Thèse de Doctorat de l'Université Pierre et Marie Curie

Contributions à la Prévision Statistique

Olivier P. Faugeras

Université Pierre et Mare Curie - Paris VI

Laboratoire de Statistique Théorique et Appliquée

28/11/2008

Outline

Part I : Parametric Statistical Prediction for a Stochastic Process.

Observe : $X_0, \ldots, X_T$ of a stochastic process (X_t) with law P_θ.
Predict : X_{T+h} a future value.

Part II : A nonparametric quantile-copula approach to conditional density estimation. Applications to prediction.

Observe : $(X_i, Y_i)_{i=1,\ldots,n}$ independent identically distributed.
Predict : Y, given that $X = x$.

Part I : Parametric Statistical Prediction for a Stochastic Process.

Outline

The Statistical Prediction Problem (1)

Let $\mathbb{X} = \{X_t, t \in \mathbb{Z}\}$ a real-valued, square integrable, stochastic process, with distribution P_θ, θ a parameter.
Observed data: $(X_0, \ldots, X_T) := X_0^T$
Aim : Forecast $Y := g(X_{T+h})$ by a function $f(X_0^T) = \hat{Y}$
Criteria : Error $\mathbb{L}^2$

Lemma : Decomposition of the prediction error

$$E_\theta(Y - \hat{Y})^2 = E_\theta(Y - E_\theta(Y|X_0^T))^2 + E_\theta(E_\theta(Y|X_0^T) - f(X_0^T))^2$$

The prediction error splits between a probabilistic prediction error term and a statistical prediction error term.

The error is thus minimised by choosing the conditional expectation as a predictor
$f(X_0^T) = E_\theta(Y|X_0^T) := Y^*$

The Statistical Prediction Problem (1)

Let $\mathbb{X} = \{X_t, t \in \mathbb{Z}\}$ a real-valued, square integrable, stochastic process, with distribution P_θ, θ a parameter.
Observed data: $(X_0, \ldots, X_T) := X_0^T$
Aim : Forecast $Y := g(X_{T+h})$ by a function $f(X_0^T) = \hat{Y}$
Criteria : Error $\mathbb{L}^2$

Lemma : Decomposition of the prediction error

$$E_\theta(Y - \hat{Y})^2 = E_\theta(Y - E_\theta(Y|X_0^T))^2 + E_\theta(E_\theta(Y|X_0^T) - f(X_0^T))^2$$

The prediction error splits between a probabilistic prediction error term and a statistical prediction error term.

The error is thus minimised by choosing the conditional expectation as a predictor
$f(X_0^T) = E_\theta(Y|X_0^T) := Y^*$

The Statistical Prediction Problem (1)

Let $\mathbb{X} = \{X_t, t \in \mathbb{Z}\}$ a real-valued, square integrable, stochastic process, with distribution P_θ, θ a parameter.
Observed data: $(X_0, \ldots, X_T) := X_0^T$
Aim : Forecast $Y := g(X_{T+h})$ by a function $f(X_0^T) = \hat{Y}$
Criteria : Error $\mathbb{L}^2$

Lemma : Decomposition of the prediction error

$$E_\theta(Y - \hat{Y})^2 = E_\theta(Y - E_\theta(Y|X_0^T))^2 + E_\theta(E_\theta(Y|X_0^T) - f(X_0^T))^2$$

The prediction error splits between a probabilistic prediction error term and a statistical prediction error term.

The error is thus minimised by choosing the conditional expectation as a predictor
$f(X_0^T) = E_\theta(Y|X_0^T) := Y^*$

The Statistical prediction problem (2)

Definition : the Probabilistic predictor

The Bayesian or Probabilistic predictor is defined as the random variable $Y^* := E_\theta(Y|X_0^T) := r_\theta(X_0^T)$

But : θ is unknown $\rightarrow$ to be estimated by $\hat{\theta}_T$ on X_0^T

Definition : The Statistical predictor

We build the plug-in Statistical predictor : $\hat{Y} := r_{\hat{\theta}_T}(X_0^T)$

2 mixed problems : on the same data

1. a probabilistic calculation problem : X_0^T as argument of r_θ
2. a statistical estimation problem : X_0^T as data to estimate θ by $\hat{\theta}_T$

$\rightarrow$ behaviour difficult to study.

The Statistical prediction problem (2)

Definition : the Probabilistic predictor

The Bayesian or Probabilistic predictor is defined as the random variable $Y^* := E_\theta(Y|X_0^T) := r_\theta(X_0^T)$

But : θ is unknown $\rightarrow$ to be estimated by $\hat{\theta}_T$ on X_0^T

Definition : The Statistical predictor

We build the plug-in Statistical predictor : $\hat{Y} := r_{\hat{\theta}_T}(X_0^T)$

2 mixed problems : on the same data

1. a probabilistic calculation problem : X_0^T as argument of r_θ
2. a statistical estimation problem : X_0^T as data to estimate θ by $\hat{\theta}_T$

$\rightarrow$ behaviour difficult to study.

The Statistical prediction problem (2)

Definition : the Probabilistic predictor

The Bayesian or Probabilistic predictor is defined as the random variable $Y^* := E_\theta(Y|X_0^T) := r_\theta(X_0^T)$

But : θ is unknown $\rightarrow$ to be estimated by $\hat{\theta}_T$ on X_0^T

Definition : The Statistical predictor

We build the plug-in Statistical predictor : $\hat{Y} := r_{\hat{\theta}_T}(X_0^T)$

2 mixed problems : on the same data

1. a probabilistic calculation problem : X_0^T as argument of r_θ
2. a statistical estimation problem : X_0^T as data to estimate θ by $\hat{\theta}_T$

$\rightarrow$ behaviour difficult to study.

The Statistical prediction problem (2)

Definition : the Probabilistic predictor

The Bayesian or Probabilistic predictor is defined as the random variable $Y^* := E_\theta(Y|X_0^T) := r_\theta(X_0^T)$

But : θ is unknown $\rightarrow$ to be estimated by $\hat{\theta}_T$ on X_0^T

Definition : The Statistical predictor

We build the plug-in Statistical predictor : $\hat{Y} := r_{\hat{\theta}_T}(X_0^T)$

2 mixed problems : on the same data

1. a probabilistic calculation problem : X_0^T as argument of r_θ
2. a statistical estimation problem : X_0^T as data to estimate θ by $\hat{\theta}_T$

$\rightarrow$ behaviour difficult to study.

