5.5. Estimatori statistici şi momente

În cea mai mare parte a cazurilor este dificil să se lucreze cu funcţia densitate de probabilitate pentru un vector aleator multidimensional. De fapt, într-un număr mare de aplicaţii practice funcţia densitate de probabilitate (sau funcţia masă de probabilitate) nici nu poate fi cunoscută. În astfel de situaţii este posibil însă să se lucreze cu momentele distribuţiei (în mod uzual cu cele de ordin I şi II, acestea fiind cantităţi definite în strânsă legătură cu funcţia densitate de probabilitate, *ale cărei trăsături şi proprietăţi le reflectă*.

După cum vom vedea, cu toate că aceste momente sunt definite, *teoretic*, în strânsă legătură cu funcţia densitate de probabilitate, *în practică* nu numai că nu este necesar să se cunoască această funcţie pentru estimarea lor dar, în plus, o estimaţie a funcţiei densitate de probabilitate poate fi obţinută folosindu-ne de estimaţiile acestor momente precum şi de presupunerea iniţială privind forma parametrică a densităţii. Această legătură dintre momentele unei v.a. x şi funcţia densitate de probabilitate a lui x ne permite ca, prin intermediul momentelor de ordin I şi II să realizăm diferite analize şi să tragem concluzii privind caracteristicile vectorului aleator fără a cunoaşte efectiv funcţia de distribuţie ce-l caracterizează. O tratare a acestor aspecte şi a semnificaţiei pe care o au aceste momente o vom face atunci când vom discuta *estimarea parametrică punctuală* (vezi în cadrul ***Subcapitolul 5.5.1.*** paragraful „***Estimarea parametrică punctuală clasică***”).

Pentru a înţelege ce implică, în general, un **proces de estimare statistică** (văzut ca formă tradiţională de inferenţă statistică[36]), precum şi, care sunt metodele aferente, de implementare, de care se dispune la ora actuală, facem în cele ce urmează o prezentare de ansamblu a problemei, urmând ca, ulterior, să abordăm doar parte dintre metodele amintite.

[36] Reamintim aici că, prin **inferenţă statistică** se înţelege obţinerea de concluzii bazate pe o evidenţă statistică (informaţii derivate dintr-un eşantion). Concluziile care se trag sunt asupra caracteristicilor populaţiei din care provine eşantionul sau, mai general, asupra procesului aleator al cărui comportament a fost observat într-o perioadă finită de timp.

Cea de a doua formă tradiţională de inferenţă statistică este *testarea ipotezelor* (numită şi testarea statistică de semnificaţie) şi ea, spre deosebire de estimarea statistică, oferă un răspuns de tipul „da”/„nu” unor întrebări statistice.

5.5.1. Procesul estimării statistice

Probabilitatea statistică este cea care furnizează *soluţii optimale*[37] în cazul datelor măsurate ce sunt afectate de zgomot, caracterizate de incertitudine sau sunt incomplete, ea extrăgând din date cea mai multă informaţie. În prezent există numeroase domenii în care *teoria estimării* îşi găseşte aplicabilitate. Astfel, *procesarea de semnal*, *controlul calităţii*, *telecomunicaţiile*, *interpretarea ştiinţifică a experimentelor*, *teoria controlului*, etc. – sunt doar câteva dintre exemplele pe care le întâlnim frecvent în practică.

Estimarea, ca *ramură a statisticii* şi a *procesării de semnal*, se ocupă cu *estimarea* (aproximarea), pe baza datelor empirice, *a valorilor unor parametri* ai modelelor statistice sau chiar a *structurii modelelor statistice* ce descriu *pattern*-urile dintr-o populaţie aflată în studiu.

Cunoaşterea şi specificarea, sau nu, *a priori* a structurii modelelor statistice în procesul de estimare împarte aceste modele statistice (respectiv, metodele de estimare) în:

- ***parametrice***;
- ***neparametrice***;
- ***semiparametrice***.

Pentru a înţelege mai bine distincţia dintre acestea (precum şi terminologia adoptată – *parametric-neparametric*) plecăm de la definiţa unui ***model statistic***.

Un ***model statistic*** constă într-un set de ecuaţii matematice ce descriu comportamentul unui obiect de studiu în termenii *variabilelor aleatoare* şi ai *distribuţiilor de probabilitate* (respectiv, densităţilor de probabilitate) asociate acestora. În consecinţă, cunoaşterea sau nu a structurii modelului se reduce, practic, la cunoaşterea sau nu a formei funcţionale a densităţii de probabilitate ce caracterizează v.a. sau vectorul aleator ce descrie populaţia. O determinare completă a acestei distribuţii de probabilitate implică estimarea parametrilor acesteia pe baza eşantionului de date empirice.

Mai general, un model statistic include însă, pe lângă caracterizarea statistică a datelor numerice (definiţia de mai sus), aspecte cum ar fi estimarea comportamentului probabilistic viitor al unui sistem pe baza comportamentului acestuia din trecut, extrapolarea sau interpolarea datelor pe baza unei aşa-zise celei mai bune aproximări sau a estimaţilor erorii calculate

[37] Un *estimator optimal* este estimatorul care extrage din datele măsurate toată informaţia disponibilă în acestea; în cazul în care în datele empirice mai există informaţie neutilizată atunci spunem că estimatorul nu este unul optimal.

pentru datele empirice precum şi analiza spectrală a datelor unui eşantion statistic.

1. Estimarea parametrică

Estimarea parametrică a densităţii de probabilitate

În statistică există definite câteva familii de repartiţii teoretice remarcabile (împreună cu funcţiile densitate de probabilitate derivate; pentru exemplificare vezi **Anexa: Repartiţii remarcabile** şi ***Subcapitolul 5.6***) iar diferenţierea în cadrul unei familii se face printr-un număr finit de parametri (unul sau mai mulţi) ce caracterizează respectiva familie de distribuţii.

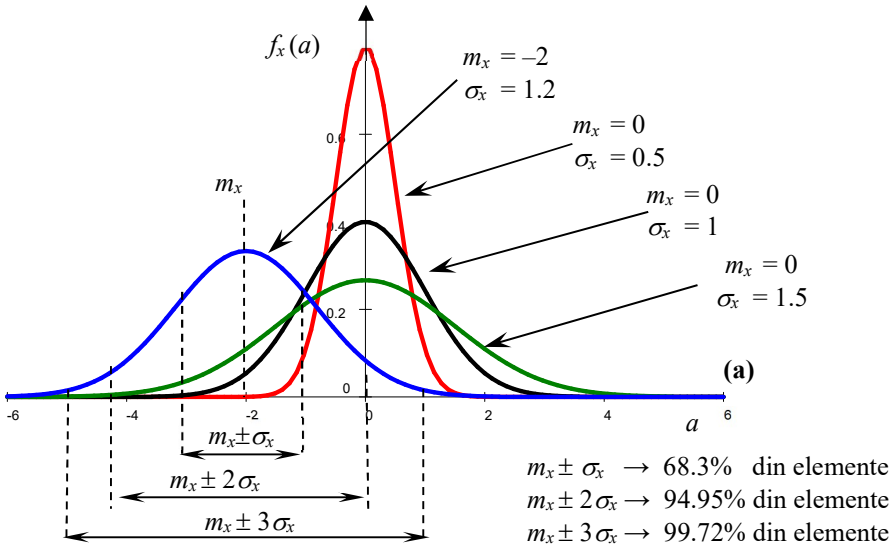


Figura 5.15. Exemple de funcţii densitate de probabilitate *Gauss*-iene din familia distribuţiei normale monodimensionale

Parametrii în statistică, la fel ca şi în matematică, sunt cantităţi referite prin simboluri ce fac parte din *definiţia* funcţiilor şi ei definesc anumite caracteristici ale acestora (de exemplu: poziţionarea sau alura funcţiilor). Astfel, spre exemplu, în cazul distribuţiei normale a unei v.a. x, cei doi parametri ai repartiţiei, m_x şi σ_x^2 (relaţia (5.115)), au semnificaţia poziţionării distribuţiei pe axa absciselor (dată de parametrul m_x), respectiv, semnificaţia caracterizării zonelor de concentrare ale punctelor (dată de parametrul σ_x^2) (vezi **Figura 5.15**, pentru exemplificare cu diverse valori particulare ale celor doi parametri):

$$f_x(a) = f_x(a; m_x, \sigma_x^2) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{\frac{(a-m_x)^2}{2\sigma_x^2}} \tag{5.115}$$

În relaţia de mai sus a este variabila generică a funcţiei densitate de probabilitate, $f_x(a)$, iar m_x şi σ_x^2 sunt parametrii acestei funcţii.

Aceşti parametri ai modelului statistic cu structură (*fdp*) predefinită nu sunt oarecare ci ei au o semnificaţie particulară, fiind chiar ***parametri ai populaţiei*** studiate şi având valori tipice (fiind, de exemplu, momente ale distribuţiei variabilei aleatoare ce caracterizează caracteristica studiată a populaţiei). În cazul nostru particular, m_x şi σ_x^2 corespund următorilor doi parametri ai populaţiei studiate (sau, momente de ordin unu, respectiv, doi ale distribuţiei):

(a) *media* (valoarea sperată a) *v.a. x*, care prin definiţie este:

$$m_x \overset{def}{=} E\{x\} = \int_{-\infty}^{+\infty} a f_x(a) da \tag{5.116}$$

(b) şi, *varianţa v.a. x*, dată de:

$$\sigma_x^2 \overset{def}{=} E\{(x - m_x)^2\} = \int_{-\infty}^{+\infty} (a - m_x)^2 f_x(a) da \tag{5.117}$$

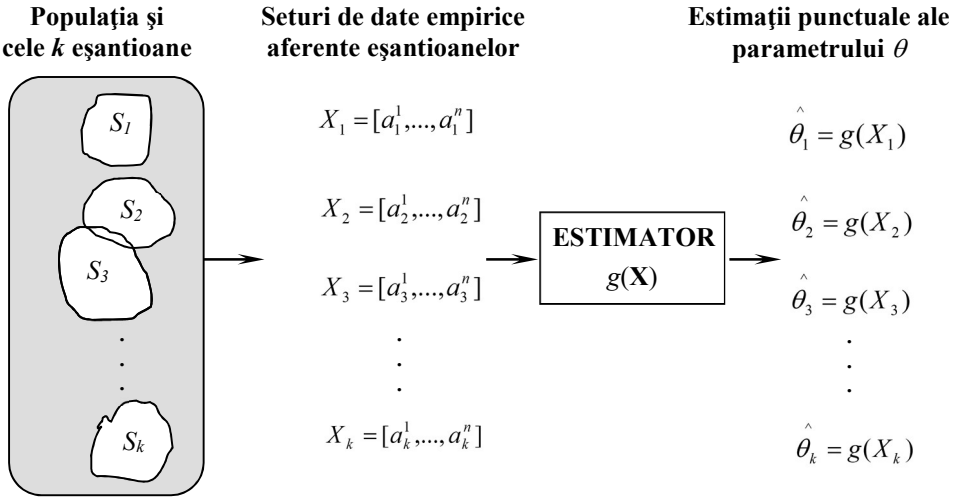


Figura 5.16. Procesul de estimare punctuală a parametrului θ al unui model statistic parametric

În acest context, funcţia densitate de probabilitate din relaţia (5.115) mai este referită ca *repartiţie normală* cu media m_x şi dispersia σ_x^2, valorile celor doi parametri determinând complet repartiţia (poziţionarea şi alura acesteia). Faptul că *v.a. x* este repartizată normal cu parametrii, m_x şi σ_x^2, se mai notează şi:

$$x \sim N(m_x, \sigma_x^2) \tag{5.118}$$

Estimarea (*punctuală* sau *de tip interval*) a ***parametrilor*** populaţiei se face pe câte un eşantion[38] (vezi **Figura 5.16**) – ales din populaţia investigată astfel încât acesta să fie reprezentativ.

