A quantile-copula approach to conditional density estimation.

Applications to prediction.

Olivier P. Faugeras

Université Toulouse 1 Capitole - GREMAQ

15/09/2009

Outline

Setup and Motivation

Objective

- observe a sample $((X_i, Y_i); i = 1, \ldots, n)$ i.i.d. of (X, Y).
- predict the output Y for an input X at location x

with minimal assumptions on the law of (X, Y) (Nonparametric setup).

Notation

- $(X, Y) \rightarrow$ joint c.d.f $F_{X,Y}$, joint density $f_{X,Y}$;
- $X \rightarrow$ c.d.f. F, density f;
- $Y \rightarrow$ c.d.f. G, density g.

Setup and Motivation

Objective

- observe a sample $((X_i, Y_i); i = 1, \ldots, n)$ i.i.d. of (X, Y).
- predict the output Y for an input X at location x

with minimal assumptions on the law of (X, Y) (Nonparametric setup).

Notation

- $(X, Y) \rightarrow$ joint c.d.f $F_{X,Y}$, joint density $f_{X,Y}$;
- $X \rightarrow$ c.d.f. F, density f;
- $Y \rightarrow$ c.d.f. G, density g.

Setup and Motivation

Objective

- observe a sample $((X_i, Y_i); i = 1, \ldots, n)$ i.i.d. of (X, Y).
- predict the output Y for an input X at location x

with minimal assumptions on the law of (X, Y) (Nonparametric setup).

Notation

- $(X, Y) \rightarrow$ joint c.d.f $F_{X,Y}$, joint density $f_{X,Y}$;
- $X \rightarrow$ c.d.f. F, density f;
- $Y \rightarrow$ c.d.f. G, density g.

Why estimating the conditional density ?

What is a good prediction ?

1. Classical approach ($\mathbb{L}_2$ theory): the conditional mean or *regression function* $r(x) = E(Y|X = x)$,
2. Fully informative approach: the *conditional density* $f(y|x)$

Why estimating the conditional density ?

What is a good prediction ?

1. Classical approach ($\mathbb{L}_2$ theory): the conditional mean or *regression function* $r(x) = E(Y|X = x)$,
2. Fully informative approach: the *conditional density* $f(y|x)$

Why estimating the conditional density ?

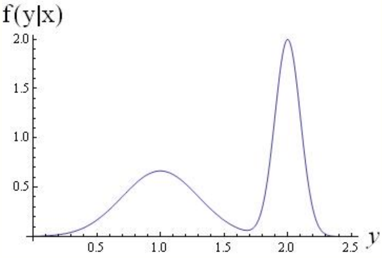
What is a good prediction ?

1. Classical approach ($\mathbb{L}_2$ theory): the conditional mean or *regression function* $r(x) = E(Y|X = x)$,
2. Fully informative approach: the *conditional density* $f(y|x)$

Why estimating the conditional density ?

What is a good prediction ?

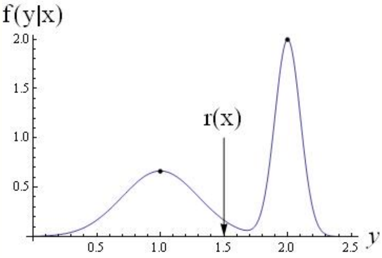
1. Classical approach ($\mathbb{L}_2$ theory): the conditional mean or *regression function* $r(x) = E(Y|X = x)$,
2. Fully informative approach: the *conditional density* $f(y|x)$



Why estimating the conditional density ?

What is a good prediction ?

1. Classical approach ($\mathbb{L}_2$ theory): the conditional mean or *regression function* $r(x) = E(Y|X = x)$,
2. Fully informative approach: the *conditional density* $f(y|x)$



Estimating the conditional density - 1

A first *density*-based approach

$$f(y|x) = \frac{f_{X,Y}(x,y)}{f(x)} \quad \leftarrow \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}(x)}$$

$\hat{f}_{X,Y}, \hat{f}$: Parzen-Rosenblatt kernel estimators with kernels K, K', bandwidths h and h'.

The double kernel estimator

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} K'_{h'}(X_i - x) K_h(Y_i - y)}{\sum_{i=1}^{n} K'_{h'}(X_i - x)} \rightarrow \textit{ratio} \text{ shaped}$$

Estimating the conditional density - 1

A first *density*-based approach

$$f(y|x) = \frac{f_{X,Y}(x,y)}{f(x)} \quad \leftarrow \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}(x)}$$

$\hat{f}_{X,Y}, \hat{f}$: Parzen-Rosenblatt kernel estimators with kernels K, K', bandwidths h and h'.

The double kernel estimator

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} K'_{h'}(X_i - x) K_h(Y_i - y)}{\sum_{i=1}^{n} K'_{h'}(X_i - x)} \to \textit{ratio} \text{ shaped}$$

Estimating the conditional density - 1

A first *density*-based approach

$$f(y|x) = \frac{f_{X,Y}(x,y)}{f(x)} \quad \leftarrow \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}(x)}$$

$\hat{f}_{X,Y}, \hat{f}$: Parzen-Rosenblatt kernel estimators with kernels K, K', bandwidths h and h'.

The double kernel estimator

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} K'_{h'}(X_i - x) K_h(Y_i - y)}{\sum_{i=1}^{n} K'_{h'}(X_i - x)} \rightarrow \textit{ratio} \text{ shaped}$$

Estimating the conditional density - 1

A first *density*-based approach

$$f(y|x) = \frac{f_{X,Y}(x,y)}{f(x)} \quad \leftarrow \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}(x)}$$

$\hat{f}_{X,Y}, \hat{f}$: Parzen-Rosenblatt kernel estimators with kernels K, K', bandwidths h and h'.

The double kernel estimator

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} K'_{h'}(X_i - x) K_h(Y_i - y)}{\sum_{i=1}^{n} K'_{h'}(X_i - x)} \rightarrow \text{\textit{ratio} shaped}$$

Estimating the conditional density - 2

A *regression* strategy

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

Conditional density estimation problem $\rightarrow$ a regression framework

1. *Transform* the data:

$$Y_i \rightarrow Y_i' := (2h)^{-1}\mathbb{1}_{|Y_i-y|\leq h}$$
$$Y_i \rightarrow Y_i' := K_h(Y_i - y) \text{ smoothed version}$$

2. Perform a nonparametric regression of Y_i' on X_is by local averaging methods (Nadaraya-Watson, local polynomial, orthogonal series,...)

