EADS Innovation Works

Uncertainty modeling and quantification

RTG12 Training, MBDA

Régis LEBRUN, regis.lebrun@eads.net

September 26, 2012

EADS

Outline

Motivation

Why probabilities?

The choice of a probabilistic framework looks natural to model uncertainties, in particular when these uncertainties are related to unpredictable natural events (e.g weather conditions). It is more questionable when dealing with uncertainties related to a lack of knowledge, and other modeling frameworks have been proposed (e.g interval arithmetic).
We promote the systematic use of a probabilistic framework because:

- it has well developed mathemathical bases;
- it allows to integrate real-life data, either through statistics or through expert knowledge (information theory);
- it allows to formulate many (all?) of the questions of interest for an engineer;
- many good quality softwares are available (forget Excel!)

Most of the technical complexity related to probability theory is hidden in the tools, but the engineer remains responsible for his results. As such, he must know what is behind the software.

Introduction

Historical view

17th Century First attempt to formalize the probability calculus by Pascal, Pierre de Fermat, Huygens. Mainly focused on gambling games, the theory was mainly a matter of counting.

20th Century A formalisation based only on counting leads to numerous paradoxes, mainly due to a fuzzy definition of the probabilistic experiment. The formalization proposed by Kolmogorov at the beginning of the 20th Century has been a successful attempt to give strong fundations to the theory of probability.

- The several paradoxes resulting from the intuitive notion of probability as a frequency have found a convincing explanation;
- The new formulation are more involved, the probability theory is no more linked to a physical experiment.

This lecture will be based on the modern view of probabilities.

Probability and statistics

Different point of views

Broadly speaking, we have the following separation between the probability theory and the statistics theory:

- The statistics theory is focused on the effective gathering of the information related to a particular topic (e.g opinion, physical measurement) and uses the probability theory to build a mathematical model of this gathering and to study the quality of the resulting conclusions from a mathematical point of view.
- The probability theory is focused on the definition of abstract concepts and on their interaction. In particular, it provides useful mathematical models for the statisticians.

These two fields are complementary, the statistics being the field that makes the link between raw data and the concepts found in the probability theory. In return, using probability theory results, one is able to justify the correctness of some data threatments.

Kolmogorov formalism

σ-field

Let Ω be a given non-empty set. A σ-field $\mathcal{F} \in \mathcal{P}(\Omega)$ defined on Ω is a collection of subsets of Ω such that:

- $\Omega \in \mathcal{F}$
- $\forall B \in \mathcal{F}$, $\Omega \; B \in \mathcal{F}$
- For all countable sequence $B_i \in \mathcal{F}$, $\bigcup_{i \in \mathbb{N}} B_i \in \mathcal{F}$

Generated σ-field

Let Ω be a non-empty set and $(A_i)_{i \in I}$ be an arbitrary collection of subsets of Ω. The σ-field generated by $(A_i)_{i \in I}$ $\mathcal{F}((A_i)_{i \in I})$ is the smallest σ-field (for the inclusion) that contains all the A_i.

Example

If $\Omega = \mathbb{R}$ (or any topological space) and $(A_i)_{i \in I}$ is the collection of its open sets, then $\mathcal{F}((A_i)_{i \in I}) = \mathcal{B}(\mathbb{R})$ is the Borel σ-field associated with $\mathbb{R}$.

Kolmogorov formalism

Measurable space

A measurable space is a couple $(\Omega, \mathcal{F})$ where Ω is a given non-empty set and $\mathcal{F}$ is a σ-field defined on Ω.

Probability space

A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ such that $(\Omega, \mathcal{F})$ is a measurable space and $\mathbb{P}$ is probability measure, it means a function defined on $\mathcal{F}$, taking values into $[0, 1]$ and such that:

- $\mathbb{P}(\Omega) = 1$;
- If A_i is a countable collection of disjoint elements of $\mathcal{F}$, then $\mathbb{P}(\bigcup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} \mathbb{P}(A_i)$.

Examples

1. Arbitrary non-empty set Ω, $\mathcal{F} = \{\emptyset, \Omega\}$. We can only observe if an experiment has been done (event Ω) or not (event $\emptyset$). By definition, $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$.
2. Coin flipping: $\Omega = \{\mathit{Tail}, \mathit{Head}\}$, $\mathcal{F} = \{\emptyset, \Omega, \{\mathit{Tail}\}, \{\mathit{Head}\}\} = \mathcal{P}(\Omega)$. All the possible outcomes are considered in this modeling. A fair coin is associated with $\mathbb{P}$ such that $\mathbb{P}(\mathit{Tail}) = \mathbb{P}(\mathit{Head}) = 1/2$.
3. Dice tossing: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \{$ $\emptyset$, Ω, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$, $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{1, 5\}$, $\{1, 6\}$, $\{2, 3\}$, $\{2, 4\}$, $\{2, 5\}$, $\{2, 6\}$, $\{3, 4\}$, $\{3, 5\}$, $\{3, 6\}$, $\{4, 5\}$, $\{4, 6\}$, $\{5, 6\}$, $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{1, 2, 5\}$, $\{1, 2, 6\}$, $\{1, 3, 4\}$, $\{1, 3, 5\}$, $\{1, 3, 6\}$, $\{1, 4, 5\}$, $\{1, 4, 6\}$, $\{1, 5, 6\}$, $\{2, 3, 4\}$, $\{2, 3, 5\}$, $\{2, 3, 6\}$, $\{2, 4, 5\}$, $\{2, 4, 6\}$, $\{2, 5, 6\}$, $\{3, 4, 5\}$, $\{3, 4, 6\}$, $\{3, 5, 6\}$, $\{4, 5, 6\}$, $\{1, 2, 3, 4\}$, $\{1, 2, 3, 5\}$, $\{1, 2, 3, 6\}$, $\{1, 2, 4, 5\}$, $\{1, 2, 4, 6\}$, $\{1, 2, 5, 6\}$, $\{1, 3, 4, 5\}$, $\{1, 3, 4, 6\}$, $\{1, 3, 5, 6\}$, $\{1, 4, 5, 6\}$, $\{2, 3, 4, 5\}$, $\{2, 3, 4, 6\}$, $\{2, 3, 5, 6\}$, $\{2, 4, 5, 6\}$, $\{3, 4, 5, 6\}$, $\{1, 2, 3, 4, 5\}$, $\{1, 2, 3, 4, 6\}$, $\{1, 2, 3, 5, 6\}$, $\{1, 2, 4, 5, 6\}$, $\{1, 3, 4, 5, 6\}$, $\{2, 3, 4, 5, 6\}$ $\} = \mathbb{P}(\Omega)$. All the possible outcomes are considered in this modeling. A fair dice is associated with $\mathbb{P}$ such that $\mathbb{P}(\{i\}) = 1/6$, $i = 1, \ldots, 6$.
4. Coin flipping using a dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \{$ $\emptyset$, Ω, $\{1, 3, 5\}$, $\{2, 4, 6\}$ $\}$. Only the parity of the face is considered in the outcome of a dice roll.
5. Real number localization: $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$. All possible outcomes made of countable union of intervals are considered. If $\mathbb{P}(\{x\}) = 0$ for all $x \in \mathbb{R}$, the probability measure $\mathbb{P}$ is continuous, and if there exists an at most countable collection of reals $(x_i)_{i \in \mathbb{N}}$ such that $\sum_{i \in \mathbb{N}} \mathbb{P}(x_i) = 1$ then $\mathbb{P}$ is discrete.

Random variable

Definition

A real-valued random variable X defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathbb{R}$ is a function defined on Ω and taking values in $\mathbb{R}$ and such that

$$\forall B \in \mathcal{B}(\mathbb{R}),\ X^{-1}(B) \in \mathcal{F} \tag{1}$$

Induced probability measure

Let X be a real-valued random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The probability measure $\mathbb{P}_X$ induced by X on the σ-field $\mathcal{B}(\mathbb{R})$ is defined by:

$$\forall B \in \mathcal{B}(\mathbb{R}),\ \mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) \tag{2}$$

From a modeling point of view, a random variable is a way to numerize through a unique number the event resulting from a random experiment.

Random vector

Definition

A real-valued n-dimensional random vector $\boldsymbol{X}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathbb{R}^n$ is a function defined on Ω and taking values in $\mathbb{R}^n$ and such that

$$\forall B \in \mathcal{B}(\mathbb{R}^n),\ X^{-1}(B) \in \mathcal{F} \tag{3}$$

From a modeling point of view, a random vector is a way to numerize through a set of real numbers the event resulting from a random experiment. A one-dimensional random vector is no more than a random variable.

Distribution function

Definition

Let $\boldsymbol{X}$ be an n-dimensional real-valued random vector. Its distribution function $F_{\boldsymbol{X}}$ (or F for short) is the real-valued function defined on $\mathbb{R}^n$ such that:

$$\forall \vec{x} \in \mathbb{R}^n,\ F(\boldsymbol{x}) = \mathbb{P}_{\boldsymbol{X}}((-\infty, x_1] \times \cdots \times (-\infty, x_n]) \tag{4}$$

The distribution function is also named the *cumulative distribution function*, abbreviated in *CDF*.

Theorem

The distribution function F of a random vector $\boldsymbol{X}$ characterizes its probability measure $\mathbb{P}_{\boldsymbol{X}}$.

The probabilistic modeling of a random vector is thus equivalent to the construction of its distribution function.

Marginal distribution functions, quantile

Definition

Let $\boldsymbol{X}$ be an n-dimensional random vector with distribution function F. Its ith marginal component X_i is the random variable obtained by projection of $\boldsymbol{X}$ on the ith dimension of $\mathbb{R}^n$. The distribution function F_i of X_i is the ith marginal distribution function of F and is such that:

$$\forall x_i \in \mathbb{R},\ F_i(x_i) = F(+\infty, \ldots, +\infty, x_i, +\infty, \ldots, +\infty) \tag{5}$$

Given an n-dimentional distribution function F, the marginal distribution functions F_i are known. The stochastic dependence associated with F (or $\boldsymbol{X}$) is the complementary information that allows to recover F from $F_1, \ldots, F_n$. It is exactly what does the *copula* concept.

Discrete random vectors

Definition

If there exists an at most countable set of points $(\boldsymbol{x}_i)_{i\in\mathbb{N}} \in \mathbb{R}^n$ such that $\sum_{i\in\mathbb{N}} \mathbb{P}(\boldsymbol{x}_i) = 1$, then $\boldsymbol{X}$ is said to be a discrete random vector.
The function $p_{\boldsymbol{X}}$ defined on the countable set $\mathcal{S} = \{\boldsymbol{x}_i, \| \, i \in \mathbb{N}\}$ by:

$$\forall \boldsymbol{x} \in \mathcal{S}, \; p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{x}) \tag{6}$$

is called the probability function of the random vector.

The probability function is also named the *probability distribution function* abbreviated in *PDF*.

Continuous random vectors

Definition

If the distribution function F of a random vector $\boldsymbol{X}$ is a continuous function, then the random vector is said to be continuous. If there exists a positive function $p_{\boldsymbol{X}}$ defined on $\mathbb{R}^n$ such that:

$$\forall \boldsymbol{x} \in \mathbb{R}^n,\ F(\boldsymbol{x}) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_n} p_{\boldsymbol{X}}(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi} \tag{7}$$

In that case, the nth cross derivative $\frac{\partial^n F}{\partial x_1 \dots \partial x_n}$ exists and we have:

$$\forall \boldsymbol{x} \in \mathbb{R}^n,\ \frac{\partial^n F}{\partial x_1 \dots \partial x_n}(\boldsymbol{x}) = p_{\boldsymbol{X}}(\boldsymbol{x}) \tag{8}$$

The function $p_{\boldsymbol{X}}$ (or p for short) is the density function of $\boldsymbol{X}$.

The density function is also named the *probability density function* abbreviated in *PDF*. A random vector can be neither discrete nor continuous.

