# Curs 10: Inegalitatea lui Markov şi inegalitatea lui Cebîşev. Covarianta. Coeficientul de coralatie. Mixuri de probabilitate

## 1.1 Inegalitatea lui Markov şi inegalitatea lui Cebîşev.

Prezentăm două inegalităţi relativ la valorile unei variabile aleatoare comparativ cu media acesteia, inegalităţi folosite în analiza algoritmilor.

**Propoziţia 1.1.1 (Inegalitatea lui Markov)** *Fie $X$ o variabilă aleatoare astfel încât $X \geq 0$, adică $X$ ia valori nenegative. Dacă $X$ are medie finită, atunci probabilitatea ca $X$ să valori mai mari sau egale decât $a > 0$ verifică inegalitatea*

$$P(X \geq a) \leq \frac{M(X)}{a}.$$

**Demonstraţie**: Fie $f$ densitatea de probabilitate a variabilei aleatoare $X$. Deoarece $X \geq 0$, rezultă că $P(X \geq 0) = 1$, deci $P(X < 0) = 0 = \int_{-\infty}^{0} f(x)\,dx$. De aici se poate deduce că funcţia $f$ este zero pe intervalul $(-\infty, 0)$ exceptând, eventual, o mulţime de măsură Lebesgue nulă (noţiunea de *mulţime de măsură Lebesgue nulă* este un concept uzual din teoria măsurii; de exemplu, orice mulţime cel mult numărabilă este de măsură Lebesgue nulă). Fără a restrânge generalitatea putem presupune că funcţia $f$ este zero pe întreg intervalul $(-\infty, 0)$. Astfel, are loc:

$$\begin{aligned} M(X) &= \int_{-\infty}^{\infty} xf(x)\,dx = \underbrace{\int_{-\infty}^{0} xf(x)\,dx}_{=0} + \underbrace{\int_{0}^{a} xf(x)\,dx}_{\geq 0} + \int_{a}^{\infty} xf(x)\,dx \\ &\geq \int_{a}^{\infty} xf(x)\,dx. \end{aligned}$$

În ultima integrală se verifică inegalitatea $x \geq a$ (deoarece limitele de integrare sunt $a$, $\infty$) şi, deci, se obţine

$$M(X) \geq \int_{a}^{\infty} af(x)\,dx = a\int_{a}^{\infty} f(x)\,dx = a\int_{[a,\infty)} f(x)\,dx = aP(X \geq a)$$

sau echivalent

$$P(X \geq a) \leq \frac{M(X)}{a}.$$

Demonstraţia este identică şi pentru variabile aleatoare discrete, doar că integralele se înlocuiesc cu sume. Presupunem, de exemplu, că variabila aleatoare $X$ ia valorile nenegative $x_1, x_2, \ldots, x_n$ cu probabilităţile $p_1, p_2, \ldots, p_n$. Atunci

$$M(X) = \sum_{i=1}^{n} x_i p_i = \sum_{x_i<a} x_i p_i + \sum_{x_i \geq a} x_i p_i \geq \sum_{x_i \geq a} x_i p_i \geq \sum_{i|x_i \geq a} a p_i = aP(X \geq a).$$

□

Inegalitatea lui Markov se foloseşte în studiul algoritmilor probabilişti pentru a estima probabilitatea ca timpul de execuţie $X$ al unui algoritm cu intrări de volum precizat să depăşească $a$ unităţi de timp, când se cunoaşte doar timpul mediu de execuţie, $M(X)$.