Nadaraya-Watson estimator

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} K'_{h'}(X_i - x)K_h(Y_i - y)}{\sum_{i=1}^{n} K'_{h'}(X_i - x)} \rightarrow \text{(same) } \textit{ratio} \text{ shape.}$$

Estimating the conditional density - 2

A *regression* strategy

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

Conditional density estimation problem $\rightarrow$ a regression framework

1. *Transform* the data:

$$Y_i \rightarrow Y_i' := (2h)^{-1}\mathbb{1}_{|Y_i-y|\leq h}$$
$$Y_i \rightarrow Y_i' := K_h(Y_i - y) \text{ smoothed version}$$

2. Perform a nonparametric regression of Y_i' on X_is by local averaging methods (Nadaraya-Watson, local polynomial, orthogonal series,...)

Nadaraya-Watson estimator

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} K'_{h'}(X_i - x)K_h(Y_i - y)}{\sum_{i=1}^{n} K'_{h'}(X_i - x)} \rightarrow \text{(same) } \textit{ratio} \text{ shape.}$$

Estimating the conditional density - 2

A *regression* strategy

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

Conditional density estimation problem $\rightarrow$ a regression framework

1. *Transform* the data:

$$Y_i \rightarrow Y_i' := (2h)^{-1}\mathbb{1}_{|Y_i-y|\leq h}$$

$$Y_i \rightarrow Y_i' := K_h(Y_i - y) \text{ smoothed version}$$

2. Perform a nonparametric regression of Y_i' on X_is by local averaging methods (Nadaraya-Watson, local polynomial, orthogonal series,...)

Nadaraya-Watson estimator

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} K'_{h'}(X_i - x)K_h(Y_i - y)}{\sum_{i=1}^{n} K'_{h'}(X_i - x)} \rightarrow \text{(same) } \textit{ratio} \text{ shape.}$$

Ratio shaped estimators

Bibliography

1. Double kernel estimator: Rosenblatt [1969], Roussas [1969], Stute [1986], Hyndman, Bashtannyk and Grunwald [1996];
2. Local Polynomial: Fan, Yao and Tong [1996], Fan and Yao [2005];
3. Local parametric and constrained local polynomial: Hyndman and Yao [2002]; Rojas, Genovese, Wasserman [2009];
4. Partitioning type estimate: Györfi and Kohler [2007];
5. Projection type estimate: Lacour [2007].

The trouble with ratio shaped estimators

Drawbacks

- quotient shape of estimator is tricky to study;
- *explosive behavior* when the denominator is small $\rightarrow$ numerical implementation delicate (trimming);
- minoration hypothesis on the marginal density $f(x) \geq c > 0$.

How to remedy these problems?

$\rightarrow$ build on the idea of using synthetic data:

find a *representation* of the data more adapted to the problem.

The trouble with ratio shaped estimators

Drawbacks

- quotient shape of estimator is tricky to study;
- *explosive behavior* when the denominator is small $\rightarrow$ numerical implementation delicate (trimming);
- minoration hypothesis on the marginal density $f(x) \geq c > 0$.

How to remedy these problems?
$\rightarrow$ build on the idea of using synthetic data:
find a *representation* of the data more adapted to the problem.

Outline

The quantile transform

What is the “best” transformation of the data in that context ?

The quantile transform theorem

- when F is arbitrary, if U is a uniformly distributed random variable on $(0, 1)$, $X \stackrel{d}{=} F^{-1}(U)$;
- whenever F is continuous, the random variable $U = F(X)$ is uniformly distributed on $(0, 1)$.

$\longrightarrow$ use the invariance property of the quantile transform to construct a pseudo-sample (U_i, V_i) with a *prescribed uniform* marginal distribution.

$$\begin{array}{cc} (X_1, \dots, X_n) & (Y_1, \dots, Y_n) \\ \downarrow & \downarrow \\ (U_1 = F(X_1), \dots, U_n = F(X_n)) & (V_1 = G(Y_1), \dots, V_n = G(Y_n)) \end{array}$$

The quantile transform

What is the "best" transformation of the data in that context ?

The quantile transform theorem

- when F is arbitrary, if U is a uniformly distributed random variable on $(0, 1)$, $X \stackrel{d}{=} F^{-1}(U)$;
- whenever F is continuous, the random variable $U = F(X)$ is uniformly distributed on $(0, 1)$.

$\longrightarrow$ use the invariance property of the quantile transform to construct a pseudo-sample (U_i, V_i) with a *prescribed uniform* marginal distribution.

$$\begin{array}{cc} (X_1, \ldots, X_n) & (Y_1, \ldots, Y_n) \\ \downarrow & \downarrow \\ (U_1 = F(X_1), \ldots, U_n = F(X_n)) & (V_1 = G(Y_1), \ldots, V_n = G(Y_n)) \end{array}$$

The quantile transform

What is the “best” transformation of the data in that context ?

The quantile transform theorem

- when F is arbitrary, if U is a uniformly distributed random variable on $(0, 1)$, $X \stackrel{d}{=} F^{-1}(U)$;
- whenever F is continuous, the random variable $U = F(X)$ is uniformly distributed on $(0, 1)$.

$\longrightarrow$ use the invariance property of the quantile transform to construct a pseudo-sample (U_i, V_i) with a *prescribed uniform* marginal distribution.

$$\begin{array}{cc} (X_1, \dots, X_n) & (Y_1, \dots, Y_n) \\ \downarrow & \downarrow \\ (U_1 = F(X_1), \dots, U_n = F(X_n)) & (V_1 = G(Y_1), \dots, V_n = G(Y_n)) \end{array}$$

The copula representation

$\longrightarrow$ leads naturally to the copula function:

Sklar's theorem [1959]

For any bivariate cumulative distribution function $F_{X,Y}$ on $\mathbb{R}^2$, with marginal c.d.f. F of X and G of Y, there exists some function $C : [0,1]^2 \to [0,1]$, called the dependence or copula function, such as

$$F_{X,Y}(x,y) = C(F(x), G(y)) \ , \ -\infty \leq x, y \leq +\infty.$$

If F and G are continuous, this representation is unique with respect to (F, G). The copula function C is itself a c.d.f. on $[0,1]^2$ with uniform marginals.

