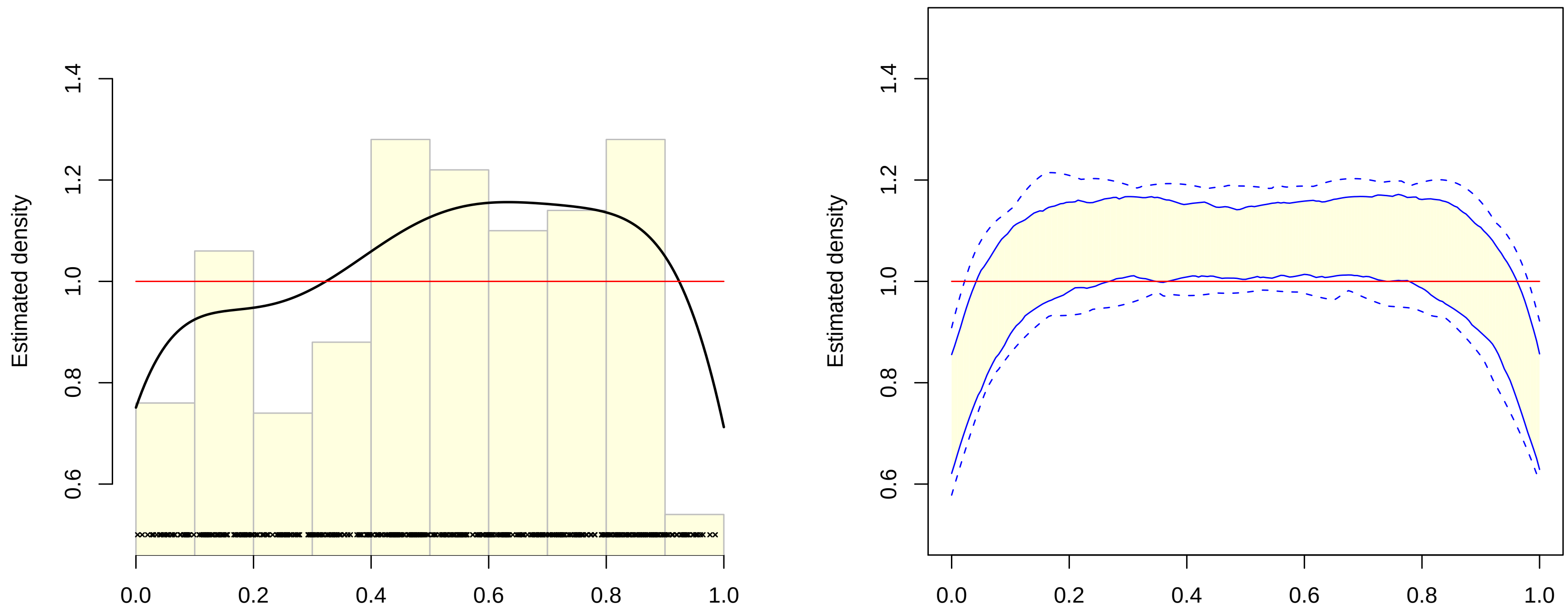


Figure 36 – Estimation of the density of the $F_\theta(X_i)$'s, $\theta < \theta_0$ (heavier tails).

