Cursul 10

Covarianţa şi coeficientul de corelaţie

Ştiind să identificăm variabile aleatoare independente, ne întrebăm în mod natural cum caracterizăm intensitatea legăturii dintre două variabile ce nu sunt independente. Această intensitate este măsurată de covarianţa, respectiv coeficientul lor de corelaţie.

10.1 Covarianţa a două variabile aleatoare

Definiţia 10.1.1 *Covarianţa* variabilelor aleatoare X şi Y, ce au mediile $m_X = M(X)$, $m_Y = M(Y)$ finite, este definită prin

$$cov(X, Y) = M((X - m_X)(Y - m_Y)). \tag{10.1}$$

În cele ce urmează vom presupune că toţi indicatorii (media, dispersia, covarianţa) ce apar în formule există şi sunt numere reale (finite).

Observaţia 10.1.1 Covarianţa unei variabile cu ea însăşi coincide cu dispersia sa, căci

$$\text{cov}(X, X) = M((X - m_X)(X - m_X)) = M((X - m_X)^2) = \sigma^2(X).$$

Propoziţia 10.1.1 *Covarianţa variabilelor aleatoare X, Y se poate calcula astfel:*

$$cov(X, Y) = M(XY) - M(X)\, M(Y). \tag{10.2}$$

Demonstraţie: Efectuând produsul

$$(X - m_X)(Y - m_Y) = XY - m_X Y - m_Y X + m_X m_Y$$

şi aplicând proprietăţile mediei, obţinem

$$\begin{aligned} cov(X, Y) &= M(XY) - m_X M(Y) - m_Y M(X) + m_X m_Y \\ &= M(XY) - M(X)\, M(Y). \end{aligned}$$

□

Definiţia 10.1.2 Două variabile aleatoare X, Y ce au covarianţa zero se numesc *variabile aleatoare necorelate*.

Deoarece am introdus covarianţa ca o măsură a intensităţii dependenţei dintre două variabile aleatoare, este natural să ne aşteptăm ca două variabile aleatoare independente să aibă covarianţa zero, adică să fie necorelate.

Propoziţia 10.1.2 *Dacă X, Y sunt independente, atunci X şi Y sunt necorelate, adică dacă X, Y sunt independente, atunci $cov(X, Y) = 0$.*

Demonstraţie: Rezultă din faptul că dacă X şi Y sunt variabile aleatoare independente, atunci $M(XY) = M(X)M(Y)$. □

Observaţia 10.1.2 Reciproca propoziţiei de mai sus nu este adevărată. Două variabile aleatoare necorelate nu sunt în mod necesar independente, adică dacă $cov(X, Y) = 0$, **nu rezultă** în general că X şi Y sunt independente.

Folosind definiţia covarianţei sau relaţia (10.2) şi proprietăţile operatorului mediei, au loc următoarele reguli de calcul:

1) $cov(X, X) = \sigma^2(X)$;

2) $cov(X, Y) = cov(Y, X)$;

3) $cov(aX, Y) = a\, cov(X, Y)$, $a \in \mathbb{R}$;

4) $cov(X + Y, Z) = cov(X, Z) + cov(Y, Z)$;

5) $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y) + 2cov(X, Y)$;

   Dacă în plus X şi Y sunt variabile aleatoare independente, atunci dispersia sumei lor este egală cu suma dispersiilor, adică

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y).$$

Demonstraţie: O să arătăm doar relaţia 5), celelalte fiind evidente. Fie m_X, m_Y mediile celor două variabile aleatoare. Conform definiţiei dispersiei, avem:

$$\begin{aligned}
\sigma^2(X+Y) &= M\left([(X+Y) - M(X+Y)]^2\right) \\
&= M\left([(X - m_X) + (Y - m_Y)]^2\right) \\
&= M\left((X - m_X)^2 + (Y - m_Y)^2 + 2(X - m_X)(Y - m_Y)\right) \\
&= M\left((X - m_X)^2\right) + M\left((Y - m_Y)^2\right) + 2M((X - m_X)(Y - m_Y)) \\
&= \sigma^2(X) + \sigma^2(Y) + 2cov(X, Y).
\end{aligned}$$

□

Ţinând seama că $\sigma^2(aX) = a^2\sigma^2(X)$, $\forall\, a \in \mathbb{R}$, al doilea rezultat din relaţia 5) de mai sus se poate generaliza astfel:

Propoziţia 10.1.3 *Dacă $X_1, X_2, \ldots, X_n$ sunt variabile aleatoare independente, atunci are loc:*

$$\sigma^2(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2\sigma^2(X_1) + a_2^2\sigma^2(X_2) + \cdots + a_n^2\sigma^2(X_n), \tag{10.3}$$

pentru orice $a_1, a_2, \ldots, a_n \in \mathbb{R}$.