$\longrightarrow$ captures the dependence structure of the vector (X, Y), irrespectively of the marginals.
$\longrightarrow$ allows to deal with the randomness of the dependence structure and the randomness of the marginals *separately*.

The copula representation

$\longrightarrow$ leads naturally to the copula function:

Sklar's theorem [1959]

For any bivariate cumulative distribution function $F_{X,Y}$ on $\mathbb{R}^2$, with marginal c.d.f. F of X and G of Y, there exists some function $C : [0,1]^2 \to [0,1]$, called the dependence or copula function, such as

$$F_{X,Y}(x,y) = C(F(x), G(y)) \ , \ -\infty \leq x, y \leq +\infty.$$

If F and G are continuous, this representation is unique with respect to (F, G). The copula function C is itself a c.d.f. on $[0,1]^2$ with uniform marginals.

$\longrightarrow$ captures the dependence structure of the vector (X, Y), irrespectively of the marginals.
$\longrightarrow$ allows to deal with the randomness of the dependence structure and the randomness of the marginals *separately*.

The copula representation

$\longrightarrow$ leads naturally to the copula function:

Sklar's theorem [1959]

For any bivariate cumulative distribution function $F_{X,Y}$ on $\mathbb{R}^2$, with marginal c.d.f. F of X and G of Y, there exists some function $C : [0,1]^2 \rightarrow [0,1]$, called the dependence or copula function, such as

$$F_{X,Y}(x,y) = C(F(x), G(y)) \ , \ -\infty \leq x, y \leq +\infty.$$

If F and G are continuous, this representation is unique with respect to (F, G). The copula function C is itself a c.d.f. on $[0,1]^2$ with uniform marginals.

$\longrightarrow$ captures the dependence structure of the vector (X, Y), irrespectively of the marginals.
$\longrightarrow$ allows to deal with the randomness of the dependence structure and the randomness of the marginals *separately*.

A product shaped estimator

Assume that the copula function $C(u, v)$ has a density $c(u, v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ i.e. $c(u, v)$ is the density of the transformed r.v. $(U, V) = (F(X), G(Y))$.

A product form of the conditional density

By differentiating Sklar's formula,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f(x)} = g(y)c(F(x), G(y))$$

A product shaped estimator

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

A product shaped estimator

Assume that the copula function $C(u, v)$ has a density $c(u, v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ i.e. $c(u, v)$ is the density of the transformed r.v. $(U, V) = (F(X), G(Y))$.

A product form of the conditional density

By differentiating Sklar's formula,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f(x)} = g(y)c(F(x), G(y))$$

A product shaped estimator

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

Construction of the estimator - 1

$\rightarrow$ get an estimator of the conditional density by plugging estimators of each quantities.

- density of Y: $g \leftarrow$ kernel estimator $\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)$
- c.d.f. $$\begin{aligned} F(x) &\leftarrow F_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j \leqslant x} \\ G(y) &\leftarrow G_n(y) := \frac{1}{n}\sum_{j=1}^{n} 1_{Y_j \leqslant y} \end{aligned}$$ empirical c.d.f.
- copula density $c(u,v) \leftarrow c_n(u,v)$ a bivariate Parzen-Rosenblatt kernel density (*pseudo*) estimator

$$c_n(u,v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K\left(\frac{u-U_i}{a_n}, \frac{v-V_i}{a_n}\right) \tag{1}$$

with kernel $K(u,v) = K_1(u)K_2(v)$, and bandwidths a_n.

Construction of the estimator - 1

$\rightarrow$ get an estimator of the conditional density by plugging estimators of each quantities.

- density of Y: $g \leftarrow$ kernel estimator $\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0 \left(\frac{y-Y_i}{h_n} \right)$
- c.d.f. $F(x) \leftarrow F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \leqslant x}$, $G(y) \leftarrow G_n(y) := \frac{1}{n} \sum_{j=1}^{n} 1_{Y_j \leqslant y}$ empirical c.d.f.
- copula density $c(u,v) \leftarrow c_n(u,v)$ a bivariate Parzen-Rosenblatt kernel density (*pseudo*) estimator

$$c_n(u,v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K \left(\frac{u-U_i}{a_n}, \frac{v-V_i}{a_n} \right) \quad (1)$$

with kernel $K(u,v) = K_1(u)K_2(v)$, and bandwidths a_n.

Construction of the estimator - 1

$\rightarrow$ get an estimator of the conditional density by plugging estimators of each quantities.

- density of Y: $g \leftarrow$ kernel estimator $\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)$
- c.d.f. $$\begin{aligned} F(x) &\leftarrow F_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j \leqslant x} \\ G(y) &\leftarrow G_n(y) := \frac{1}{n}\sum_{j=1}^{n} 1_{Y_j \leqslant y} \end{aligned}$$ empirical c.d.f.
- copula density $c(u,v) \leftarrow c_n(u,v)$ a bivariate Parzen-Rosenblatt kernel density (*pseudo*) estimator

$$c_n(u,v) := \frac{1}{na_n^2}\sum_{i=1}^{n} K\left(\frac{u-U_i}{a_n}, \frac{v-V_i}{a_n}\right) \tag{1}$$

with kernel $K(u,v) = K_1(u)K_2(v)$, and bandwidths a_n.

Construction of the estimator - 1

$\rightarrow$ get an estimator of the conditional density by plugging estimators of each quantities.

- density of Y: $g \leftarrow$ kernel estimator $\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)$
- c.d.f. $$\begin{aligned} F(x) &\leftarrow F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \leqslant x} \\ G(y) &\leftarrow G_n(y) := \frac{1}{n} \sum_{j=1}^{n} 1_{Y_j \leqslant y} \end{aligned}$$ empirical c.d.f.
- copula density $c(u,v) \leftarrow c_n(u,v)$ a bivariate Parzen-Rosenblatt kernel density (*pseudo*) estimator

$$c_n(u,v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K\left(\frac{u-U_i}{a_n}, \frac{v-V_i}{a_n}\right) \tag{1}$$

with kernel $K(u,v) = K_1(u)K_2(v)$, and bandwidths a_n.

