Funcții de Cuplare cu aplicații în probabilități și statistică matematică

Student: Ciulavu George-Adrian

Facultatea de Matematică și Informatică
Universitatea ,,Ovidius'' din Constanța
Sesiunea de Comunicări Matematice

9.decembrie.2023

Cuprins

Introducere

Multe situații din viața reală pot fi modelate printr-un număr mare de variabile care joacă un rol semnificativ și astfel de variabile nu sunt în general independente.

În teoria probabilităților și statistică, o funcție de cuplare este o funcție de repartiție cumulativă multivariată pentru care distribuția de probabilitate marginală a fiecărei variabile este uniformă pe intervalul $[0, 1]$.

Funcțiile de cuplare sunt folosite pentru a descrie/modela dependența dintre variabilele aleatoare.

Funcțiile de cuplare sunt populare în aplicațiile statistice, deoarece permit modelarea și estimarea cu ușurință a repartiției vectorilor aleatori prin estimarea separată a marginalelor și a funcțiilor de cuplare.

Funcția de repartiție

$F_X : \mathbb{R} \to [0,1]$ se numește **funcție de repartiție** asociată variabilei aleatoare X dacă $F_X(x) = P(X < x) \stackrel{\text{not.}}{=} P(X^{-1}(-\infty, x))$

Exemplu funcție de repartiție asociată unei variabile aleatoare discrete

$$X \begin{pmatrix} x_1 & x_2 & .... & x_n & ..... \\ p_1 & p_2 & .... & p_n & ..... \end{pmatrix}$$

Funcția de repartiție este:

$$F_X(x) = \begin{cases} 0, x \leq x_1 \\ p_1, x_1 < x < x_2 \\ ... \\ p_1 + p_2 + ... + p_n, x_n < x < x_{n+1} \\ ... \end{cases}$$

Funcția de repartiție pentru variabile aleatoare de tip continuu

O funcție de repartiție F este de tip **absolut continuu** dacă $\exists f : \mathbb{R} \to \mathbb{R}$ măsurabilă Borel astfel încât $F(x) = \int_{-\infty}^{x} f(u)du$. Atunci f se numește **densitate de repartiție**.

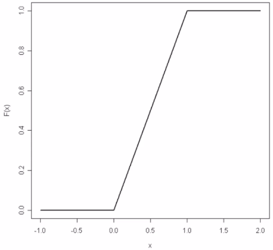
Observații:

- $P(X < b) = F(b) = \int_{-\infty}^{b} f(x)dx$
- $P(X \geq a) = 1 - P(X < a) = 1 - F(a) = \int_{a}^{\infty} f(x)dx$
- $P(a \leq x \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$

Fie X o variabilă aleatoare continuă uniformă pe $[0,1]$, funcția de repartiție este:

$$F(x) = \begin{cases} 0, x < 0 \\ x, 0 < x < 1 \\ 1, x = 1 \end{cases}$$

Reprezentarea grafică a funcției de repartiție:



Funcția de repartiție multidimensională $F_X : \mathbb{R}^d \to [0, 1]$ este:

$$F_X(a_1, ..., a_d) = P(X_1 < a_1, ..., X_d < a_d)$$

Proprietăți caracteristice funcției de repartiție:

- Este nedescrescătoare în raport cu fiecare argument.
- Este continuă la stânga în raport cu fiecare argument.
- $\lim\limits_{x_1 \to +\infty ... x_d \to +\infty} F(x_1, x_2, ..., x_d) = 1$
- $\lim\limits_{x_d \to -\infty} F(x_1, x_2, ..., x_d) = 0, \ \forall x_i$

Noțiuni introductive

Definiția matematică a funcției de cuplare

Fie vectorul aleatoriu $(X_1, X_2, ..., X_d)$ având marginalele continue; adică funcția de repartiție cumulativă (FRC) $F_i(x)$ este o funcție continuă.

