Multivariate Continuous Random Variables

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Goal

Model interactions between multiple uncertain continuous quantities

Notation

Deterministic variables: a, b, x, y

Random variables: $\tilde{a}$, $\tilde{b}$, $\tilde{x}$, $\tilde{y}$

What is a random variable?

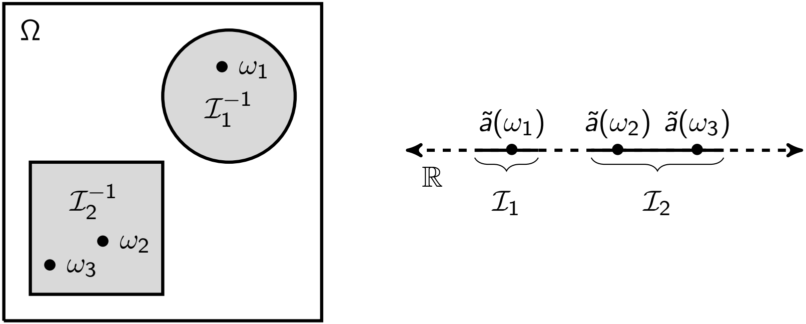
Data scientist:

An uncertain variable described by probabilities estimated from data

Mathematician:

A function mapping outcomes in a probability space to real numbers

Continuous random variables



Continuous random variable

Probability space $(\Omega, \mathcal{C}, \mathrm{P})$

Function $\tilde{a} : \Omega \to \mathbb{R}$

The function $\tilde{a}$ is a valid random variable if for any interval $\mathcal{I} := [a, b] \subseteq \mathbb{R}$, $a \leq b$

$$\mathcal{I}^{-1} := \{\omega \mid \tilde{a}(\omega) \in \mathcal{I}\}$$

is in the collection $\mathcal{C}$, so

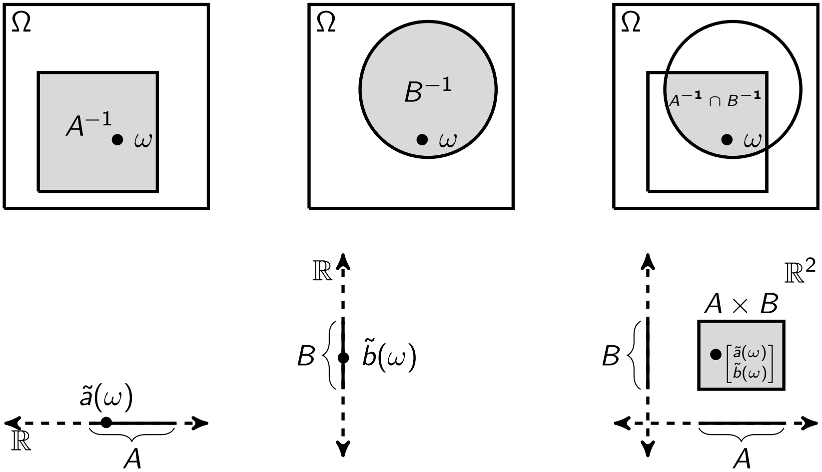
$$\mathrm{P}(\tilde{a} \in \mathcal{I}) = \mathrm{P}(\mathcal{I}^{-1}) \quad \text{is well defined}$$

Continuous random variables

We describe continuous random variables in terms of the probability that they belong to any interval

What about multiple continuous random variables defined on the same probability space?

Two continuous random variables



Two continuous random variables

$$\mathrm{P}\left(\begin{bmatrix}\tilde{a}\\ \tilde{b}\end{bmatrix} \in A \times B\right) := \mathrm{P}\left(\left\{\omega \mid \tilde{a}(\omega) \in A \text{ and } \tilde{b}(\omega) \in B\right\}\right)$$
$$= \mathrm{P}\left(A^{-1} \cap B^{-1}\right)$$

where

$$A^{-1} := \{\omega \mid \tilde{a}(\omega) \in A\},$$
$$B^{-1} := \left\{\omega \mid \tilde{b}(\omega) \in B\right\}.$$

Higher dimensions

Let $\tilde{x} : \Omega \to \mathbb{R}^d$ be a d-dimensional vector containing d continuous random variables $\tilde{x}[1]$, $\tilde{x}[2]$, ..., $\tilde{x}[d]$

Defined on the same probability space $(\Omega, \mathcal{C}, \mathrm{P})$

For any d Borel sets X_1, X_2, ..., $X_d \subseteq \mathbb{R}$, the probability of the event

$$\{\omega \mid \tilde{x}(\omega) \in X_1 \times X_2 \times \cdots \times X_d\} = \cap_{i=1}^{d} \{\omega \mid \tilde{x}[i](\omega) \in X_i\}$$

is well defined

Cumulative distribution function

The cumulative distribution function (cdf) of a random variable $\tilde{a}$ is

$$F_{\tilde{a}}(a) := \mathrm{P}(\tilde{a} \leq a)$$

It encodes the probability that $\tilde{a}$ is less than or equal to a

Joint cdf

The joint cdf of $\tilde{a}$ and $\tilde{b}$ is

$$F_{\tilde{a},\tilde{b}}(a, b) := \mathrm{P}\left(\tilde{a} \leq a, \tilde{b} \leq b\right)$$