Pentru problematica tratată în această carte de interes particular este estimarea punctuală, despre care vom vorbi în cele ce urmează.

Estimarea parametrică punctuală clasică

Procesul estimării punctuale a parametrilor unei populaţiei sau ai unui proces se implementează cu ajutorul unui ***estimator***[39]; acesta foloseşte la intrare datele măsurate ale eşantionului şi furnizează la ieşire *valoarea estimată* (aproximată) a parametrilor necunoscuţi ai populaţiei. Exprimat matematic, aceasta s-ar traduce astfel:

- Presupunem că dorim să estimăm parametrul θ (scalar sau vectorial) al unei fdp cunoscute, $f_x(a; \theta)$, a *v.a. x* (discuţia se poate generaliza la vectori aleatori). Pentru aceasta extragem dintr-o populaţie *k* eşantioane de volum *n*, respectiv, eşantioanele S_1,..., S_k. (similar ca în **Figura 5.16**). Fiecare eşantion, S_i, generează un vector al datelor observabile, $X_i=[a_i^1,\ldots, a_i^n]$.
- *Estimatorul* $\hat{\boldsymbol{\theta}}$ al parametrului θ al populaţiei este atunci definit ca o funcţie $g(\cdot)$ ce se aplică vectorului aleator de date empirice $\mathbf{X}=[\mathbf{a}^1,\ldots, \mathbf{a}^n]$[40]; dacă în loc de v.a. *x* am fi avut vectorul aleator *x*, atunci am fi vorbit de o matrice aleatoare de date empirice.

  Pentru a înţelege mai bine cele ce urmează, dăm şi următoarele definiţii pentru o *v.a.* $h(x)$ a lui *x*, pentru media acesteia şi pentru noţiunea de statistică:

  Fie x o v.a. şi $h(x)$ *o funcţie a v.a. x. Expresia*:

---

38 Submulţime a populaţiei statistice considerate.

39 Formulă sau procedură de estimare.

40 Notaţiile adiacente folosite în acest subcapitol au fost a_i^j, $\bar{x}$, $\hat{\theta}_i$, X_i = valori particulare ale variabilelor aleatoare/ vectorilor aleatori/ matricelor aleatoare $\mathbf{a}^j$, $\bar{\mathbf{x}}$, $\hat{\boldsymbol{\theta}}$, respectiv, $\mathbf{X}$.

$$y = h(x) \tag{5.119}$$

este ***o nouă v.a.*** *definită astfel: pentru un* ζ *dat*, $x(\zeta)$ *este un număr iar* $h[x(\zeta)]$ – valoarea funcţei $h(\cdot)$ calculată pentru acest număr – *este valoarea lui* $y(\zeta) = h[x(\zeta)]$ *pentru care mulţimea domeniului este mulţimea S a realizărilor experimentale.* (definiţia de mai sus este valabilă şi pentru x vector aleator, cu menţiunea că de această dată y poate fi *v.a.*, vector aleator sau chiar o matrice aleatoare).

Orice funcţie care se aplică vectorului eşantion $\mathbf{X}$ *se mai numeşte şi* ***statistică***. Astfel, statisticile sunt caracteristici matematice ale eşantioanelor, şi ele pot fi utilizate ca valori estimate ale parametrilor (ce sunt, la rândul lor, caracteristici matematice ale populaţiilor din care au fost luate eşantioanele).

Din relaţia (5.116) *rezută că* ***media v.a y*** *este şi ea dată de*:

$$E\{y\} = \int_{-\infty}^{+\infty} b f_y(b) db \tag{5.120}$$

Conform unei teoreme de bază (lăsăm ca exerciţiu demonstrarea acesteia), $E\{y\}$ *poate fi exprimat nu neapărat în termenii fdp asociate lui y,* $f_y(b)$*, ci direct în termenii funcţiei* $f_x(a)$ *a lui x, ca în relaţia de mai jos*:

$$E\{h(x)\} = \int_{-\infty}^{+\infty} h(x) f_x(a) da \tag{5.121}$$

Din cele de mai sus reiese că şi estimatorul punctual $\hat{\boldsymbol{\theta}} = g(\mathbf{X})$ este tot o *v.a.* (sau, funcţie de situaţie, un vector aleator/o matrice aleatoare), iar din modul cum a fost definit, el se mai numeşte şi statistică.

- *Estimatul punctual* este rezultatul (valoarea) funcţiei $g(\cdot)$ aplicată unui singur eşantion, $\hat{\theta}_i = g(X_i)$. În eşantioane diferite, statisticile calculate au valori diferite. Prin urmare, se poate vorbi despre o distribuţie a valorilor statisticii $\hat{\boldsymbol{\theta}}$ în mulţimea eşantioanelor de acelaşi volum, n, numită şi *distribuţie de sondaj*[41] a statisticii respective.

[41] Operaţiunea de formare a unui eşantion (selecţie) se numeşte sondaj.

Din cele prezentate până acum reiese că, atunci când facem inferenţe privind o caracteristică studiată din populaţie operăm, în fapt, cu *trei distribuţii asociate acesteia*:

- distribuţia populaţiei – este acea distribuţie (în general, necunoscută) pe care o are caracteristica studiată (sau *v.a.* asociată ei) în populaţie şi pe care dorim să o determinăm;
- distribuţia eşantionului – este distribuţia pe care o are *v.a.* în datele empirice (eşantionul) de care dispunem şi ea poate fi determinată complet din aceste date;
- •. distribuţia de sondaj – este distribuţia pe care o are statistica calculată în mulţimea tuturor eşantioanelor de volum, n, dat.

Aparent complicat, lucrurile se simplifică însă foarte mult atunci când, datorită unor teoreme de limită centrală, se demonstrează că forma distribuţiei de sondaj este una cunoscută atunci când volumul eşantionului, n, creşte (vezi **Anexa: Teorema limită centrală**). Mai mult, din considerente teoretice, relaţia repartiţiei de sondaj a statisticii cu valoarea tipică (parametrul) din populaţie este, de asemenea, una bine precizată.

Fără a detalia prea mult din statistica matematică ce stă la baza acestor aspecte teoretice, vom enunţa în continuare numai acele rezultate ale teoriei statistice de care ne vom folosi în mod deosebit în această carte, ca parte a statisticii aplicate. Astfel:

I. Pentru orice parametru al populaţiei, θ, dat pot fi implementaţi unul sau mai mulţi estimatori, $\hat{\boldsymbol{\theta}} = g(\mathbf{X})$. Obiectivul estimării este acela de a selecta o funcţie $g(X)$ care să minimizeze într-un anumit sens eroarea de estimare $g(\mathbf{X}) - \theta$. Atunci când minimizarea se face în sensul erorii pătratice medii, respectiv:

$$e = E\{[g(\mathrm{X}) - \theta]^2\} = \int_{\Re} [g(X) - \theta]^2 f_x(X, \theta) dX \qquad (5.122)$$

atunci estimatorul $\hat{\boldsymbol{\theta}} = g(\mathbf{X})$ se numeşte ***cel mai bun estimator***.

II. În ***Subcapitolul*** **Error! Not a valid link.** s-a făcut menţiunea că *mediile calculate pe un număr de observaţii se apropie de o valoare constantă pe măsură ce numărul observaţiilor creşte*. Aceasta reprezintă, în fapt, *interpretarea frecvenţială* a probabilităţii, conform căreia, media

aritmetică $\bar{x}$ a valorilor observate a^i ale v.a. x tinde către integrala din (5.116), ce reprezintă media v.a. x, atunci când $n \to \infty$:

$$\bar{x} = \frac{a^1 + \cdots + a^n}{n} \to E\{x\} = m_x \tag{5.123}$$

III. În contextul notaţiilor de mai sus, v.a. $\bar{\mathrm{x}} = \hat{\boldsymbol{\theta}} = g(\mathbf{X})$, reprezentând statistica *media aritmetică a eşantionului* a v.a. x:

$$\bar{\mathrm{x}} = \frac{\mathrm{a}^1 + \cdots + \mathrm{a}^n}{n} \tag{5.124}$$

este o v.a. cu media m_x şi varianţa σ_x^2/n (mai exact, se acceptă conform unor teoreme de limită centrală (vezi **Anexa: Teorema limită centrală**) că distribuţia de sondaj a mediei este caracterizată de o distribuţie $N(m_x, \sigma_x^2/n)$); această afirmaţie este valabilă, în general, pentru valori ale lui $n > 10$, atunci când distribuţia lui x este aproape simetrică, şi pentru $n > 30$, atunci când distribuţia lui x are asimetrie pronunţată sau necunoscută.

Folosindu-ne de toate informaţiile prezentate mai sus se poate arăta că statistica *media aritmetică a eşantionului*, $\bar{\mathrm{x}}$, a *v.a. x* este, din punct de vedere teoretic, un:

- *estimator nedeplasat* ($E\{\hat{\boldsymbol{\theta}}\} = \theta$) al parametrului *media*, m_x, al populaţiei, adică $E\{\bar{\mathrm{x}}\} = m_x$, şi un
- *estimator consistent* (eroarea de estimare $[\hat{\boldsymbol{\theta}} - \theta]$ tinde către 0 în probabilitate atunci când $n \to \infty$); mai exact, varianţa lui $\bar{\mathrm{x}}$, $\sigma_x^2/n \to 0$ atunci când $n \to \infty$ de unde rezultă că $\bar{\mathrm{x}} \to m_x$ în sensul mediei pătratice, şi, deci, de asemenea, în probabilitate.

Observaţia 5.25: Un estimator am spus că este definit ca o funcţie a eşantionului (setului de date empirice) însă, pentru ca el să fie un estimator de încredere este necesar să posede următoarele proprietăţi: să fie nedeplasat, să fie consistent şi să aibă varianţă minimă.

Întrucât decizia de estimator consistent se poate lua doar atunci când $n \to \infty$, ceea ce nu este cazul în practică întrucât n poate lua cel mult valori finite, rezultă că cel mult putem căuta şi, în cel mai bun caz, selecta empiric, cel mai bun estimator.

În cele ce urmează, alegerea empirică a estimatorului o vom face ţinând cont de două lucruri:

1. statistica *media aritmetică a eşantionului* a lui x, din relaţia (5.124),

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} a^i \qquad (5.125)$$

este, aşa după cum am văzut din punct de vedere teoretic, un estimator consistent al parametrului $\theta = m_x = E\{x\}$, ce reprezintă media populaţiei pentru caracteristica studiată a populaţiei;

2. din definiţia dată în relaţia (5.119) avem că pentru o *v.a.* x, o funcţie $h(x) = y$ este tot o *v.a.*

Combinând cele două informaţii ajungem la **construirea empirică a următorului estimator** pentru parametrul *media* $\theta = E\{h(x)\}$ $(= m_y = E\{y\})$ a funcţiei $h(x)$ a lui x, respectiv:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} h(a^i) \qquad (5.126)$$

Acest estimator reprezintă media aritmetică a eşantionului pentru *v.a.* y, deci pentru funcţia $h(x)$, şi el este:

- din *punct* de vedere teoretic, un estimator consistent iar
- din punct de vedere practic, de cele mai multe ori, cel mai bun estimator (asta cel puţin pentru n de valori mari).