Construction of the estimator - 2

But, F and G are unknown: the random variables $(U_i = F(X_i), V_i = G(Y_i))$ are not observable.
$\Rightarrow c_n$: is not a true statistic.
$\rightarrow$ approximate the pseudo-sample $(U_i, V_i), i = 1, \ldots, n$ by its empirical counterpart $(F_n(X_i), G_n(Y_i)), i = 1, \ldots, n$.

A genuine estimator of $c(u, v)$

$$\hat{c}_n(u, v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K_1\left(\frac{u - F_n(X_i)}{a_n}\right) K_2\left(\frac{v - G_n(Y_i)}{a_n}\right).$$

Construction of the estimator - 2

But, F and G are unknown: the random variables $(U_i = F(X_i), V_i = G(Y_i))$ are not observable.
$\Rightarrow c_n$: is not a true statistic.
$\rightarrow$ approximate the pseudo-sample $(U_i, V_i), i = 1, \ldots, n$ by its empirical counterpart $(F_n(X_i), G_n(Y_i)), i = 1, \ldots, n$.

A genuine estimator of $c(u, v)$

$$\hat{c}_n(u, v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K_1\left(\frac{u - F_n(X_i)}{a_n}\right) K_2\left(\frac{v - G_n(Y_i)}{a_n}\right).$$

Construction of the estimator - 2

But, F and G are unknown: the random variables $(U_i = F(X_i), V_i = G(Y_i))$ are not observable.
$\Rightarrow c_n$: is not a true statistic.
$\rightarrow$ approximate the pseudo-sample $(U_i, V_i), i = 1, \dots, n$ by its empirical counterpart $(F_n(X_i), G_n(Y_i)), i = 1, \dots, n$.

A genuine estimator of $c(u, v)$

$$\hat{c}_n(u, v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K_1\left(\frac{u - F_n(X_i)}{a_n}\right) K_2\left(\frac{v - G_n(Y_i)}{a_n}\right).$$

Construction of the estimator - 2

But, F and G are unknown: the random variables $(U_i = F(X_i), V_i = G(Y_i))$ are not observable.
$\Rightarrow c_n$: is not a true statistic.
$\rightarrow$ approximate the pseudo-sample $(U_i, V_i), i = 1, \ldots, n$ by its empirical counterpart $(F_n(X_i), G_n(Y_i)), i = 1, \ldots, n$.

A genuine estimator of $c(u, v)$

$$\hat{c}_n(u, v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K_1\left(\frac{u - F_n(X_i)}{a_n}\right) K_2\left(\frac{v - G_n(Y_i)}{a_n}\right).$$

Construction of the estimator - 2

But, F and G are unknown: the random variables $(U_i = F(X_i), V_i = G(Y_i))$ are not observable.
$\Rightarrow c_n$: is not a true statistic.
$\rightarrow$ approximate the pseudo-sample $(U_i, V_i), i = 1, \ldots, n$ by its empirical counterpart $(F_n(X_i), G_n(Y_i)), i = 1, \ldots, n$.

A genuine estimator of $c(u, v)$

$$\hat{c}_n(u, v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K_1\left(\frac{u - F_n(X_i)}{a_n}\right) K_2\left(\frac{v - G_n(Y_i)}{a_n}\right).$$

The quantile-copula estimator

Recollecting all elements, we get,

The quantile-copula estimator

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

that is to say,

$$\hat{f}_n(y|x) := \left[\frac{1}{nh_n}\sum_{i=1}^{n} K_0\left(\frac{y - Y_i}{h_n}\right)\right] \cdot \left[\frac{1}{na_n^2}\sum_{i=1}^{n} K_1\left(\frac{F_n(x) - F_n(X_i)}{a_n}\right) K_2\left(\frac{G_n(y) - G_n(Y_i)}{a_n}\right)\right]$$

Outline

Hypothesis

Assumptions on the densities

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 differentiable with continuous bounded second derivatives on their support.

Assumptions on the kernels

(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: $m_0(K) = 1$, $m_1(K) = 0$ and $m_2(K) < +\infty$, and the same for K_0.

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

$\rightarrow$ classical regularity assumptions in nonparametric literature.

Hypothesis

Assumptions on the densities

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 differentiable with continuous bounded second derivatives on their support.

Assumptions on the kernels

(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: $m_0(K) = 1$, $m_1(K) = 0$ and $m_2(K) < +\infty$, and the same for K_0.

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

$\rightarrow$ classical regularity assumptions in nonparametric literature.

Asymptotic results - 1

Under the above regularity assumptions, with $h_n \to 0$, $a_n \to 0$,

Pointwise Consistency

- weak consistency $h_n \simeq n^{-1/5}$, $a_n \simeq n^{-1/6}$ entail

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

- strong consistency $h_n \simeq (\ln\ln n/n)^{1/5}$ and $a_n \simeq (\ln\ln n/n)^{1/6}$

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

- asymptotic normality $h_n \simeq n^{-1/5}$ and $a_n = o(n^{-1/6})$ entail

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

Asymptotic results - 2

Uniform Consistency

Under the above regularity assumptions, with $h_n \to 0$, $a_n \to 0$, for x in the interior of the support of f and $[a, b]$ included in the interior of the support of g,

- weak consistency $h_n \simeq (\ln n/n)^{1/5}$, $a_n \simeq (\ln n/n)^{1/6}$ entail

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

- strong consistency $h_n \simeq (\ln n/n)^{1/5}$, $a_n \simeq (\ln n/n)^{1/6}$ entail

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

Asymptotic Mean square error

Asymptotic Bias and Variance for the quantile-copula estimator

- Bias:

$$E(\hat{f}_n(y|x)) - f(y|x) = g(y)m_2(K).\nabla^2 c(F(x), G(y))\frac{a_n^2}{2} + o(a_n^2)$$

with $m_2(K) = (m_2(K_1), m_2(K_2))$, $\nabla^2 c(u,v) = (\frac{\partial^2 c(u,v)}{\partial u^2}, \frac{\partial^2 c(u,v)}{\partial v^2})$.