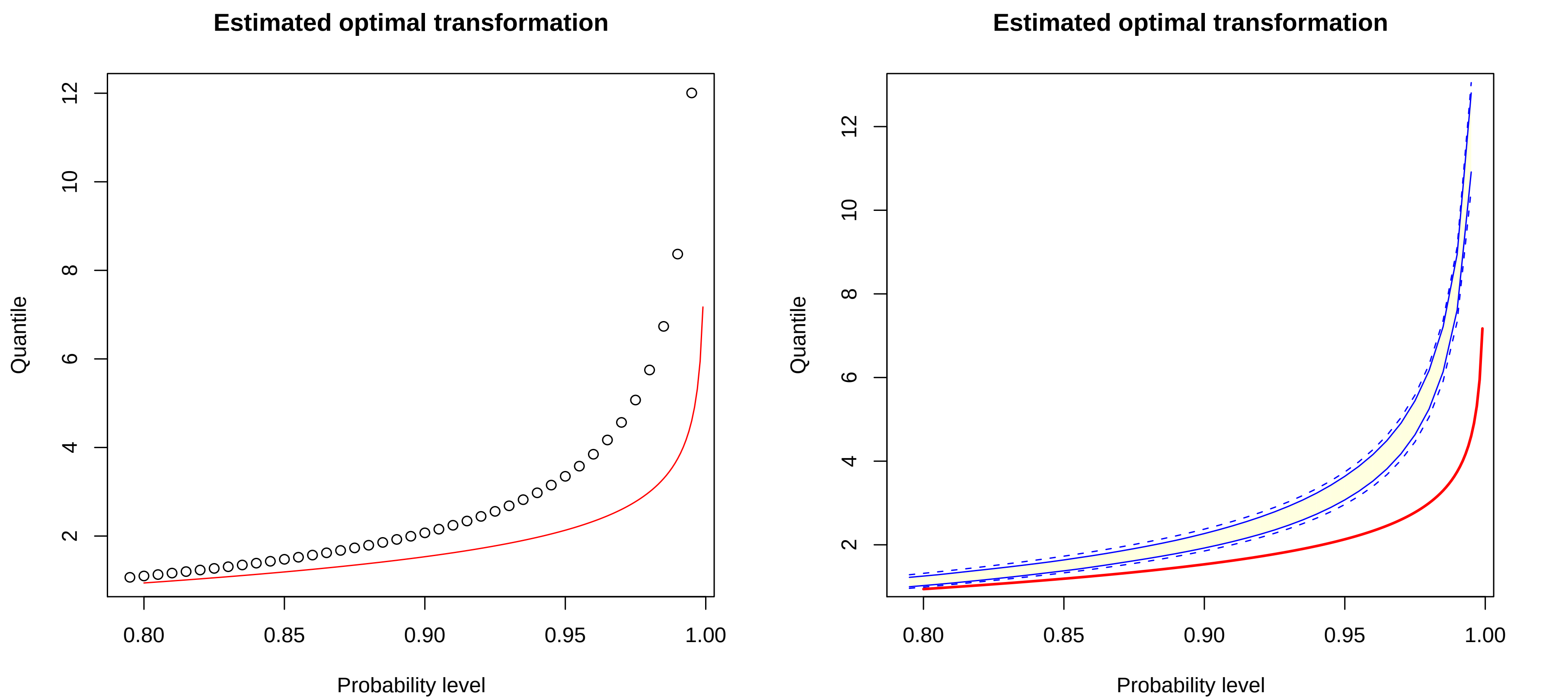


Figure 37 – Estimation of quantile function, $F_\theta^{-1}(q)$, $\theta < \theta_0$ (heavier tails).

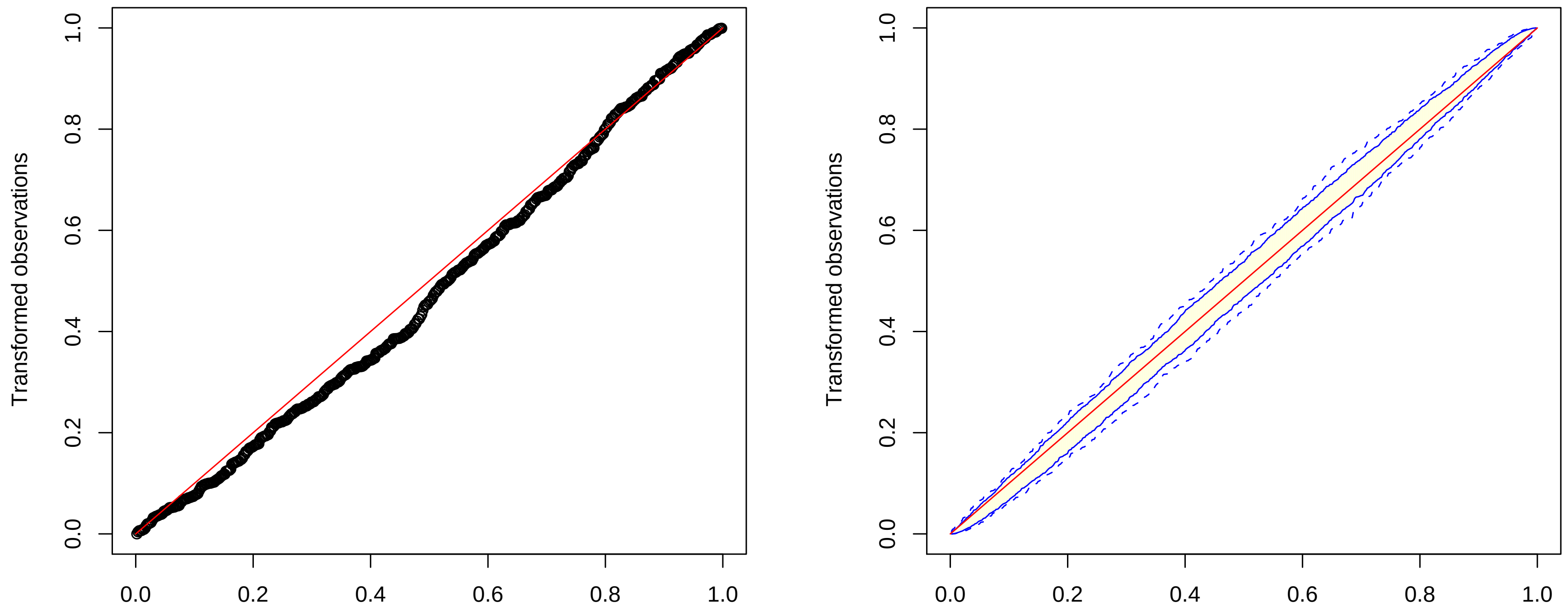
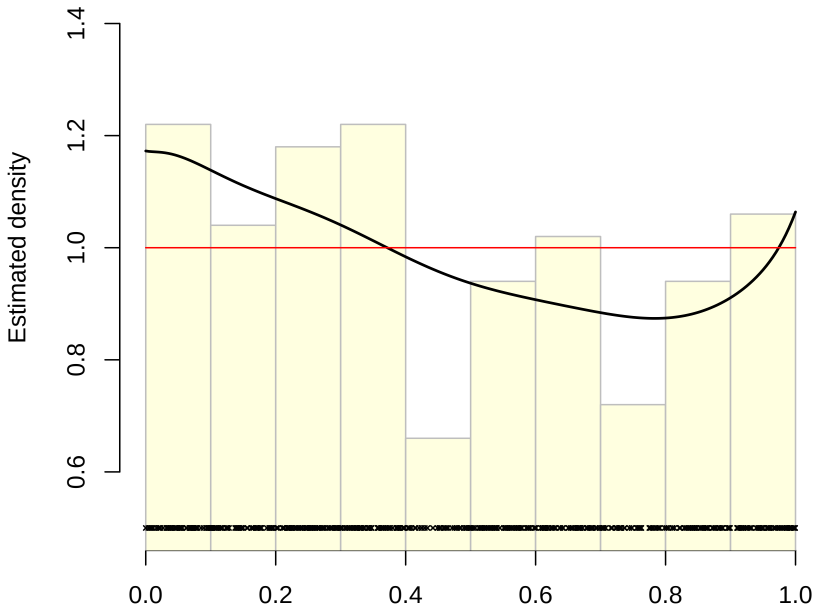