10.2 Coeficientul de corelaţie a două variabile aleatoare

Deoarece covarianţa a două variabile aleatoare este un număr real oarecare, deci dificil de interpretat datorită nemărginirii mulţimii valorilor posibile, definim o altă măsură a dependenţei lor, care ia valori într-un interval mărginit.

Definiţia 10.2.1 *Coeficientul de corelaţie* a două variabile aleatoare X şi Y, de abateri standard nenule, este un număr real, notat cu $\rho(X, Y)$, definit prin

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}, \tag{10.4}$$

unde σ_X, σ_Y sunt abaterile standard ale variabilelor aleatoare X, respectiv Y.

Observaţia 10.2.1 Coeficientul de corelaţie a două variabile aleatoare X şi Y este, de fapt, covarianţa standardizatelor variabilelor X şi Y, adică a variabilelor aleatoare

$$Z_1 = \frac{X - m_X}{\sigma_X}, \quad Z_2 = \frac{Y - m_Y}{\sigma_Y}.$$

Într-adevăr, ţinând seama că $M(Z_1) = M(Z_2) = 0$, avem

$$\begin{aligned} cov(Z_1, Z_2) = M(Z_1\, Z_2) = M\left(\frac{X - m_X}{\sigma_X} \cdot \frac{Y - m_Y}{\sigma_Y}\right) = \frac{1}{\sigma_X\sigma_Y} M((X - m_X)(Y - m_Y)) \\ = \frac{cov(X, Y)}{\sigma_X\sigma_Y} = \rho(X, Y). \end{aligned}$$

Propoziţia 10.2.1 *Coeficientul de corelaţie a două variabile aleatoare X, Y are valoarea absolută subunitară, adică*

$$\rho(X, Y) \in [-1, 1]. \tag{10.5}$$

Demonstraţie: Are loc

$$\sigma^2(tX+Y) \geq 0,\ \forall\, t \in \mathbb{R}.$$

Dar

$$\begin{aligned}\sigma^2(tX+Y) &= \sigma^2(tX) + \sigma^2(Y) + 2cov(tX,Y)\\ &= t^2\sigma^2(X) + 2tcov(X,Y) + \sigma^2(Y),\end{aligned}$$

deci $\sigma^2(tX+Y) \geq 0$ pentru orice $t \in \mathbb{R}$ dacă şi numai dacă discriminantul

$$\Delta = 4\left[cov(X,Y)\right]^2 - 4\sigma^2(X)\sigma^2(Y) \leq 0,$$

ceea ce este echivalent cu

$$\left|\frac{cov(X,Y)}{\sigma(X)\sigma(Y)}\right| \leq 1,$$

adică ceea ce trebuia demonstrat. □

Observăm că pentru două variabile aleatoare independente, coeficientul de corelaţie este 0. Este natural să ne întrebăm în ce caz coeficientul de corelaţie a două variabile aleatoare ia valorile extreme ± 1. Răspunsul este dat de următorul rezultat:

Propoziţia 10.2.2 *Dacă între variabilele aleatoare X şi Y există o relaţie liniară de forma*

$$Y = aX + b,\ a, b \in \mathbb{R},\ a \neq 0,$$

atunci coeficientul de corelaţie al variabilelor aleatoare X şi Y este ± 1 şi anume:

$$\rho(X,Y) = \begin{cases} -1, & \text{dacă } a < 0,\\ 1, & \text{dacă } a > 0.\end{cases}$$

Reciproc, dacă modulul coeficientului de corelaţie a două variabile aleatoare X, Y este 1, atunci între ele există o relaţie liniară, $Y = aX + b$, $a \neq 0$.