- Variance:

$$Var(\hat{f}(y|x)) = 1/(na_n^2)g(y)f(y|x)||K||_2^2 + o(1/(na_n^2)).$$

Sketch of the proofs

Decomposition diagram

$$
\begin{array}{ccccc}
\hat{g}(y)\hat{c}_n(F_n(x), G_n(y)) & & & & \\
\downarrow & & & & \\
g(y)\hat{c}_n(F_n(x), G_n(y)) & \rightarrow & g(y)\hat{c}_n(F(x), G(y)) & \rightarrow & g(y)c_n(F(x), G(y)) \\
 & & & & \downarrow \\
 & & & & g(y)c(F(x), G(y))
\end{array}
$$

$\downarrow$: consistency results of the kernel density estimators

$\rightarrow$: two approximation lemmas

1. $\hat{c}_n$ from $(F_n(x), F_n(y)) \rightarrow (F(x), G(y))$
2. $\hat{c}_n \rightarrow c_n$.

Tools: results for the K-S statistics $||F - F_n||_\infty$ and $||G - G_n||_\infty$.
$\rightarrow$ Heuristic: rate of convergence of density estimators $<$ rate of approximation of the K-S Statistic.

Outline

Theoretical asymptotic comparison - 1

Competitor: e.g. Local Polynomial estimator, $\hat{f}_n^{(LP)}(y|x) := \hat{\theta}_0$ with

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

where $\hat{\theta}_{xy} := (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_r)$ is the value of θ which minimizes $R(\theta, x, y)$.

Comparative Bias

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

$$B_{QC} = g(y) m_2(K) . \nabla_2 c(F(x), G(y)) \frac{a_n^2}{2} + o(a_n^2)$$

Theoretical asymptotic comparison - 2

Asymptotic bias comparison

- All estimators have bias of the same order $\approx h^2 \approx n^{-1/3}$;
- Distribution dependent terms:
  - difficult to compare
  - sometimes less unknown terms for the quantile-copula estimator
- c of compact support : the "classical" kernel method to estimate the copula density induces bias on the boundaries of $[0,1]^2$ $\longrightarrow$ techniques to reduce the bias of the kernel estimator on the edges (boundary kernels, beta kernels, reflection and transformation methods,...)

Theoretical asymptotic comparison - 2

Asymptotic bias comparison

- All estimators have bias of the same order $\approx h^2 \approx n^{-1/3}$;
- Distribution dependent terms:
  - difficult to compare
  - sometimes less unknown terms for the quantile-copula estimator
- c of compact support : the "classical" kernel method to estimate the copula density induces bias on the boundaries of $[0, 1]^2$
  $\rightarrow$ techniques to reduce the bias of the kernel estimator on the edges (boundary kernels, beta kernels, reflection and transformation methods,...)

Theoretical asymptotic comparison - 2

Asymptotic bias comparison

- All estimators have bias of the same order $\approx h^2 \approx n^{-1/3}$;
- Distribution dependent terms:
  - difficult to compare
  - sometimes less unknown terms for the quantile-copula estimator
- c of compact support : the "classical" kernel method to estimate the copula density induces bias on the boundaries of $[0, 1]^2$
  $\rightarrow$ techniques to reduce the bias of the kernel estimator on the edges (boundary kernels, beta kernels, reflection and transformation methods,...)

Theoretical asymptotic comparison - 3

Asymptotic Variance comparison

Main terms in the asymptotic variance:

- Ratio shaped estimators: $Var(LP) := \frac{f(y|x)}{f(x)} \rightarrow$ explosive variance for small value of the density $f(x)$, e.g. in the tail of the distribution of X.
- Quantile-copula estimator: $Var(QC) := g(y)f(y|x) \rightarrow$ does not suffer from the unstable nature of competitors.
- Asymptotic relative efficiency: ratio of variances

$$\frac{Var(QC)}{Var(LP)} := f(x)g(y)$$

$\rightarrow$ the QC has a lower asymptotic variance for a large amount of x,y values.

Theoretical asymptotic comparison - 3

Asymptotic Variance comparison

Main terms in the asymptotic variance:

- Ratio shaped estimators: $Var(LP) := \frac{f(y|x)}{f(x)} \rightarrow$ explosive variance for small value of the density $f(x)$, e.g. in the tail of the distribution of X.
- Quantile-copula estimator: $Var(QC) := g(y)f(y|x) \rightarrow$ does not suffer from the unstable nature of competitors.
- Asymptotic relative efficiency: ratio of variances

$$\frac{Var(QC)}{Var(LP)} := f(x)g(y)$$

$\rightarrow$ the QC has a lower asymptotic variance for a large amount of x,y values.

Theoretical asymptotic comparison - 3

Asymptotic Variance comparison

Main terms in the asymptotic variance:

- Ratio shaped estimators: $Var(LP) := \frac{f(y|x)}{f(x)} \rightarrow$ explosive variance for small value of the density $f(x)$, e.g. in the tail of the distribution of X.
- Quantile-copula estimator: $Var(QC) := g(y)f(y|x) \rightarrow$ does not suffer from the unstable nature of competitors.
- Asymptotic relative efficiency: ratio of variances

$$\frac{Var(QC)}{Var(LP)} := f(x)g(y)$$

$\rightarrow$ the QC has a lower asymptotic variance for a large amount of x,y values.

Theoretical asymptotic comparison - 3

Asymptotic Variance comparison

Main terms in the asymptotic variance:

- Ratio shaped estimators: $Var(LP) := \frac{f(y|x)}{f(x)} \rightarrow$ explosive variance for small value of the density $f(x)$, e.g. in the tail of the distribution of X.
- Quantile-copula estimator: $Var(QC) := g(y)f(y|x) \rightarrow$ does not suffer from the unstable nature of competitors.
- Asymptotic relative efficiency: ratio of variances

$$\frac{Var(QC)}{Var(LP)} := f(x)g(y)$$

$\rightarrow$ the QC has a lower asymptotic variance for a large amount of x,y values.