The Statistical prediction problem (2)

Definition : the Probabilistic predictor

The Bayesian or Probabilistic predictor is defined as the random variable $Y^* := E_\theta(Y|X_0^T) := r_\theta(X_0^T)$

But : θ is unknown $\longrightarrow$ to be estimated by $\hat{\theta}_T$ on X_0^T

Definition : The Statistical predictor

We build the plug-in Statistical predictor : $\hat{Y} := r_{\hat{\theta}_T}(X_0^T)$

2 mixed problems : on the same data

1. a probabilistic calculation problem : X_0^T as argument of r_θ
2. a statistical estimation problem : X_0^T as data to estimate θ by $\hat{\theta}_T$

$\longrightarrow$ behaviour difficult to study.

The Statistical prediction problem (2)

Definition : the Probabilistic predictor

The Bayesian or Probabilistic predictor is defined as the random variable $Y^* := E_\theta(Y|X_0^T) := r_\theta(X_0^T)$

But : θ is unknown $\longrightarrow$ to be estimated by $\hat{\theta}_T$ on X_0^T

Definition : The Statistical predictor

We build the plug-in Statistical predictor : $\hat{Y} := r_{\hat{\theta}_T}(X_0^T)$

2 mixed problems : on the same data

1. a probabilistic calculation problem : X_0^T as argument of r_θ
2. a statistical estimation problem : X_0^T as data to estimate θ by $\hat{\theta}_T$

$\longrightarrow$ behaviour difficult to study.

Prevision versus Regression

Régression

1. estimation step : on the data $D_n := \{(X_i, Y_i), i = 0, \ldots, n\}$, estimate $r(x) = E[Y|X = x]$ by $\hat{r}(x, D_n)$
2. prediction step : for a new (X, Y), predict Y by $\hat{r}(X, D_n)$

if (X, Y) were independent of D_n, then $E[Y|X, D_n] = E[Y|X]$ and

$$E_\theta[r(X) - \hat{r}(X, D_n)]^2 = \int E_\theta\left[(r(X) - \hat{r}(X, D_n))^2 | X = x\right] dP_X(x)$$
$$= \int E_\theta\left[(r(x) - \hat{r}(x, D_n))^2\right] dP_X(x)$$

$\rightarrow$ The Prediction error is the same as the MISE regression error.

Prediction

For a Markov process, $(X_i, Y_i) = (X_i, X_{i+1})$ et $(X, Y) = (X_T, X_{T+1})$
$\Rightarrow D_n$ not independent of X

Prevision versus Regression

Régression

1. estimation step : on the data $D_n := \{(X_i, Y_i), i = 0, \ldots, n\}$, estimate $r(x) = E[Y|X = x]$ by $\hat{r}(x, D_n)$
2. prediction step : for a new (X, Y), predict Y by $\hat{r}(X, D_n)$

if (X, Y) were independent of D_n, then $E[Y|X, D_n] = E[Y|X]$ and

$$E_\theta[r(X) - \hat{r}(X, D_n)]^2 = \int E_\theta\left[(r(X) - \hat{r}(X, D_n))^2 | X = x\right] dP_X(x)$$
$$= \int E_\theta\left[(r(x) - \hat{r}(x, D_n))^2\right] dP_X(x)$$

$\rightarrow$ The Prediction error is the same as the MISE regression error.

Prediction

For a Markov process, $(X_i, Y_i) = (X_i, X_{i+1})$ et $(X, Y) = (X_T, X_{T+1})$
$\Rightarrow D_n$ not independent of X

Prevision versus Regression

Régression

1. estimation step : on the data $D_n := \{(X_i, Y_i), i = 0, \ldots, n\}$, estimate $r(x) = E[Y|X = x]$ by $\hat{r}(x, D_n)$
2. prediction step : for a new (X, Y), predict Y by $\hat{r}(X, D_n)$

if (X, Y) **were independent of** D_n, then $E[Y|X, D_n] = E[Y|X]$ and

$$\begin{aligned} E_\theta[r(X) - \hat{r}(X, D_n)]^2 &= \int E_\theta\left[(r(X) - \hat{r}(X, D_n))^2 | X = x\right] dP_X(x) \\ &= \int E_\theta\left[(r(x) - \hat{r}(x, D_n))^2\right] dP_X(x) \end{aligned}$$

$\rightarrow$ The Prediction error is the same as the MISE regression error.

Prediction

For a Markov process, $(X_i, Y_i) = (X_i, X_{i+1})$ et $(X, Y) = (X_T, X_{T+1})$
$\Rightarrow D_n$ not independent of X

Prevision versus Regression

Régression

1. estimation step : on the data $D_n := \{(X_i, Y_i), i = 0, \ldots, n\}$, estimate $r(x) = E[Y|X = x]$ by $\hat{r}(x, D_n)$
2. prediction step : for a new (X, Y), predict Y by $\hat{r}(X, D_n)$

if (X, Y) **were independent of** D_n, then $E[Y|X, D_n] = E[Y|X]$ and

$$\begin{aligned} E_\theta[r(X) - \hat{r}(X, D_n)]^2 &= \int E_\theta\left[(r(X) - \hat{r}(X, D_n))^2 | X = x\right] dP_X(x) \\ &= \int E_\theta\left[(r(x) - \hat{r}(x, D_n))^2\right] dP_X(x) \end{aligned}$$

$\longrightarrow$ The Prediction error is the same as the MISE regression error.

Prediction

For a Markov process, $(X_i, Y_i) = (X_i, X_{i+1})$ et $(X, Y) = (X_T, X_{T+1})$
$\Rightarrow D_n$ not independent of X

Prevision versus Regression

Régression

1. estimation step : on the data $D_n := \{(X_i, Y_i), i = 0, \ldots, n\}$, estimate $r(x) = E[Y|X = x]$ by $\hat{r}(x, D_n)$
2. prediction step : for a new (X, Y), predict Y by $\hat{r}(X, D_n)$

if (X, Y) **were independent of** D_n, then $E[Y|X, D_n] = E[Y|X]$ and

$$\begin{aligned} E_\theta[r(X) - \hat{r}(X, D_n)]^2 &= \int E_\theta\left[(r(X) - \hat{r}(X, D_n))^2 | X = x\right] dP_X(x) \\ &= \int E_\theta\left[(r(x) - \hat{r}(x, D_n))^2\right] dP_X(x) \end{aligned}$$

$\longrightarrow$ The Prediction error is the same as the MISE regression error.