Se știe că vectorul aleatoriu $(U_1, U_2, ..., U_d) = (F_1(X_1), F_2(X_2), ..., F_d(X_d))$ are marginalele repartizate uniform pe intervalul $[0, 1]$.

Funcția de cuplare a lui $(X_1, X_2, ..., X_d)$ este definită ca funcția de repartiție cumulativă comună a $(U_1, U_2, ..., U_d)$:
$C(u_1, u_2, ...u_d) = P(U_1 \leq u_1, U_2 \leq u_2, ..., U_d \leq u_d)$
$= P(X_1 \leq F_1^{-1}(u_1), X_2 \leq F_2^{-1}(u_2), ..., X_d \leq F_d^{-1}(u_d))$.

1. Funcția de cuplare C conține toate informațiile despre structura de dependență dintre componentele lui $(X_1, X_2, ..., X_d)$, în timp ce funcția cumulativă F_i conține toate informațiile despre repartiția marginală a lui X_i.
2. Funcția de cuplare se poate folosi și pentru a genera eșantioane pseudo-aleatoare din clasele generale de repartiții de probabilitate multivariate: se generează un eșantion $(U_1, U_2, ..., U_d)$ din funcția de cuplare, din care se construiește eșantionul necesar astfel: $(X_1, X_2, ..., X_d) = (F_1^{-1}(U_1), F_2^{-1}(U_2), ..., F_d^{-1}(U_d))$.

Funcția de cuplare bidimensională

$$C(x, y) = P(U_1 \leq x, U_2 \leq y)$$

Din această funcție rezultă:

- $C(x, 0) = P(U_1 \leq x, U_2 \leq 0) = 0$
- $C(0, y) = P(U_1 \leq 0, U_2 \leq y) = 0$
- $C(x, 1) = P(U_1 \leq x, U_2 \leq 1) = P(U_1 \leq x) = x$
- $C(1, y) = P(U_1 \leq 1, U_2 \leq y) = P(U_2 \leq y) = y$

Se demonstrează că:

$$max(x + y - 1) \leq C(x, y) \leq min(x, y)$$

fiind un caz special pentru cele mai importante rezultate ale statisticii bidimensionale.

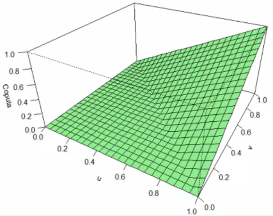
Teoremă (Limita Fréchet-Hoeffding)

Teorema Fréchet-Hoeffding

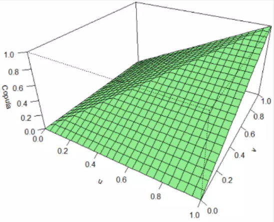
Presupunem $F_1, ..., F_d$ funcții de distribuție marginale și F o funcție de distribuție cu marginale date, atunci pentru $\forall x \in \mathbb{R}^d$

$$(\sum_{k=1}^{d} F_k(x_k) + 1 - d)^+ \leq F(x) \leq min(F_1(x_1), F_2(x_2), .., F_d(x_d))$$

Minimul funcției de cuplare



Maximul funcției de cuplare



Teorema Sklar

Se consideră un vector aleatoriu $(X_1, X_2, ..., X_d)$ și funcțiile de distribuție cumulative marginale $F_1, ..., F_d$, atunci există o unică funcție de cuplare C (definită pe $[0, 1]^d$ cu margini uniforme) pentru $\forall$x$\in$ $\mathbf{R}^d$:

$$F(x_1, x_2, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)),$$

iar pentru cazul bivariat avem:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)).$$

Densitatea funcțiilor de cuplare

Dacă C este absolut continuă, acesta poate fi scris ca:

$$C(u) = \int_{[0,u]^d} c(s)ds$$

unde funcția c se numește **densitatea** lui C.