Finite sample simulation

Model

Sample of $n = 100$ i.i.d. variables (X_i, Y_i), from the following model:

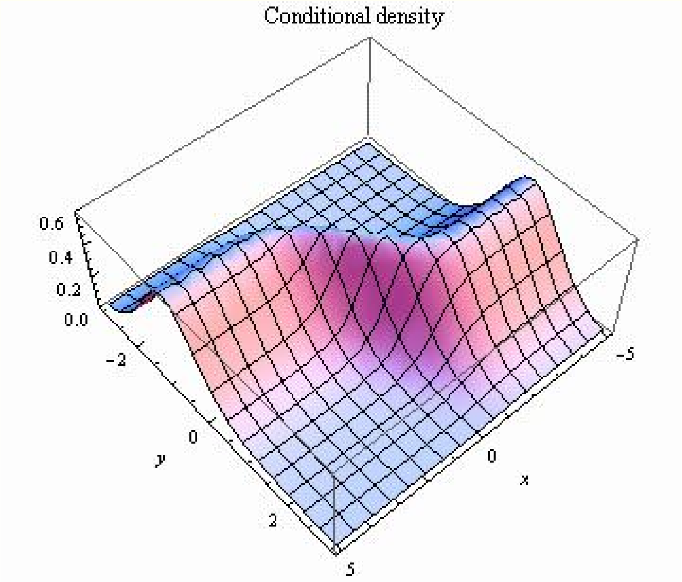
- X, Y is marginally distributed as $\mathcal{N}(0, 1)$
- X, Y is linked via Frank Copula .

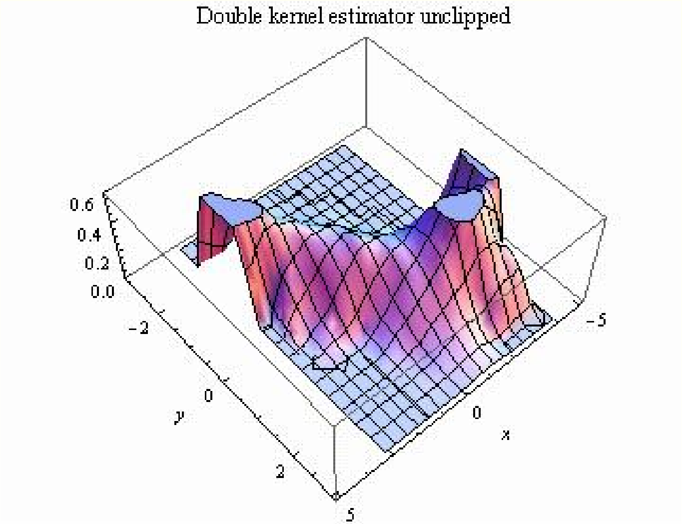
$$C(u, v, \theta) = \frac{\ln[(\theta + \theta^{u+v} - \theta^u - \theta^v)/(\theta - 1)]}{\ln \theta}$$

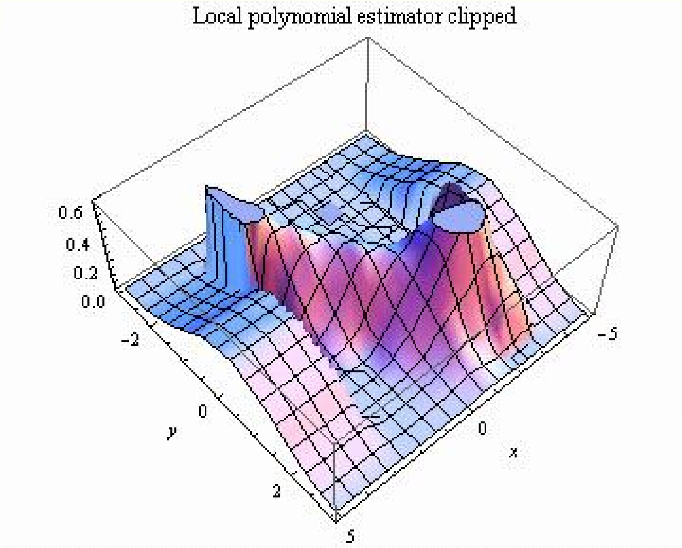
with parameter $\theta = 100$.

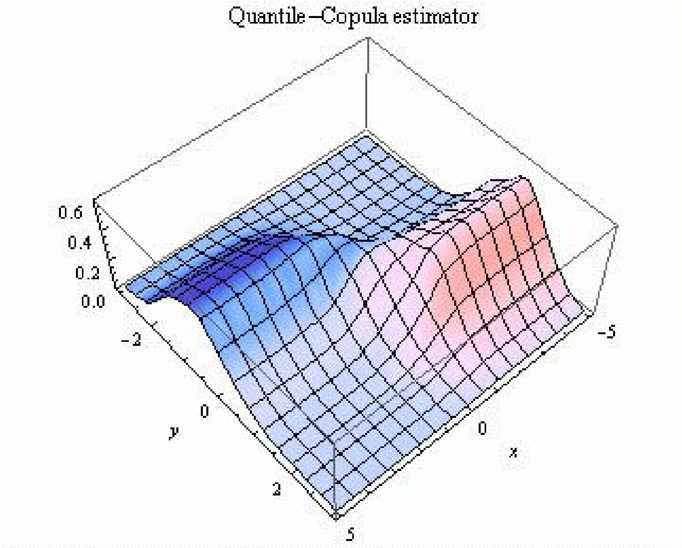
Practical implementation:

- Beta kernels for copula estimator, Epanechnikov for other.
- simple Rule-of-thumb method for the bandwidths.


Conditional density
0.6
0.4
0.2
0.0
-2
0
y
2
5
-5
0
x


Double kernel estimator unclipped
0.6
0.4
0.2
0.0
-2
0
y
2
5
0
x
-5


Local polynomial estimator clipped
0.6
0.4
0.2
0.0
-2
0
y
2
5
-5
0
x


Quantile –Copula estimator
0.6
0.4
0.2
0.0
-2
0
y
2
5
-5
0
x

Outline

Application to prediction - definitions

Point predictors: Conditional mode predictor

Definition of the mode: $\theta(x) := \arg\sup_y f(y|x)$
$\rightarrow$ plug in predictor : $\hat{\theta}(x) := \arg\sup_y \hat{f}_n(y|x)$

Set predictors: Level sets

Predictive set $\mathcal{C}_\alpha(x)$ such as $P(Y \in \mathcal{C}_\alpha(x)|X = x) = \alpha$
$\rightarrow$ Level set or Highest density region $\mathcal{C}_\alpha(x) := \{y : f(y|x) \geq f_\alpha\}$ with f_α the largest value such that the prediction set has coverage probability α.
$\rightarrow$ plug-in level set: $\mathcal{C}_{\alpha,n}(x) := \{y : \hat{f}_n(y|x) \geq \hat{f}_\alpha\}$ where $\hat{f}_\alpha$ is an estimate of f_α.