**Propoziţia 1.1.2 (Inegalitatea lui Cebîşev)** *Fie $X$ o variabilă aleatoare arbitrară de medie $M(X)$ şi dispersie $\sigma^2(X)$ finite. Atunci are loc inegalitatea:*

$$P(|X - M(X)| \geq a) \leq \frac{\sigma^2(X)}{a^2},\ a > 0.$$

**Demonstraţie**: Considerăm variabila aleatoare $Y = (X - M(X))^2 \geq 0$. Media lui $Y$ este dispersia lui $X$, $M(Y) = \sigma^2(X)$. Inegalitatea enunţată revine la a aplica inegalitatea lui Markov pentru variabila aleatoare $Y$:

$$P(|X - M(X)| \geq a) = P((X - M(X)^2) \geq a^2) = P(Y \geq a^2) \overset{\text{Markov}}{\leq} \frac{M(Y)}{a^2}.$$

Ţinând seama că $M(Y) = \sigma^2(X)$, rezultă inegalitatea lui Cebîşev. □

Inegalitatea lui Cebîşev afirmă de fapt că probabilitatea ca valoarea absolută a abaterii variabilei $X$ de la media sa să fie mai mare decât o constantă $a > 0$ este mai mică decât dispersia sa supra $a^2$.

O versiune mai utilă în aplicaţii se obţine luând constanta $a = k\sigma(X)$, $k \in \mathbb{N}^*$:

$$P(|X - M(X)| \geq a) \leq \frac{\sigma^2(X)}{k^2\sigma^2(X)} = \frac{1}{k^2}.$$

Ţinând cont de faptul că evenimentul $(|X - M(X)| \geq k\sigma)$ este contrarul evenimentului $(|X - M(X)| < k\sigma)$, se obţine

$$P(|X - M(X)| < k\sigma) = 1 - P(|X - M(X)| \geq k\sigma) \geq 1 - \frac{1}{k^2}. \quad (1)$$

Să interpretăm inegalitatea (1). În acest scop o particularizăm pentru $k = 2$, $k = 3$ şi $k = 4$. Notând $m = M(X)$, $\sigma = \sigma(X)$, se obţine

- pentru $k = 2$ inegalitatea devine

$$P(m - 2\sigma < X < m + 2\sigma) \geq 1 - \frac{1}{4} = 0.75.$$

- pentru $k = 3$ avem

$$P(m - 3\sigma < X < m + 3\sigma) \geq 1 - \frac{1}{9} = \frac{8}{9} = 0.88.$$

- pentru $k = 4$ are loc

$$P(m - 4\sigma < X < m + 4\sigma) \geq 1 - \frac{1}{16} = \frac{15}{16} = 0.9375.$$

În concluzie, probabilitatea ca variabila aleatoare $X$ să ia valori in intervale centrate în valoarea sa medie şi de lungime $4\sigma$, $6\sigma$, $8\sigma$ este mai mare decât $0.75, 0.88$, respectiv $0.9375$.

## 1.2 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.

### 1.2.1 Covarianţa a două variabile aleatoare

**Definiţia 1.2.1** *Covarianţa* variabilelor aleatoare $X$ şi $Y$, ce au mediile $m_X = M(X)$, $m_Y = M(Y)$ finite, este definită prin

$$cov(X, Y) = M((X - m_X)(Y - m_Y)). \tag{2}$$

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

**Observaţia 1.2.1** 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 1.2.1** *Covarianţa variabilelor aleatoare $X, Y$ se poate calcula astfel:*

$$cov(X, Y) = M(XY) - M(X)\, M(Y). \tag{3}$$

**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 1.2.2** 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 1.2.2** *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 1.2.2** 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 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 1.2.3** *Dacă $X_1, X_2, \ldots, X_n$ sunt variabile aleatoare independente, atunci are loc:*

$$\sigma^2(a_1X_1+a_2X_2+\cdots+a_nX_n)=a_1^2\sigma^2(X_1)+a_2^2\sigma^2(X_2)+\cdots+a_n^2\sigma^2(X_n), \qquad (4)$$

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

### 1.2.2 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 1.2.3** *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

$$\rho(X,Y)=\frac{cov(X,Y)}{\sigma_X\sigma_Y}, \qquad (5)$$

unde $\sigma_X, \sigma_Y$ sunt abaterile standard ale variabilelor aleatoare $X$, respectiv $Y$.