În particular,

$$c(u) = \frac{\partial^d C(u)}{\partial u_1 \ldots . \partial u_d}$$

Pentru cazul bivariat avem:

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}$$

Coeficientul de corelație liniară

Definiție:

Fie $(X, Y)^T$ un vector de variabile aleatoare. Coeficientul de corelație liniară este:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

unde $Cov(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$ și $Var(X)$, $Var(Y)$ sunt varianțele lui X și Y.
De asemenea,

$$\rho(X, Y) = \frac{1}{\sigma_1\sigma_2}\iint_{\mathbf{R}^2}[F_{12}(x, y) - F_1(x)F_2(y)]\,dxdy$$

Coeficientul de corelație liniară și funcția de cuplare

Fie X_1 și X_2 două variabile aleatoare.

Definiție (Spearman):

$$\rho_s(X_1, X_2) = \rho(F_1(x_1), F_2(x_2)) = 12 \iint_{[0,1]^2} C(u, v) du dv - 3$$

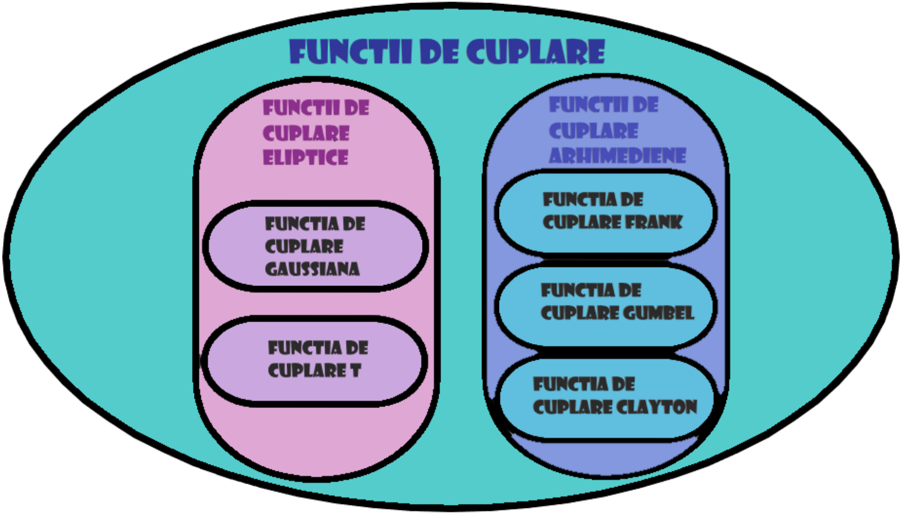
unde C este funcția de cuplare pentru (X_1, X_2).

Definiție (Kendall):

$$\tau(X_1, X_2) = P[(\hat{x}_1 - \tilde{x}_1)(\hat{x}_2 - \tilde{x}_2) > 0] - P[(\hat{x}_1 - \tilde{x}_1)(\hat{x}_2 - \tilde{x}_2) < 0]$$

$$= 4 \iint_{[0,1]^2} C(u, v) dC(u, v) - 1$$

unde $(\hat{x}_1 \hat{x}_2)$ și $(\tilde{x}_1 \tilde{x}_2)$ sunt copiile independente și identice distribuite ale lui (X_1, X_2).


FUNCTII DE CUPLARE
FUNCTII DE CUPLARE ELIPTICE
FUNCTIA DE CUPLARE GAUSSIANA
FUNCTIA DE CUPLARE T
FUNCTII DE CUPLARE ARHIMEDIENE
FUNCTIA DE CUPLARE FRANK
FUNCTIA DE CUPLARE GUMBEL
FUNCTIA DE CUPLARE CLAYTON

Funcții de cuplare eliptice

Funcția de cuplare Gaussiană

$$\int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{\sqrt{2\pi(1-\theta^2)}} \exp\left(-\frac{1}{2(1-\theta^2)}(s^2 - 2\theta st + t^2)\right) dsdt$$

unde $\theta \in [-1, 1]$.