Some classical distribution

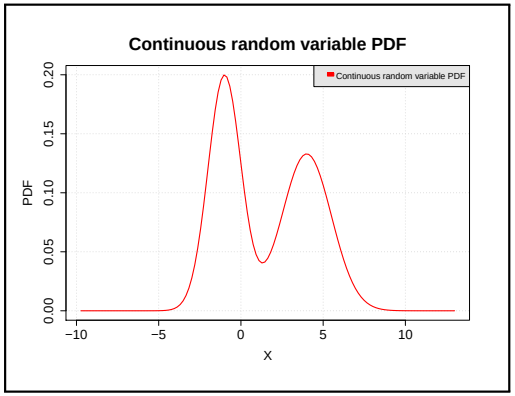
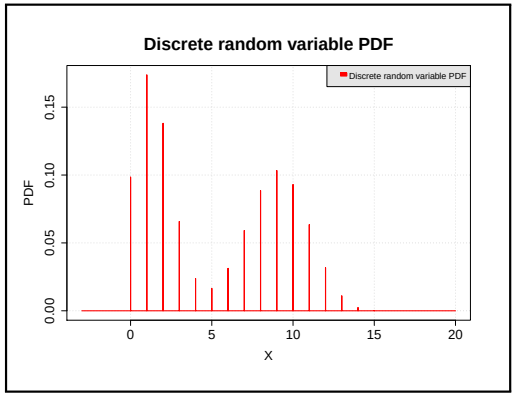
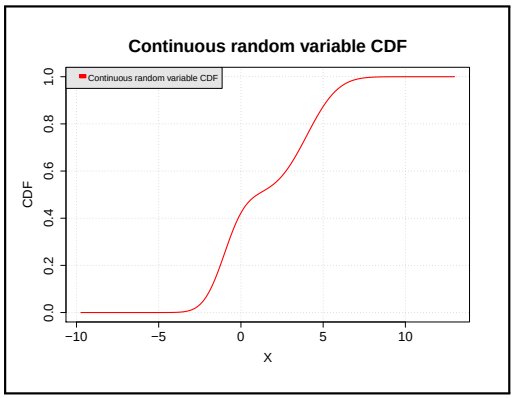
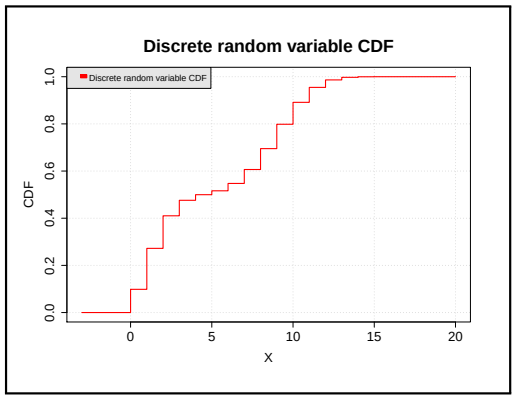
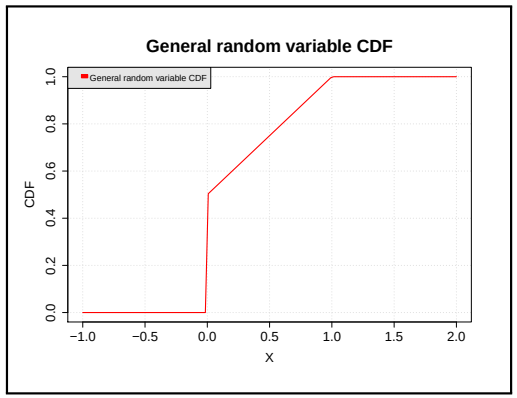
Continuous distributions

Name	probability density function
Exponential	$\lambda e^{-\lambda(x-\gamma)}\mathbf{1}_{[\gamma,+\infty[}(x),\ x \in \mathbb{R}$
Normal	$\frac{1}{(2\pi)^{\frac{n}{2}}(\det\boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)'\boldsymbol{\Sigma}^{-1}(x-\mu)},\ x \in \mathbb{R}^n$
Uniform	$\frac{1}{b-a}\mathbf{1}_{[a,b]}(x),\ x \in \mathbb{R}$

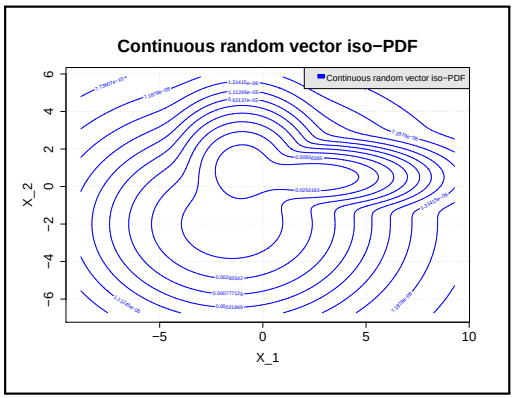
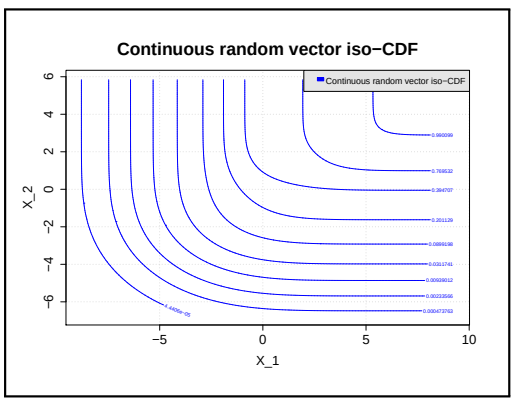
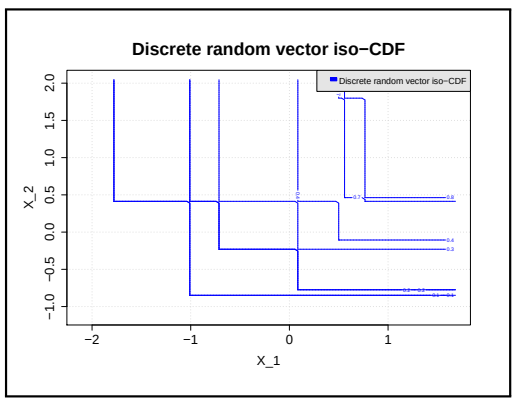
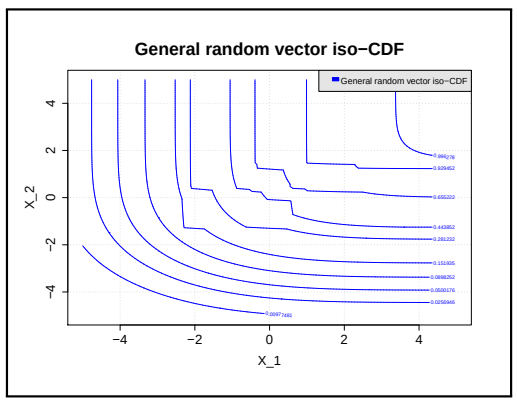
Discrete distributions

Name	probability distribution function
Bernoulli	$P(X=1)=p, P(X=0)=1-p,\ p \in [0,1]$
Multinomial	$P(\boldsymbol{X}=\boldsymbol{x}) = \frac{N!}{x_1!\ldots x_n!(N-s)!} p_1^{x_1}\ldots p_n^{x_n}(1-q)^{N-s}$ with $0 \leq p_i \leq 1, x_i \in \mathbb{N}, q=\sum_{k=1}^n p_k \leq 1, s=\sum_{k=1}^n x_k \leq N$

PDF and CDF, random variables

	Continuous	Discrete	General
PDF			
CDF			

PDF and CDF, random vectors

	Continuous	Discrete	General
PDF			
CDF			

What is the stochastic dependence?

The stochastic dependence between the components of a random vector $\boldsymbol{X}$ is the interaction between these components that does not depend on the marginal distribution functions of $\boldsymbol{X}$.

In terms of distribution functions, the stochastic dependence is described by the part of the distribution function F of $\boldsymbol{X}$ that does not depend on its marginal distribution functions $F_1, \ldots, F_n$. This part corresponds to the concept of *copula*

Copula

Definition

An *n*-dimensional copula C is a function defined on $[0,1]^n$ and taking values in $[0,1]$ which is the restriction to $[0,1]^n$ of an n-dimensional distribution function with uniform marginal distibutions on $[0,1]$.

Sklar's theorem

Let F be an n-dimensional distribution function with marginal distribution functions $F_1,\ldots,F_n$. Then there exists an n-dimensional copula C such that:

$$\forall \boldsymbol{x} \in \mathbb{R}^n,\ F(x_1,\ldots,x_n) = C(F_1(x_1),\ldots,F_n(x_n)) \tag{9}$$

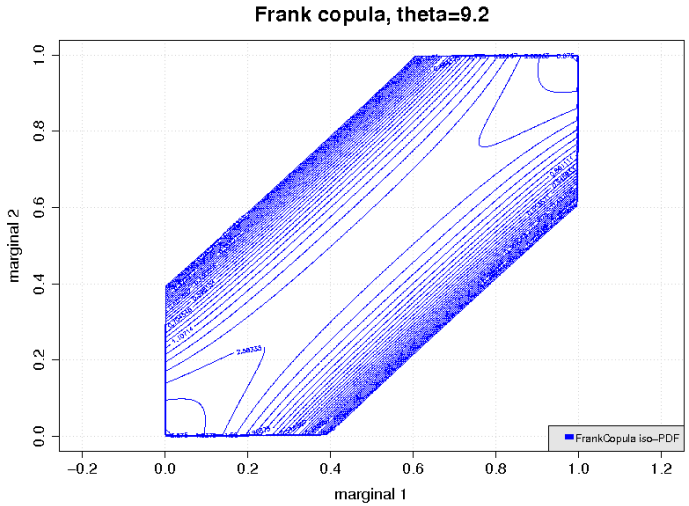
If the marginal distribution functions are continuous, then C is unique, else it is uniquely defined on $Im(F_1) \times \ldots \times Im(F_n)$.
When the marginal distribution functions are continuous, then

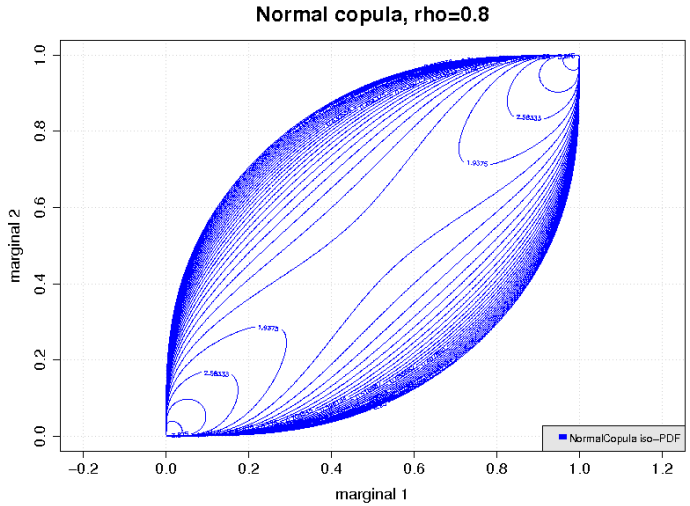
$$\forall \boldsymbol{u} \in [0,1]^n,\ C(u_1,\ldots,u_n) = F(F_1^{-1}(u_1),\ldots,F_n^{-1}(u_n)) \tag{10}$$

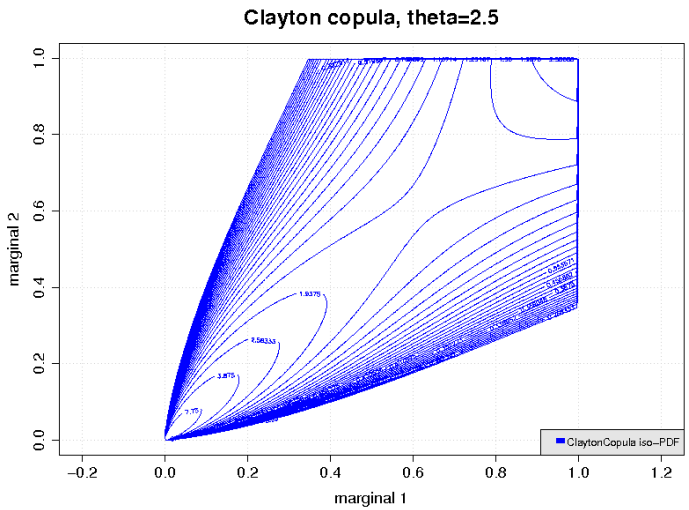
Some classical bi-dimensional copulas

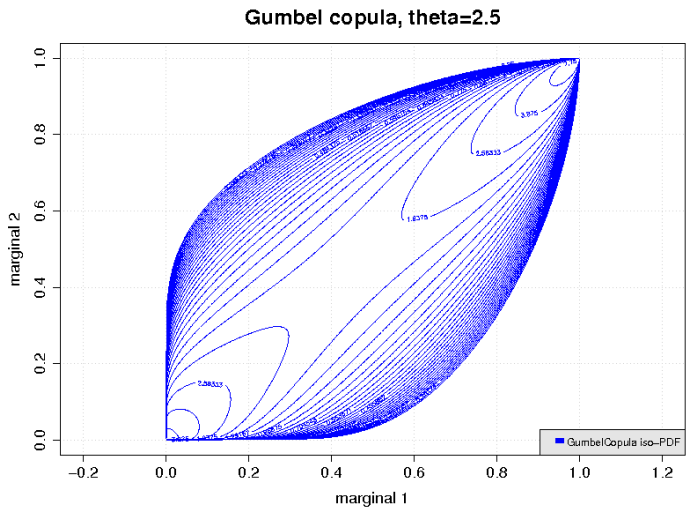
Name	$C(u_1, u_2)$
Independent	$u_1 u_2$
Normal	$\int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right) \mathrm{d}s\,\mathrm{d}t,\ \rho \in [-1,1]$
Frank	$-\frac{1}{\theta}\log\left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1}{e^{-\theta} - 1}\right),\ \theta \neq 0$
Clayton	$\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta},\ \theta \geq 0$
Gumbel	$\exp\left(-\left((-\log(u_1))^\theta + (-\log(u_2))^\theta\right)^{1/\theta}\right),\ \theta \geq 1$

Classical bi-dimensional iso-density









Measures of association

Definition

A measure of association r between the two components X_1 and X_2 of a bi-dimensional random vector $\boldsymbol{X}$ is a scalar function of X_1 and X_2 with the following properties:

1. $-1 \leq r(X_1, X_2) \leq 1$
2. If X_1 and X_2 are independent, then $r(X_1, X_2) = 0$
3. If g and h are strictly increasing functions, $r(X_1, X_2) = r(g(X_1), h(X_2))$.