Demonstraţie: Dacă $Y = aX + b$, atunci

$$\rho(X,Y) = \frac{cov\,(X, aX+b)}{\sqrt{\sigma^2(X)\sigma^2(aX+b)}}. \tag{10.6}$$

Dar $cov(X, aX+b) = a\,cov(X,X) + cov(X,b)$, iar

$$cov(X,b) = M(b\,X) - M(b)\,M(X) = b\,M(X) - b\,M(X) = 0,$$

deci $cov(X, aX+b) = a\,\sigma^2(X)$. Pe de altă parte, $\sigma^2(aX+b) = a^2\sigma^2(X)$. Astfel, rezultă

$$\rho(X,Y) = \frac{a\,\sigma^2(X)}{\sqrt{\sigma^2(X)\,a^2\,\sigma^2(X)}} = \frac{a}{|a|}. \tag{10.7}$$

Prin urmare, pentru $a > 0$, $\rho(X, Y) = 1$, iar pentru $a < 0$, $\rho(X, Y) = -1$.

Reciproc, considerăm funcţia $g(a, b) = M\left((Y - aX - b)^2\right)$ (variabilele aleatoare X şi Y sunt fixate). Determinăm $a, b \in \mathbb{R}$ astfel încât g să fie minimă, adică determinăm parametrii a, b astfel încât media abaterii la pătrat a lui Y faţă de o funcţie de gradul întâi după X să fie minimă.

Pentru a arăta că funcţia g are un punct de minim local o descompunem astfel:

$$\begin{aligned} g(a, b) &= M\left((Y - aX - b)^2\right) \\ &= M(Y^2 + a^2X^2 + b^2 - 2aXY - 2bY + 2abX) \\ &= M(Y^2) + a^2M(X^2) + b^2 - 2aM(XY) - 2bM(Y) + 2abM(X). \end{aligned}$$

Evident că $M(Y^2), M(X^2), M(XY), M(Y), M(X)$ sunt constante. Rezolvând sistemul

$$\begin{cases} \frac{\partial g}{\partial a} = 0, \\ \frac{\partial g}{\partial b} = 0, \end{cases}$$

obţinem punctul critic (a_0, b_0), unde

$$a_0 = \frac{\operatorname{cov}(X, Y)}{\sigma^2(X)}, \quad b_0 = M(Y) - a_0M(X).$$

Se arată că (a_0, b_0) este punct de minim local al lui g, adică matricea

$$\begin{bmatrix} \frac{\partial^2 g}{\partial a^2}(a_0, b_0) & \frac{\partial^2 g}{\partial a \partial b}(a_0, b_0) \\ \frac{\partial^2 g}{\partial a \partial b}(a_0, b_0) & \frac{\partial^2 g}{\partial b^2}(a_0, b_0) \end{bmatrix} = \begin{bmatrix} 2M(X^2) & 2M(X) \\ 2M(X) & 2 \end{bmatrix}$$

este pozitiv definită. Într-adevăr, avem $\Delta_1 = 2M(X^2) > 0$ (dacă $M(X^2) = 0$, atunci, cum $M(X^2) \geq M(X)^2$, rezultă că $M(X) = 0$, deci $\sigma^2(X) = 0$, ceea ce este fals), respectiv

$$\Delta_2 = \begin{vmatrix} 2M(X^2) & 2M(X) \\ 2M(X) & 2 \end{vmatrix} = 4\left[M(X^2) - M(X)^2\right] = 4\sigma^2(X) > 0.$$

Punctul (a_0, b_0) este chiar punct de minim global al funcţiei g. Calculând $g(a_0, b_0)$, obţinem valoarea minimă a funcţiei g. Astfel,

$$g(a_0, b_0) = \min_{a,b} g(a, b) = \min_{a,b} M\left((Y - aX - b)^2\right) = \sigma^2(Y)(1 - \rho^2(X, Y)).$$

Dar cum $\rho(X, Y) = \pm 1$, obţinem

$$g(a_0, b_0) = M\left((Y - a_0X - b_0)^2\right) = 0.$$

Deoarece $(Y - a_0X - b_0)^2 \geq 0$, media sa este zero dacă şi numai dacă $Y - a_0X - b_0 = 0$ sau, echivalent, $Y = a_0X + b_0$. □

În concluzie:

• când coeficientul de corelaţie a două variabile aleatoare este apropiat de zero, variabilele sunt slab corelate (intensitatea legăturii dintre ele este redusă);

• când valoarea absolută a coeficientului de corelaţie este apropiată de 1, relaţia dintre variabilele aleatoare este "aproape liniară", adică valorile (x, y) ale vectorului aleator (X, Y) sunt uşor dispersate în jurul unei drepte de ecuaţie $y = ax + b$.