Funcţie de forma pe care o îmbracă funcţia $h(x)$ identificăm, de exemplu, următoarele cazuri particulare:

a) $h(x)=x$:	$\theta = E\{x\} = m_x$ şi $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} a^i$	→ media aritmetică a eşantionului este estimator pentru media populaţiei;
b) $h(x)=(x-m_x)^2$:	$\theta = E\{(x-m_x)^2\} = \sigma_x^2$ şi $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} (a^i - m_x)^2$	→ varianţa eşantionului este estimator pentru varianţa populaţiei;
c) etc.		

unde, **parametrul θ al populaţiei are,** aşa cu cum se poate remarca, **valori tipice**, respectiv, **valori asociate diverselor momente statistice ale v.a. x**.

Observaţia 5.26: Aşa după cum am prezentat şi în ***Subcapitolul 5.2***, ***elemente fundamentale de statistică*** – spre deosebire de *variabilele statistice* ce desemnează o anumită caracteristică a populaţiei –, prin *parametru* al

aceleiaşi populaţiei am spus că se înţelege doar *o proprietate numerică a acesteia*. În acest moment, cu informaţiile prezentate până acum, suntem în măsură să dăm următorul exemplu simplu care să ne ajute să înţelegem (în caz că nu am reuşit să o facem deja) distincţia dintre cele două noţiuni. Astfel, dacă luăm în considerare parametrul *media* (valoarea speratăă) a unei *v.a. x* (fie aceasta de tip continuu) ce descrie o caracteristică a populaţiei, avem că ea este teoretic dată de relaţia (5.116):

$$m_x \overset{def}{=} E\{x\} = \int_{-\infty}^{+\infty} a f_x(a) da \tag{5.127}$$

iar empiric, ea este estimată prin media aritmetică a eşantionului (relaţia (5.123)):

$$\bar{x} = \frac{a^1 + \cdots + a^n}{n} \tag{5.128}$$

care, la limită, pentru $n \to \infty$, tinde spre valoarea reală a parametrului m_x. Cu alte cuvinte, atunci când ea există, *media* lui *x* este, aproape sigur, limita mediei eşantionului obţinută pentru o mărime a eşantionului ce creşte la infinit.

În acest context, termenul echivalent de "*valoare sperată*" dat parametrului *media* este unul care ne poate induce în eroare, el putând fi, în mod eronat, asociat cu „*valoarea cea mai probabilă*" a respectivei variabile statistice. La fel ca şi pentru media eşantionului, şi pentru media populaţiei putem obţine însă o valoare „*imposibilă*" în sensul că ea nu reprezintă o valoare tipică dintre valorile posibile pe care le poate lua efectiv *v.a. x*.

Astfel, dacă, de exemplu, considerăm eşantionul reprezentat de totalitatea pacientelor dintr-un spital de obstretică-ginecologie, pentru care observăm statistic *variabila statistică x* ce desemnează numărul de naşteri realizate de către fiecare pacientă (unitate statistică), atunci un estimat al mediei acestei variabile statistice, dat de media aritmetică a eşantionului, s-ar putea să ne conducă la un rezultat de tipul **3.7** naşteri/pacientă; acest rezultat statistic, care nu poate reprezenta nicicum o valoare posibilă în mod real, ne dă doar o informaţie orientativă şi anume aceea că, în medie, fiecare pacientă a avut circa 3.7 naşteri până la data realizării studiului.

Observaţia 5.27: O abordare similară a problemei de estimare parametrică punctuală o întâlnim şi în cazul în care avem de-a face cu **statistici**

comune pentru două variabile aleatoare, x şi y, ce caracterizează două trăsături distincte ale populaţiei. Aceasta implică următoarele:

- date două variabile aleatoare, x şi y, şi o funcţie $h(a,b)$, formăm *v.a.* $z = h(x,y)$;
- parametrul media $\theta = E\{h(x,y)\} = E\{z\}$ a funcţiei $h(x,y)$ a variabilelor aleatoare x şi y poate fi exprimat şi el direct în termenii funcţiei $h(x,y)$ şi ai funcţiei densitate comună, $f_{xy}(a,b)$, astfel:

$$E\{h(x,y)\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(a,b) f_{xy}(a,b)\,da\,db \qquad (5.129)$$

- se poate, de asemenea, demonstra că un *estimator empiric*, $\hat{\boldsymbol{\theta}}$, al parametrului θ de mai sus este statistica *media aritmetică a eşantionului* pentru v.a. z, deci pentru funcţia $h(x,y)$, a lui x şi y, respectiv:

$$\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} h(\mathrm{a}^{\mathrm{i}}, \mathrm{b}^{\mathrm{i}}) \qquad (5.130)$$

Funcţie de forma particulară pe care o îmbracă funcţia $h(x,y)$ identificăm, de exemplu, următoarele cazuri, deosebit de utile în aplicaţiile practice, pentru parametrul θ şi estimatorul său, $\hat{\boldsymbol{\theta}}$:

a) $h(x,y) = xy$:	$\theta = E\{xy\}$ $\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{a}^{\mathrm{i}}\mathrm{b}^{\mathrm{i}}$
b) $h(x,y) = (x-m_x)(y-m_y)$:	$\theta = E\{(x-m_x)(y-m_y)\} = cov(x,y)$ $\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} (\mathrm{a}^{\mathrm{i}} - m_x)(\mathrm{b}^{\mathrm{i}} - m_y)$
c) *etc.*	

- în cazul particular când variabilele aleatoare x şi y au o distribuţie comună normală funcţia densitate comună, $f_{xy}(a,b)$ este determinată în mod unic de următoarele momente de ordinul întâi şi doi: m_x, m_y, σ_x, σ_y şi $cov(x,y)$ (vezi şi ***Subcapitolul 5.6***).

Până în acest punct al prezentării noastre parametrii populaţiei au fost trataţi ca fiind nişte *constante necunoscute* – aceasta reprezintă, de fapt, ipoteza de lucru de bază a ***estimării clasice***. În estimarea parametrică mai există însă şi o abordare în care parametrii populaţiei sunt ei înşăşi văzuţi nu ca nişte cantităţi necunoscute şi *fixe* ci ca nişte *variabile aleatoare*, având

propriile lor distribuţii de probabilitate. Această nouă abordare poartă numele de ***estimare bayesiană*** a parametrilor populaţiei.

Estimarea parametrică punctuală Bayes-iană

În estimarea *Bayes*-iană a parametrilor unei populaţii se pleacă de la următoarele ***două premize***:

(1) parametrul θ al populaţiei nu este în totalitate unul necunoscut (altfel spus, dispunem de anumite informaţii *a priori* fie dintr-o cunoaştere parţială a procesului studiat, fie din alte realizări anterioare ale aceluiaşi experiment etc.), şi

(2) întrucât valoarea parametrului este necunoscută atunci are sens să specificăm o distribuţie de probabilitate care să descrie valorile posibile pentru acest parametru, la fel ca şi probabilităţile asociate acestor valori.

Informaţiile *a priori* despre parametrul θ sunt valorificate în momentul în care construim şi precizăm funcţia de densitate *a priori*, $f_{\theta}(\theta)$, a vectorului aleator $\boldsymbol{\theta}$. În această abordare**,** parametrul necunoscut al populaţiei este doar o valoare a vectorului aleator $\boldsymbol{\theta}$ iar funcţia de distribuţie a vectorului aleator x, ce descrie caracteristica populaţiei, este interpretată ca fiind funcţia de distribuţie condiţionată, $F_x(a|\theta)$, a lui x pentru $\boldsymbol{\theta} = \theta$.

În aceste ipoteze de lucru, *metoda bayesiană de estimare statistică* apare ca o alternativă la *metoda clasică de estimare statistică* de care se diferenţiază, însă, în principal, prin următoarele două elemente cheie:

i) *parametrul populaţiei* este el însuşi văzut ca un *vector aleator*, $\boldsymbol{\theta}$, *valoarea sa necunoscută*, *fiind o valoare a acestui vector aleator*;

ii) *informaţia furnizată de setul de date empirice*, de care dispunem la un moment dat, este *combinată*, *în plus*, *cu informaţia a priori pe care o avem despre parametrul θ al populaţiei*. Reamintim că, în estimarea parametrică clasică, informaţia folosită în ghidarea procesului de inferenţă statistică era doar cea furnizată de setul de date empirice.

O concluzie imediată ce poate fi trasă din cele de mai sus este aceea că în abordarea *Bayes*-iană, **problema estimării parametrului necunoscut al populaţiei revine la a estima valoarea** θ a lui $\boldsymbol{\theta}$ **în termenii**:

- eşantionului X (constând în valorile observate a^i ale lui x) şi, respectiv,
- funcţiei de densitate de probabilitate a lui $\boldsymbol{\theta}$.

Observaţia 5.28: Atenţie! Prin transformarea problemei dintr-o estimare a parametrului necunoscut θ (întâlnită în estimarea clasică) într-o problemă de estimare a valorii θ a vectorului aleator $\boldsymbol{\theta}$ (în cazul abordării *bayes*-iene), ***estimarea*** devine în această ultimă abordare, practic, ***o predicţie*** (vezi diferenţa dintre acestea în **Anexa: Predicţie versus estimare**).

Ţinând cont de faptul că problema de estimare s-a transformat într-o problemă de predicţie (vezi ***Observaţia 5.28***, respectiv, relaţia (A.71)), atunci putem afirma că soluţia pe care o căutăm – fie aceasta constanta $\hat{\theta}$:

- **era**, înainte de a obţine orice evidenţă statistică (date observabile), **dată de**,

$$\hat{\theta} = E\{\boldsymbol{\theta}\} = \int_{-\infty}^{+\infty} \theta f_{\boldsymbol{\theta}}(\theta) d\theta \tag{5.131}$$

- iar după obţinerea de noi probe (respectiv, obţinerea unui set de date empirice pentru vectorul aleator x, $X=[a^1,..., a^n]$), ea **este dată de** (vezi relaţia (5.187)),

$$\hat{\theta} = E\{\boldsymbol{\theta} \mid X\} = \int_{-\infty}^{+\infty} \theta \cdot f_{\boldsymbol{\theta}|\mathbf{X}}(\theta \mid X) d\theta \tag{5.132}$$

În formula de mai sus $f_{\boldsymbol{\theta}|\mathbf{X}}(\theta|X)$ este tot funcţia densitate de probabilitate a lui $\boldsymbol{\theta}$ însă **revizuită** în sensul că funcţia de densitate *a priorică*, $f_{\boldsymbol{\theta}}(\theta)$ – cunoscută în totalitate şi construită în baza tuturor informaţiilor de care dispuneam înainte de a obţine noi rezultate experimentale – este modificată corespunzător, în contextul şi conform noilor dovezi statistice obţinute experimental.