The property (3) insures that r is a function of the copula C of $\boldsymbol{X}$ only.

It is the most general way to synthetize the full dependence information between two random variables into a single scalar.

Measures of concordance

Definition

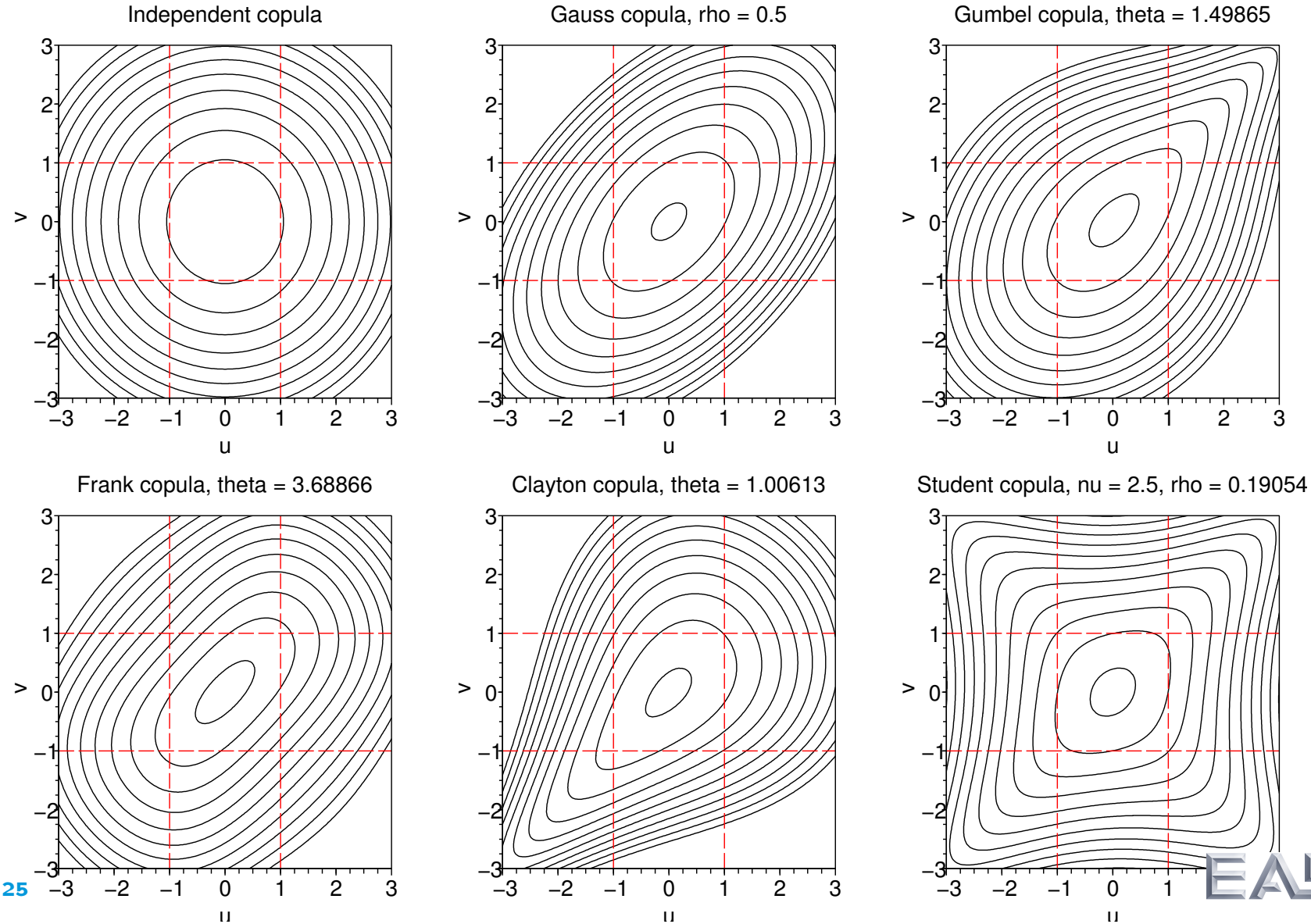
A measure of concordance κ between the two components X_1 and X_2 of a bi-dimensional random vector $\boldsymbol{X}$ is a scalar function of X_1 and X_2 that has the following properties:

1. κ is defined for all continuous bi-dimensional random vectors $\boldsymbol{X}$,
2. $\kappa(X_1, X_2) = \kappa(X_2, X_1)$,
3. κ is monotone in the copula $C_{\boldsymbol{X}}$ of $\boldsymbol{X}$, it means that if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two bi-dimensional random vectors with respective copulas $C_{\boldsymbol{X}}$ and $C_{\boldsymbol{Y}}$ and if $\forall \boldsymbol{u} \in [0,1]^2$, $C_{\boldsymbol{X}}(\boldsymbol{u}) \geq C_{\boldsymbol{Y}}(\boldsymbol{u})$, then $\kappa(X_1, X_2) \geq \kappa(Y_1, Y_2)$.
4. $\kappa(X_1, X_2) \in [-1, 1]$, $\kappa(X_1, X_1) = 1$, $\kappa(X_1, -X_1) = -1$,
5. if X_1 and X_2 are independent, then $\kappa(X_1, X_2) = 0$,
6. $\kappa(X_1, -X_2) = \kappa(-X_1, X_2) = -\kappa(X_1, X_2)$,
7. if C_n is a sequence of copulas that converges pointwise to the copula C, then $\kappa(C_n)$ converges pointwise to $\kappa(C)$, where $\kappa(C)$ is a shorthand for $\kappa(X_1, X_2)$, the support of $\boldsymbol{X}$ is $[0,1]^2$ and its distribution function restricted to this support is C.

Is a scalar measure enough to quantify the dependence?

- The short answer is NO. It ease quite easy to build bi-dimensional distribution functions with common marginal distribution functions and a common value for a given measure of association, but with very diffrent tail behaviour for example.
- If we are able to combine different such measures, then the answer is MAYBE.
- In any case, these measures are usefull if one is interested in a global quantification of the dependence, and it is also of first importance for the statistical parametric estimation of copulas.

A common measure of concordance is not enough to share the same dependence structure



A common measure of concordance is not enough to share the same dependence structure

Échantillon aleatoire

Un echantillon aleatoire de taille n est un vecteur aleatoire de dimension n tel que les variables aleatoires $X_1, \ldots, X_n$ soient independantes et aient même loi. Cela signifie:

$$F_{X_1} \equiv \cdots \equiv F_{X_n} \text{ et } F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1}(x_1) \times \cdots \times F_{X_n}(x_n) \tag{11}$$

Expectation of a random vector

Definition

Let $\boldsymbol{X}$ be an n-dimensional random vector with distribution function F. Its expectation $\mathbb{E}[\boldsymbol{X}]$, if it exists, is given by:

$$\mathbb{E}[\boldsymbol{X}] = \int_{\mathbb{R}^{\boldsymbol{n}}} \boldsymbol{x}\, F_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}) \tag{12}$$

In the case of a discret random vector, it rewrites:

$$\mathbb{E}[\boldsymbol{X}] = \sum_{i \in \boldsymbol{I}} \boldsymbol{x}_i\, \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}_i) \tag{13}$$

and in the case of an absolutely continuous one:

$$\mathbb{E}[\boldsymbol{X}] = \int_{\mathbb{R}^{\boldsymbol{n}}} \boldsymbol{x}\, p_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{14}$$

General moments

Definition

Let $\boldsymbol{X}$ be an n-dimensional random vector and ϕ a measurable function from $\mathbb{R}^n$ into $\mathbb{R}^p$, i.e such that $\phi(\boldsymbol{X})$ is a p-dimensional random vector defined on the same probability space than $\boldsymbol{X}$.
The general moment of $\boldsymbol{X}$ with respect to ϕ is the expectation of $\boldsymbol{Y} = \phi(\boldsymbol{X})$.

The variance of a random variable is obtained with $\phi(u) = (u - \mathbb{E}[X])^2$, and the generic element of the covariance matrix of a random vector is given by $\phi(\boldsymbol{u}) = (u_i - \mathbb{E}[X_i])(u_j - \mathbb{E}[X_j])$.

Almost sure convergence

Definition

A sequence of n-dimensional random vectors $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ all defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ converges almost surely if the set:

$$\{\omega \in \Omega \,|\, (\boldsymbol{X}_n(\omega))_{n\in\mathbb{N}} \text{ converges}\} \tag{15}$$

has a probability 1. If we note by $\boldsymbol{X}_\infty(\omega)$ the limit of $(\boldsymbol{X}_n(\omega))_{n\in\mathbb{N}}$ when it exists, it defines a random vector on $(\Omega, \mathcal{F}, \mathbb{P})$ called the almost sure limit of $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$:

$$\boldsymbol{X}_n \xrightarrow{a.s} \boldsymbol{X}_\infty \tag{16}$$

This convergence means that for n large enough, both $\boldsymbol{X}_n$ and $\boldsymbol{X}_\infty$ will take the same value on a given random experiment.

Convergence in distribution

Definition

A sequence of n-dimensional random vectors $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ possibly defined on different probability spaces $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ converges in distribution if for all g in the set $C_b^0(\mathbb{R}^n, \mathbb{R})$ of bounded continuous functions defined on $\mathbb{R}^n$ and taking value into $\mathbb{R}$ we have:

$$\lim_{n\to\infty} \mathbb{E}[g(\boldsymbol{X}_n)] = \mathbb{E}[g(\boldsymbol{X}_\infty)] \tag{17}$$

This convergence is denoted by $\boldsymbol{X}_n \xrightarrow{\mathcal{D}} \boldsymbol{X}_\infty$.

Theorem

A sequence of n-dimensional random vectors $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ converges in distribution if and only if the sequence of distribution functions $(F_n)_{n\in\mathbb{N}}$ converges pointwise to a function F for all points $\boldsymbol{x} \in \mathbb{R}^n$ at which F is continuous.

Strong law of large numbers

Theorm

Let $(\boldsymbol{X})_{n\in\mathbb{N}}$ be a sequence of n-dimensional random vectors defined over the same probability space, independent and with the same distribution. For all measurable functions $f \in \{(\mathbb{R}^n, \mathbb{R}^p)$ such that $\mathbb{E}[|f(\boldsymbol{X}_1)|] < \infty$ exists, we have:

$$\frac{1}{n}\sum_{k=1}^{n} f(\boldsymbol{X}) \xrightarrow{a.s} \mathbb{E}[f(\boldsymbol{X}_1)] \tag{18}$$

Central Limit Theorem

Theorem

Under the hypotheses of the strong law of large numbers, if the covariance matrix of $\boldsymbol{X}_1$ axists and is finite, then:

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} f(\boldsymbol{X}) - \mathbb{E}[f(\boldsymbol{X}_1)]\right) \xrightarrow{\mathcal{D}} \boldsymbol{X}_\infty \tag{19}$$

where the random vector $\boldsymbol{X}_\infty$ is distributed according to the n-dimensional normal distribution $\mathcal{N}(\mathbf{0}, \mathrm{Cov}\,[\boldsymbol{X}_1])$.