Prediction

For a Markov process, $(X_i, Y_i) = (X_i, X_{i+1})$ et $(X, Y) = (X_T, X_{T+1})$
$\Rightarrow D_n$ not independent of X

Prevision versus Regression

Régression

1. estimation step : on the data $D_n := \{(X_i, Y_i), i = 0, \ldots, n\}$, estimate $r(x) = E[Y|X = x]$ by $\hat{r}(x, D_n)$
2. prediction step : for a new (X, Y), predict Y by $\hat{r}(X, D_n)$

if (X, Y) **were independent of** D_n, then $E[Y|X, D_n] = E[Y|X]$ and

$$\begin{aligned} E_\theta[r(X) - \hat{r}(X, D_n)]^2 &= \int E_\theta\left[(r(X) - \hat{r}(X, D_n))^2 | X = x\right] dP_X(x) \\ &= \int E_\theta\left[(r(x) - \hat{r}(x, D_n))^2\right] dP_X(x) \end{aligned}$$

$\rightarrow$ The Prediction error is the same as the MISE regression error.

Prediction

For a Markov process, $(X_i, Y_i) = (X_i, X_{i+1})$ et $(X, Y) = (X_T, X_{T+1})$
$\Rightarrow D_n$ not independent of X

Towards asymptotic independence

Issue

How to let X be independent of D_n ?

A solution : temporal separation

Let $\varphi(T) \to \infty$ and $k_T \to \infty$ such that $k_T - \varphi(T) \to \infty$.
Split the data $(X_0, \ldots, X_T)$:

1. estimate θ on $[0, \varphi(T)]$: $\hat{\theta}_{\varphi(T)}$
2. predict on $[T - k_T, T]$: $\hat{Y} := r_{\hat{\theta}_{\varphi(T)}}(X_{T-k_T}^T)$

by using an assumption of asymptotic independence (short memory) on the process.

Towards asymptotic independence

Issue

How to let X be independent of D_n ?

A solution : temporal separation

Let $\varphi(T) \to \infty$ and $k_T \to \infty$ such that $k_T - \varphi(T) \to \infty$.
Split the data $(X_0, \ldots, X_T)$:

1. estimate θ on $[0, \varphi(T)]$: $\hat{\theta}_{\varphi(T)}$
2. predict on $[T - k_T, T]$: $\hat{Y} := r_{\hat{\theta}_{\varphi(T)}}(X_{T-k_T}^T)$

by using an assumption of asymptotic independence (short memory) on the process.

Outline

Some notions on α-mixing

Definition : α-mixing coefficients, Rosenblatt [1956]

Let $(\Omega, \mathcal{A}, P)$ a probability space and $\mathcal{B}$, $\mathcal{C}$ two sub-sigma fields of $\mathcal{A}$. The α-mixing coefficient between $\mathcal{B}$ and $\mathcal{C}$ is defined by

$$\alpha(\mathcal{B}, \mathcal{C}) = \sup_{\substack{B \in \mathcal{B} \\ C \in \mathcal{C}}} |P(B \cap C) - P(B)P(C)|$$

and the α-mixing coefficient of order k for the stochastic process $\mathbb{X} = \{X_t, t \in \mathbb{N}\}$ defined on the probability space $(\Omega, \mathcal{A}, P)$ as

$$\alpha(k) = \sup_{t \in \mathbb{N}} \alpha(\sigma(X_s, s \leq t), \sigma(X_s, s \geq t + k))$$

Model

Let $\mathbb{X} = (X_t, t \in \mathbb{N})$ a stochastic process. We assume that :

1. $\mathbb{X}$ is a second order, square integrable, α-mixing process.
2. the regression function $r_\theta(.)$ depends approximately of the last k_T values $(X_{T-i}, i = 1, \ldots, k_T)$:

$$X^*_{T+1} := E_\theta \left[X_{T+1} \middle| X_0^T \right] := \sum_{i=0}^{k_T} r_i(X_{T-i}, \theta) + \eta_{k_T}(\mathbb{X}, \theta).$$

Assumptions $\mathbf{H_0}$ on the process

(i) $\lim_{T \to \infty} E_\theta(\eta^2_{k_T}(\mathbb{X}, \theta)) = 0$;

(ii) for all $i \in \mathbb{N}$, $\|r_i(X_{T-i}, \theta_1) - r_i(X_{T-i}, \theta_2)\| \leq H_i(X_{T-i}) \|\theta_1 - \theta_2\|$, $\forall \theta_1, \theta_2$;

(iii) there exists a $r > 1$ such that $\sup_{i \in \mathbb{N}} \left(E_\theta H_i^{2r}(X_{T-i}) \right)^{1/r} < \infty$.

This additive model is an extension of a model studied by Bosq [2007].

Statistical Prediction and assumptions

We assume we have an estimator $\hat{\theta}_T$ of θ.

Assumptions $\mathbf{H_1}$ on the estimator $\hat{\theta}_T$

(i) $\limsup_{T\to\infty} T.E_\theta(\hat{\theta}_T - \theta)^2 < \infty$;

(ii) there exists $q > 1$ such that $\limsup_{T\to\infty} T^q E(\hat{\theta}_T - \theta)^{2q} < \infty$.

We build a statistical predictor : $\hat{X}_{T+1} := \sum_{i=0}^{k_T} r_i(X_{T-i}, \hat{\theta}_{\varphi(T)})$

Assumptions $\mathbf{H_2}$ on the coefficients

(i) $\frac{k_T^2}{\varphi(T)} \underset{T\to\infty}{\to} 0$;

(ii) $(T - k_T - \varphi(T)) \underset{T\to\infty}{\to} \infty$.

Statistical Prediction and assumptions

We assume we have an estimator $\hat{\theta}_T$ of θ.

Assumptions $\mathbf{H_1}$ on the estimator $\hat{\theta}_T$

(i) $\limsup_{T\to\infty} T.E_\theta(\hat{\theta}_T - \theta)^2 < \infty$;

(ii) there exists $q > 1$ such that $\limsup_{T\to\infty} T^q E(\hat{\theta}_T - \theta)^{2q} < \infty$.