Derivarea noii funcţii de densitate pentru vectorul aleator $\boldsymbol{\theta}$ are ca regulă de calcul regula lui Bayes (vezi relaţia (5.93)),

$$f_x(a \mid B) = \frac{P(B \mid x = a) f_x(a)}{P(B)} = \frac{P(B \mid x = a) f_x(a)}{\int_{-\infty}^{+\infty} P(B \mid x = a) f_x(a) da} \tag{5.133}$$

particularizată pentru $x=\boldsymbol{\theta}$ şi, respectiv, pentru evenimentul $B = \mathbf{X}$, ce constă în obţinerea unui eşantion de dimensiune n pentru vectorul aleator x. Astfel:

$$f_{\theta|\mathrm{X}}(\theta \mid X) = \frac{f_{\mathrm{X}|\theta}(X \mid \theta) f_{\theta}(\theta)}{f_{\mathrm{X}}(X)} \tag{5.134}$$

În relaţia de mai sus notaţiile folosite sunt aceleaşi cu cele folosite în estimarea parametrică clasică (vezi ***Subcapitolul 5.5.1.***), respectiv:

- $\boldsymbol{\theta}$ este vectorul aleator ce desemnează parametrul populaţiei,
- θ este o realizare particulară a acestui vector aleator,
- **X** desemnează vectorul aleator sau matricea aleatoare (după cum x este un scalar, respectiv, un vector) a eşantionului vectorului aleator x (vezi **Figura 5.17**) iar
- X reprezintă o realizare particulară a lui **X**, respectiv, setul de date empirice pe baza căruia se realizează inferenţa statistică.

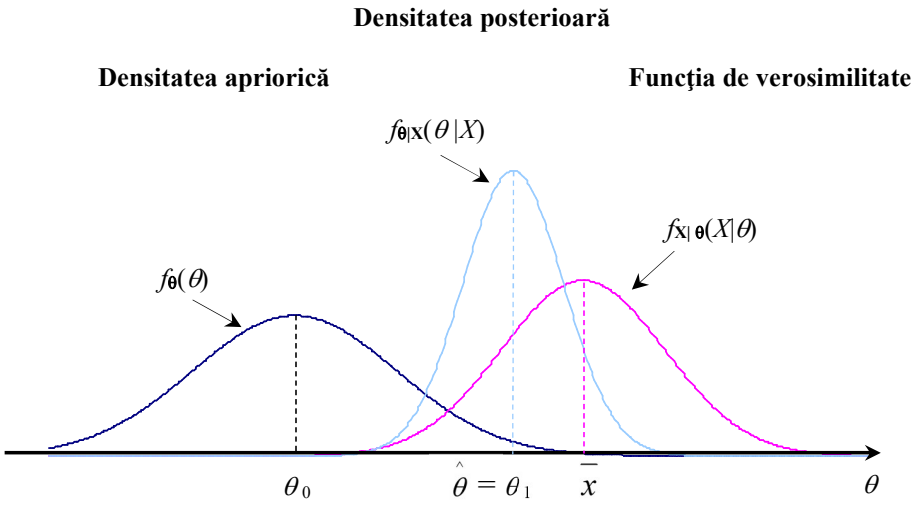


Figura 5.17. Exemplu de estimare bayesiană punctuală parametrică în ipoteza unor distribuţii normale ale lui $\boldsymbol{\theta}$ şi, respectiv, x.

Această formulă de calcul, ilustrată grafic în **Figura 5.17**, ne arată cum, *pornind de la informaţia disponibilă* – specificată prin *funcţia de densitate a priori*, $f_{\boldsymbol{\theta}}(\theta)$, pe care o presupunem cunoscută –, *în lumina unei noi evidenţe statistice* – respectiv, datele observabile ale eşantionului, a căror informaţie furnizată este reflectată în *funcţia de verosimilitate*, $f_{\mathbf{X}|\boldsymbol{\theta}}(X|\theta)$ – *derivăm o funcţie de densitate de probabilitate* ***actualizată*** pentru vectorul aleator $\boldsymbol{\theta}$, $f_{\boldsymbol{\theta}|\mathbf{X}}(\theta|X)$, funcţie denumită, în mod sugestiv, *funcţie densitate de probabilitate posterioară* sau *funcţie de densitate de probabilitate condiţionată* a lui $\boldsymbol{\theta}$ dată de evidenţa **X**.

Observaţia 5.29.: Facem observaţia aici că pentru funcţia $f_{\mathbf{X}|\boldsymbol{\theta}}(X|\theta)$ – ce reprezintă densitatea condiţionată comună a n vectori aleatori, $\mathbf{a}^i$, atunci

când $\hat{\theta} = \theta$ –, este adevărată următoarea relaţie de calcul pentru cazul când aceşti vectori aleatori sunt independenţi:

$$f_{\mathrm{X}|\theta}(X \mid \theta) = f_{\mathrm{a}^1|\theta}(a^1 \mid \theta) \cdots f_{\mathrm{a}^n|\theta}(a^n \mid \theta) \tag{5.135}$$

Aici, funcţia generică $f_{\mathbf{a}|\boldsymbol{\theta}}(a|\theta)$ reprezintă funcţia densitate condiţionată a vectorului aleator **a** atunci când $\boldsymbol{\theta} = \theta$.

În situaţia particulară când funcţia $f_{\mathbf{X}|\boldsymbol{\theta}}(X|\theta)$ este considerată ca o funcţie de θ, ea poartă numele de funcţie de verosimilitate; acesta este şi cazul funcţiei $f_{\mathbf{X}|\boldsymbol{\theta}}(X|\theta)$ din relaţia (5.198), unde pentru o valoare particulară a lui **X** (un eşantion X_k oarecare, vezi **Figura 5.16**) avem:

$$f_{\mathrm{X}|\theta}(X_k \mid \theta) = f_{\mathrm{a}^1|\theta}(a_k^1 \mid \theta) \cdots f_{\mathrm{a}^n|\theta}(a_k^n \mid \theta) \tag{5.136}$$

Problema 5.22: Să se estimeze punctual, prin metoda *bayes*-iană, parametrul scalar *media*, $\theta = m_x$, al unei populaţii a cărei caracteristică studiată este descrisă de vectorul aleator unidimensional, x, despre care se ştie că este distribuit conform unei legi de distribuţie normale,

$$x \sim N(\theta, \sigma_x^2) \tag{5.137}$$

şi pentru care se cunoaşte doar varianţa, σ_x^2. De asemenea, mai dispunem:

- de o serie de informaţii *a priori* despre parametrul θ care ne fac să credem că valorile posibile pentru acesta respectă, de asemenea, o distribuţie normală, de medie şi varianţă, cunoscute, respectiv,

$$\boldsymbol{\theta} \sim N(\theta_0, \sigma_0^2) \tag{5.138}$$

- şi de un eşantion, de dimensiune n, de date empirice.

Să se comenteze rezultatul obţinut.

Notă Se arată că estimatul $\hat{\theta} = \theta_1$ pentru parametrul necunoscut al populaţiei este unul dat de formula:

$$\theta_1 = \frac{n\sigma_1^2}{\sigma_x^2}\bar{x} + \frac{\sigma_1^2}{\sigma_0^2}\theta_0 \tag{5.139}$$

unde $\bar{x}$ reprezintă statistica media aritmetică a eşantionului pentru v.a. x iar,

$$\sigma_1^2 = \frac{\sigma_x^2 \sigma_0^2}{\sigma_x^2 + n\sigma_0^2} \tag{5.140}$$

Observaţia 5.30: În problema de mai sus se poate remarca faptul că, atât distribuţia parametrului θ, cât şi distribuţia caracteristicii populaţiei (reprezentată prin *v.a.* x), sunt distribuţii normale iar distribuţia posterioară obţinută pentru θ în urma aplicării regulei lui *Bayes* este, de asemenea, o distribuţie normală. Rezultatul nu este, aşa după cum s–ar putea crede, unul total întâmplător. Mai exact, dacă în general în estimarea parametrică bayesiană familia de distribuţii căreia îi aparţine funcţia de verosimilitate, $f_{\mathbf{X}|\theta}(X|\theta)$, se consideră a fi una bine-determinată (fixată), în ceea ce priveşte distribuţia *a priorică* a parametrului populaţiei, $f_{\theta}(\theta)$, avem doar presupunerea crucială că aceasta este o distribuţie cunoscută, rămânând ca alegerea uneia sau a alteia dintre familiile de distribuţii teoretice existente să se facă, în parte pe baza informaţiilor *a priori* disponibile, iar în parte, pe baza intuiţiei noastre. Este însă evident faptul că alegeri diferite pentru densitatea *a priori* $f_{\theta}(\theta)$ poate face ca produsul $f_{\mathbf{X}|\theta}(X|\theta) \times f_{\theta}(\theta)$ să ia o formă algebrică arbitrară sau alta (de cele mai multe ori, destul de greu de descris), formă care îngreuiază foarte mult, dacă nu chiar, face imposibil calculul distribuţiei posterioare, $f_{\theta|\mathbf{X}}(\theta|X)$.

Soluţia în astfel de situaţii o reprezintă folosirea aşa – numitelor **densităţi** *a priori* **conjugate**.

O clasă de distribuţii de probabilitate *a priori*, $f_{\theta}(\theta)$, se numeşte **conjugata** unei clase de funcţii de verosimilitate, $f_{\mathbf{X}|\theta}(X|\theta)$, doar atunci când distribuţiile posterioare rezultante, $f_{\theta|\mathbf{X}}(\theta|X)$, fac parte din aceeaşi familie ca şi $f_{\theta}(\theta)$. Practic, o probabilitate *a priori* conjugată este o convenienţă algebrică care simplifică foarte mult scrierea matematică, oferind, în acest mod o formulă simplă de calcul ce leagă valorile hyper-parametrilor[42] densităţii posterioare de valorile hyper-parametrilor densităţii *a priori* [Gelman, 2003], [Fink, 1995]. În acest fel, şi calculul densităţii posterioare devine unul foarte uşor, densitatea posterioară fiind, în final, o densitate din aceeaşi familie de distribuţie ca şi densitatea *a priori* însă cu hiperparametri diferiţi ce reflectă efectul, suplimentar, al informaţiei furnizate de datele empirice.

[42] Un **hyper-parametru** reprezintă un parametru al distribuţiei *a priori* iar termenul este utilizat pentru a distinge acest parametru de parametrii modelului propus pentru distribuţia ce descrie caracteristica populaţiei.

Revenind la ***Problema 5.22*** putem acum bănui că presupunerea conform căreia distribuţia *a priori* pentru θ_0 este o distribuţie normală a fost una de natură să simplifice scrierea matematică. Uitându-ne în tabelul cu funcţii densitate *a priori* conjugate, prezentat în **Anexa: Distribuţii de verosimilitate**, observăm că, într-adevăr, familia de distribuţii *gauss*-iene este conjugată cu ea însăşi (auto-conjugată). Mai mult, corespunzător celor prezentate mai sus şi ţinând cont de soluţia dată problemei identificăm:

Funcţia de verosimilitate	**Parametrii necunoscuţi ai modelului**	**Distribuţia *a priori* conjugată**	**Hyper-parametrii *a priori***	**Hyper-parametrii posteriori**
Normală cu varianţă, σ_x^2, cunoscută	$\theta = m_x$ (media)	Normală	θ_0, σ_0^2	$\theta_1 = \left(\frac{n\sigma_0^2}{\sigma_x^2 + n\sigma_0^2}\right)\bar{x} + \frac{\sigma_x^2}{\sigma_x^2 + n\sigma_0^2}\theta_0$ $\sigma_1^2 = \frac{\sigma_x^2\sigma_0^2}{\sigma_x^2 + n\sigma_0^2}$.