Outline

Sampling model

Definition

Let $\boldsymbol{X}$ be an n-dimensional random vector. A statistical sample S_N of size N associated with $\boldsymbol{X}$ is a collection $(\boldsymbol{X}_i)_{i\in\{1,\dots,N}$ of independent random vectors all defined on the same probability space than $\boldsymbol{X}$, independent and sharing the same distribution than $\vec{X}$ (they are N independent copies of $\boldsymbol{X}$).
The distribution function of S_N is the distribution function of the nN dimensional random vector obtained by stacking $\boldsymbol{X}_1,\dots,\boldsymbol{X}_N$:

$$F_{S_N}(\boldsymbol{X}_1,\dots,\boldsymbol{X}_N) = \prod_{i=1}^{N} F_{\boldsymbol{X}_i}(\boldsymbol{x}_i) \tag{20}$$

Parametric model

Definition

A parametric model $\mathcal{L}_{\boldsymbol{\theta}}$ is a collection of probability distributions indexed by a vector $\boldsymbol{\theta} \in \mathcal{O} \subset \mathbb{R}^q$. If different values of $\boldsymbol{\theta}$ lead to different distributions (i.e the function $\boldsymbol{\theta} \mapsto \mathcal{L}_{\boldsymbol{\theta}}$ is one-to-one), then the model is said to be identifiable.

Likelihood function

Let S_N be a statistical sample S_N of size N associated with a random vector $\boldsymbol{X}$ whose distribution is a member of the parametric model $\mathcal{L}_{\boldsymbol{\theta}}$. Its likelihood function is the function L defined on $\mathcal{O}$ and taking value in $\mathbb{R}^+$ such that:

$$L_x(\boldsymbol{\theta}) = \prod_{i=1}^{N} p_{\boldsymbol{X}_i}(x_i; \boldsymbol{\theta}) \tag{21}$$

where $p_{\boldsymbol{X}_i}$ is the probability density function or the probability distribution function depending on the continuous or the discrete nature of $\boldsymbol{X}$.

Estimation of distributions

Depending on the nature and the quantity of the available information about the distribution of a random vector, the estimation of its distribution can be based on:

- a statistical treatment of the available data if thay are in sufficient quantity and quality,
- a judgement of experts, who fully prescribe the distribution,
- a maximum entropy principle, to build the less informative distribution that integrate the partial knowledge available on the target distribution.

but in real-life, things are less clear-cut:

- Even a large amount of data can be usefully completed by an expert judgement, leading to the parametric estimation approach,
- There is no judgement of expert emanating from nowhere. The associated data, even scarce, are precious and should be integrated into the estimation process.

Estimation with data

A two-steps approach: estimation and validation

If data $(x_1, \ldots, x_N)$ are available, one can use the classical statistical tools to quantify the distribution of interest. The data are seen as the realization of a sample of size N of the random vector of interest $\boldsymbol{X}$ and we perform the following two steps:

Step 1: estimation of the distribution, which can be either:

- a parametric estimation: given the hypothesis that the target distribution is a member of a parametric family of distributions $\mathcal{L}_\theta$, one uses the data to compute the best estimation of θ. The main methods are the maximum likelihood estimator and the moment-based estimator.
- a non-parametric estimation: the whole shape of the distribution function is obtained without a priori hypothesis on the target distribution. The main methods are the empirical distribution function, the histogram-based estimation and the kernel smoothing estimation.

Step 2: validation of the fitting, which can be either:

- A fitting test checking the hypothesis made on the distribution,
- A graphical validation

Parametric estimation

Estimator

An estimator $\hat{\boldsymbol{\Theta}}_N$ of the parameters of a given parametric model $\mathcal{L}_{\boldsymbol{\theta}}$ is a random vector build as a function of the sample model S_N: $\hat{\boldsymbol{\Theta}}_N = \psi(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$.
The estimated value $\hat{\boldsymbol{\theta}}_N$ of the parameter $\boldsymbol{\theta}$ is the value taken by the estimator $\hat{\boldsymbol{\Theta}}_N$ when the realization of S_N is equal to the observed data:

$$\hat{\boldsymbol{\theta}}_N = \psi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \tag{22}$$

If $\forall N$, $\mathbb{E}[\hat{\boldsymbol{\theta}}_N] = \boldsymbol{\theta}$ the estimator is unbiased.

Consistent estimator

An estimator $\hat{\boldsymbol{\Theta}}_N$ of the parameters of a given parametric model $\mathcal{L}_{\boldsymbol{\theta}}$ is consistent if and only if it converges almost surely to the value $\boldsymbol{\theta}$ of the target distribution:

$$\hat{\boldsymbol{\Theta}}_N \xrightarrow{a.s} \boldsymbol{\theta} \tag{23}$$

A consistent estimator is such that for N large enough, any realization of $\hat{\boldsymbol{\Theta}}_N$ will have a value close to the target value $\boldsymbol{\theta}$, in particular when this realization is equal to the observed data.

Maximum likelihood estimator

Definition

Let $\mathcal{L}_\theta$ be a parametric model and S_N the associated sample of size N. The maximum likelihood estimator of θ is the value of θ (supposed to be unique) that maximizes the likelihood function of S_N given S_N:

$$\hat{\Theta}_N = \underset{\theta \in \mathcal{O}}{\text{Argmax}}\, L_{X_1,\ldots,X_N}(\theta) \tag{24}$$

Theorem

Under the following hypotheses:

1. The model is identifiable;
2. The function $(x, \theta) \mapsto L_x(\theta)$ is bounded;
3. The function $\theta \mapsto L_x(\theta)$ is continuous;
4. The expectation $\mathbb{E}[\log L_{X_1,\ldots,X_N}(\theta)]$ exists for all $\theta \in \mathcal{O}$;

the maximum likelihood estimator is consistent.

Moments estimator

Definition

Let $\mathcal{L}_{\boldsymbol{\theta}}$ be a parametric model, m be a continuous invertible function from $\mathcal{O}$ into $\mathcal{O}$ and ϕ be a measurable function such that $\mathbb{E}_{\boldsymbol{\theta}}[|\phi(\boldsymbol{X}_1)|] < \infty$ and $m(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\phi(\boldsymbol{X}_1)]$. Then the moments estimator of $\boldsymbol{\theta}$ is defined by:

$$\hat{\boldsymbol{\Theta}}_N = m^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\phi(X_i)\right) \tag{25}$$

In practice, one choose ϕ and check that the resulting function m is continuous and invertible. A common choice is to take $\phi(\boldsymbol{x}) = (x_i, x_i x_j, x_i x_j x_k, \ldots)$.

Theorem

The moments estimator is always convergent.

Examples

Parametric model	Maximum likelihood	Moments
$\mathcal{N}(\mu, \sigma)$	$\mu_N = 1/N \sum x_i$	identical
	$\sigma_N^2 = 1/N \sum (x_i - \mu_N)^2$	identical
$\mathcal{E}xp(\lambda)$	$\lambda_N = N/\sum x_i$	identical
$Unif(a, b)$	$a_N = min(x_i)$	$a_N = \mu_N - \sqrt{3}\sigma_N$
	$b_N = max(x_i)$	$b_N = \mu_N + \sqrt{3}\sigma_N$

Asymptotic normality

Definition

An estimator $\hat{\boldsymbol{\Theta}}_N$ is asymptoticaly normal if it converges in distribution when $N \to +\infty$ to a normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$:

$$\sqrt{N}\left(\hat{\boldsymbol{\Theta}}_N - \boldsymbol{\theta}\right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{26}$$

Confidence region

Let $\hat{\boldsymbol{\Theta}}_N$ be a consistent and asymptoticaly normal estimator of $\boldsymbol{\theta}$. Then the following random ellipsoid:

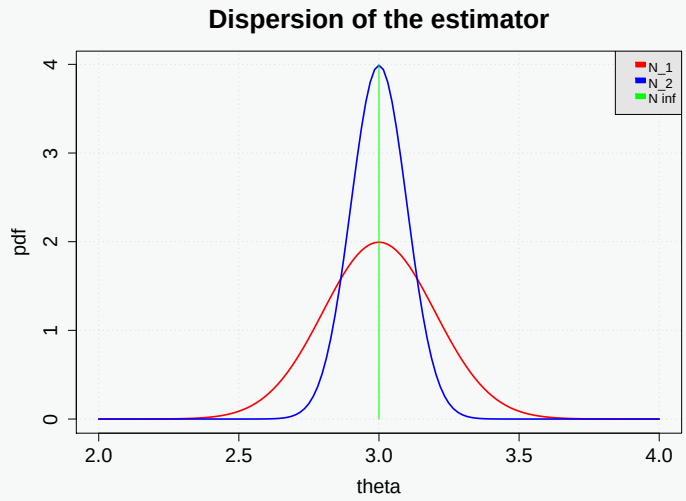
$$E_\alpha = \{\boldsymbol{u} \in \mathbb{R}^q \,|\, (\boldsymbol{u} - \hat{\boldsymbol{\Theta}}_N)' \boldsymbol{\Sigma} (\boldsymbol{u} - \hat{\boldsymbol{\Theta}}_N) \leq a_\alpha / N\} \tag{27}$$

where a_α is the $(1-\alpha)$-quantile of the $\chi^2(q)$ distribution is such that:

$$\mathbb{P}(\theta \in E_\alpha) \to 1 - \alpha \tag{28}$$

Interpretation

Dispersion of the estimator as a function of the sample size N

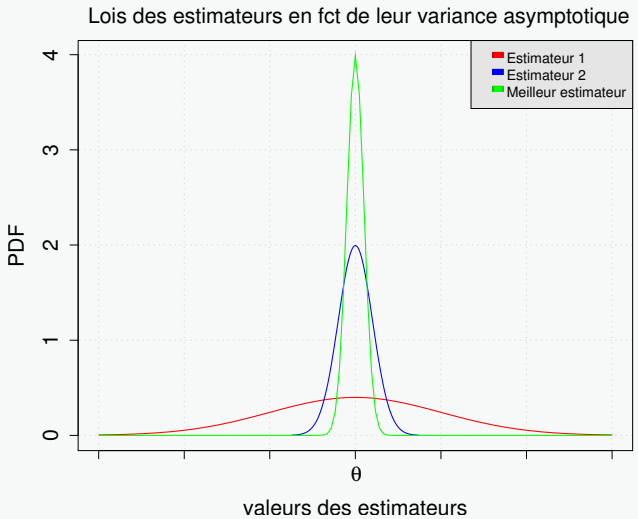


$$\begin{cases} N_1 \text{ large}, & \hat{\Theta}_{N_1} \sim \mathcal{N}(\theta, \sigma^2/N_1) \\ N_2 \geq N_1, & \hat{\Theta}_{N_2} \sim \mathcal{N}(\theta, \sigma^2/N_2) \\ N = \infty, & \hat{\Theta}_{\infty} \equiv \theta \end{cases}$$

A larger sample leads to less dispersion for the estimator: the estimated value is more likely close to the true value of the parameter

Interpretation

Precision of the estimators for a fixed sample size N



$$\begin{cases} \hat{\Theta}_N^1 \sim \mathcal{N}(\theta, \sigma_1^2/N) \\ \hat{\Theta}_N^2 \sim \mathcal{N}(\theta, \sigma_2^2/N) \\ \hat{\Theta}_N^{opt} \sim \mathcal{N}(\theta, \sigma_{opt}^2/N) \end{cases}$$

For a given sample size, different estimators can have significant differences in their asymptotic variance.

Theorem

For unbiased estimators, the minimal possible variance is $\boldsymbol{I}^{-1}(\theta)$ where the matrix $\boldsymbol{I}(\theta) = -\mathbb{E}_\theta[\partial^2 p_\theta(x)/\partial\theta^2]$ is Fisher's information matrix.

An unbiased estimator with such a covariance matrix is an Asymptoticaly efficient estimator. For a given sample size, it will give the smallest confidence region.

Estimators comparison

Main properties

Estimator	Asympt. Normality	Asympt. Efficiency	Unbiased
Mx. Likelihood*	yes	yes	no in general
Moments	yes	no in general	no in general

* Under additional regularity conditions (support independent of θ).

Confidence interval associated with the maximum likelihood estimator

Parametric model	Asymptotic distribution	Confidence interval
$\mathcal{N}(\mu,\sigma)$	$\sqrt{N}(\lambda_N-\lambda)\rightarrow\mathcal{N}(0,I(\lambda)^{-1}=\lambda^2)$	$[\lambda_N-\frac{a_\alpha}{\sqrt{N}}\lambda_N\,;\,\lambda_N+\frac{a_\alpha}{\sqrt{N}}\lambda_N]$
$\mathcal{E}xp(\lambda)$	$\sqrt{N}(\theta_N-\theta)\rightarrow\mathcal{N}(0,I(\theta)^{-1}$	$[\mu_N-\frac{a_\alpha}{\sqrt{N}}\mu_N\,;\,\mu_N+\frac{a_\alpha}{\sqrt{N}}\mu_N]$
		$[\sigma^2_N-\frac{a_\alpha}{2\sqrt{N}}\sigma^2_N\,;\,\sigma^2_N+\frac{a_\alpha}{2\sqrt{N}}\sigma^2_N]$
$Unif(a,b)$	Regularity conditions not fulfilled	
	$N(a_N-a)\rightarrow\mathcal{E}xp(1/(b-a))$	$[a_N\,;\,a_N+\frac{\log(1-\alpha)}{N}(b_N-a_N)]$
	$N(b-b_N)\rightarrow\mathcal{E}xp(1/(b-a))$	$[b_N-\frac{\log(1-\alpha)}{N}(b_N-a_N)\,;\,b_N]$

where a_α is the $(1+\alpha)/2$ quantile of $\mathcal{N}(0,1)$.