The joint cdf of a d-dimensional vector $\tilde{x}$ is

$$F_{\tilde{x}}(x) := \mathrm{P}\left(\tilde{x}[1] \leq x[1], \tilde{x}[2] \leq x[2], \ldots, \tilde{x}[d] \leq x[d]\right)$$

Properties of the joint cdf

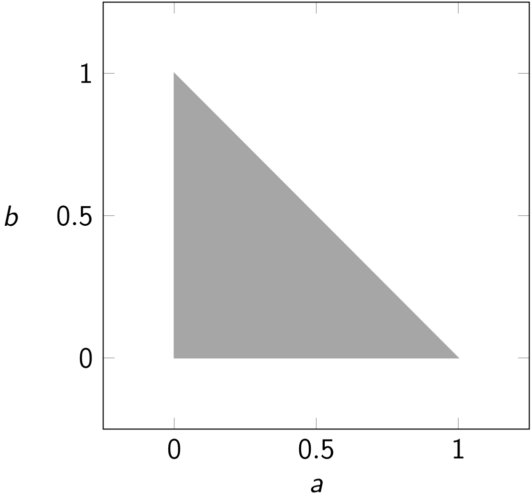
$$\lim_{a \to -\infty} F_{\tilde{a},\tilde{b}}(a,b) = 0$$

$$\lim_{b \to -\infty} F_{\tilde{a},\tilde{b}}(a,b) = 0$$

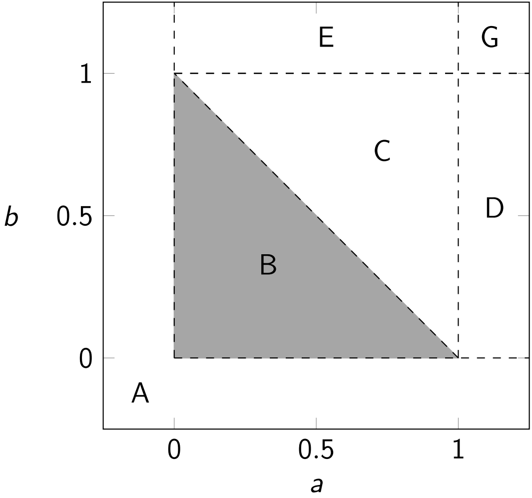
$$\lim_{a \to \infty, b \to \infty} F_{\tilde{a},\tilde{b}}(a,b) = 1$$

Can $F_{\tilde{a},\tilde{b}}(a_1, b_1) > F_{\tilde{a},\tilde{b}}(a_2, b_2)$ if $a_2 \geq a_1,\ b_2 \geq b_1$?   No

Triangle lake



Triangle lake



Triangle lake

$$F_{\tilde{a},\tilde{b}}(a,b) = \begin{cases} 0 & \text{if } a < 0 \text{ or } b < 0, \\ 2ab, & \text{if } a \geq 0, b \geq 0, a+b \leq 1, \\ 2a + 2b - b^2 - a^2 - 1, & \text{if } a \leq 1, b \leq 1, a+b \geq 1, \\ 2b - b^2, & \text{if } a \geq 1, 0 \leq b \leq 1, \\ 2a - a^2, & \text{if } 0 \leq a \leq 1, b \geq 1, \\ 1, & \text{if } a \geq 1, b \geq 1. \end{cases}$$

Computing probabilities

$$
\begin{aligned}
&\mathrm{P}\left(a_1 < \tilde{a} \leq a_2, b_1 < \tilde{b} \leq b_2\right) \\
&= \mathrm{P}\left(\tilde{a} \leq a_2, \tilde{b} \leq b_2\right) - \mathrm{P}\left(\tilde{a} \leq a_1, \tilde{b} \leq b_2\right) \\
&\quad - \mathrm{P}\left(\tilde{a} \leq a_2, \tilde{b} \leq b_1\right) + \mathrm{P}\left(\tilde{a} \leq a_1, \tilde{b} \leq b_1\right) \\
&= F_{\tilde{a},\tilde{b}}\left(a_2, b_2\right) - F_{\tilde{a},\tilde{b}}\left(a_1, b_2\right) - F_{\tilde{a},\tilde{b}}\left(a_2, b_1\right) + F_{\tilde{a},\tilde{b}}\left(a_1, b_1\right)
\end{aligned}
$$

What have we learned?

How to jointly model multiple continuous quantities

Definition and properties of joint cdf