**Observaţia 1.2.3** 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 1.2.4** *Coeficientul de corelaţie a două variabile aleatoare $X, Y$ are valoarea absolută subunitară, adică*

$$\rho(X,Y) \in [-1,1]. \tag{6}$$

**Demonstraţie**: **Metoda 1**. Fie $m_X, m_Y$ valorile medii ale variabilelor aleatoare $X, Y$, iar $\sigma_X, \sigma_Y$ abaterile lor standard. Evident, valoarea medie a variabilei aleatoare

$$Z = \left(\frac{X - m_X}{\sigma_X} \pm \frac{Y - m_Y}{\sigma_Y}\right)^2 \tag{7}$$

este nenegativă, deoarece $Z$ ia valori nenegative. Exploatând proprietăţile operatorului valoare medie avem:

$$\begin{aligned} M(Z) &= M\left(\left(\frac{X - m_X}{\sigma_X}\right)^2\right) + M\left(\left(\frac{Y - m_Y}{\sigma_Y}\right)^2\right) \pm 2M\left(\frac{(X - m_X)(Y - m_Y)}{\sigma_X \sigma_Y}\right) \\ &= \frac{1}{\sigma_X^2} M\left((X - m_X)^2\right) + \frac{1}{\sigma_Y^2} M\left((Y - m_Y)^2\right) \pm \frac{2}{\sigma_X \sigma_Y} cov(X,Y) \\ &= \frac{\sigma_X^2}{\sigma_X^2} + \frac{\sigma_Y^2}{\sigma_Y^2} \pm 2\rho(X,Y). \end{aligned}$$

Astfel, $2 \pm 2\rho(X,Y) \geq 0$, ceea ce este echivalent cu $-1 \leq \rho(X,Y) \leq 1$.
**Metoda 2**. 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 $\pm 1$. Răspunsul este dat de următorul rezultat:

**Propoziţia 1.2.5** *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 $\pm 1$ 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

$$\rho(X,Y) = \frac{cov\,(X, aX+b)}{\sqrt{\sigma^2(X)\sigma^2(aX+b)}}. \tag{8}$$

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

$$\rho(X,Y) = \frac{a\,\sigma^2(X)}{\sqrt{\sigma^2(X)\,a^2\,\sigma^2(X)}} = \frac{a}{|a|}. \tag{9}$$

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

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

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

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

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 $cov(X, Y) = -2$.

□

### 1.2.3 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 1.2.4** *Matricea de covarianţă* a vectorului aleator $\mathbf{X}$ este matricea notată cu $\Sigma$, 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

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

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 1.2.6** *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 $X_1, X_2, \ldots, X_n$. Matricea de covariantă $\Sigma$ 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 $\Sigma$, unde $D$ este matricea diagonală a valorilor proprii $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ ale lui $\Sigma$, iar $Q$ este matricea ortogonală (i.e. $Q^TQ = I_n$) ce are pe coloane coordonatele vectorilor proprii ortonormaţi $u_1, u_2, \ldots, u_n$, corespunzători valorilor proprii, $\Sigma u_i = \lambda_i u_i$, $i = \overline{1,n}$.

## 1.3 Mixturi de probabilitate

Fie $p_1, p_2, \ldots, p_n \in (0,1)$ astfel încât $p_1 + p_2 + \cdots + p_n = 1$. Dacă $F_1, F_2, \ldots, F_n$ sunt funcţiile de repartiţie ale variabilelor aleatoare $X_1, X_2, \ldots, X_n$, atunci funcţia

$$F = p_1F_1 + p_2F_2 + \cdots + p_nF_n$$

este o funcţie de repartiţie, numită repartiţie compusă.

Analog, dacă $f_1, f_2, \ldots, f_n$ sunt densităţile de probabilitate ale variabilelor aleatoare $X_1, X_2, \ldots, X_n$, atunci

$$f = p_1f_1 + p_2f_2 + \cdots + p_nf_n$$

este o densitate de probabilitate, numită densitate compusă.

**Definiţia 1.3.1** O variabilă aleatoare $X$ ce are densitatea de probabilitate compusă $f$ sau funcţia de repartiţie compusă $F$, se numeşte *mixtură de distribuţii de probabilitate* sau, mai simplu, *mixtură de probabilitate*.