Probabilitatea a priori

Deşi subiectivă în esenţă (a se vedea şi **Anexa: Interpretarea noţiunii de probabilitate**), probabilitatea *Bayes*-iană este interpretată uneori ca fiind ***obiectivă.*** Cele două puncte de vedere diferite – varianta **subiectivă** şi, respectiv, varianta **obiectivă** ale probabilităţii bayesiene – diferă, în principal, prin modul cum este **interpretată şi construită probabilitatea *a priori*** din relaţia lui Bayes. Astfel:

- conform punctului de vedere **subiectiv**, probabilitatea este interpretată ca fiind "*gradul de încredere*" pe care un individ îl are în adevărul unui enunţ/eveniment; de aici rezultă şi caracterul personal sau subiectiv; cu alte cuvinte, nivelul de cunoaştere corespunde unei "convingeri personale";
- conform punctului de vedere **obiectiv**, nivelul *a priori* al cunoaşterii defineşte – pentru problemele bine puse – o distribuţie de probabilitate *a priori* unică. În general, problemele de inferenţă statistică, aşa cum se întâlnesc ele în investigarea ştiinţifică, sunt rezolvate, în mod uzual, plecând de la presupunerea că un anume model statistic este un descriptor adecvat al mecanismului probabilistic ce a generat datele; însăşi alegerea

acelui model implică, însă, în continuare, un anume grad de subiectivitate. Cu toate acestea, a devenit o practică standard ca orice analiză statistică ce depinde doar de modelul propus şi de datele observate să fie descrisă ca fiind "obiectivă". În acest sens (şi doar în acest sens) inferenţa bayesiană poate fi privită ca fiind "obiectivă". În construirea obiectivă a distribuţiilor *a priori*, au fost propuse utilizarea mai multor principii dintre care amintim doar: *entropia maximă* (Edwin T. Jayne), *analiza de referinţă* (Jose M. Bernardo) şi *analiza frecvenţială* (în această ultimă situaţie distribuţia de frecvenţă are parametri ce reflectă o analiză şi o sinteză a unor date existente).

În ceea ce priveşte **densităţile *a priori* de referinţă**, acestea nu sunt, aşa cum s-ar putea crede, descriptori ai gradelor subiective de încredere; ele sunt propuse ca funcţii *a priori* formale, unanim acceptate spre a fi utilizate ca standarde în comunicările ştiinţifice.

Ceea ce trebuie să reţinem în final, de aici, este faptul că **multe dintre metodele moderne de învăţare maşină şi de clasificare se bazează, în general, pe principiile bayesiene obiective**.

Estimarea parametrică a funcţiei de regresie

Cele două metode de estimare parametrică discutate până acum reprezintă, aşa după cum ştim *metode de estimare a funcţiilor densitate de probabilitate* aferente diferitelor variabile aleatoare (mai general, vectori aleatori). Reluând puţin, vom spune că, în cadrul acestor metode se pleacă de la un eşantion, X, de date observate ale variabilei aleatoare x, se presupune o anumită formă pentru densitatea de probabilitate a lui x, $f_x(a)$ (suplimentar, în cazul estimării *Bayes*-iene, se precizează şi densitatea *a priorică* a parametrului θ al modelului, $f_\theta(\theta)$) şi, pe baza acestora, se estimează parametrul θ, ce implicit determină complet estimatul funcţiei densitate, $f_x(a)$, a lui x.

Un alt tip de aplicaţii statistice cu care ne întâlnim destul de frecvent în practică şi care implică, de asemenea, o estimare parametrică – de această dată, însă, a unei funcţii numită de regresie – este şi următorul, în care: se cunoaşte un eşantion, X, de perechi de date (a^i, b^i), ale unei perechi de variabile aleatoare (x, y), între care ştim că există o *relaţie funcţională*, se cunoaşte forma acestei relaţii funcţionale şi, pe baza acestora, se estimează vectorul parametru, θ, ce descrie complet relaţia funcţională de mai sus. În acest caz particular, vorbim de o estimare parametrică a unei funcţii de regresie. Funcţia de regresie are, după cum vom vedea, o utilitate foarte largă, ea putând fi utilizată în aplicaţii diverse, începând cu aplicaţiile de predicţie,

inferenţă statistică, testare de ipoteze şi terminând, nu în cele din urmă, cu aplicaţiile de modelare a relaţiilor cauzale.

Estimarea unei funcţii de regresie constă în găsirea acelui model care aproximează cel mai bine relaţia funcţională care există între o variabilă aleatoare dependentă, *y*, (numită şi *variabilă de răspuns*) şi, respectiv, una sau mai multe variabile aleatoare independente (numite şi predictori). Formal, ecuaţia de regresie pentru cazul unei singure *v.a.* independente, *x* (generalizarea la mai mulţi predictori este una imediată), se scrie astfel:

$$b = \phi(a;\theta) + e \tag{5.141}$$

În această relaţie *v.a.* dependentă *y* este modelată ca o funcţie de predictorul *x*, de vectorul parametru corespunzător θ (ce reprezintă un vector de componente „constante") la care se adaugă un termen de eroare ce reprezintă acea variaţie din *y* ce nu a putut fi explicată (modelată). Termenul de eroare este aici tratat ca o variabilă aleatoare. Reamintim că variabilele *a* şi *b* desemnează variabilele generice aferente variabilelor aleatoare *x* şi, respectiv, *y*.

Înainte de a merge mai departe cu prezentarea acestui tip de estimare, vom introduce în cele ce urmează câteva noţiuni noi. Scopul este acela de a uşura înţelegerea modului cum vom obţine un estimator pentru funcţia de mai sus – funcţie numită şi *funcţie* sau *linie de regresie*.

Media condiţionată a unei variabile aleatoare. Funcţia mediei condiţionate

Prin ***definiţie***, **media condiţionată** a unei variabile aleatoare *y*, presupunând evenimentul *A*, este dată de integrala:

$$E\{y \mid A\} \overset{def}{=} \int_{-\infty}^{+\infty} b f_y(b \mid A) db \tag{5.142}$$

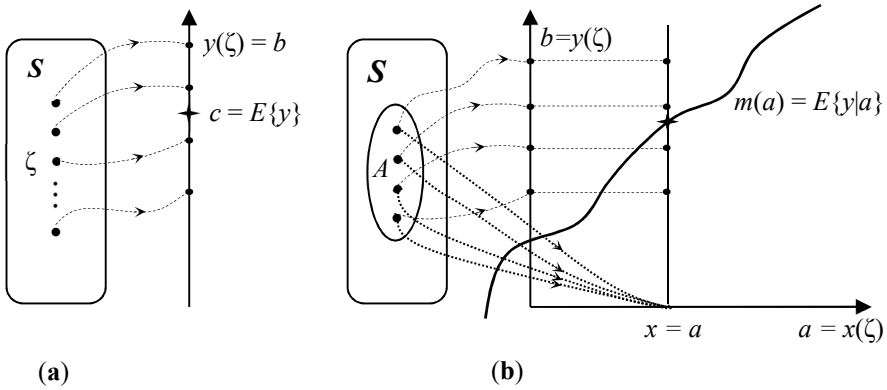
În contextul problemei de estimare de mai sus vom considera pentru evenimentul *A* din relaţia (5.142), următorul eveniment $\{x = a\}$. Ţinând cont şi de relaţia (5.78) obţinem:

$$E\{y \mid x = a\} = \int_{-\infty}^{+\infty} b f_{y|x}(b \mid a) db = \int_{-\infty}^{+\infty} b \frac{f_{xy}(a,b)}{f_x(a)} db = m(a) \tag{5.143}$$

unde $f_{y|x}(b|a)$ este funcţia densitate de probabilitate condiţionată a lui *y* dată de $x = a$, $f_y(b)$ este *fdp* marginală a lui *y* iar $m(a)$ este o notaţie pe care o folosim pentru a desemna funcţia rezultantă ce depinde doar de variabila *a*.

Dacă în relaţia de mai sus baleiem toată mulţimea valorilor posibile, *a*, pentru *v.a. x* atunci obţinem o funcţie de la *a* la $b = m(a)$, numită ***funcţia mediei condiţionate*** sau ***funcţia de regresie***; această funcţie, desemnată adesea prin ***E{y|x}***, ne spune cum sunt legate între ele „în medie" variabilele aleatoare *x* şi *y* (vezi **Figura 5.18**). Pentru a înţelege reprezentarea din **Figura 5.18.(a)** vezi problema de predicţie tratată în **Anexa: Predicţie versus estimare**, şi rezumată prin formula (A.71). O înţelegere completă (şi nu doar intuitivă) a reprezentării din **Figura 5.18.(b)** o vom obţine abia după

prezentarea teoriei *estimării neliniare, în sensul erorii medii pătratice* minime (*mean square* - MS), pe care o redăm în acest subcapitol.



(**a**) (**b**)

Figura 5.18. Reprezentarea grafică a estimării variabilei aleatoare *y*: **(a)** printr-o constantă *c* şi, respectiv, **(b)** printr-o funcţie $m(x)$ a unei v.a. *x*.

Funcţia mediei condiţionate poate fi o funcţie neliniară chiar şi pentru densităţi comune, $f_{xy}(a,b)$, şi densităţi marginale aparent simple la prima vedere.

Problema 5.23: Fie următoarea funcţie densitate de probabilitate comună pentru variabilele aleatoare *x* şi *y*:

$$f_{xy}(a,b) = \begin{cases} a+b & , \textit{pentru } a \in [0,2] \textit{ şi } b \in [0,2] \\ 0 & , \textit{în rest}. \end{cases} \tag{5.144}$$

Să se deducă funcţia mediei condiţionate pentru *v.a. y*, funcţie de *v.a. x*.

Rezolvare:

Din definiţia lui $f_{xy}(a,b)$ obţinem, conform relaţiei (5.43), funcţia densitate de probabilitate marginală a lui *x* astfel:

$$f_x(a) = \begin{cases} \int_0^2 f_{xy}(a,b)db = \left(ab + \frac{1}{2}b^2 \right)\Big|_0^2 = 2a+2 & , \textit{pentru } a \in [0,2] \\ 0 & , \textit{în rest}. \end{cases} \tag{5.145}$$

Pe baza relaţiei (5.207) deducem mai departe funcţia mediei condiţionate a lui *y* funcţie de *x*:

$$\begin{aligned} E\{y|x=a\} &= \int_0^2 b \frac{f_{xy}(ab)}{f_x(a)} db = \int_0^2 b \frac{a+b}{2a+2} db = \int_0^2 \frac{a}{2a+2} bdb + \int_0^2 \frac{1}{2a+2} b^2 db \\ &= \frac{a}{2a+2} \cdot \frac{b^2}{2}\Big|_0^2 + \frac{1}{2a+2} \cdot \frac{b^3}{3}\Big|_0^2 = \frac{a}{a+1} + \frac{4/3}{a+1} = \\ &= \frac{3a+4}{3a+3} \end{aligned} \tag{5.146}$$

Aşa cum arată şi rezultatul de mai sus, funcţia obţinută în acest caz este, evident, o funcţie neliniară (vezi figura **Figura 5.19**); acest lucru vine şi confirmă afirmaţia că, chiar şi pentru forme simple ale densităţilor comune şi marginale, funcţia $m(a)$ poate fi o funcţie neliniară.