Non-parametric estimation

The objective is to estimate the density function from a given sample S_N of $\boldsymbol{X}$.

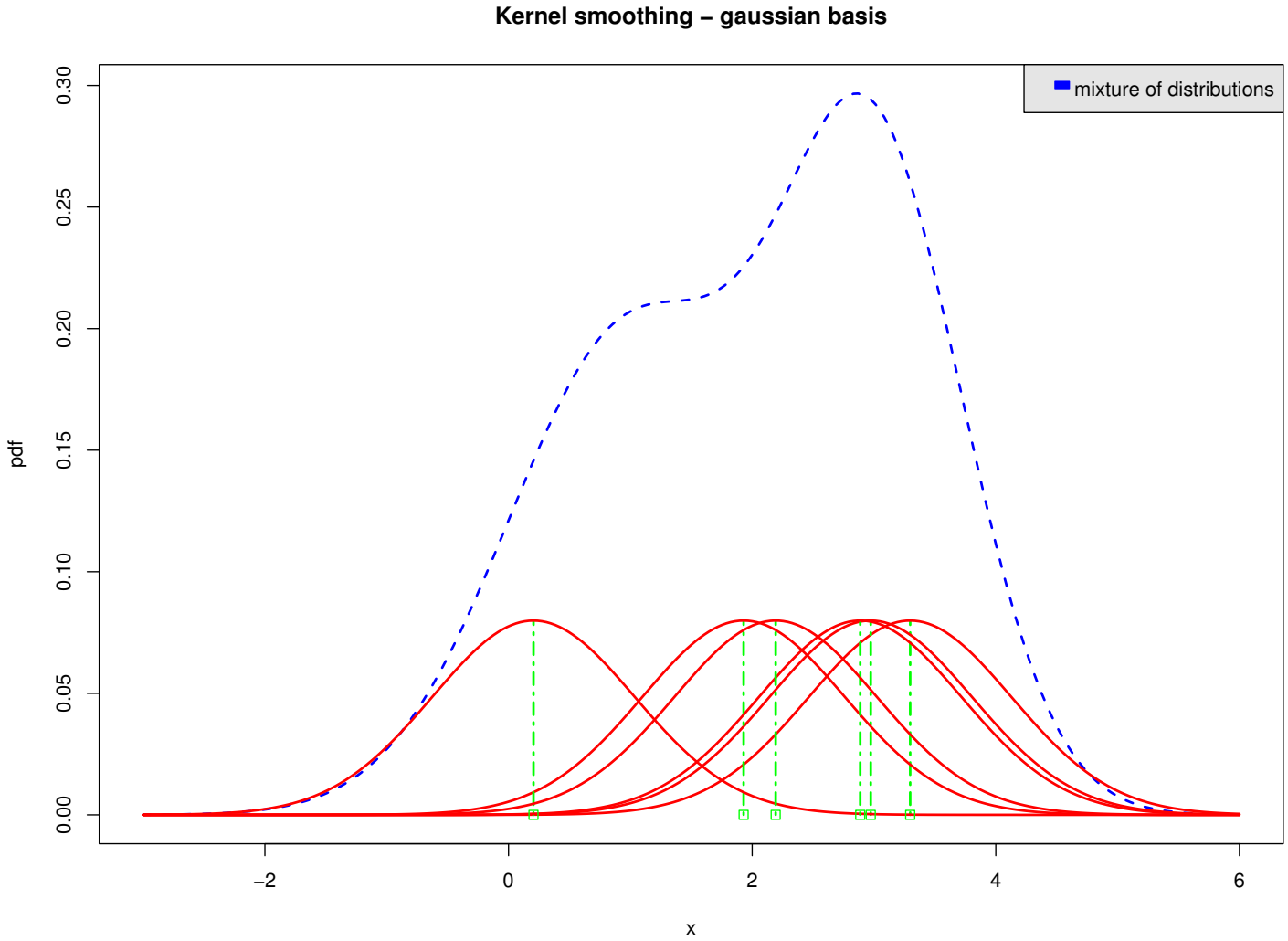
Kernel smoothing: the uni-dimensional case

The density function p is estimated by the random function $\hat{p}_N$ given by:

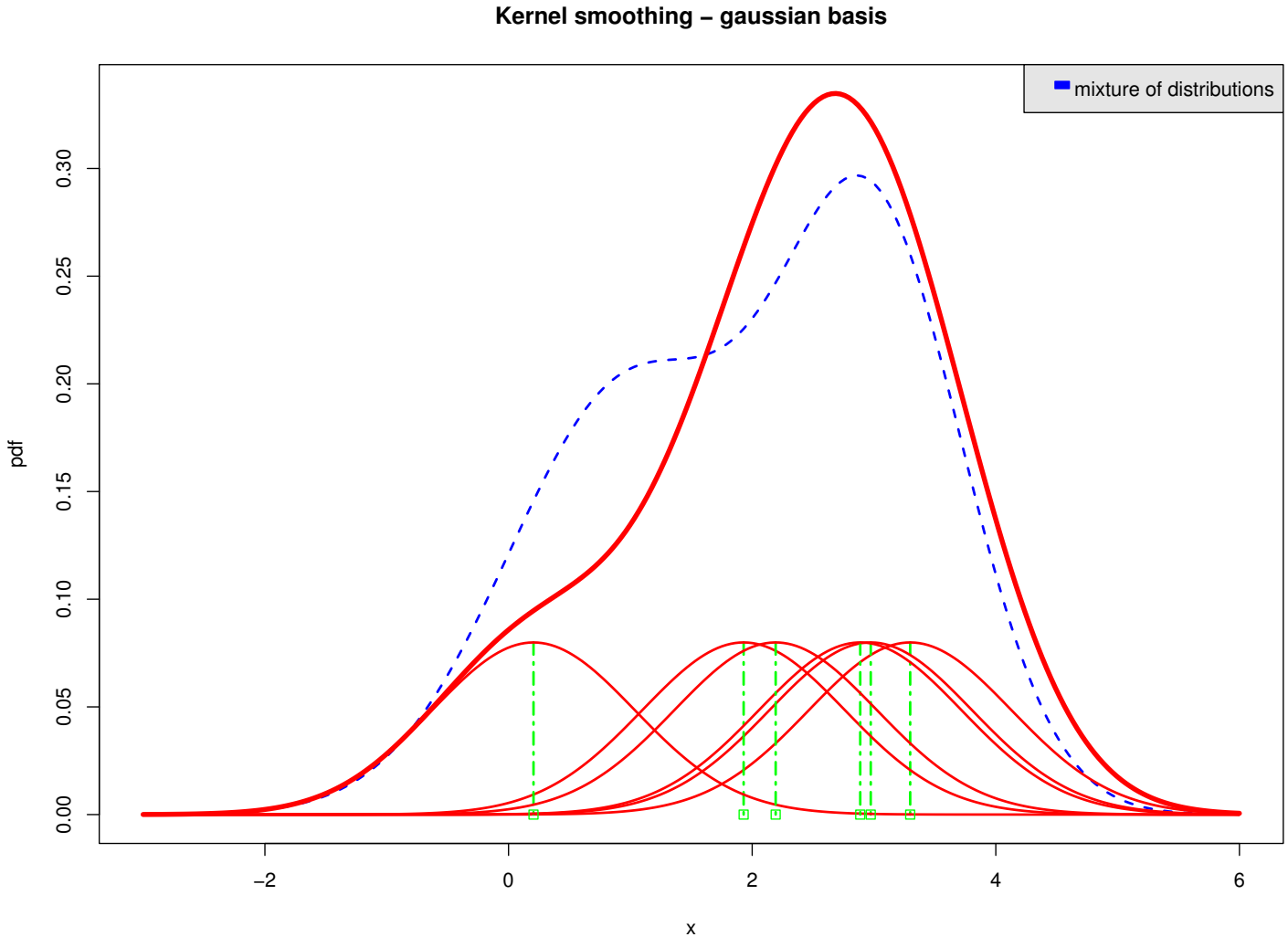
$$\hat{p}_N(x) = \frac{1}{Nh} \sum_{i=1}^{i=N} K\left(\frac{x - X_i}{h}\right) \tag{29}$$

where K is a symmetric density function called the kernel (e.g $\mathcal{N}(0,1)$, $\mathcal{U}(-1,1)...$) and $h > 0$ is a scalar parameter called the bandwidth.

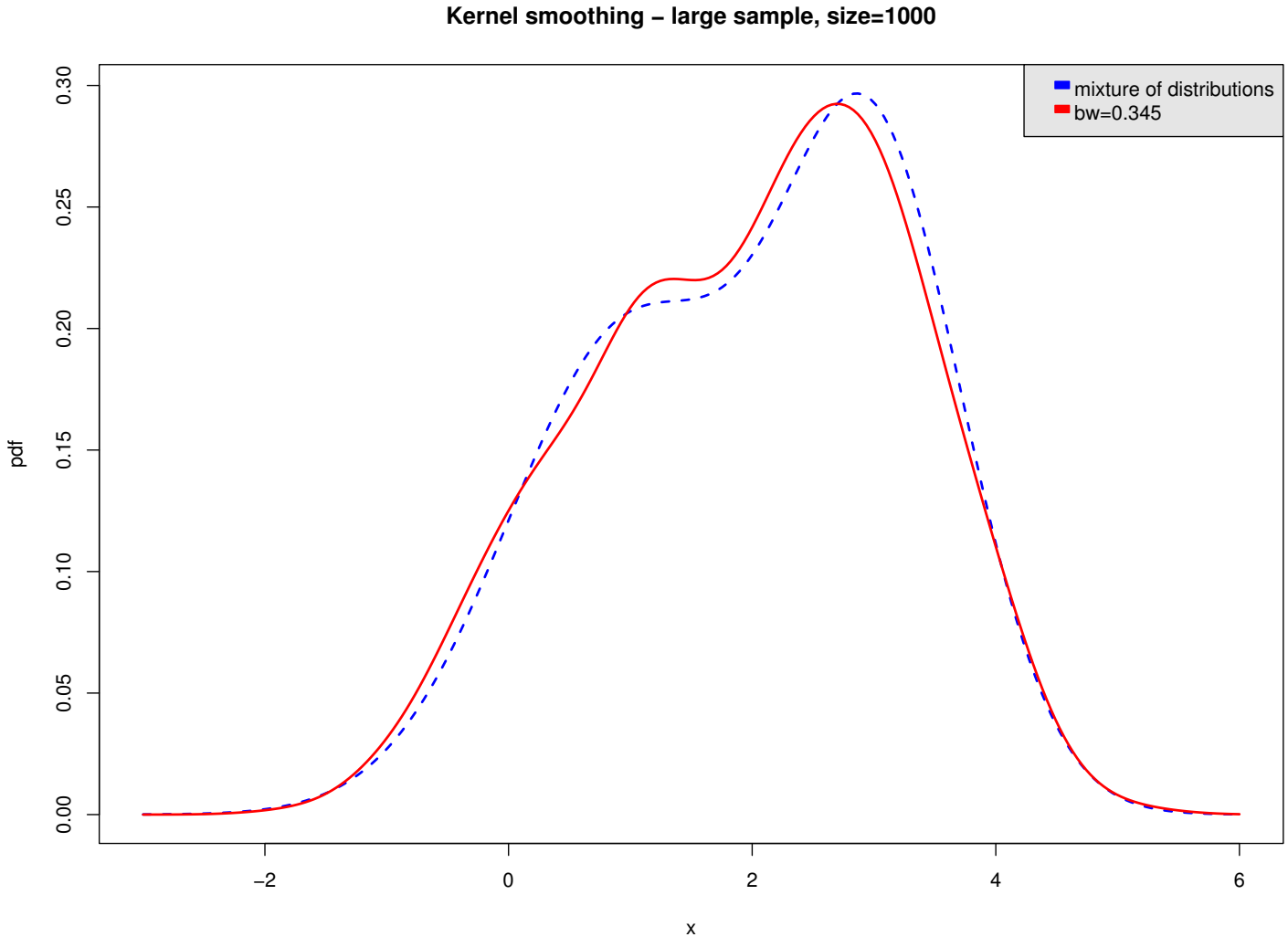
Principle of the kernel smoothing: put a scaled version of the kernel at each sample point