Dacă variabila aleatoare $X$ are densitatea $f = p_1f_1+p_2f_2+\ldots+p_nf_n$, aceasta înseamnă că $X$ are densitatea $f_1$ cu probab. $p_1$, ..., $X$ are densitatea $f_n$ cu probab. $p_n$. Reprezentând fiecare densitate $f_k$ prin indicele său $k$, asociem unei densităţi compuse o variabilă aleatoare discretă:

$$H = \begin{pmatrix} 1 & 2 & \ldots & k & \ldots & n \\ p_1 & p_2 & \ldots & p_k & \ldots & p_n \end{pmatrix}. \tag{10}$$

Selectând la întâmplare o valoare a lui $H$ este echivalent cu a selecta la întâmplare, cu aceeaşi probabilitate, o densitate de probabilitate din cele $n$.

**Exemplul 2**. O variabilă aleatoare $X$ a cărei densitate de probabilitate

$$f = p_1f_1 + p_2f_2 + \cdots + p_nf_n, \text{unde } f_i(x) = \begin{cases} \dfrac{1}{\theta_i}e^{-x/\theta_i}, & \text{dacă } x \geq 0, \\ 0, & \text{dacă } x < 0, \end{cases} \tag{11}$$

este compusa a $n$ densităţi ale distribuţiei exponenţiale de parametrii $\theta_1, \theta_2, \ldots, \theta_n$ se numeşte variabilă aleatoare hiperexponenţială. Variabilele aleat. hiperexponenţiale modelează durata serviciului procesorului. Acestea se folosesc în simularea reţelelor de cozi.

Dăm în continuare un exemplu de aplicare a mixturii în analiza algoritmilor:

**Exemplul 3**. Considerăm instrucţiunea `if-else`:

```
if(B) then I1 else I2;
```

Fie $X_1, X_2$ variabilele aleatoare exponenţial distribuite care dau timpul de execuţie al grupului de instrucţiuni $I_1$, respectiv $I_2$. Fie $p \in (0,1)$ probabilitatea ca expresia booleană `B` să fie adevărată. Densitatea de probabilitate a variabilei ce dă timpul total de execuţie al blocului `if-else` este $f = pf_1 + (1-p)f_2$, unde $f_1, f_2$ sunt densităţi exponenţiale.

Dacă $p = 0.75$ şi $M(X_1) = 20$ milisecunde, $M(X_2) = 40$ milisecunde, să determinăm media timpului de execuţie pentru `if-else`.

Se ştie că dacă $X \sim \text{Exp}(\theta)$, atunci $M(X) = \theta$. Prin urmare, $\theta_1 = 20$, respectiv $\theta_2 = 40$ şi, deci, densitatea de probabilitate a timpului de execuţie a lui `if-else` este

$$f(x) = \begin{cases} 0.75\dfrac{1}{20}e^{-x/20} + 0.25\dfrac{1}{40}e^{-x/40}, & \text{dacă } x \geq 0, \\ 0, & \text{dacă } x < 0. \end{cases}$$

Astfel, media timpului de execuţie este

$$\begin{aligned} M(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_{0}^{\infty} x\left(0.75\frac{1}{20}e^{-x/20} + 0.25\frac{1}{40}e^{-x/40}\right)dx \\ &= 0.75M(X_1) + 0.25M(X_2). \end{aligned}$$

Mixturile de densităţi normale, numite şi *mixturi Gaussiene*, se folosesc, de exemplu, în analiza şi procesarea imaginilor, recunoaşterea formelor, în clusterizare etc.

Pe lângă mixturile de distribuţii de probabilitate continue se folosesc şi mixturi de distribuţii discrete.