Revenind acum la problema estimării relaţiei funcţionale, $\phi(\cdot)$, ce leagă între ele cele două variabile aleatoare, x şi y, prezentăm în continuare estimatorul pe care-l obţinem în estimarea neliniară în sensul MS a acestei funcţii.

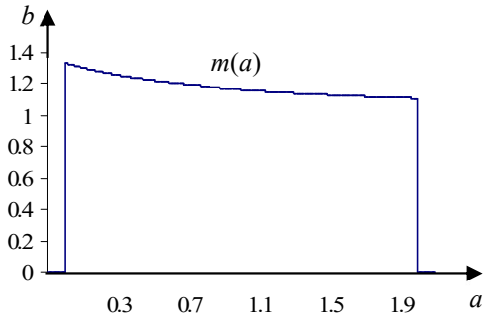


Figura 5.19. Funcţia mediei condiţionate, $m(a)$.

Estimarea neliniară în sensul erorii medii pătratice minime

Această problemă de estimare revine la a găsi acea funcţie $\phi(\cdot)$ pentru care eroarea medie pătratică:

$$e = E\{[y-\phi(x)]^2\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[b-\phi(a)]^2 f_{xy}(a,b)dadb \qquad (5.147)$$

este minimă.

Pentru a obţine estimatorul corespunzător pornim de la relaţia de mai sus în care rescriem densitatea comună (vezi şi relaţia (5.69)) astfel:

$$\begin{aligned} e &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[b-\phi(a)]^2 f_{y|x}(b\,|\,a)f_x(a)dadb = \\ &= \int_{-\infty}^{+\infty} f_x(a)\int_{-\infty}^{+\infty}[b-\phi(a)]^2 f_{y|x}(b\,|\,a)dbda \end{aligned} \qquad (5.148)$$

În relaţia (5.212) toţi termenii care se integrează sunt pozitivi ceea ce înseamnă că, eroarea e astfel calculată va fi minimă doar dacă integrala,

$$\int_{-\infty}^{+\infty}[b-\phi(a)]^2 f_{y|x}(b\,|\,a)db \qquad (5.149)$$

este minimă pentru fiecare a. Soluţia care se obţine în acest caz este (a se vedea **Anexa: Predicţie versus estimare**, relaţia (A.68), în care c este

înlocuit cu $\phi(a)$ iar $f_x(a)$ este înlocuit cu $f_{y|x}(b,a)$):

$$\phi(a) = \int_{-\infty}^{+\infty} b f_{y|x}(b \mid a) db = E\{y \mid a\} = m(a) \tag{5.150}$$

Din relaţia de mai sus reiese că, estimatul obţinut pentru funcţia ce leagă între ele cele două variabile aleatoare este chiar *funcţia mediei condiţionate a variabilei dependente y dată de variabila independentă x*. Cu alte cuvinte, rescriind relaţia **Error! Not a valid link.** în lumina noilor rezultate obţinute, deducem:

$$b = m(a;\theta) + e \tag{5.151}$$

Numele acestei metode de estimare, respectiv, acela de ***estimare parametrică neliniară*** a regresiei se datorează faptului că forma funcţiei de regresie este una ce se presupune a fi cunoscută, ea este dată de o combinaţie neliniară a parametrilor modelului şi, în plus, parametrii modelului au o semnificaţie particulară (uneori chiar fizică) pentru procesul studiat. Metoda estimează aceşti parametri în sensul găsirii acelor valori pentru care se obţine cea mai bună aproximare a datelor observate în condiţiile îndeplinirii unui criteriu de eroare anterior stabilit. Estimarea parametrilor se face printr-o *metodă a aproximărilor successive* şi, din acest motiv, regresia neliniară este cunoscută ca fiind una intens computaţională.

Observaţia 5.31: Funcţia de regresie din ***Problema 5.23***,

$$m(a) = 1 + \frac{1}{3a+3} = \alpha_0 + \frac{\alpha_1}{\alpha_2 a + \alpha_3} \tag{5.152}$$

este o funcţie neliniară întrucât ea nu poate fi exprimată ca o combinaţie liniară a parametrilor α_i. Exemple de funcţii neliniare mai uzual folosite sunt: funcţiile gaussiene, funcţiile putere, funcţiile exponenţiale, funcţiile trigonometrice etc.

O alternativă mai simplă dar, în general, mai puţin eficientă decât metoda parametrică neliniară de estimare a regresiei este şi metoda estimării parametrice liniare, în sens MS, a acesteia.

Estimarea liniară în sensul erorii medii pătratice minime

În cazul liniar univariat, problema estimării parametrice a funcţiei de regresie dintre variabilele aleatoare x şi y se reduce la estimarea, în sensul erorii medii pătratice, a *v.a. y* în termenii unei funcţii liniare a lui x, funcţie de tipul $Ax+B$. Aşa după cum vom vedea mai jos, parametrii cei mai buni de aproximare au, în estimarea liniară a regresiei, expresii bine determinate; în

cazul nostru vom arăta că parametrii estimaţi se calculează pe baza momentelor de ordin unu şi doi ale distribuţiilor lui *x* şi, respectiv, *y*.

Pentru aceasta, scriem, mai întâi formula pentru eroarea medie pătratică, ce trebuie minimizată:

$$e = E\{[y-(Ax+B)]^2\} \tag{5.153}$$

Pentru un *A* fixat eroarea *e* de mai sus este minimă numai dacă:

$$\frac{de}{dB} = 0 \Leftrightarrow \frac{E\{y^2 - 2(Ax+B)y + [(Ax)^2 + 2AxB + B^2]\}}{dB} = \\ = E\{-2y + 2Ax + 2B\} = 0 \tag{5.154}$$

de unde rezultă mai departe că,

$$E\{-y + Ax\} + E\{B\} = 0 \Rightarrow B = E\{y - Ax\} = m_y - Am_x \tag{5.155}$$

Înlocuind în relaţia **Error! Not a valid link.** pe *B* cu valoarea sa dedusă în relaţia (5.219) obţinem:

$$\begin{aligned} &e = E\{[y-(Ax+m_y-Am_x)]^2\} = E\{[(y-m_y)-A(x-m_x)]^2\} \Rightarrow \\ &e = E\{(y-m_y)^2\} - 2AE\{(y-m_y)(x-m_x)\} + A^2E\{(x-m_x)^2\}\} \Rightarrow \\ &e = \sigma_y^2 - 2A\rho\sigma_x\sigma_y + A^2\sigma_x^2 \end{aligned} \tag{5.156}$$

unde: σ_x^2, σ_y^2 sunt varianţele celor două variabile aleatoare, *x* şi *y*, iar ρ este coeficientul de corelaţie Pearson (vezi ***Subcapitolul 5.5.4***, relaţia (5.245)).

Mai departe relaţia (5.156), ce reprezintă eroarea *e* ca o funcţie de *A*, este minimizată în raport cu *A* dacă:

$$\frac{de}{dA} = 0 \Leftrightarrow -2\rho\sigma_x\sigma_y + 2A\sigma_x^2 = 0 \Rightarrow A = \frac{\rho\sigma_y}{\sigma_x} \tag{5.157}$$

În concluzie, parametrii *A* şi *B* pentru care se obţine cel mai bun estimat liniar, în sensul erorii medii pătratice, al funcţiei de regresie a două variabile aleatoare *x* şi *y* sunt daţi de relaţiile:

$$A = \frac{\rho\sigma_y}{\sigma_x} \text{ şi } B = m_y - Am_x \tag{5.158}$$

Observaţia 5.32: În general, funcţia de regresie neliniară nu este o linie dreaptă iar eroarea medie pătratică este mai mică decât eroarea în sens mediu pătratic din estimarea liniară regresiei.

Problema 5.24: Să se demonstreze că dacă variabilele aleatoare x şi y au o densitate comună de tip *Gauss*-iană atunci, estimaţiile, în sensul erorii medii pătratice minime, determinate prin cele două metode (liniară, respectiv, neliniară) pentru funcţia de regresie sunt identice.

În cazul în care există mai mulţi predictori, x_i, în funcţie de care se doreşte exprimarea unei *v.a.* dependente y, atunci avem de-a face cu o regresie liniară multiplă iar modelul în acest caz este dat de:

$$y = \alpha_0 + \alpha_1 x_1 + ... + \alpha_k x_k + e \qquad (5.159)$$

unde: x_i, cu $i = \overline{1,k}$ sunt predictorii (regresorii) iar α_i sunt coeficienţii de regresie. Formula de determinarea a coeficienţilor de regresie este una care diferă funcţie de criteriul de eroare folosit. Cele mai multe însă din aplicaţiile de estimare parametrică a regresiei liniare folosesc metoda celor mai mici pătrate.

În mod echivalent, ca în cazul estimării neliniare, şi în cazul estimării liniare a regresiei numele metodei, respectiv, acela de ***estimare parametrică liniară*** se datorează faptului că forma funcţiei de regresie este una cunoscută, ea este dată de o combinaţie liniară a parametrilor modelului şi, în plus, parametrii modelului au, aşa cum am văzut, o semnificaţie particulară pentru procesul studiat.

Observaţia 5.33: În cazul particular discutat anterior am avut de-a face cu o regresie liniară simplă (o singură *v.a.* independentă) iar funcţia de regresie estimată a fost o linie dreaptă. Nu trebuie să facem însă confuzie între numele metodei (estimare liniară) şi această linie dreaptă care nu este decât un caz particular al funcţiei de regresie liniare. Mai mult chiar, inclusiv o funcţie de forma:

$$y = \alpha_0 + \alpha_1 x_1^3 + ... + \alpha_k x_k^2 + e \qquad (5.160)$$

care este evident o funcţie neliniară în variabilele independente x_i, este o funcţie de regresie liniară întrucât, să ne reamintim – caracterul liniar al metodei este dat de liniaritatea exprimată în termenii parametrilor modelului – ceea ce se verifică şi pentru relaţia noastră.

Estimarea parametrică a unui semnal

Pe lângă aplicaţiile de modelare a densităţilor de probabilitate sau a funcţiilor de regresie, estimarea parametrică îşi mai găseşte aplicabilitate şi în domenii cum ar fi modelarea unui semnal (pentru definiţia unui semnal vezi **Anexa: Noţiunea de semnal. Semnal staţionar**).

Estimarea unui semnal presupune găsirea unui model matematic care să fie, pe de o parte, cât mai apropiat de realitatea fizică care a generat semnalul observat iar pe de altă parte, acelaşi model trebuie să prezinte şi un grad mare de generalitate.

Considerăm în continuare problema standard a unui semnal S peste care s-a suprapus un zgomot Z:

$$x_i = s_i + z_i,\ 1 \le i \le M \tag{5.161}$$

În ecuaţia de mai sus z_i sunt variabile aleatoare i.i.d., distribuite normal, după o distribuţie $N(0, \sigma^2)$ iar x_i sunt datele observate ale semnalului.

Estimarea, în acest caz, presupune că se cunoaşte modelul care ar descrie procesul ce a generat datele semnalului şi se doreşte, pe baza datelor observate să se estimeze parametrii modelului, care descriu modelul, având totodată o anumită semnificaţie pentru acesta.

Observaţia 5.34: Un exemplu ilustrativ de estimare parametrică a unui semnal este şi aplicaţia în care se urmăreşte modelarea unui semnal discret despre care se ştie că este sinusoidal şi pentru care se dispune de date observate. Precizăm aici că, o sinusoidă este orice funcţie de forma $A\cdot\sin(\omega t + \phi)$, unde t este variabila independentă iar A, ω şi ϕ sunt parametrii ficşi ai sinusoidei, numiţi amplitudinea, frecvenţa şi, respectiv, faza semnalului.