We build a statistical predictor : $\hat{X}_{T+1} := \sum_{i=0}^{k_T} r_i(X_{T-i}, \hat{\theta}_{\varphi(T)})$

Assumptions $\mathbf{H_2}$ on the coefficients

(i) $\frac{k_T^2}{\varphi(T)} \underset{T\to\infty}{\to} 0$;

(ii) $(T - k_T - \varphi(T)) \underset{T\to\infty}{\to} \infty$.

Consistency of the predictor

Theorem 2.5

Under the assumptions $\mathbf{H_0}$,$\mathbf{H_1}$,$\mathbf{H_2}$, we have that

$$\limsup_{T\to\infty} E_\theta(\hat{X}_{T+1} - X^*_{T+1})^2 = 0$$

Tool : Davydov's covariance inequality

Let $X \in L^q(\mathbb{P})$ and $Y \in L^r(\mathbb{P})$, if $q > 1$, $r > 1$ and $\frac{1}{r} + \frac{1}{q} = 1 - \frac{1}{p}$, then

$$|Cov(X,Y)| \leq 2p\big(2\alpha(\sigma(X),\sigma(Y))\big)^{\frac{1}{p}} \|X\|_q \|Y\|_r.$$

Consistency of the predictor

Theorem 2.5

Under the assumptions $\mathbf{H_0}$,$\mathbf{H_1}$,$\mathbf{H_2}$, we have that

$$\limsup_{T\to\infty} E_\theta(\hat{X}_{T+1} - X^*_{T+1})^2 = 0$$

Tool : Davydov's covariance inequality

Let $X \in L^q(\mathbb{P})$ and $Y \in L^r(\mathbb{P})$, if $q > 1$, $r > 1$ and $\frac{1}{r} + \frac{1}{q} = 1 - \frac{1}{p}$, then

$$|Cov(X,Y)| \leq 2p\big(2\alpha(\sigma(X),\sigma(Y))\big)^{\frac{1}{p}} \|X\|_q \|Y\|_r.$$

Example of process

For a linear, weakly stationary, centered, non deterministic, inversible process in discrete time, its Wold decomposition writes:

$$X_T = e_T + \sum_{i=1}^{k_T} \varphi_i(\theta) X_{T-i} + \sum_{i>k_T} \varphi_i(\theta) X_{T-i}$$

with $\sum_{i=1}^{\infty} \varphi_i^2(\theta) < \infty$. Set $\eta_{k_T}(\mathbb{X}, \theta) = \sum_{i>k_T+1} \varphi_i(\theta) X_{T+1-i}$

Proposition

If $\mathbb{X}$ verifies the assumptions

1. $\forall i$, φ_i is differentiable and $\|\varphi_i'(.)\|_\infty < \infty$;
2. there exists a $r > 1$ such as (X_t) has a moment of order $2r$;
3. $\mathbb{X}$ is α-mixing and such that $\sum_{i,j} \varphi_{i+1}(\theta)\varphi_{j+1}(\theta)\alpha^{1/p}(|i-j|) < \infty$.

Then, $\mathbb{X}$ verifies the assumptions of theorem 2.5.

Outline

Assumptions for the limit law

Assumptions $\mathbf{H}_0'$ on the process

(i) $\theta \mapsto r_i(X_{T-i}, \theta)$ is twice differentiable w.r.t. θ;

(ii) $\sup_i \left\| \partial_\theta^2 r_i(X_{T-i}, .) \right\|_\infty = O_P(1)$;

(iii) $\eta_{k_T}(\mathbb{X}, \theta) = o_P\left(\sqrt{\frac{1}{\varphi(T)}}\right)$;

(iv) $\sum_{i=0}^{+\infty} \partial_\theta r_i(X_{T-i}; \theta)$ exists and converge a. s. to a vector V as $T \to +\infty$.

Assumption $\mathbf{H}_1'$ on the estimator $\hat{\theta}_T$

(i) $\sqrt{T}(\hat{\theta}_T - \theta) \overset{\mathcal{L}}{\rightsquigarrow} N(0, \sigma^2(\theta))$.

Assumption $\mathbf{H}_2'$ on the coefficients

(i) $k_T = o(\sqrt{\varphi(T)})$;

(ii) $(T - k_T - \varphi(T)) \underset{T \to \infty}{\to} \infty$.

Limit law of the predictor

Theorem 2.10

If the assumptions $\mathbf{H}_0'$,$\mathbf{H}_1'$,$\mathbf{H}_2'$ are verified, then

$$\sqrt{\varphi(T)}(\hat{X}_{T+1} - X^*_{T+1}) \overset{\mathcal{L}}{\rightsquigarrow} < U, V >$$

where U and V are two independent random variables, U with law $\mathcal{N}(0, \sigma^2(\theta))$ and V is the limit of $\sum_{i=0}^{+\infty} \partial_\theta r_i(X_{T-i}; \theta)$ as $T \to \infty$

Tool

An asymptotic independence lemma

Let (X'_n) and (X''_n) two sequences of real-valued random variables with laws P'_n and P''_n respectively, defined on the probability space $(\Omega, \mathcal{A}, P)$. Assume that (X'_n) and (X''_n) are asymptotically mixing w.r.t. each other, in the sense that there exists a sequence of coefficients $\alpha(n)$ with $\alpha(n) \underset{n\to\infty}{\to} 0$ such that, for all Borel set A and B of $\mathcal{R}$,

$$\left|P(X'_n \in A, X''_n \in B) - P(X'_n \in A)P(X''_n \in B)\right| \leq \alpha(n)$$

Then, if

1. $X'_n \overset{\mathcal{L}}{\rightsquigarrow} X'$ with law P';
2. $X''_n \overset{\mathcal{L}}{\rightsquigarrow} X''$ with law P'';

$(X'_n, X''_n) \overset{\mathcal{L}}{\rightsquigarrow} (X', X'')$, and the law (X', X'') is $P' \otimes P''$.

Conclusions

Some limits of the temporal decoupling method

1. heuristically under-efficient : gap in the data ;
2. the mixing coefficients = a real number which reduces the dependence structure of the process to a property of asymptotic independence ;
3. practical applications are difficult to undertake.