Exemplul 1. Fie X o variabilă aleatoare ce are media $M(X) = 3$ şi dispersia $\sigma^2(X) = 1$, iar $Y = -2X + 5$. Să se calculeze covarianţa şi coeficientul de corelaţie pentru variabilele X, Y.

Rezolvare: Deoarece între X şi Y există o relaţie liniară de forma $Y = aX + b$ cu $a < 0$, coeficientul de corelaţie este

$$\rho(X, Y) = -1.$$

Dar cum

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma(X)\,\sigma(Y)},$$

calculând $\sigma^2(Y) = \sigma^2(-2X + 5) = 4\sigma^2(X) = 4$, rezultă că $-1 = \dfrac{cov(X, Y)}{2}$, deci $cov(X, Y) = -2$.

□

10.3 Matricea de covarianţă a unui vector aleator

Fie $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ un vector aleator ale cărui componente nu sunt neapărat variabile aleatoare independente. Pentru a cuantifica intensitatea legăturii dintre două câte două componente, se asociază vectorului $\mathbf{X}$ *matricea de covarianţă*. Înainte de a o defini, notăm prin

$$M(\mathbf{X}) = (M(X_1), M(X_2), \ldots, M(X_n))^T$$

vectorul ce are drept coordonate mediile corespunzătoare variabilelor X_i, $i = \overline{1, n}$, sau $\mathbf{m} = (m_1, m_2, \ldots, m_n)^T$, $m_i = M(X_i)$.

Definiţia 10.3.1 *Matricea de covarianţă* a vectorului aleator $\mathbf{X}$ este matricea notată cu Σ, ale cărei elemente sunt $\sigma_{ij} = \text{cov}(X_i, X_j)$, $i, j = \overline{1, n}$.

Remarcăm că $\sigma_{ii} = cov(X_i, X_i) = \sigma^2(X_i)$. Cu alte cuvinte, elementele de pe diagonala principală ale matricei de covarianţă a unui vector aleator sunt dispersiile componentelor vectorului.

Notând cu $\mathbf{Y} = \mathbf{X} - \mathbf{m} = (X_1 - m_1, X_2 - m_2, \ldots, X_n - m_n)^T$, matricea de covarianţă se poate exprima astfel:

$$\Sigma = M(\mathbf{Y}\mathbf{Y}^T),$$

unde

$$\mathbf{Y}\mathbf{Y}^T = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} = \begin{bmatrix} Y_1Y_1 & Y_1Y_2 & \dots & Y_1Y_n \\ Y_2Y_1 & Y_2Y_2 & \dots & Y_2Y_n \\ \vdots & \vdots & \dots & \vdots \\ Y_nY_1 & Y_nY_2 & \dots & Y_nY_n \end{bmatrix},$$

iar media matricei $\mathbf{Y}\mathbf{Y}^T$ este matricea mediilor elementelor sale, $M(Y_iY_j) = cov(X_i, X_j)$, $i, j = \overline{1, n}$.

Propoziţia 10.3.1 *Matricea de covarianţă a unui vector aleator* $\mathbf{X} = (X_i)$, $i = \overline{1, n}$, *este simetrică şi semipozitiv definită.*

În *recunoasterea formelor* se studiază intensitatea legăturii dintre un număr imens de variabile $X_1, X_2, \dots, X_n$. Matricea de covariantă Σ este supusă analizei PCA (Principal Component Analysis) din care se extrage informaţie valoroasă despre corelaţiile dintre variabile. Informaţia se extrage din descompunerea $\Sigma = QDQ^T$ a matricei simetrice Σ, unde D este matricea diagonală a valorilor proprii $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ ale lui Σ, iar Q este matricea ortogonală (i.e. $Q^TQ = I_n$) ce are pe coloane coordonatele vectorilor proprii ortonormaţi $u_1, u_2, \dots, u_n$, corespunzători valorilor proprii, $\Sigma u_i = \lambda_i u_i$, $i = \overline{1, n}$.