Principle of the kernel smoothing: average all the scaled versions of the kernel



Large sample approximation



Kernel smoothing

Kernel and bandwidth selection

The modeling error is quantified using the Asymptotic Mean Integrated Squared Error (AMISE) built this way:

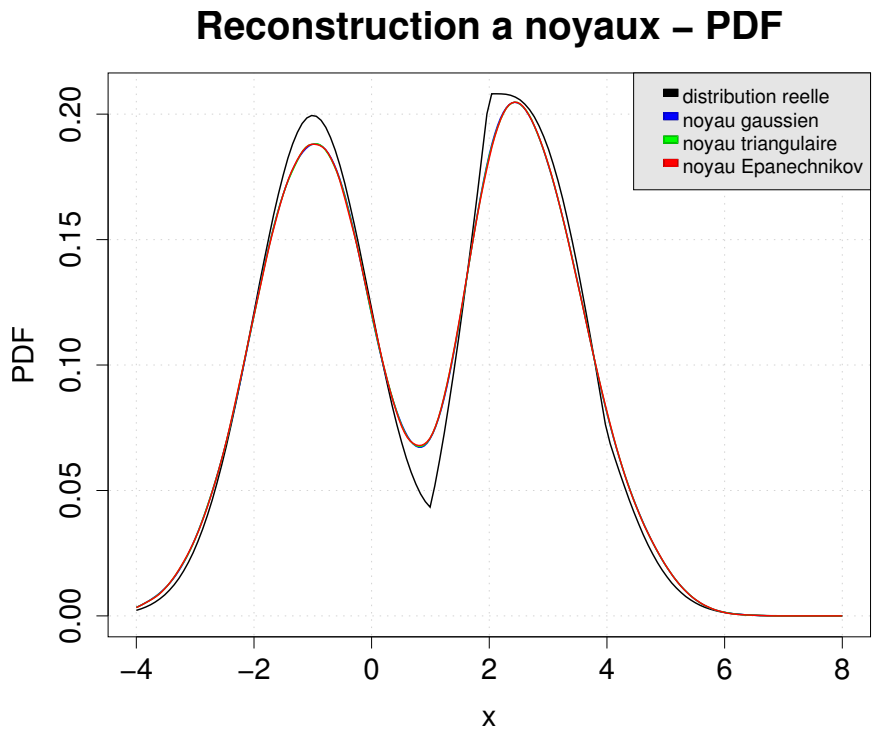
- The Mean Squared Error is defined by $MSE(\hat{p}_N, x) = (\mathbb{E}[\hat{p}_N(x)] - p(x))^2 + Var[\hat{p}_N(x)]$, also called the quadratic risk of $\hat{p}_N(x)$;
- The Mean Integrated Square Error is defined by $MISE(\hat{p}_N) = \int_{\mathbb{R}} MSE(\hat{p}_N, x)\,\mathrm{d}x$
- The Asymptotic Mean Integrated Squared Error is defined as being equal to the two first terms of the asymptotic expansion of $MISE(\hat{p}_N)$ with respect to N.

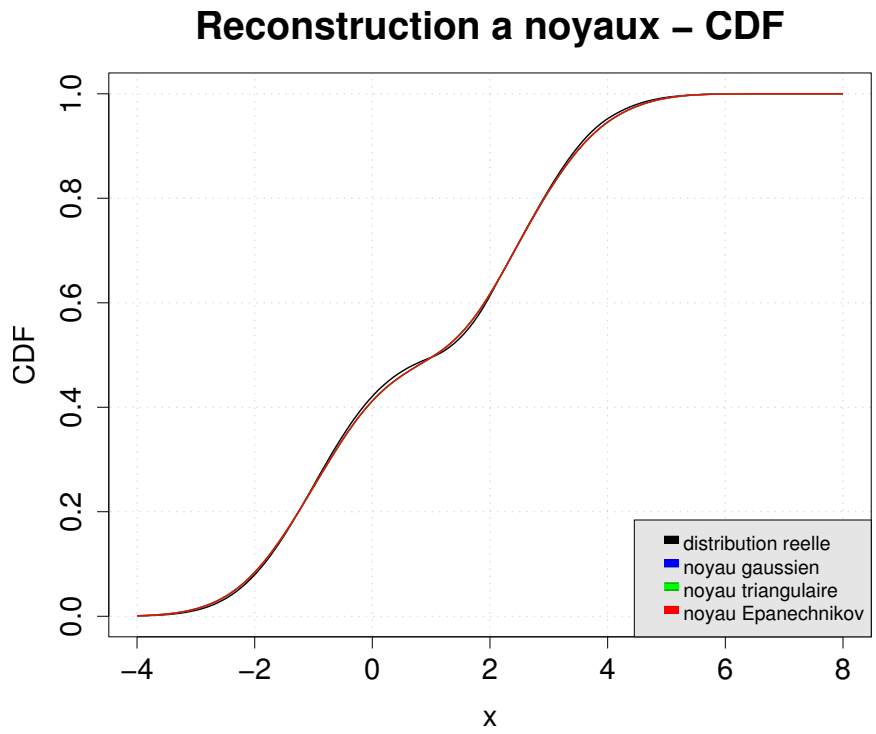
The choice of K is not crucial, but the choice of h is crucial. For large h, the data are oversmoothed while for small values of h they are undersmoothed. In the uni-dimensional case, the optimal choice for h is, according to the AMISE minimization:

$$h_{AMISE}(K) = \left[\frac{R(K)}{\mu_2(K)^2 R(p'')}\right]^{\frac{1}{5}} N^{-\frac{1}{5}} \tag{30}$$

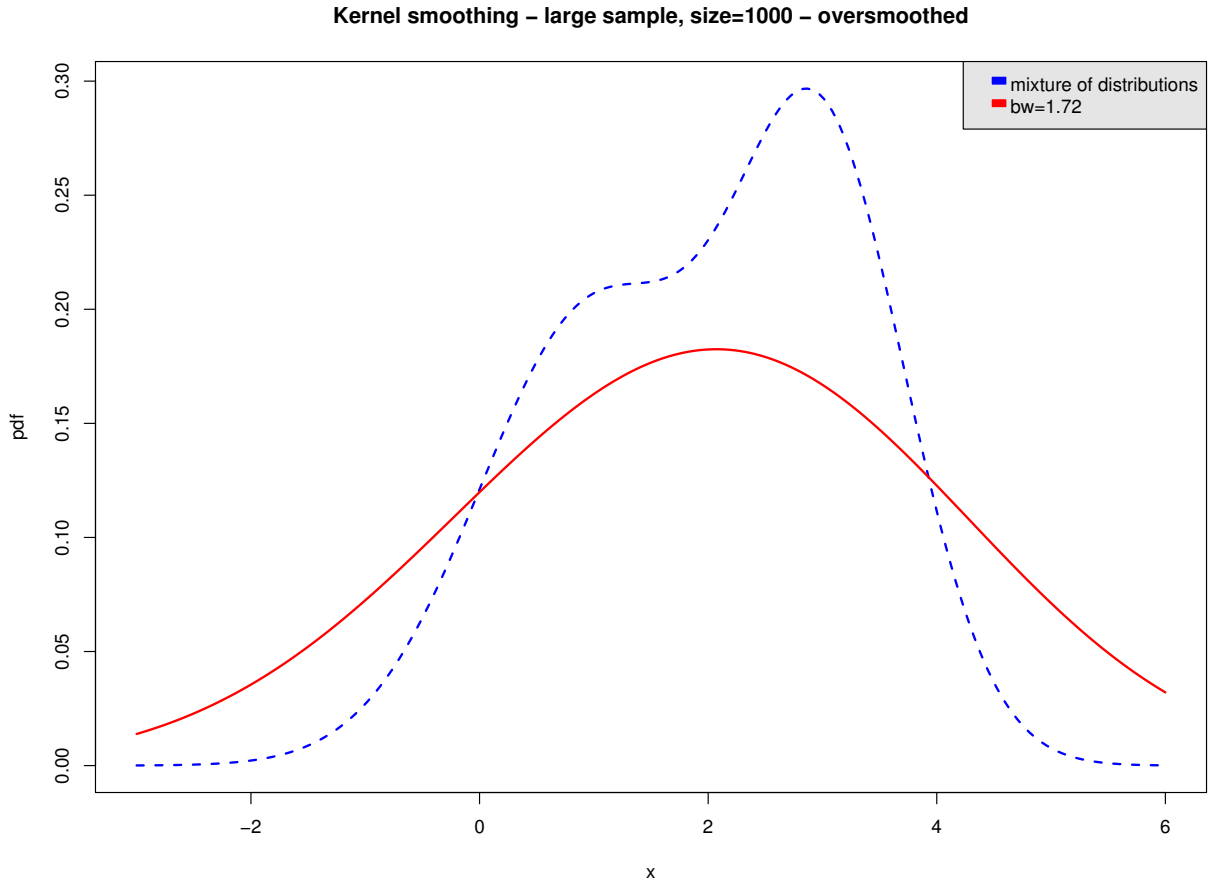
where $R(\psi) = \int_{\mathbb{R}} \psi^2(x)\,\mathrm{d}x$ and $\mu_2(K) = \int x^2 K(x)\,\mathrm{d}x = \sigma_K^2$.
The value of $R(p'')$ is unknown, and the different bandwidth selection rules correspond to different ways to estimate this quantity: Silverman's rule, Scott's rule, Solve-the-equation plug-in rule.
The main interest of the kernel smoothing approach is to be model-free (i.e non-parametric): even exotic density shapes can be consistently approximated such as multimodal densities.

Impact of the kernel

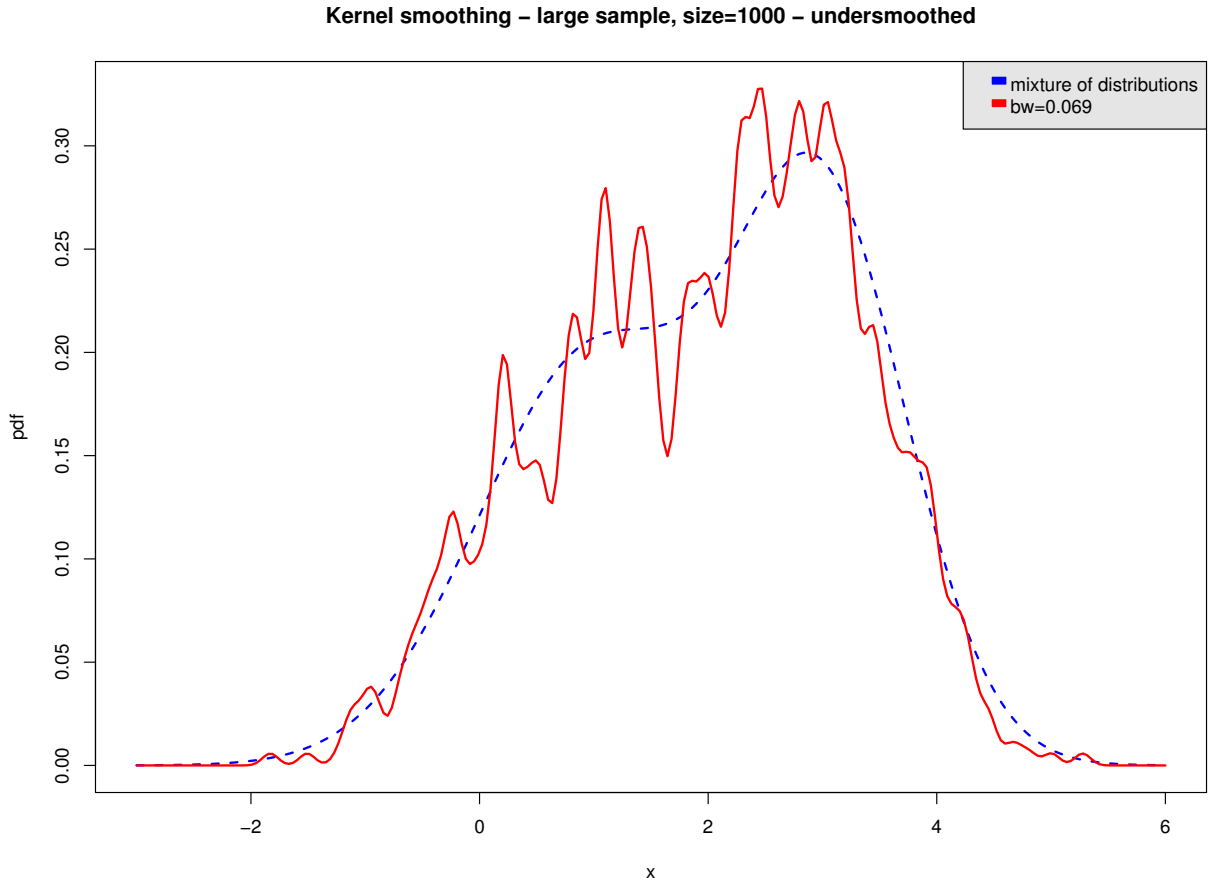




Impact of the bandwidth



Impact of the bandwidth



Silverman's rule

When p is the density of a normal distribution $\mathcal{N}(0, \sigma^2)$, $R(p)$ is explicitely known and we get:

$$h_{AMISE}^{p=normal}(K) = \left[\frac{8\sqrt{\pi}R(K)}{3\mu_2(K)^2}\right]^{\frac{1}{5}} \sigma N^{-\frac{1}{5}} \tag{31}$$

An estimator $\hat{h}$ of $h_{AMISE}^{p=normal}(K)$ is obtained using an estimator $\hat{\sigma}_N^2$ of σ^2 using $(X_1, \ldots, X_n)$.
Silverman's rule is to choose $h = \hat{h}$ even if p is not normal:

$$h^{Silver}(K) = \left[\frac{8\sqrt{\pi}R(K)}{3\mu_2(K)^2}\right]^{\frac{1}{5}} \hat{\sigma}_N N^{-\frac{1}{5}} \tag{32}$$

This rule is an heuristic that gives good results as soon as the target distribution is almost symmetric and unimodal.