References

Faugeras, O. (2007) Prévision statistique paramétrique par séparation temporelle. *Accepted to Annales de l'ISUP.*

Part II : A nonparametric quantile-copula approach to conditional density estimation.

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 ?

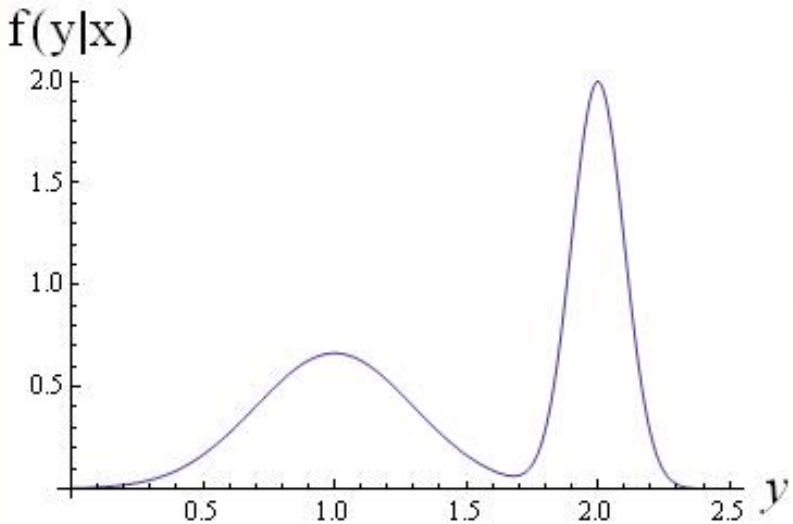
What is a good prediction ?

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 ?

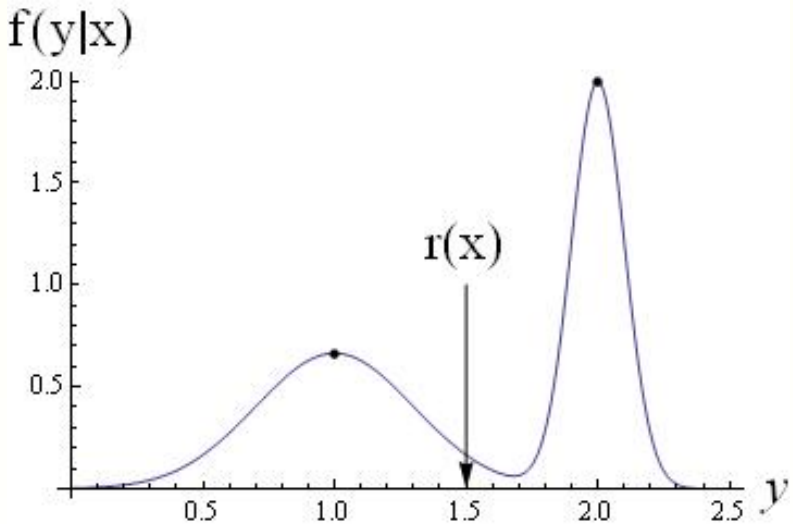
1. Classical approach ($\mathbb{L}_2$ theory): the conditional mean or *regression function* $r(x) = E(Y|X = x)$,
2. Fully informative approach: the *conditional density* $f(y|x)$



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

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

$\hat{f}_{X,Y}, \hat{f}$: 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

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

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

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

$\hat{f}_{X,Y}, \hat{f}$: 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 - 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 $\longrightarrow$ a regression framework

1. *Transform* the data:

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

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

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

Nadaraya-Watson estimator

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

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 $\longrightarrow$ numerical implementation delicate (trimming);
- minoration hypothesis on the marginal density $f(x) \geq c > 0$.

How to remedy these problems?

$\longrightarrow$ 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 $\longrightarrow$ numerical implementation delicate (trimming);
- minoration hypothesis on the marginal density $f(x) \geq c > 0$.

How to remedy these problems?
$\longrightarrow$ 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.

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

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.

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

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.

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

The 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 $C : [0,1]^2 \longrightarrow [0,1]$, 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 $C : [0,1]^2 \longrightarrow [0,1]$, 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 $C : [0,1]^2 \longrightarrow [0,1]$, 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

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

- density of Y: $g \leftarrow$ kernel estimator $\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)$

$$F(x) \leftarrow F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \leqslant x}$$

- c.d.f. $G(y) \leftarrow G_n(y) := \frac{1}{n} \sum_{j=1}^{n} 1_{Y_j \leqslant y}$ empirical c.d.f.

- copula density $c(u,v) \leftarrow c_n(u,v)$ a bivariate Parzen-Rosenblatt kernel density *(pseudo)* estimator

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

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

Construction of the estimator - 1

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

- density of Y: $g \leftarrow$ kernel estimator $\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)$

$$F(x) \quad \leftarrow \quad F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \leqslant x}$$

- c.d.f. $G(y) \quad \leftarrow \quad G_n(y) := \frac{1}{n} \sum_{j=1}^{n} 1_{Y_j \leqslant y}$ empirical c.d.f.

- copula density $c(u,v) \leftarrow c_n(u,v)$ a bivariate Parzen-Rosenblatt kernel density *(pseudo)* estimator

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

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

Construction of the estimator - 1

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

- density of Y: $g \leftarrow$ kernel estimator $\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)$

$$F(x) \quad \leftarrow \quad F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \leqslant x}$$

- c.d.f. $G(y) \quad \leftarrow \quad G_n(y) := \frac{1}{n} \sum_{j=1}^{n} 1_{Y_j \leqslant y}$ empirical c.d.f.

- copula density $c(u,v) \leftarrow c_n(u,v)$ a bivariate Parzen-Rosenblatt kernel density (*pseudo*) estimator

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

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

Construction of the estimator - 1

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

- density of Y: $g \leftarrow$ kernel estimator $\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)$

$$F(x) \quad \leftarrow \quad F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j \leqslant x}$$

- c.d.f. $G(y) \quad \leftarrow \quad G_n(y) := \frac{1}{n} \sum_{j=1}^{n} 1_{Y_j \leqslant y}$ empirical c.d.f.

- copula density $c(u,v) \leftarrow c_n(u,v)$ a bivariate Parzen-Rosenblatt kernel density (*pseudo*) estimator

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

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 $c(u, v)$

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

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 $c(u, v)$

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

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 $c(u, v)$

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

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 $c(u, v)$

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

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 $c(u, v)$

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

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.