Figure 38 – $\widehat{F}(X_i)$ versus $F_\theta(X_i)$, i.e. PP plot, $\theta > \theta_0$ (lighter tails).



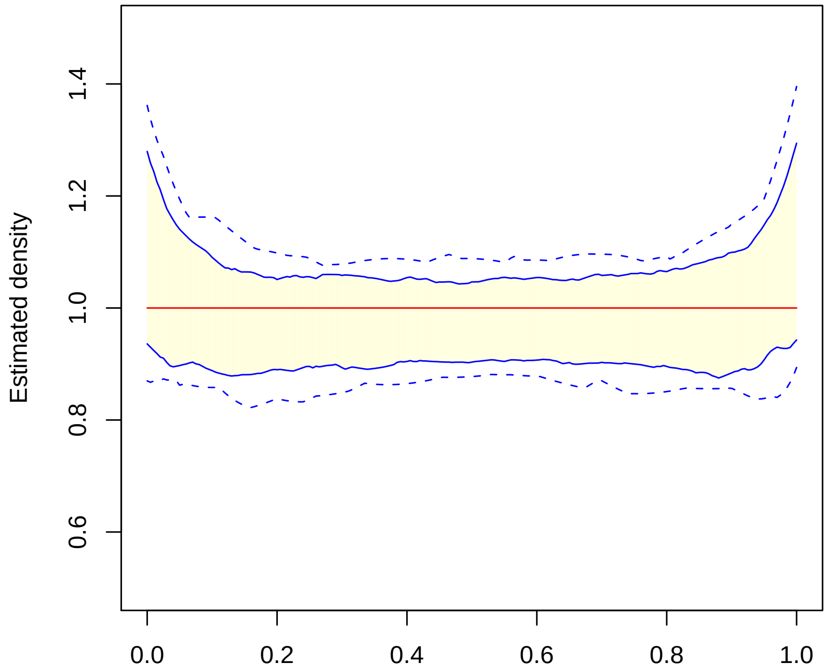


Figure 39 – Estimation of density of $F_\theta(X_i)$'s, $\theta > \theta_0$ (lighter tails).