În acest caz, cel mai simplu model propus pentru semnalul s sinusoidal este dat de o singură componentă sinusoidală complexă, $A\exp(j\omega_0 n)$. Rescriind relaţia (B036) pentru această aplicaţie particulară obţinem:

$$x[n] = Ae^{j\omega_0 n} + z[n] \tag{5.162}$$

cu $1 \le n \le M$, iar $x[n]$ sunt datele observate ale semnalului, iar $\theta = [A, \omega, \phi]^T$ este vectorul parametru al modelului propus, vector ce trebuie estimat. Estimarea parametrilor modelului matematic (parametri care, aşa cum am văzut, caracterizează procesul analizat, având o semnificaţie particulară) presupune folosirea analizei statistice, respectiv, definirea unui criteriu de eroare. La ora actuală există propuse mai multe criterii de eroare însă dintre acestea, de departe cea mai folosită este metoda celor mai mici pătrate, în care se urmăreşte minimizarea sumei pătratelor erorii dintre datele observate şi datele generate de model:

$$J(\theta) = \sum_{n=0}^{M-1} |x[n] - \hat{x}[n]|^2 \tag{5.163}$$

Minimizarea lui $J(\theta)$ se face în raport cu vectorul parametru.

În estimarea parametrilor modelului unui semnal se pot folosi atât metode clasice de estimare statistică (de exemplu, metoda celor mai mici pătrate, metoda momentului, etc.), cât şi metode *Bayes*-iene (de exemplu, metoda probabilităţii posterioare maxime), filtrul Kalman ş.a.m.d.

2. Estimarea neparametrică şi semiparametrică

Aşa după cum am amintit deja, este posibil să se facă inferenţe statistice şi fără să se presupună *a priori* o anumită familie parametrică a distribuţiilor de probabilitate sau un anumit model pentru funcţia de regresie sau semnalul analizat. În acest caz vorbim de modele neparametrice (respectiv, de statistici neparametrice[43]) şi de modele semiparametrice.

Modelele neparametrice

Modelele neparametrice diferă de cele parametrice prin aceea că structura modelului este una mult mai flexibilă, ea nefiind specificată *a priori* ci ea este determinată direct din date. Din acest motiv, modelele neparametrice mai sunt numite şi *modele ce nu depind de distribuţie*.

Anticipând puţin, multe dintre metodele neparametrice folosesc în cursul procesului de estimare, aşa după cum vom vedea, funcţii de forma:

$$f(a) = \sum_{i=1}^{\infty} \alpha_i \varphi_i(a) \tag{5.164}$$

unde α_i sunt consideraţi „*parametrii*" estimatorului neparametric iar $\varphi_i(\cdot)$ reprezintă nişte *funcţii bază* (vezi **Anexa: Vectori bază şi funcţii bază**).

După cum se poate vedea termenul ***neparametric***, asociat oarecum impropriu acestor modele statistice, nu semnifică faptul că astfel de modele sunt lipsite complet de parametri ci, faptul că:

- pe de o parte, numărul şi natura parametrilor sunt fexibile şi nu fixate dinainte (cu alte cuvinte, dacă cantitatea datelor creşte este posibil să avem o funcţie mai complexă şi, deci, mai flexibilă), iar
- pe de altă parte, parametrii acestor modele nu au asociată o anumită semnificaţie, respectiv, aceea de parametri ai populaţiei; în consecinţă, metodele neparametrice de estimare *nu estimează parametrii populaţiei*.

[43] Funcţii aplicate unui eşantion, a căror interpretare nu depinde de aproximarea vreunei distribuţii de probabilitate parametrizate.

Prin estimare neparametrică înţelegem, în general, una din următoarele situaţii:

- *estimare de semnale* (vezi **Anexa: Noţiunea de semnal. Semnal staţionar**),
- *estimare de funcţii de regresie* şi, respectiv,
- *estimare de distribuţii de probabilitate*.

În mare vorbind, există două tipuri de metode de estimare neparametrică:

Tabel 5.10. Metode de estimare neparametrică

Metode de estimare neparametrică	***Comentarii:***
Locală	- În această abordare estimarea unei funcţii de valori reale[44], $f(\cdot)$, se face punctual, în fiecare punct $a = a^0$. Din această perspectivă, estimarea funcţiei se reduce la estimarea succesivă a numerelor reale $f(a)$, unde variabila a baleiază tot domeniul de valori pe care este definită funcţia (sau doar un domeniu de interes). - O altă estimare, tot punctuală, a funcţiei $f(\cdot)$, aplicabilă însă doar pentru cazul când aceasta este o funcţie continuu diferenţiabilă[45], este aceea care evaluează funcţia nu doar în punctul generic a^0 ci într-o vecinătate a acestuia, caz în care aproximarea punctuală a funcţiei se face prin valoarea $f(a^0) + f'(a^0)(a - a^0)$, ceea ce presupune, suplimentar, utilizarea observaţiilor din vecinătate. La fel ca şi mai sus, estimarea funcţiei presupune estimări punctuale succesive ce acoperă domeniul de interes. - Exemple de metode ce se bazează pe acest gen de abordare locală a estimării neparametrice sunt: histograma, estimarea cu funcţii nucleu, metoda

[44] Aici, prin funcţie înţelegem oricare dintre următoarele variante: funcţie densitate, funcţie de regresie sau un semnal oarecare.

[45] Aşa cum după cum ştim, metodele neparametrice nu impun, în general, restricţii asupra funcţiilor ce se estimează, cu excepţia anumitor metode pentru care este necesară îndeplinirea de către funcţia de estimat a condiţiei, nu foarte restrictive, de continuitate a funcţiei.

	celui mai apropiat eşantion (*NN* – Nearest Neighbour), metoda celor mai apropiate *k*-eşantioane (metoda *k*-NN) etc.
Globală	- Aceste metode introduc un sistem de coordonate (în particular, o bază) într-un spaţiu de funcţii (vezi **Anexa: Vectori bază şi funcţii bază**); în acest mod problema estimării unei funcţii se reduce la estimarea unui set de numere reale (este vorba de scalarii α_i ce reprezintă coordonatele funcţiei în raport cu baza de funcţii considerată). - Cu alte cuvinte, folosind o mulţime adecvată de funcţii liniar independente $\{\varphi_i(a)\}_{i=1}^{\infty}$ ca şi coordonate ale spaţiului de funcţii atunci, orice funcţie de valori reale şi de energie finită poate fi, în mod unic, exprimată printr-o combinaţie liniară a funcţiilor bazei, respectiv, prin setul de coeficienţi scalari α_i astfel: $$f(a)=\sum_{i=1}^{\infty}\alpha_i\varphi_i(a) \qquad (5.165)$$ - Alegerea bazei (care, aşa cum am mai spus în Anexa.aaa, nu este unică într-un spaţiu de funcţii considerat) reflectă presupunerile noastre despre caracteristicile semnalului (de exemplu, putem avea de-a face cu un semnal lent variabil, continuu, semnal oscilator ş.a.m.d.). • nişte baze bine cunoscute sunt *seriile polinomiale* şi *seriile Fourier* (aceste funcţii bază au proprietatea că sunt infinit diferenţiabile în orice punct); • alte funcţii larg folosite sunt şi funcţiile bază spline polinomiale şi funcţiile bază wavelet etc.

Pentru a înţelege următorul exemplu pe care îl vom discuta, facem aici precizarea că funcţiile nucleu folosite în estimarea neparametrică de tip local permit, prin însăşi modul cum sunt definite (vezi ***Subcapitolul 8.2.2***), o „analiză" a vecinătăţii fiecărui punct în care are loc aproximarea funcţiei dorite. Această analiză presupune de fapt asignarea de ponderi fiecărei

observaţii particulare aflate în vecinătatea punctului în care se face estimarea; în acest caz, vecinătatea este definită de un parametru al funcţiei nucleu, numit lăţime de bandă sau parametrul de netezire al funcţiei iar ponderile sunt o măsură a proximităţii acestor observaţii faţă de punctul ales, ele fiind întotdeauna zero pentru observaţiile din afara vecinătăţii.

Exemplu 5.9: Problema estimării neparametrice a unei funcţii de regresie (cazul univariat) folosind pentru aceasta funcţiile nucleu, presupune că, în formula dată de relaţia (5.215):

$$b = m(a;\theta) + e,$$

pentru $m(a;\theta)$ – reprezentând media condiţionată a v.a. y (b, variabilă generică) dată de v.a. $x = a$ –, nu se precizează *a priori* nici o formă parametrică. Cele N perechi de observaţii (a_i, b_i) ale eşantionului sunt folosite pentru a estima funcţia densitate comună pentru variabilele aleatoare x şi y astfel: densitatea în punctul (a^0, b^0) este estimată prin observarea, din cele N date empirice, a proporţiei de date care se află "aproape" de punctul (a^0, b^0). Această procedură implică utilizarea unei funcţii nucleu ce asignează ponderi corespunzătoare observaţiilor din vecinătatea punctului în care se face estimarea.

Modelele neparametrice, deşi prezintă avantajul unei foarte mari flexibilităţi, ele au şi o serie de dezavantaje:

(a) Pentru a aproxima, în mod adecvat, o funcţie, modelul folosit (vezi, de exemplu, relaţia (5.164)) are un număr practic infinit de parametri necesari pentru a putea "captura" caracteristicile complexe, presupuse, ale funcţiei de estimat. Un număr mare al parametrilor modelului (parametri ce trebuie estimaţi) duce la o rată de convergenţă a metodei foarte mică. Dependenţa ratei de convergenţă a unui estimator de numărul parametrilor modelului este adesea referită sub numele de "blestemul dimensionalităţii".

(b) Pe de altă parte, pentru a estima într-un mod corespunzător parametrii de mai sus avem nevoie de un eşantion de date suficient de mare (condiţie ce în practică arareori poate fi îndeplinită).

(c) Un alt dezavantaj constă în aceea că, în mod deosebit pe seturi mari de date, aceşti estimatori flexibili presupun un volum de calcul foarte mare.

Modele semiparametrice

Modelarea semiparametrică a apărut ca un compromis între modelarea parametrică şi cea neparametrică, compromis concretizat prin introducerea

parţială în model a unor componente parametrice. Din acest punct de vedere, modelele semiparametrice sunt mai restrictive decât modelele complet nespecificate dar mai flexibile însă decât modelele parametrice. Modelele semiparametrice sunt construite, în general, pornind de la modele neparametrice în care s-au introdus, parţial, componente parametrice (parametri ce au o anumită semnificaţie pentru populaţia sau procesul analizat).

Metodele semiparametrice prezintă, ca un avantaj major, rate de convergenţă mult mai mari decât ratele de convergenţă obţinute în abordările neparametrice ale aceloraşi probleme. Mai mult chiar, în multe cazuri rata de convergenţă obţinută este una similară cu cea obţinută folosind metode parametrice. Câteva exemple de modele semiparametrice sunt şi: regresia generalizată, modelele liniare parţial generalizate, modelele aditive şi modelele aditive generalizate etc.