$\longrightarrow$ 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.

$\longrightarrow$ 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 $nh_n \to \infty$, $na_n^4 \to \infty$, $na_n^6 \to 0$, and $\sqrt{\ln\ln n}/(na_n^3) \to 0$ 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 $h_n \longrightarrow 0$, $a_n \longrightarrow 0$, 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

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

- 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

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

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 $[0,1]^2$
  $\longrightarrow$ 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 $[0,1]^2$
  $\longrightarrow$ 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:

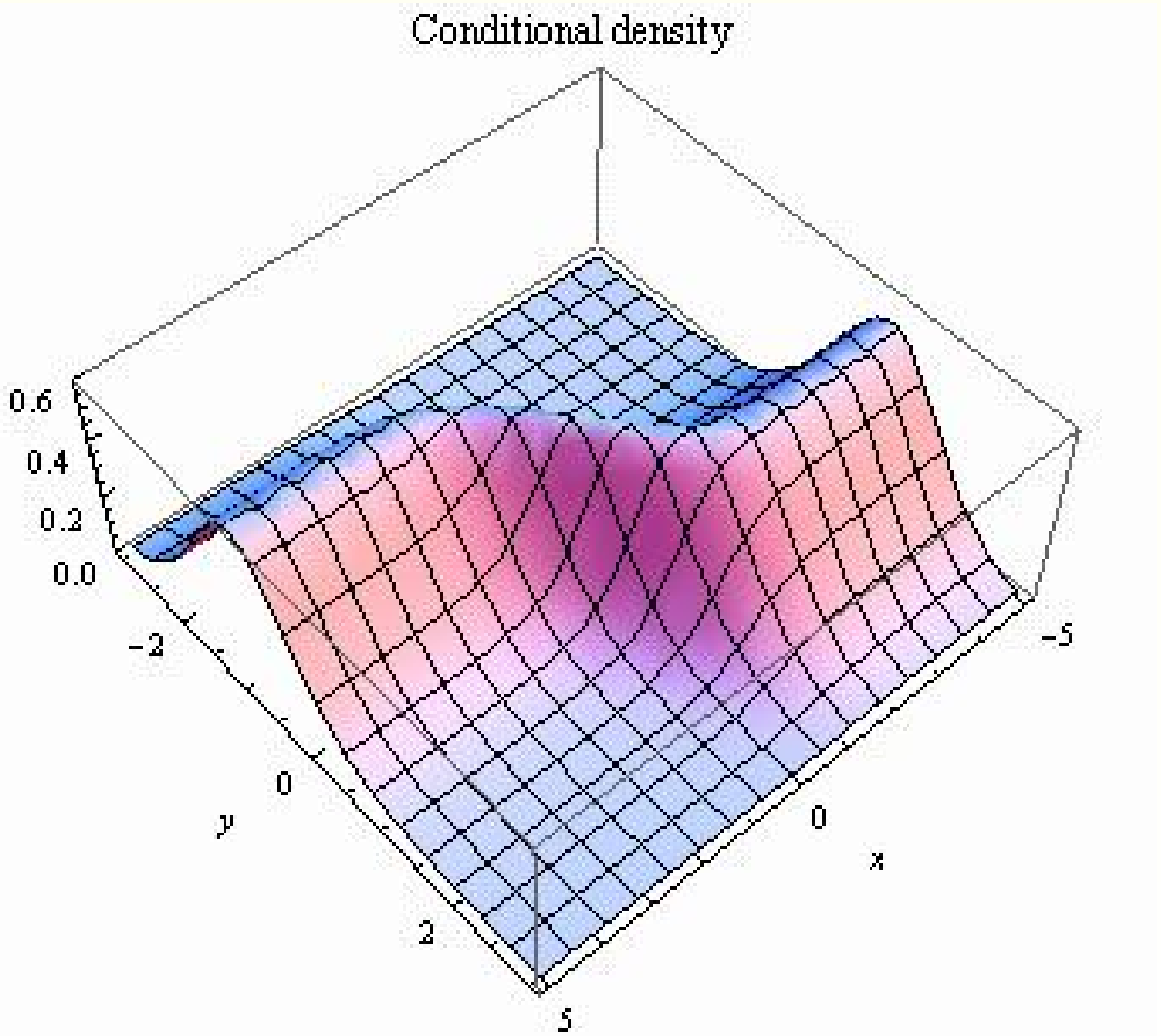
- X, Y is marginally distributed as $\mathcal{N}(0, 1)$
- X, Y is linked via Frank Copula .