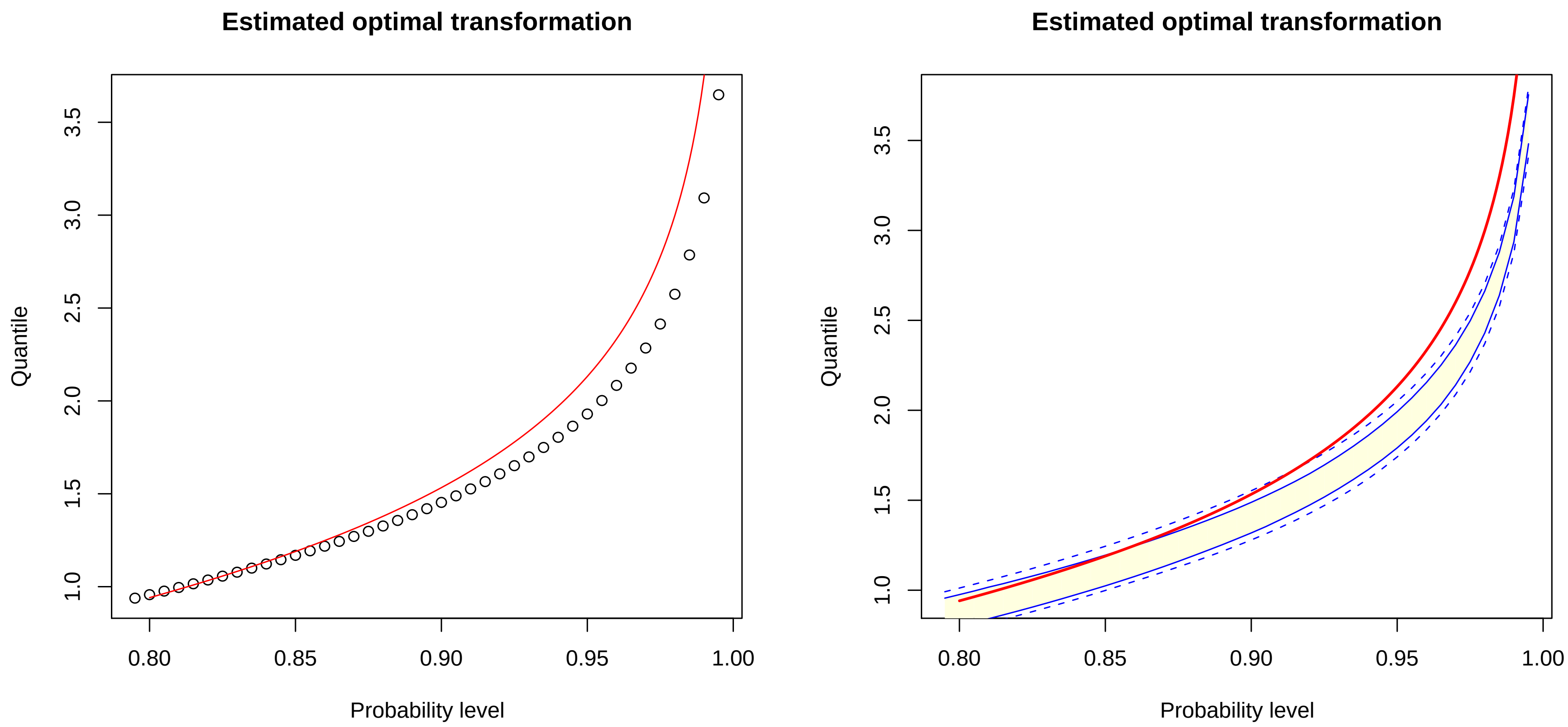


FIGURE 40 – Estimation of quantile function, $F_\theta^{-1}(q)$, $\theta > \theta_0$ (lighter tails).

A *universal* distribution for losses

Buch-Larsen,Nielsen, Guillen, & Bolancé (2005) considered the Champernowne generalized distribution to model insurance claims, i.e. positive variables,

$$F_{\alpha,M,c}(y) = \frac{(y+c)^{\alpha} - c^{\alpha}}{(y+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}} \text{ where } \alpha > 0, c \geq 0 \text{ and } \; M > 0.$$

The associated density is then

$$f_{\alpha,M,c}(y) \;=\; \frac{\alpha (y+c)^{\alpha-1}\left((M+c)^{\alpha} - c^{\alpha}\right)}{\left((y+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}\right)^2}.$$

A Monte Carlo study to compare those nonparametric estimators

As in Buch-Larsen,Nielsen, Guillen, & Bolancé (2005), 4 distributions were considered

- normal distribution,
- Weibull distribution,
- log-normal distribution,
- mixture of Pareto and log-normal distributions,