**Exemplul 4**. Facebook monitorizează atitudinea unui user faţă de postările pe wall-uri şi îi asociază un număr de reacţii ce sunt modelate de o mixtură Poisson. Pentru a înţelege mai uşor, discutăm cazul cel mai simplu când userul are reacţia $R_1$ cu probabilitatea $p_1$ şi reacţia $R_2$ cu probabilitatea $p_2$, $p_1 + p_2 = 1$. Reacţia $R_1$ constă din linkuri la articole din *Times New Roman* cu rata $\lambda_1$/oră şi reacţia $R_2$, ce constă din like-uri la pozele amicilor, cu rata $\lambda_2$/oră. Astfel, numărul de reacţii/manifestări ale userului pe oră este o variabilă aleatoare $X$ ce are ca distribuţie de probabilitate mixtura Poisson:

$$P_X(k) := P(X = k) = p_1 e^{-\lambda_1}\frac{\lambda_1^k}{k!} + p_2 e^{-\lambda_2}\frac{\lambda_2^k}{k!}.$$

**ATENŢIE!!!** Mixtura de distribuţii de probabilitate **nu** înseamnă că variabila $X$ este de forma $X = p_1X_1 + p_2X_2$, cu $X_i \sim Poiss(\lambda_i)$, $i = 1, 2$, ci că $X$ are distribuţia Poisson de rată $\lambda_1$ cu probab. $p_1$, respectiv $X$ are distribuţia Poisson de rată $\lambda_2$ cu probab. $p_2$.

O mixtura modelează foarte bine comportamentul uman, care nu este constant, ci în funcţie de diverse circumstanţe "umanul" are o reacţie sau alta.

**Exercitiu:** Dacă $p_1 = 0.6$ şi $p_2 = 0.4$, iar $\lambda_1 = 8$, $\lambda_2 = 10$, să se calculeze probabilitatea ca $(X < 3)$ (probabilitatea evenimentului ca într-o oră userul să înregistreze mai puţin de 3 manifestări din cele monitorizate) şi numărul mediu de manifestări pe oră.

*Mixturile de distribuţii $m_i$-Erlang*, $i = \overline{1, n}$, de parametrii $\theta_1, \theta_2, \ldots, \theta_n$,

$$f(x) = p_1\frac{x^{m_1-1}e^{-x/\theta_1}}{\theta^{m_1}(m_1-1)!} + p_2\frac{x^{m_2-1}e^{-x/\theta_2}}{\theta^{m_2}(m_2-1)!} + \cdots + p_n\frac{x^{m_n-1}e^{-x/\theta_n}}{\theta^{m_n}(m_n-1)!}$$

se folosesc ca modele pentru aplicaţii în reţele wireless şi sisteme de calcul mobil. O variabilă aleatoare $X$ ce are o astfel de densitate de probabilitate se zice că are distribuţie hyper-Erlangen. Exemplu de astfel de model: reţelele wireless de generaţia a treia oferă servicii integrate de telefonie, date, multimedia etc. Dacă o reţea wireless cu structură celulară oferă $n$ tipuri de servicii şi rata medie a sosirii apelurilor de tip $i$ într-o celulă, în unitatea de timp, este $\lambda_i$, iar $f_i(x)$ este densitatea de probabilitate a duratei de servire a cererii de tip $i$ într-o celulă, atunci

$$f(x) = \underbrace{\frac{\lambda_1}{\sum_{j=1}^{n} \lambda_j}}_{p_1} f_1(x) + \underbrace{\frac{\lambda_2}{\sum_{j=1}^{n} \lambda_j}}_{p_2} f_2(x) + \cdots + \underbrace{\frac{\lambda_n}{\sum_{j=1}^{n} \lambda_j}}_{p_n} f_n(x)$$

este densitatea de probabilitate a duratei de servire a celulei respective.

Exploatând interpretarea dată mai sus distribuţiei compuse, putem da următoarea modalitate de generare a $N$ valori de observaţie asupra variabilei aleatoare $X$ ce are distribuţie de probabilitate compusă:

`for i = 1 : N {`
`k` $\leftarrow$ simulatorul variabilei aleatoare discrete `H`;
$\mathtt{x_i} \leftarrow$ simulatorul distribuţiei $\mathtt{f_k}$;
}

În şirul generat $(x_i)$, $i = \overline{1, N}$, o proporţie de aproximativ $100\,p_1\%$ valori vor fi din legea $f_1$, $100\,p_2\%$ valori vor fi din legea $f_2$ etc.