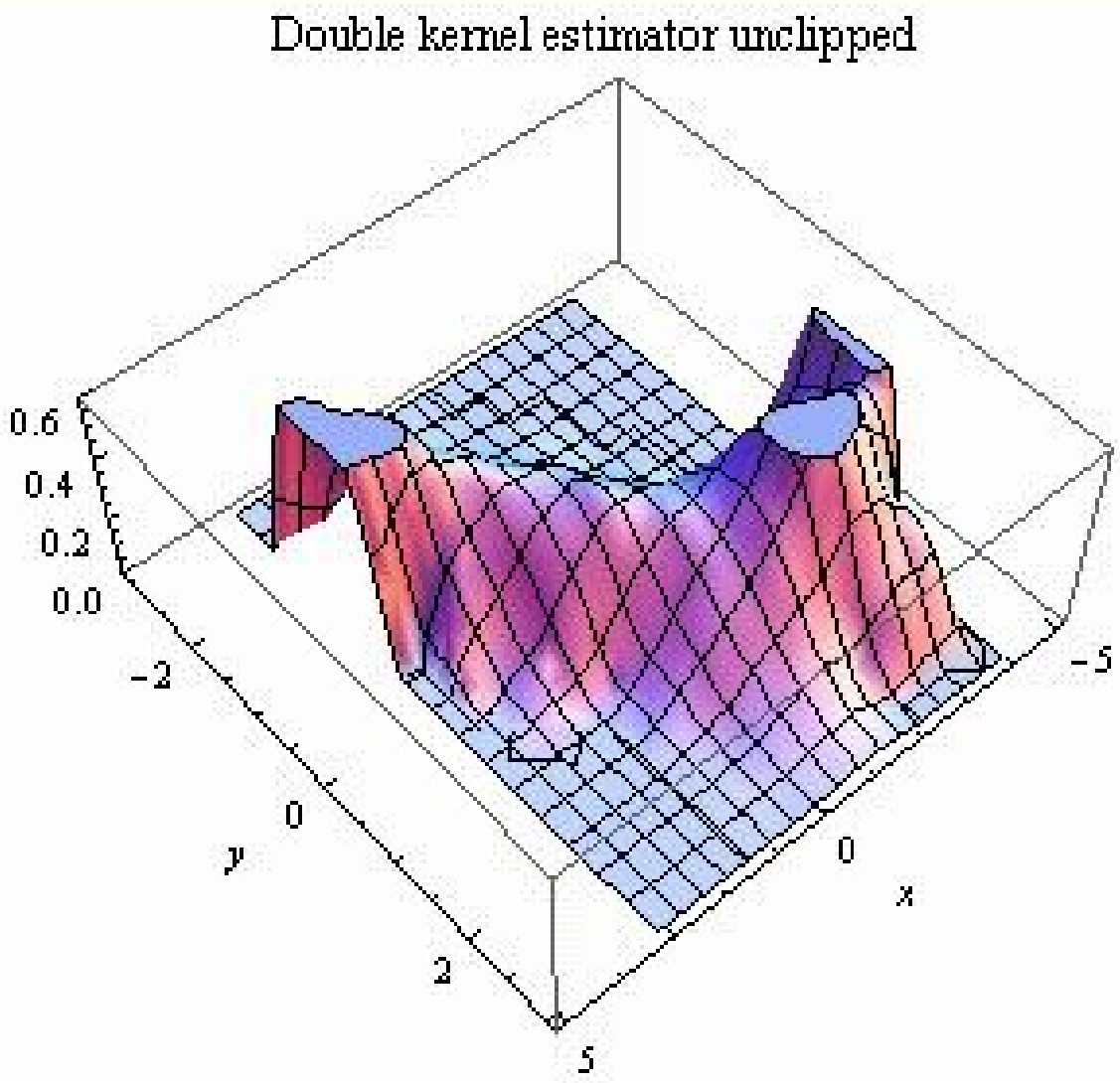
$$C(u, v, \theta) = \frac{\ln[(\theta + \theta^{u+v} - \theta^u - \theta^v)/(\theta - 1)]}{\ln \theta}$$

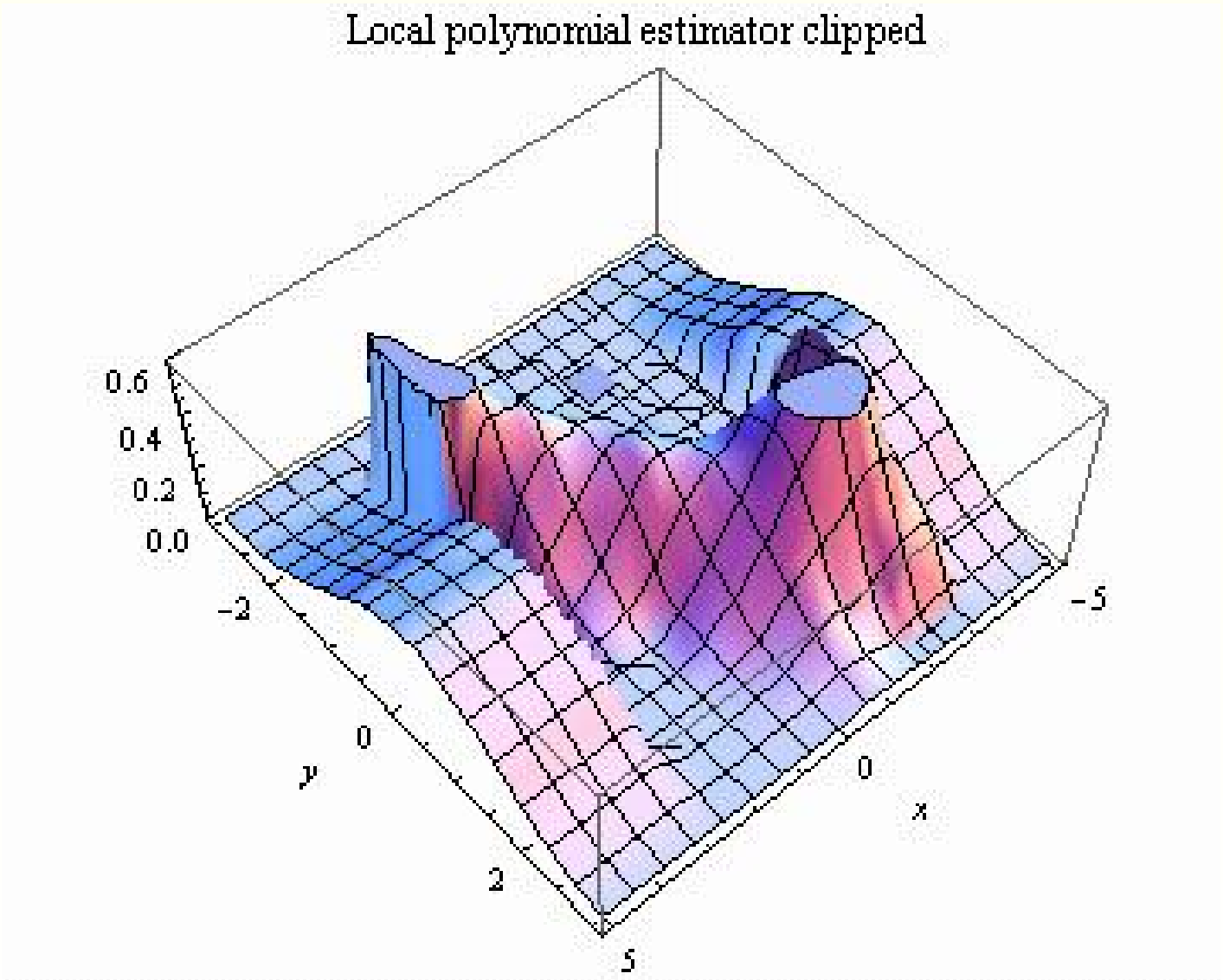
with parameter $\theta = 100$.