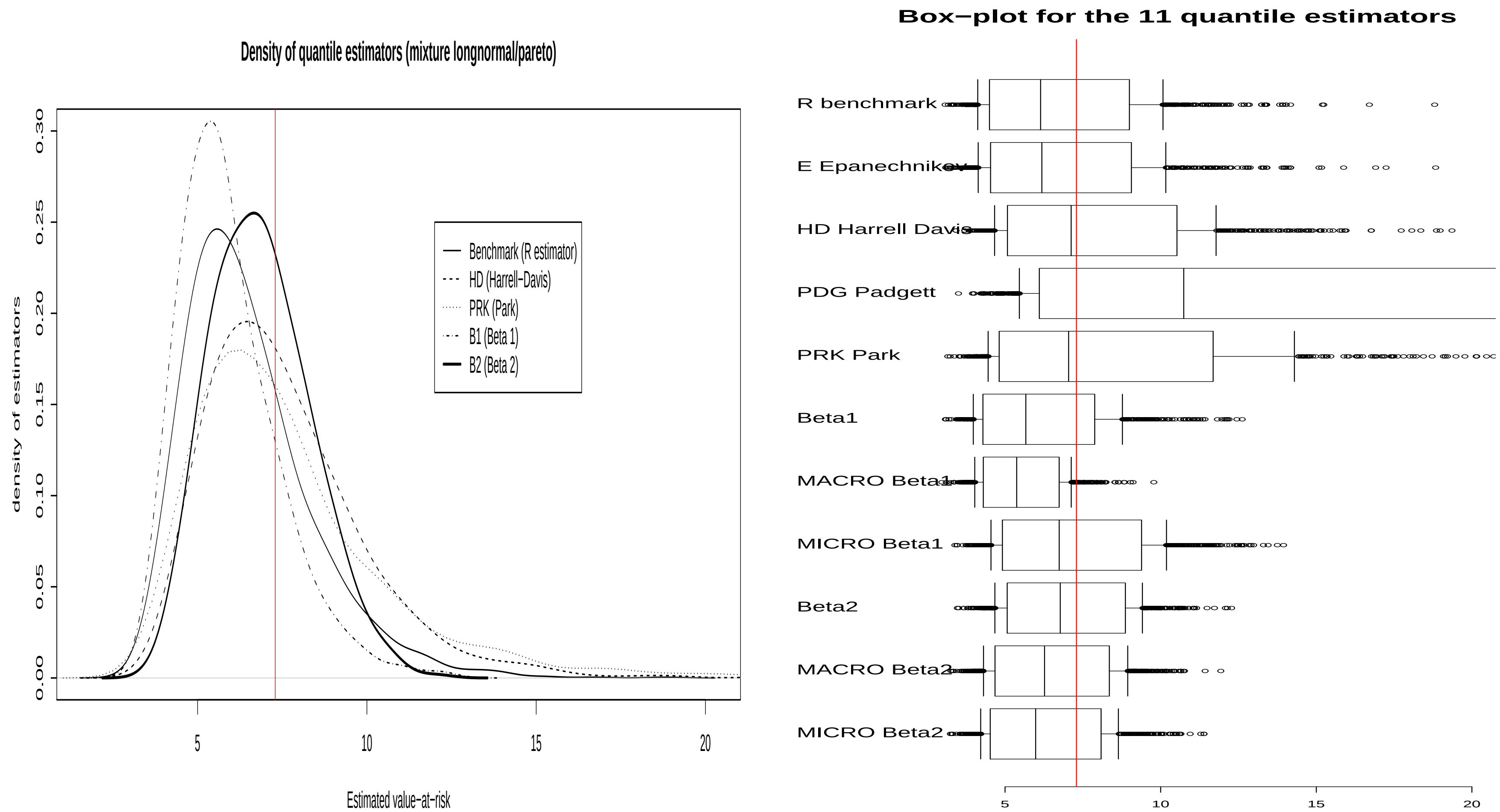


Figure 41 – Distribution of the 95% quantile of the mixture distribution, $n = 200$, and associated box-plots.