Funcția de cuplare T

$$\int_{-\infty}^{t_{\theta_2}^{-1}(u_1)} \int_{-\infty}^{t_{\theta_2}^{-1}(u_2)} \frac{\Gamma(\frac{(\theta_2+2)}{2})}{\Gamma(\frac{\theta_2}{2})\pi\theta_2\sqrt{1-\theta_1^2}} \left(1 + \frac{x^2 - 2\theta_1 xy + y^2}{\theta_2}\right)^{\frac{(\theta_2+2)}{2}} dxdy$$

unde $\theta_1 \in [-1, 1], \quad \theta_2 \in (0, \infty)$.

Funcții de cuplare Arhimediene

Funcția de cuplare Frank

$$-\frac{1}{\theta}\ln\left(1+\frac{(\exp(-\theta x)-1)}{(\exp(-\theta x)-1)\exp(-\theta x)-1}\right)$$

unde $\theta \in \mathbb{R} \setminus \{0\}$.

Funcția de cuplare Gumbel

$$exp\left(-\left(\left((-\ln(u_1))^{\theta}\right)+(-\ln(u_2))^{\theta}\right)^{\frac{1}{\theta}}\right)$$

unde $\theta \in [1, \infty)$.

Funcția de cuplare Clayton

$$max\left(u_1^{-\theta}+u_2^{-\theta}-1, 0\right)^{\frac{-1}{\theta}}$$

unde $\theta \in [-1, \infty) \setminus \{0\}$

Funcțiile de cuplare în diverse domenii

Ingineria civilă

Funcțiile de cuplare au fost aplicate cu succes la formularea bazei de date pentru analiza de fiabilitate a podurilor pentru autostrăzi și la diferite studii de simulare multivariată în inginerie civilă, fiabilitatea ingineriei eoliene și a cutremurelor.

Medicină

Funcțiile de cuplare au fost folosite în domeniul imagisticii, de exemplu, pentru a segmenta imagini și a produce grafice în genetica imagistică într-un studiu asupra schizofreniei pentru a distinge pacienții normali de pacienții cu Alzheimer.

Funcțiile de cuplare în diverse domenii

Hidrologie

Funcțiile de cuplare sunt folosite atât în analizele teoretice, cât și în cele practice ale datelor hidroclimatice. Studiile teoretice au adoptat metodologia bazată pe funcțiile de cuplare, de exemplu, pentru a obține o mai bună înțelegere a structurilor de dependență a temperaturii și precipitațiilor, în diferite părți ale lumii.

Astronomie

Funcțiile de cuplare sunt aplicate în studiul galaxiilor, inclusiv interacțiunile dintre galaxii și modul în care aceste interacțiuni pot afecta structura și evoluția lor.
Funcțiile de cuplare pot fi utilizate pentru a analiza modul în care planetele influențează mișcarea reciprocă, stabilind orbitele lor și prezicând evenimente astronomice precum conjuncții și opoziții.

Aplicație intuitivă și vizuală a funcțiilor de cuplare

În această problemă se dorește analizarea riscului asociat nivelurilor maxime ale râurilor și frecvenței inundațiilor în diverse regiuni. Pentru fiecare râu studiat, se dispune de două variabile măsurate, care nu au o distribuție normală și sunt corelate.

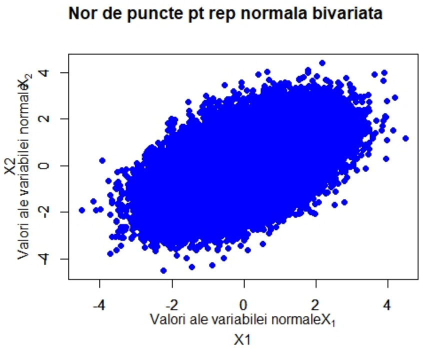
- **Variabila 1:** Conform Teoriei Valorilor Extreme, se presupune că nivelurile maxime ale râului urmează o repartiție Gumbel.
- **Variabila 2:** Frecvența inundațiilor este modelată printr-o distribuție Beta, care descrie probabilitatea de inundație în funcție de numărul de luni cu inundații versus numărul de luni fără inundații.