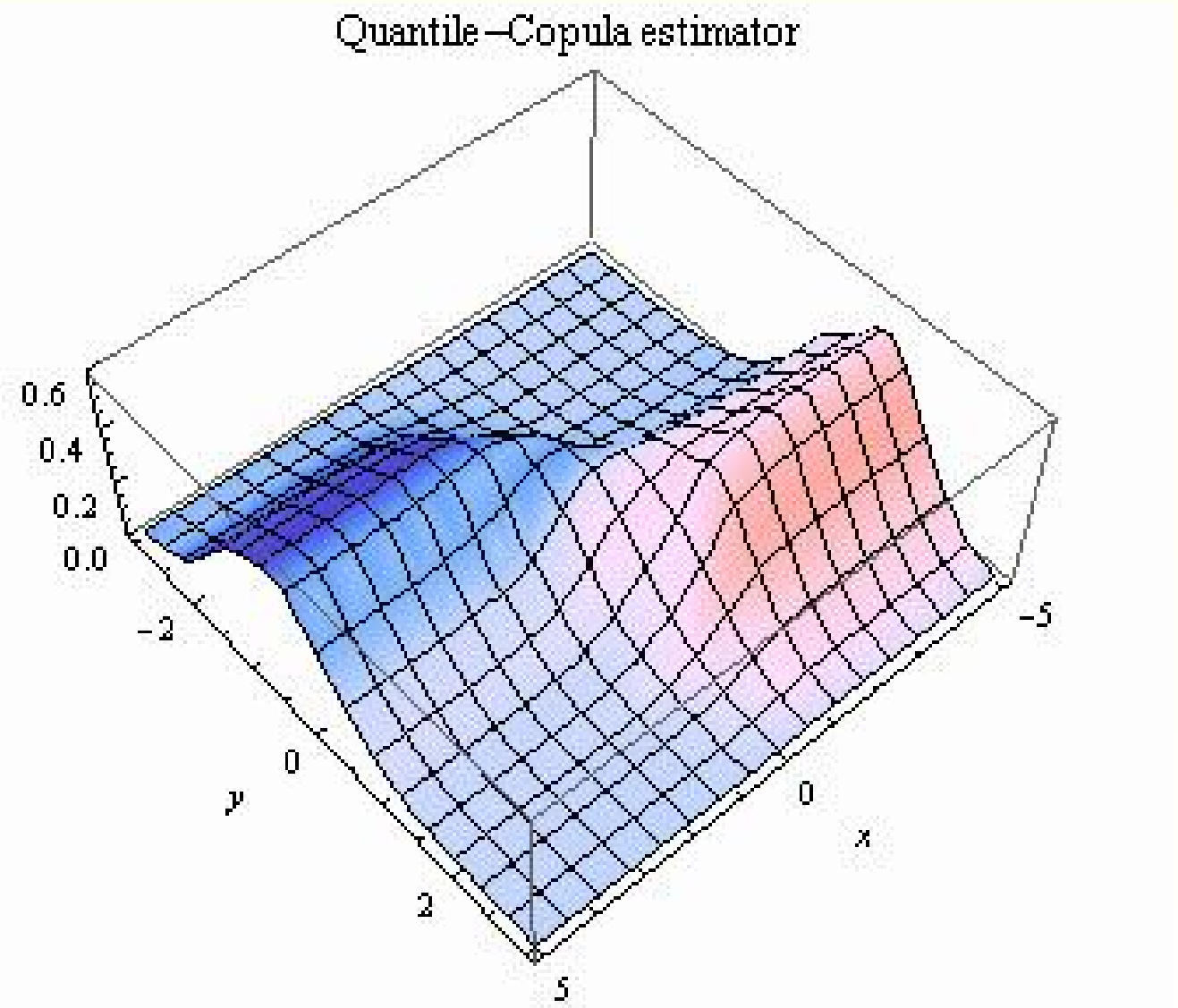
Practical implementation:

- 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
2
y
−5
0
5
x


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


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


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

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

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

Set predictors: Level sets

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

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

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

$\longrightarrow$ 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 $\hat{G}$,

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

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

Implication for the quantile copula estimator

The doubly smoothed quantile copula conditional density estimator

$\longrightarrow$ 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 $\hat{G}$,

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

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

Connection with the variable bandwidth kernel estimators

Connection with the variable bandwidth kernel estimators

Therefore, the copula density part of the estimator writes

$$\hat{c}_n(\hat{F}(x),\hat{G}(y)) = (na_nb_n)^{-1}\sum_{i=1}^{n} K_1\left(\frac{\hat{F}(X_i)-\hat{F}(x)}{a_n}\right) K_2\left(\ldots\right)$$

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

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

Outline

Conclusions

Summary

- ratio type into the product $\longrightarrow$ consistency and limit results where obtained by combination of the previous known ones on (unconditional) density estimation,
- nonexplosive behavior in the tails of the marginal density,
- no need for trimming or clipping.

Conclusions

Some perspectives and work-in-progress

- Adaptive bandwidth choices to the regularity of the model with an efficient kernel estimation of the copula density by Boundary-corrected kernels (with A. Leblanc).
- To design applications-specific conditional estimators:
  - estimation in the tail of the marginal distribution, to relate with extreme value theory, with applications in insurance, risk analysis, environmental sciences.
  - estimation for censored data with Kaplan-Meier estimators of the marginals.
- Extension to time series by coupling arguments for Markovian models.
- Alternative nonparametric methods of estimation by wavelets and minimax analysis with K.Tribouley, E. Masiello.

Bibliography

Reference

O. P. Faugeras. A quantile-copula approach to conditional density estimation. *Submitted, accepted upon minor revision*, 2008. Available on http://hal.archives-ouvertes.fr/hal-00172589/fr/.

Related work:

- J. Fan and Q. Yao. *Nonlinear time series*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2005. Nonparametric and parametric methods.
- J. Fan, Q. Yao, and H. Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.
- L. Györfi and M. Kohler. Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory*, 53(5):1872–1879, 2007.
- R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald. Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336, 1996.

References

- R. J. Hyndman and Q. Yao. Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3):259–278, 2002.
- C. Lacour. Adaptive estimation of the transition density of a markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(5):571–597, 2007.
- M. Rosenblatt. Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York, 1969.
- M. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.
- W. Stute. On almost sure convergence of conditional empirical distribution functions. *Ann. Probab.*, 14(3):891–901, 1986.

Thank you !