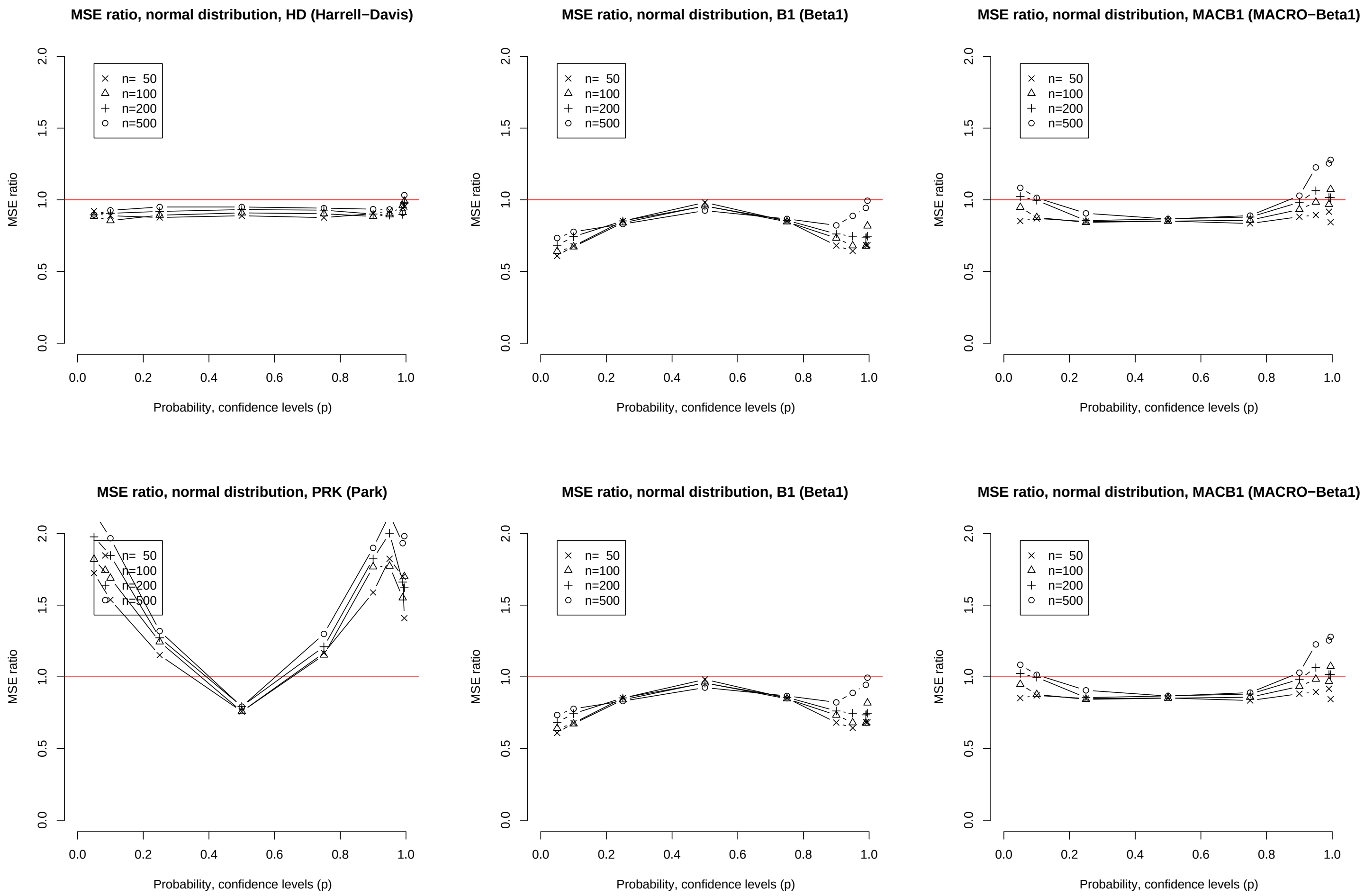


Figure 42 – Comparing MSE for 6 estimators, the normal distribution case.