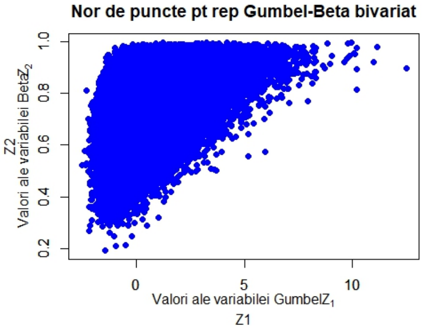
Obiectivul problemei noastre este să construim un model care să descrie repartiția comună a ambelor variabile (nivel maxim al râului și numărul de inundații) folosind o funcție de cuplare Gaussiană. Acest model ar putea oferi informații utile pentru evaluarea și gestionarea riscului asociat evenimentelor hidrologice extreme în diverse regiuni.

În primul rând, transformăm variabile aleatoare uniforme în variabile de alt tip și reciproc (În softul R).

- Simularea valorilor distribuite uniform între 0 si 1.
- Transformarea estantioanelor astfel încât, în loc să fie uniforme, acum să fie distribuite după cum dorim: spre exemplu, conform repartiției Beta, aplicând transformarea Beta la datele uniforme.
- Pentru a realiza transformarea inversă de la o distribuție arbitrară la distribuția uniformă (0, 1), aplicăm pur și simplu funcția de repartiție cumulativă (FRC).

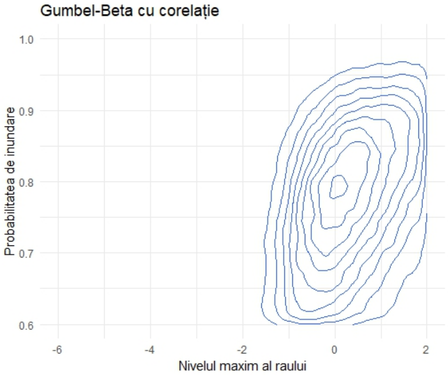
În pasul următor, creăm o distribuție de probabilitate comună bivariată. Prin simularea datelor dintr-o distribuție Gaussiană cu corelații specifice, transformăm marginile la distribuția uniformă și apoi adaptăm aceste margini uniforme la forma dorită a distribuției finale.

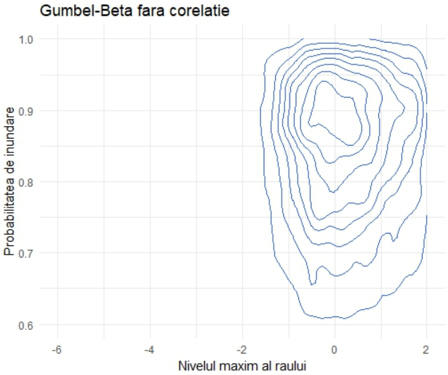




Datele simulate din repartiția Gumbel-Beta (dreapta) au fost obținute conform procedurii descrise din datele simulate din repartiția normală bivariata (stânga) cu corelatia=0.5.

Grafice tip contur pentru datele Gumbel-Beta cu corelație și fără corelație





În concluzie: Există diferențe importante între datele simulate fără funcție de cuplare (independente) și cele cu funcții de cuplare, ce surprind dependența. Modelarea dependenței cu funcții de cuplare poate fi extrem de importantă în analizele statistice.

Bibliografie

- https://twiecki.io/blog/2018/05/03/copulas/
- https://slideplayer.com/slide/3270700/
- https://www.slideshare.net/GiovanniDellaLunga/copule-slides
- https://cermics.enpc.fr/ lelievre/CEMRACS/slides/Lebrun.pdf
- https://www.sciencedirect.com/topics/mathematics/copula

Multumesc pentru atentie!