Scott's rule

Scott's rule is an approximation of Silverman's rule resulting from the remark that for a normal kernel K, one has :

$$h^{Silver}(K) \simeq \hat{\sigma}_N N^{-\frac{1}{5}} \tag{33}$$

and for all the efficient kernels, $\sigma_K R(K) \simeq 1 \ \forall K$, which keads to:

$$\frac{h_{AMISE}(K_1)}{h_{AMISE}(K_2)} = \frac{\sigma_{K_2}}{\sigma_{K_1}} \left[\frac{\sigma_{K_1} R(K_1)}{\sigma_{K_2} R(K_2)}\right]^{\frac{1}{5}} \simeq \frac{\sigma_{K_2}}{\sigma_{K_1}} \tag{34}$$

Taking $K_2 = N(0,1)$, one get :

$$h^{Silver}(K) \simeq h^{Silver}(K_2) \frac{1}{\sigma_K} \tag{35}$$

Scott's rule is to take $h^{Silver}(K)$ with the approximations (33) and (35) even if p is not a normal density:

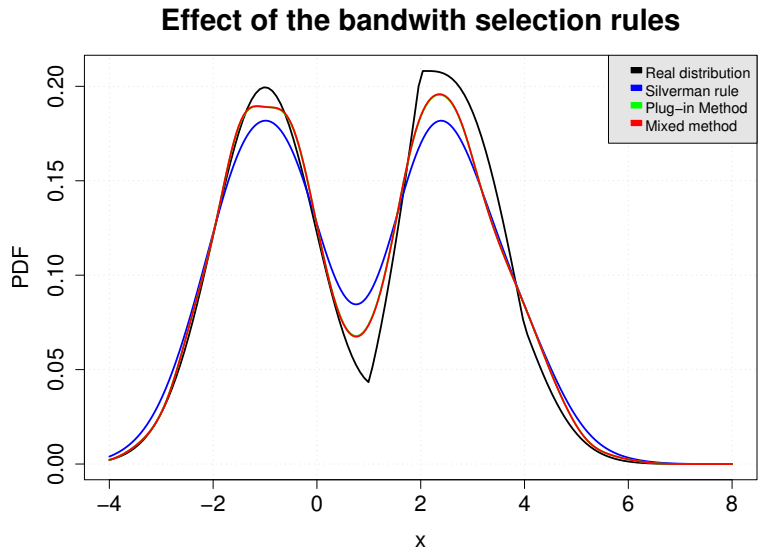
$$h^{Scott} = \frac{\hat{\sigma}_N}{\sigma_K} N^{-\frac{1}{5}} \tag{36}$$

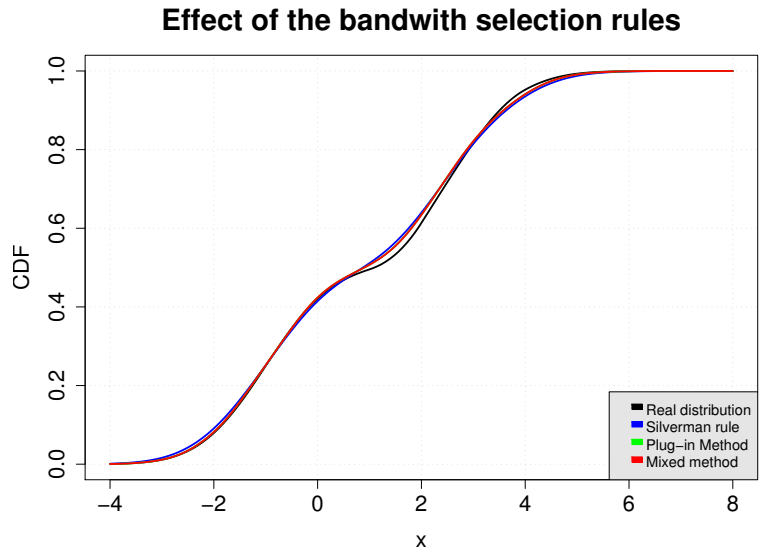
Scott's rule has the same efficiency than Silverman's rule, while being simpler to use.

Solve-the-equation plug-in rule

The method is based on a non-parametric estimation of $R(p'')$ using a further step of kernel smoothing. The key point is that the optimal bandwidth for a non-parametric estimation of $R(p'')$ is different from the optimal bandwidth for the estimation of p.
This new optimal bandwidth is computed assuming a normal density for p, and the AMISE criterion is replaced by a sampling version that involves to consider all the pairs (X_i, X_j) in the sample. The cost of this method is significantly higher than the cost of the preceeding rules, but it's performances are largely superior.

Comparison of the bandwidth selection rules



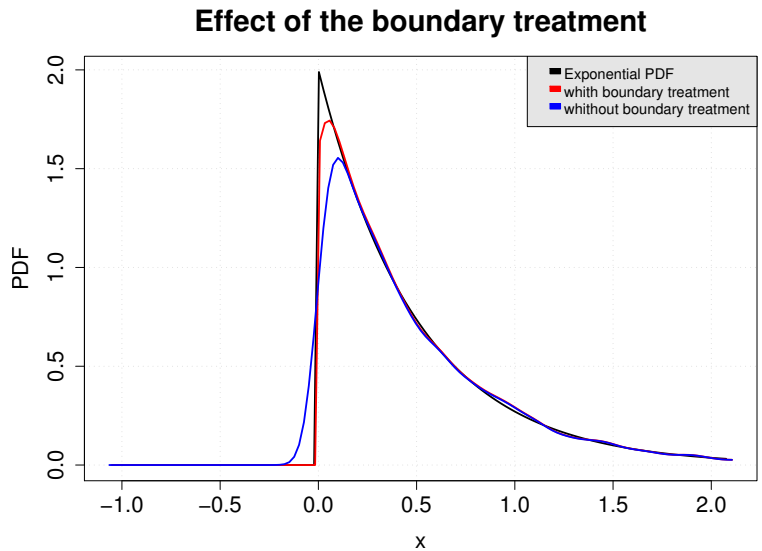


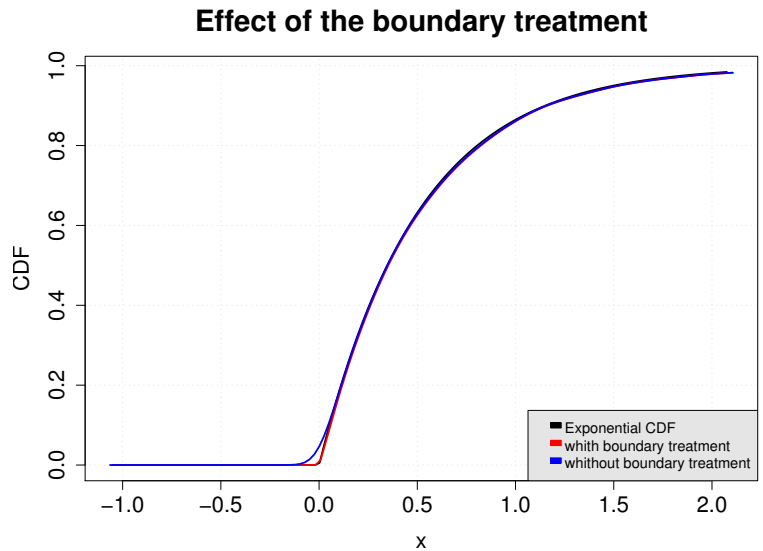
Border effect

When the target density p or its derivative p' have discontinuities, which occurs frequently when p has a bounded support (i.e when p is zero outside of a compact interval), then the kernel smoothing approximation converges to the mid-point of the discontinuity and the local rate of convergence of the approximation is reduced.
A cheap first order correction, the mirroring technique, allows to make up this loss of performance:

- The bandwidth h is estimated using the initial sample;
- The data at a distance less than h to a boundary are reflected with respect to the boundary;
- The density is estimated using the enlarged sample;
- The final estimation is obtained by truncation of the previous estimator to the suppport.

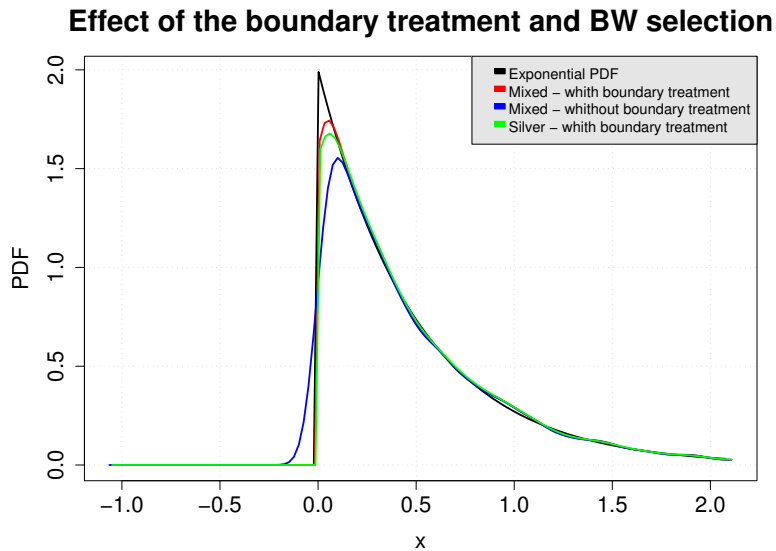
Border effect

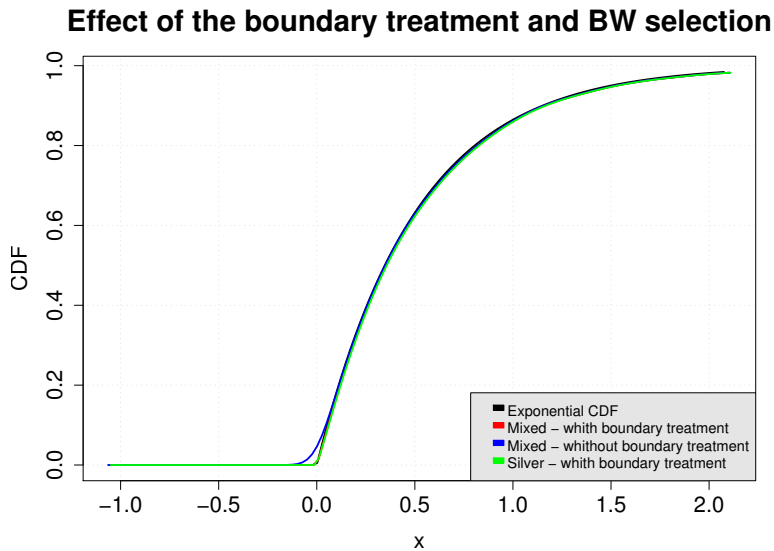




Ranking of the kernel smoothing parameters

The border effect, if present, is most important than the bandwidth selection rule, which is in turn more important than the kernel choice.