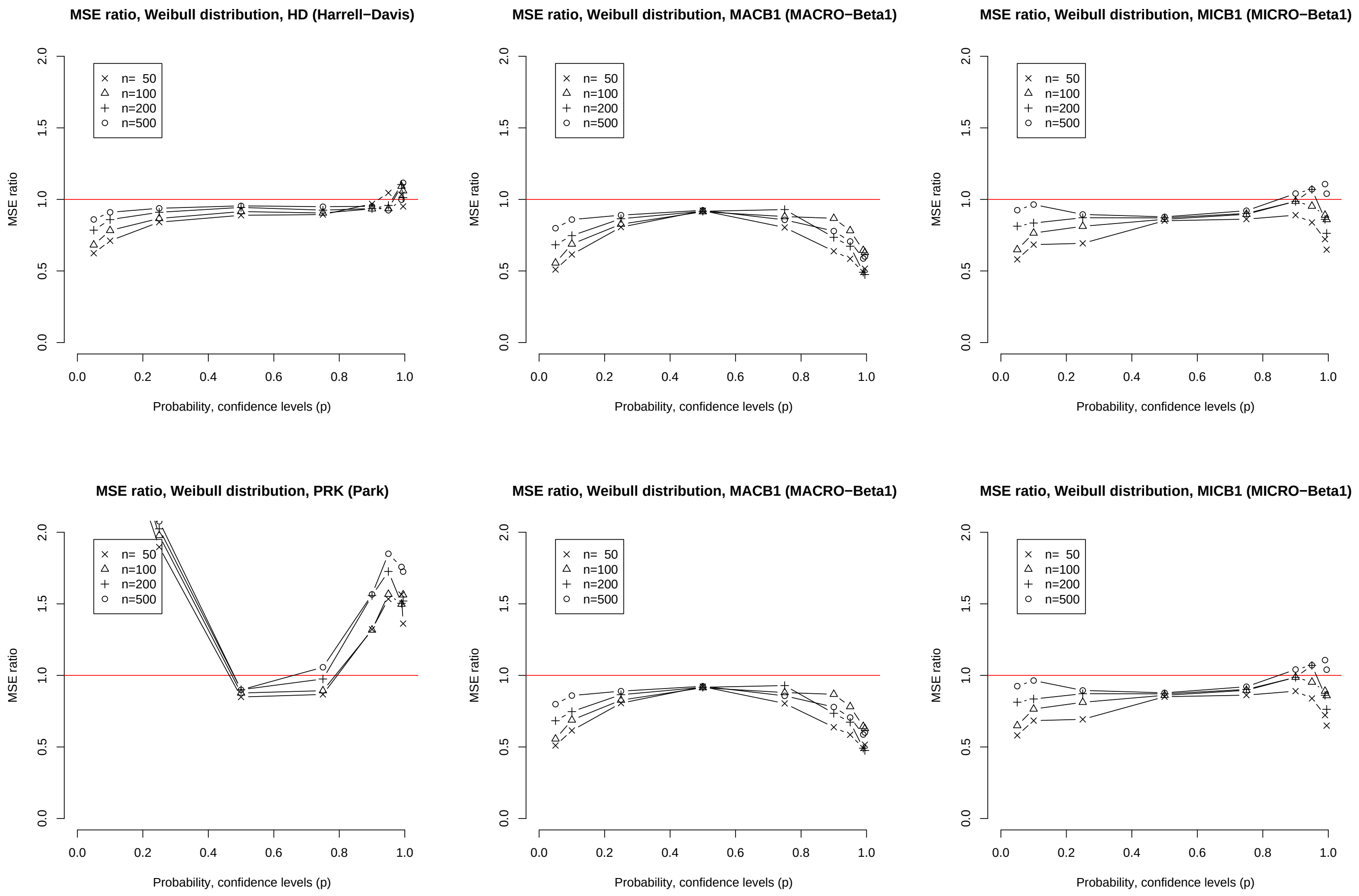


Figure 43 – Comparing MSE for 6 estimators, the Weibull distribution case.

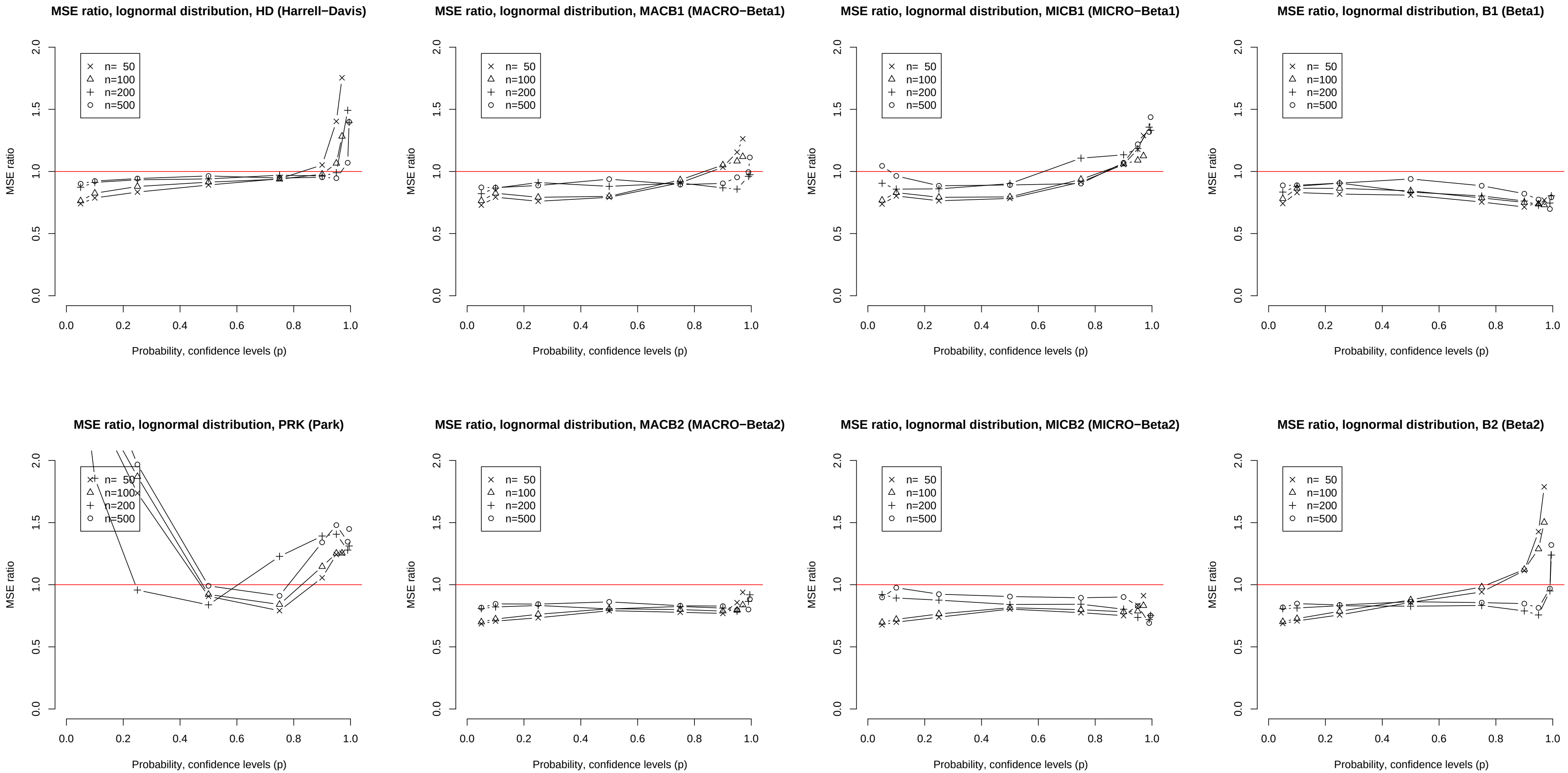


Figure 44 – Comparing MSE for 9 estimators, the lognormal distribution case.