Application to prediction - definitions

Point predictors: Conditional mode predictor

Definition of the mode: $\theta(x) := \arg\sup_y f(y|x)$

$\rightarrow$ plug in predictor : $\hat{\theta}(x) := \arg\sup_y \hat{f}_n(y|x)$

Set predictors: Level sets

Predictive set $\mathcal{C}_\alpha(x)$ such as $P(Y \in \mathcal{C}_\alpha(x)|X = x) = \alpha$

$\rightarrow$ Level set or Highest density region $\mathcal{C}_\alpha(x) := \{y : f(y|x) \geq f_\alpha\}$ with f_α the largest value such that the prediction set has coverage probability α.

$\rightarrow$ plug-in level set: $\mathcal{C}_{\alpha,n}(x) := \{y : \hat{f}_n(y|x) \geq \hat{f}_\alpha\}$ where $\hat{f}_\alpha$ is an estimate of f_α.

Application to prediction - definitions

Point predictors: Conditional mode predictor

Definition of the mode: $\theta(x) := \arg\sup_y f(y|x)$

$\rightarrow$ plug in predictor : $\hat{\theta}(x) := \arg\sup_y \hat{f}_n(y|x)$

Set predictors: Level sets

Predictive set $\mathcal{C}_\alpha(x)$ such as $P(Y \in \mathcal{C}_\alpha(x)|X = x) = \alpha$

$\rightarrow$ Level set or Highest density region $\mathcal{C}_\alpha(x) := \{y : f(y|x) \geq f_\alpha\}$ with f_α the largest value such that the prediction set has coverage probability α.

$\rightarrow$ plug-in level set: $\mathcal{C}_{\alpha,n}(x) := \{y : \hat{f}_n(y|x) \geq \hat{f}_\alpha\}$ where $\hat{f}_\alpha$ is an estimate of f_α.

Application to prediction - definitions

Point predictors: Conditional mode predictor

Definition of the mode: $\theta(x) := \arg\sup_y f(y|x)$
$\rightarrow$ plug in predictor : $\hat{\theta}(x) := \arg\sup_y \hat{f}_n(y|x)$

Set predictors: Level sets

Predictive set $\mathcal{C}_\alpha(x)$ such as $P(Y \in \mathcal{C}_\alpha(x)|X = x) = \alpha$
$\rightarrow$ Level set or Highest density region $\mathcal{C}_\alpha(x) := \{y : f(y|x) \geq f_\alpha\}$ with f_α the largest value such that the prediction set has coverage probability α.
$\rightarrow$ plug-in level set: $\mathcal{C}_{\alpha,n}(x) := \{y : \hat{f}_n(y|x) \geq \hat{f}_\alpha\}$ where $\hat{f}_\alpha$ is an estimate of f_α.

Application to prediction - definitions

Point predictors: Conditional mode predictor

Definition of the mode: $\theta(x) := \arg\sup_y f(y|x)$
$\rightarrow$ plug in predictor : $\hat{\theta}(x) := \arg\sup_y \hat{f}_n(y|x)$

Set predictors: Level sets

Predictive set $\mathcal{C}_\alpha(x)$ such as $P(Y \in \mathcal{C}_\alpha(x)|X = x) = \alpha$
$\rightarrow$ Level set or Highest density region $\mathcal{C}_\alpha(x) := \{y : f(y|x) \geq f_\alpha\}$ with f_α the largest value such that the prediction set has coverage probability α.
$\rightarrow$ plug-in level set: $\mathcal{C}_{\alpha,n}(x) := \{y : \hat{f}_n(y|x) \geq \hat{f}_\alpha\}$ where $\hat{f}_\alpha$ is an estimate of f_α.

Application to prediction - results

Point predictors: Conditional mode predictor

Under regularity conditions, uniform convergence on a compact set of the conditional density estimator entails that

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

Set predictors: Level sets

Under regularity conditions, uniform convergence on a compact set of the conditional density estimator entails that

$$\lambda(\Delta(\mathcal{C}_{\alpha,n}(x), \mathcal{C}_{\alpha}(x))) \stackrel{a.s.}{\to} 0$$

where $\Delta(.,.)$ stands for the symmetric difference, and λ for Lebesgue measure.

On the efficiency estimation of the empirical margins

Deficiency of the empirical distribution functions

- the order statistics $X_{1,n} < \ldots < X_{n,n}$ is complete sufficient for estimating F with a density f.
  $\rightarrow F_n$ is the UMVU estimator of F.
- its smoothed version $\hat{F}(x) = n^{-1} \sum_{i=1}^{n} L\left(\frac{X_i - x}{b_n}\right)$ where b_n bandwidth and $L(x) = \int_{-\infty}^{x} l(t)dt$, with l density kernel, is such that

$$\left| E(\hat{F}(x) - F(x))^2 - E(F_n(x) - F(x))^2 + 2h/nF'(x) \int tl(t)L(t)dt \right|$$
$$\leq h^4 AC^2 + O(h^2/n)$$

  $\rightarrow F_n$ is deficient w.r.t $\hat{F}$.

On the efficiency estimation of the empirical margins

Deficiency of the empirical distribution functions

- the order statistics $X_{1,n} < \ldots < X_{n,n}$ is complete sufficient for estimating F with a density f.
  $\rightarrow F_n$ is the UMVU estimator of F.
- its smoothed version $\hat{F}(x) = n^{-1} \sum_{i=1}^{n} L\left(\frac{X_i - x}{b_n}\right)$ where b_n bandwidth and $L(x) = \int_{-\infty}^{x} l(t)dt$, with l density kernel, is such that

$$\left| E(\hat{F}(x) - F(x))^2 - E(F_n(x) - F(x))^2 + 2h/nF'(x) \int tl(t)L(t)dt \right|$$
$$\leq h^4 AC^2 + O(h^2/n)$$

$\rightarrow F_n$ is deficient w.r.t $\hat{F}$.