Multi-dimensional kernel smoothing

In dimension n, one uses a product kernel K_n associated with a given one-dimensional kernel K:

$$K_n(\boldsymbol{x}) = \prod_{j=1}^{n} K(x_j)$$

which leads to the following density estimation:

$$\hat{p}_N(\boldsymbol{x}) = \frac{1}{N \prod_{j=1}^{n} h_j} \sum_{i=1}^{N} K_n\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_n - X_{in}}{h_n}\right)$$

The multi-dimensional bandwidth $\boldsymbol{h} = (h^1, \cdots, h^n)$ can be estimated using the **multi-dimensional Scott's rule** :

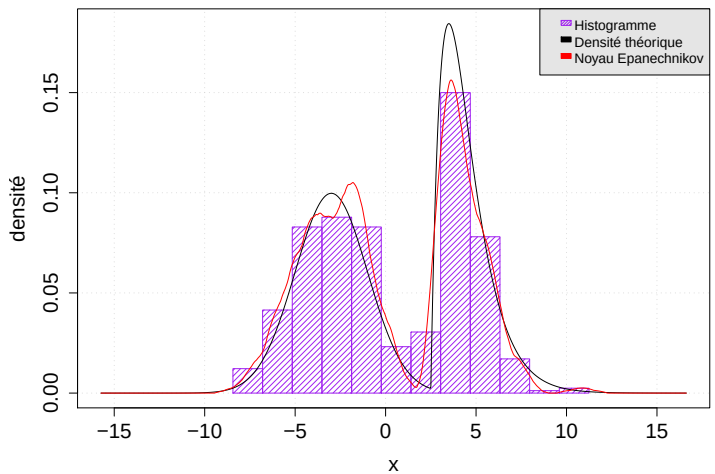
$$h^i_{Scott} = \frac{\hat{\sigma}^i_N}{\sigma_K} N^{-1/(d+4)}$$

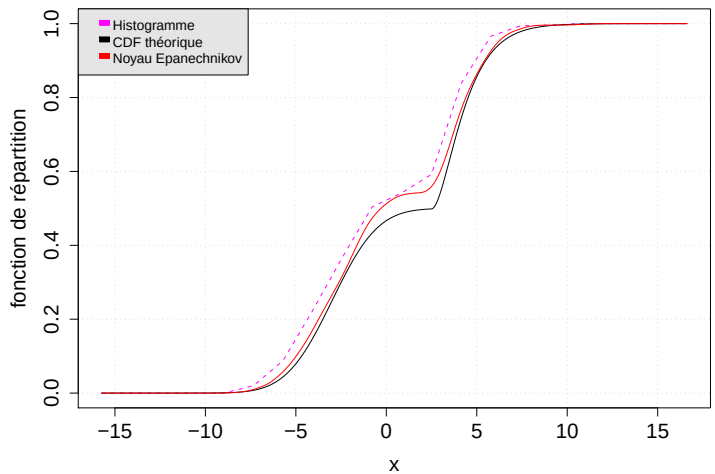
where $\hat{\sigma}^i_N$ is the standard deviation of the ith marginal of the sample.

Histogram/kernel smoothing comparison, uni-dimensional case

	h opt. according to AMISE	AMISE value
Histogram	$\propto \frac{1}{N^{1/3}}$	$\propto \frac{1}{N^{2/3}}$
Kernel smoothing	$\propto \frac{1}{N^{1/5}}$	$\propto \frac{1}{N^{4/5}}$

The kernel smoothing technique is asymptotically better than the histogram





Parametric vs non-parametric

Effect of the sample size

- A small sample size leads to a fragile estimation: the variability of the estimation increase.
- The fitting tests are based on the asymptotic distribution of the estimators, which is questionable for small values of N.
- If the model hypothesis is correct, a parametric estimation will always be better than a non-parametric estimation, but if the hypothesis is wrong, there is no way to fix it by increasing the sample size.
- Increasing the sample size will always improve a non-parametric estimation.

Tests

Several tests are available:

- Qualitative tests (visual tests)
- Quantitative tests

Qualitative tests

The main visual tests are:

- **QQ-plot test**, where the empirical quantiles are plotted versus the estimated ones. If the curve is close to the main diagonal, the estimated distribution is credible.
- **Visual comparison** of the density resulting from a parametric estimation and a non-parametric estimation

Tests

Quantitative tests

Such tests are based on two antagonist hypotheses H_O and H_1, leading to two sources of error:

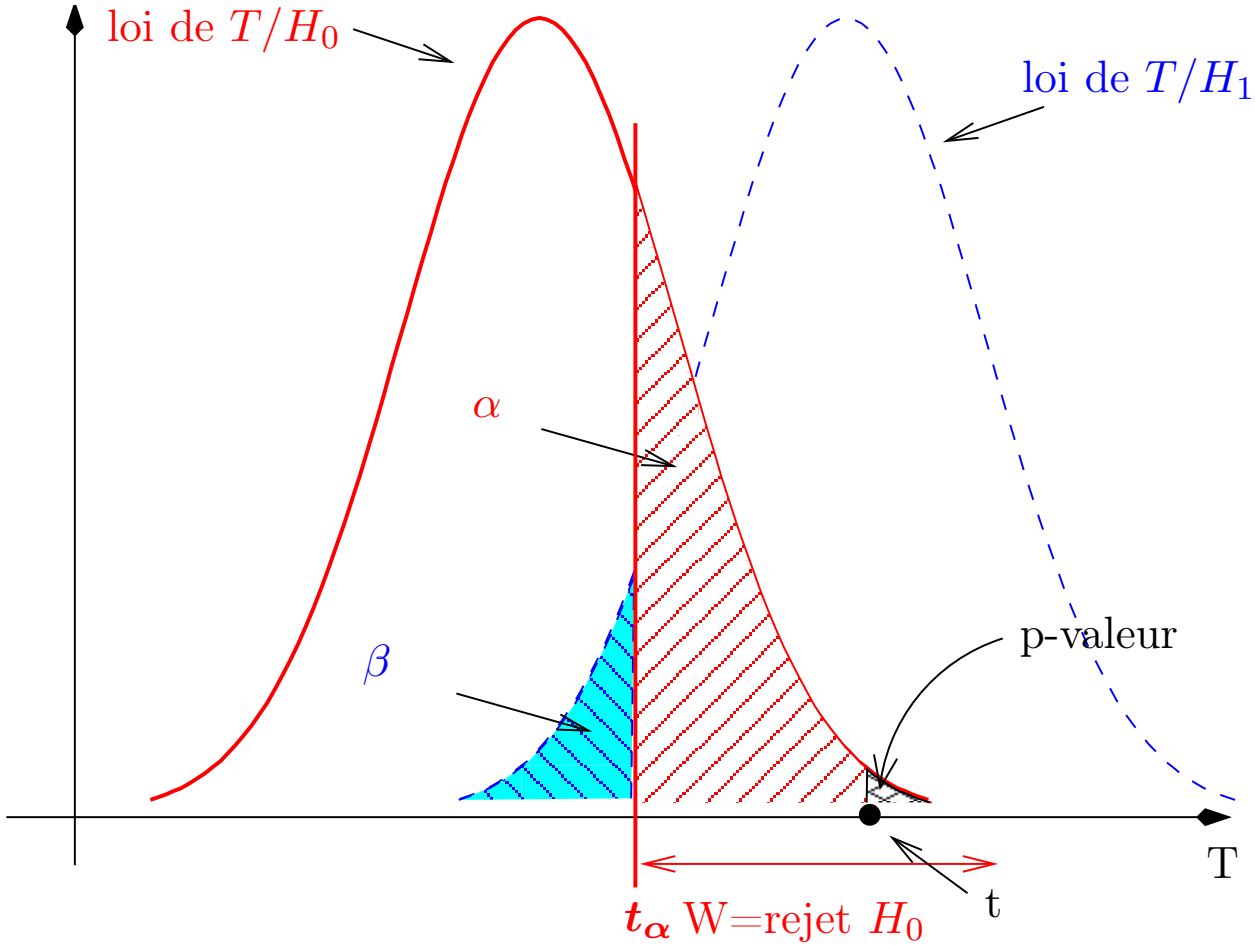
- **First kind error** : α : probability that H_0 is wrongly rejected.
- **Second kind error** : β : probability that H_1 is wrongly rejected.

These two errors are non-symmetrical in general. One want to control the first kind error while minimizing the second kind error. The first kind error is controlled but not the second kind error. As such, H_0 will be the hypothesis that the parametric model is correct.

Test methodology

- One define a test statistics T, which is a random variable built from the sample. Its distribution is known under hypothesis H_0.
- The **critical region** W of rejecting H_O is given by : $\mathbb{P}_{H_0}(W) \geq \alpha$, which leads to $W = \{\boldsymbol{x} \in \mathbb{R} \,|\, T(\boldsymbol{x}) > t_\alpha\}$ where t_α is a threshold on T.
- The test statistics is evaluated on the sample realization at hand, and compared to the threshold. H_0 is rejected if T is in W.

Graphical interpretation of the errors



Classical tests

- Tests focused on the central part of the distribution: chi-square test, Kolmogorov-Smirnov test...
- Tests focused on the tails of the distribution: Anderson-Darling test, Cramer's test...

The test statistics is the normalized gap between the candidate distribution function and the empirical distribution function obtained as the uniform discrete distribution over the sample:

- Kolmogorov-Smirnov : $\sqrt{N}\sup_{y\in\mathbb{R}}|F_N(y)-F(y)| \to W$, where W is a tabulated distribution.
- chi-square : $\zeta_N^{(2)} = N\sum_{i=1}^{m}\frac{(\hat{p}_i - p_i^0)^2}{p_i^0} \to \chi_2(m-1)$.

If several parametric models are accepted?

One can rank the parametric models according to an information criterion:

- the *p*-value: compare $P(T(\boldsymbol{X}) \geq T(\boldsymbol{x_1},\ldots,\boldsymbol{x_n}))$ to α,
- the Bayesian Information Criteria (BIC): that mitigates the log-likelihood of the sample, the dimension of the parameter space of the parametric model and the sample size.

How to question an expert?

The lack of data, or the scarcity of data are the main reason to resort to judgement of experts. One has to question them and to interpret their answer as a probability distribution.
These 3 questions are of uttermost importance:

- **Question 1 : Is there an historical reason for the choice of a specific parametric model?**
- **Question 2 : Is there a specific range in which a given variable must stay?**
- **Question 3 : Are there remarkable values for the variable?**

Choice of the parametric model

Several strategies are possible to turn the expert knowledge into a probability distribution:

- Choice based on an organigram built upon simple alternatives that result from ground experience.
- Use of the **maximum entropy principle**

Organigram

Ans. Q1	Ans. Q2	Ans. Q3	Parametric model
No	Yes : $[a, b]$	No	Uniform(a,b)
		Yes : mode m	Triangular(a,m,b)
	Yes : $[a, +\infty[$	Yes : mean and standard deviation	LogNormal
		Yes : 2 values v_1, v_2	LogNormal
	No	Yes : mean and standard deviation	Normal
		Yes : 2 values v_1, v_2	Normal
Yes	-	-	Historical distribution

Maximum entropy principle

Statistical entropy and information

The statistical entropy is a measure of the lack of knowledge of the state of a complex system.
When all the N possible states of the system are not equi-probable but weighted by a discrete probabity distribution function p, the entropy $\mathcal{S}$ is defined by

$$\mathcal{S} = -k \sum_{i=1}^{N} p_i \, log \, p_i$$

The probability p_i is linked to a level of disorder of the system, which reflect our lack of knowledge on the state of the system. Shannon has extended this definition to the case of a continuous number of states:

$$\mathcal{S} = -\int p(\boldsymbol{x}) \log \, p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

Maximum entropy principle

The principle

The maximum entropy principle is to choose the distribution that maximize the statistical entropy while being compatible with the knowledge we have on the system. Any other choice would implicitely suppose that additional information about the system is available, so the entropy should be smaller.

Some examples

Available information	Resulting distribution
Support : $[a, b]$	$Uniform(a, b)$
Mean m and support : $[a, \infty[$	$Exp(a, \lambda = 1/(m-a))$
Mean m, variance σ^2, support : $\mathbb{R}$	$Normal(m, \sigma^2)$
Mean m, variance σ^2, bounded or half-bounded support	$p(x) \propto e^{ax+bx^2}$ on the support

Validation of the distribution resulting from expert knowledge

Qualitative and quantitative validation

The distribution can be check by:

- a visual inspection of the density and a confirmation from the expert,
- using synthetic quantities derived from the distribution:
  1. median : value under which the variable must stay with probability $1/2$,
  2. 90% quantile: value under which the variable must stay with probability $9/10$,
  3. standard deviation : global measure of dispersion for uni-modal distributions
  4. ...

Outline


Open TURNS Reference Guide, www.openturns.org


Dutfoy, A. and Lebrun, R., "Practical approach to dependence modelling using copulas", Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, Vol. 223 N. 4, pp. 347–361, 2009.


Sklar, M., "Fonctions de répartition à n dimensions et leurs marges", Publication de l'Institut Statistique Universitaire Paris, Vol. 8, pp. 229–231, 1959.


Kallenberg, Olav, "Foundations of modern probability 2nd Ed.", Springer-Verlag, New York, 2002.


Scott, D. W., "Multivariate Density Estimation: Theory, Practice, and Visualization", Wiley, 1992.