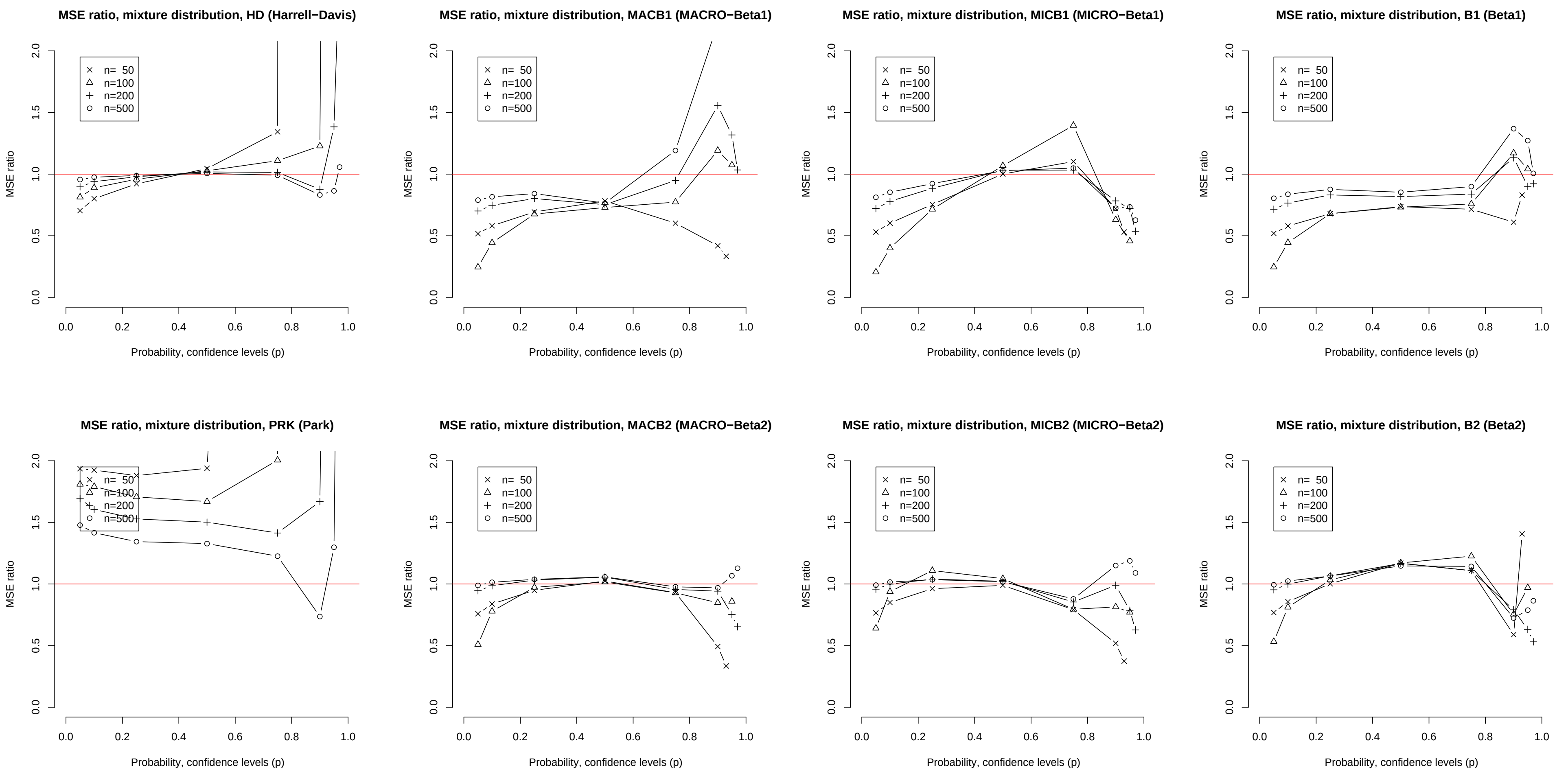


Figure 45 – Comparing MSE for 9 estimators, the mixture distribution case.

Some references

Charpentier, A. , Fermanian, J-D. & Scaillet, O. (2006). The estimation of copulas : theory and practice. In : Copula Methods in Derivatives and Risk Management : From Credit Risk to Market Risk. Risk Book

Charpentier, A. & Oulidi, A. (2008). Beta Kernel estimation for Value-At-Risk of heavy-tailed distributions. *in revision* Journal of Computational Statistics and Data Analysis.

Charpentier, A. & Oulidi, A. (2007). Estimating allocations for Value-at-Risk portfolio optimzation. *to appear in* Mathematical Methods in Operations Research.

Chen, S. X. (1999). A Beta Kernel Estimator for Density Functions. Computational Statistics & Data Analysis, **31**, 131-145.

Devroye, L. & Györfi, L. (1981). Nonparametric Density Estimation : The L1 View. Wiley.

Gouriéroux, C., & Montfort, A. 2006. (Non) Consistency of the Beta Kernel Estimator for Recovery Rate Distribution. *CREST-DP* 2006-32.