Implication for the quantile copula estimator

The doubly smoothed quantile copula conditional density estimator

$\rightarrow$ replace F_n and G_n by $\hat{F}$ and $\hat{G}$

- beneficial for small samples
- graphically more appealing: less wiggly behaviour

Consequence for local averaging

With smooth margin estimators $\hat{F}$ and $\hat{G}$,

$$\text{or } \hat{F}(x) - \hat{F}(X_i) \approx \hat{f}(X_i)(x - X_i) \quad (2)$$

$$\hat{F}(X_i) - \hat{F}(x) \approx \hat{f}(x)(X_i - x) \quad (3)$$

Implication for the quantile copula estimator

The doubly smoothed quantile copula conditional density estimator

$\rightarrow$ replace F_n and G_n by $\hat{F}$ and $\hat{G}$

- beneficial for small samples
- graphically more appealing: less wiggly behaviour

Consequence for local averaging

With smooth margin estimators $\hat{F}$ and $\hat{G}$,

$$\text{or } \hat{F}(x) - \hat{F}(X_i) \approx \hat{f}(X_i)(x - X_i) \tag{2}$$

$$\hat{F}(X_i) - \hat{F}(x) \approx \hat{f}(x)(X_i - x) \tag{3}$$

Connection with the variable bandwidth kernel estimators

Connection with the variable bandwidth kernel estimators

Therefore, the copula density part of the estimator writes

$$\begin{aligned}\hat{c}_n(\hat{F}_n(x), \hat{G}_n(y)) &= (na_nb_n)^{-1}\sum_{i=1}^{n} K_1\left(\frac{\hat{F}_n(X_i)-\hat{F}_n(x)}{a_n}\right) K_2\left(\dots\right)\\ &\approx (na_nb_n)^{-1}\sum_{i=1}^{n} K_1\left(\frac{X_i-x}{a_n/\hat{f}(X_i)}\right) K_2\left(\frac{Y_i-y}{b_n/\hat{g}(Y_i)}\right)\end{aligned}$$

with approximation (2), and

$$\approx (na_nb_n)^{-1}\sum_{i=1}^{n} K_1\left(\frac{X_i-x}{a_n/\hat{f}(x)}\right) K_2\left(\frac{Y_i-y}{b_n/\hat{g}(y)}\right)$$

with approximation (3).

Connection with the variable bandwidth kernel estimators

Connection with the variable bandwidth kernel estimators

$\longrightarrow$ the copula density estimator with smoothed margin estimates is like a kernel estimator with an adaptive local bandwidth

- $a_n/\hat{f}(X_i)$: sample smoothing bandwidth
- $a_n/\hat{f}(x)$: balloon smoothing bandwidth

$\longrightarrow$ Work-in-Progress: use the CDF transform approach to improve on kernel density estimation methods

The beta kernel is an unbiased estimate of the uniform density

- Parametric start: set a family of CDF $\mathcal{F} = \{F(,\theta)\}$.
- Estimate $\theta \leftarrow \hat{\theta}$
- Transform $Z_i = F(X_i, \hat{\theta})$ to get a semiparametric estimator with rate nearly parametric on $\mathcal{F}$

Connection with the variable bandwidth kernel estimators

Connection with the variable bandwidth kernel estimators

$\longrightarrow$ the copula density estimator with smoothed margin estimates is like a kernel estimator with an adaptive local bandwidth

- $a_n/\hat{f}(X_i)$: sample smoothing bandwidth
- $a_n/\hat{f}(x)$: balloon smoothing bandwidth

$\longrightarrow$ Work-in-Progress: use the CDF transform approach to improve on kernel density estimation methods

The beta kernel is an unbiased estimate of the uniform density

- Parametric start: set a family of CDF $\mathcal{F} = \{F(,\theta)\}$.
- Estimate $\theta \leftarrow \hat{\theta}$
- Transform $Z_i = F(X_i, \hat{\theta})$ to get a semiparametric estimator with rate nearly parametric on $\mathcal{F}$

Perspectives and work-in-progress - 1

Prediction for extreme values. Application in Insurance and Finance, with S. Loisel and E. Masiello, ISFA, Lyon 1.

Aim: make inference on Y for large values of X.
$\longrightarrow$ design an extreme-specific estimator by combining

- estimators based on Extreme value theory or the Peak-Over Threshold approach for F_n;
- the wavelet estimator $\hat{c}_n$ of C. Genest, E. Masiello, K. Tribouley (2009).

on a problem related to insurance and finance.

Perspectives and work-in-progress - 2

A semi-parametric multivariate extension via a Single Index model, with O. Lopez, Paris 6.

The issues when $X \in \mathbb{R}^d$, $d > 1$:

- the product shape of the estimator breaks down;
- curse of dimensionality.

$\longrightarrow$ use a dimension-reduction hypothesis such as a Single-Index-Model:

$$\exists \theta_0 \in \mathbb{R}^d, \quad f_{Y|X} = f_{Y|\theta_0^T.X}$$

$\longrightarrow$ use the quantile copula estimator on the auxiliary data $Z = \theta^T.X$ to estimate θ_0 by maximum pseudo-likelihood method.

Perspectives and work-in-progress - 3

Conditional Cumulative distribution function, conditional quantile, and regression estimation.

$\longrightarrow$ use the same quantile-copula approach to estimate

$$F_{Y|X}(x,y) = \int_{-\infty}^{y} g(t)c(F(x),G(t))dt = E[1_{Y\leq y}c(F(x),G(Y))]$$

by

$$\hat{F}_{Y|X}(x,y) := \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i\leq y}\hat{c}_n(F_n(x),G_n(Y_i))$$

$\longrightarrow$ get conditional quantiles (VaR) and confidence intervals.
$\longrightarrow$ proceed similarly for the regression.

$$\hat{r}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i\hat{c}_n(F_n(x),G_n(Y_i))$$

Perspectives and work-in-progress - 4

Extension to time series by coupling arguments for Markovian models. Application to tests of adequacy.

$\rightarrow$ Estimate $\pi(x,y)dy = P(\zeta_{i+1} \in dy | \zeta_i = x)$, the transition density of a Markov chain (ζ_t), $t \in \mathbb{N}$.
$\rightarrow$ Extend the results from the i.i.d. case to the mixing framework, by coupling and blocks arguments, and the substitution

$$(X_i, Y_i) \longleftarrow (\zeta_i, \zeta_{i+1})$$

$\rightarrow$ use the estimate as a proxy to test the adequacy of parametric models.

Bibliography

Reference

- O. P. Faugeras. (2009), A quantile-copula approach to conditional density estimation. *Journal of Multivariate Analysis*, in press.
- O. P. Faugeras. (2009) Modal regression and prediction via the quantile-copula estimator. *Submitted*