Histograma

În practică, ori de câte ori dispunem de un set de date empirice pentru un vector aleator ne este foarte dificil să tragem concluzii cu privire la proprietăţile acestuia din urmă – şi aceasta folosindu-ne doar de simpla inspecţie a valorilor observate. Într-un astfel de moment ne vine în ajutor statistica descriptivă care, fie prin ***reprezentări grafice***, fie prin ***măsuri cantitative***, extrage informaţii asupra caracteristicilor datelor la nivel de eşantion, informaţii pe care le putem utiliza ulterior pentru a face inferenţe la nivel de populaţie sau pentru a lansa anumite ipoteze de lucru.

Exemplu 5.10: O astfel de ipoteză de lucru este, spre exemplu, şi cea folosită în cadrul procesului de estimare parametrică în care forma funcţională a densităţii de probabilitate a caracteristicii la nivel de populaţie, deşi necunoscută, se presupune a fi una cunoscută. În acest caz, la baza alegerii *a priori* a uneia sau alteia dintre densităţile teoretice existente stau deopotrivă: **(a)** o cunoaştere prealabilă a familiilor de distribuţii teoretice şi **(b)** informaţiile culese fie printr-o analiză ştiinţifică a fenomenului fizic studiat, fie printr-o analiză empirică a datelor observate (vorbim aici deci, de statistica descriptivă aplicată la nivel de eşantion).

Cea mai directă modalitate de investigare a caracteristicilor unui eşantion este dată de prezentarea, într-o formă grafică convenabilă, a datelor observate. O astfel de afişare grafică – larg utilizată doar pentru ***datele univariate şi bivariate*** – este şi histograma.

Histograma este, prin definiţie, o reprezentare grafică, sub formă de bare, a frecvenţelor relative sau absolute de apariţie a unei variabile aleatoare, pe intervale de valori. Mai exact, ea ilustrează proporţia realizărilor particulare ale *v.a. x* care se regăsesc în fiecare dintre cele câteva intervale disjuncte de valori, predefinite, ale lui *x*. Aceste intervale de valori sunt adiacente şi, în mod ideal, de lăţime egală.

Dacă lăţimea unui interval de valori tinde, la limită, către o singură valoare iar numărul de intervale (respectiv, puncte) pe care se face reprezentarea histogramei tinde la infinit atunci histograma devine, aşa cum vom vedea mai jos, chiar funcţia densitate de probabilitate ce caracterizează setul de date empirice.

În realitate, nu vom dispune niciodată de un număr infinit de date empirice şi, în consecinţă, histograma, aşa cum a fost ea definită mai sus doar *aproximează funcţia densitate de probabilitate* asociată, fiind, prin modul de abordare, un ***estimator neparametric*** al acesteia. Estimatul obţinut pe un eşantion poartă, în acest caz, numele de ***distribuţie empirică***.

În **Figura 5.20(a)** este reprezentată grafic o posibilă formă a unei funcţii densitate de probabilitate unidimensională $f_x(a)$, necunoscută, ce este modelată prin histograma din figura **Figura 5.20(b)**.

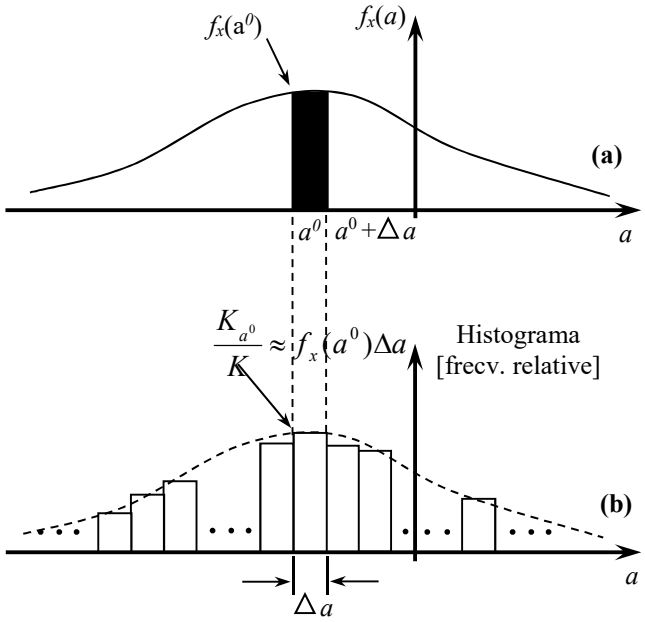


Figura 5.20. (a). Funcţia densitate de probabilitate a unei variabile aleatoare *x* versus **(b)** histograma normalizată corespunzătoare

Histograma setului de date din **Figura 5.20(b)** a fost obţinută printr-o segmentare pe lăţimi Δa a intervalului de valori ale lui x. Înălţimea fiecărui dreptunghi este una proporţională cu numărul de elemente din respectivul interval de lăţime Δa. Fie K_{a^0} numărul de valori ale v.a. x aflate în intervalul $(a^0, a^0 + \Delta a]$, respectiv $a^0 < a \le a^0 + \Delta a$. În cazul general, când numărul total de date empirice, K, din eşantion este suficient de mare iar lăţimea Δa a intervalelor este suficient de îngustă, ambele cantităţi $K_{a^0}/\boldsymbol{K}$ cât şi $f_x(a^0)\Delta a$ sunt, de fapt, aceleaşi estimări ale probabilităţii de poziţionare a unei realizări particulare a v.a. x în intervalul anterior prezentat (vezi **Anexa: Media aritmetică**(a)). Astfel:

$$f_x(a^0)\Delta a \approx \frac{K_{a^0}}{K} \tag{5.166}$$

În esenţă vorbim aici de *interpetarea frecvenţială* a probabilităţii (pentru alte interpretări a se vedea **Anexa: Interpretarea noţiunii de probabilitate**).

O discuţie ce s-ar impune a fi făcută aici ţine de alegerea optimă a numărului de intervale de valori,

$$N_{\Delta a} = \text{int}\left(\frac{\max(a) - \min(a)}{\Delta a}\right) \tag{5.167}$$

pe care se reprezintă histograma[46] (aici int(·) reprezintă partea întreagă). Lăţimi diferite ale acestor intervale pun, în general, în evidenţă caracteristici diferite ale datelor. Nu există la ora actuală metode care să ne ofere o soluţie optimă la această problemă şi, de aceea, alegerea finală se va face prin încercări succesive; respectiv, se dau pe rând valori (nu foarte mari) lui Δa şi se reţine acea valoare care pune în evidenţă anumite particularităţi ascunse ale datelor. Dintre formulele propuse pentru calculul acestui parametru important al histogramei prezentăm aici, ca informaţie orientativă în alegerea iniţială a lui Δa, formula lui Scott:

$$\Delta a = 3.5 \cdot \sigma \cdot K^{-1/3} \tag{5.168}$$

[46] O astfel de discuţie îşi are sensul dacă avem mereu reprezentat în minte faptul că histograma se realizează, în general, pe un eşantion de date ales în mod aleator din populaţia de interes. În consecinţă, frecvenţele de apariţie în cadrul eşantionului ale anumitor valori s-ar putea să difere uneori considerabil de mult faţă de frecvenţele de apariţie în populaţie ale aceloraşi valori; concret, vorbim de valori cu mult mai mici ale acestor frecvenţe. În acest context, o histogramă realizată pe intervale foarte mici s-ar putea să genereze o reprezentare foarte zgomotoasă care, practic, nu ne furnizează foarte multă informaţie privind distribuţia reală a datelor.

unde σ reprezintă deviaţia standard a v.a. *x*.

Informaţiile pe care ni le furnizează histograma sunt informaţii utile privind caracteristicile datelor, cum ar fi **tendinţa centrală**, **dispersia datelor**, precum şi **forma (alura) generală a distribuţiei**. Vom vedea în subcapitolul următor ce importanţă majoră poate juca histograma în inferenţa statistică, în special prin această ultimă informaţie furnizată.

Observaţia 5.35: Revenind în contextul ***Exemplu 5.10***, o inspecţie vizuală a histogramei eşantionului ne poate sugera, de exemplu, o anumită formă funcţională pentru densitatea de probabilitate necunoscută la nivel de populaţie, formă pe care o folosim, mai departe, ca *ipoteză de lucru* în estimarea parametrică. Ulterior fixării acestei ipoteze de lucru, în continuare – în cadrul procesului de inferenţă – se vor mai estima doar parametrii asociaţi respectivei funcţii densitate de probabilitate.

Observaţia 5.36: Aşa după cum am prezentat la început, histograma este, în principal, un instrument grafic de vizualizare pentru domeniul univariat şi bivariat de valori. Această caracteristică nu trebuie privită ca fiind una limitativă în ceea ce priveşte utilitatea metodei în cazul vectorilor aleatori multidimensionali. Şi în acest caz, o inspecţie vizuală pe fiecare variabilă aleatoare componentă, combinată eventual cu alte informaţii statistice (de exemplu, de independenţă statistică între componentele vectorului etc.), ne poate conduce la nişte concluzii la fel de utile privind forma funcţională a fdp caracteristică vectorului.

***Observaţie 5.37*:** Aşa după cum am prezentat la început, histograma este, în principal, un instrument grafic de vizualizare pentru domeniul univariat de valori însă există şi histograme multivariate.

În prezent, în neurobiologie, ca şi în multiple alte domenii, histograma este în continuare folosită cu succes ca un instrument tradiţional de estimare a funcţiei densitate de probabilitate a procesului aleator studiat, deşi alte metode neparametrice de estimare a densităţii s-au dovedit a fi mai eficiente. Această situaţie se explică în principal prin simplitatea metodei şi prin informaţia mai intuitivă, de natură vizuală, pe care ne-o furnizează.

5.5.2. Momentele unui vector aleator

Pe lângă tehnicile de vizualizare grafică a trăsăturilor unui set de date empirice, statistica descriptivă oferă şi o serie de ***măsuri cantitative*** ce

sumarizează, într-o manieră exactă, caracteristicile respectivului eşantion de date.

Astfel, informaţiile pe care ni le furnizează histograma pot fi extrase şi exprimate şi în termeni pur cantitativi, prin următoarele măsuri:

- tendinţa centrală, prin *medie* şi *mediană*, respectiv,
- dispersia datelor, prin *rang* şi *deviaţia standard*.

Parametrii[47] tendinţei centrale reprezintă cele mai importante măsuri folosite în caracterizarea unei distribuţii empirice. Valorile furnizate de acestea ne ajută să localizăm datele pe o scală liniară, oferindu-ne totodată o imagine asupra celei mai bune valori ce descrie datele. Dintre aceşti parametri cei mai uzuali sunt *media aritmetică* a eşantionului (vezi relaţia **Error! Not a valid link.**) şi, respectiv, *mediana* – aceasta reprezentând acea valoare a vectorului aleator x pentru care 50% din datele eşantionului au valori sub această valoare iar 50% au valori peste această valoare. Mediana este folosită, de obicei, ca alternativă a mediei aritmetice care este un parametru sensibil la valorile eronate ale eşantionului (valori extreme care diferă foarte mult de majoritatea datelor). Acest ultim neajuns nu constituie însă o problemă în sine, întrucât, în general, respectivele valori eronate sunt eliminate într-o aşa-zisă fază de preprocesare a datelor.

Dintre parametrii ce măsoară gradul de împrăştiere al datelor am amintit mai sus *rangul*, calculat ca diferenţă între valoarea maximă şi valoarea minimă din setul de date (parametru extrem de sensibil în cazul unor date extreme, eronate), respectiv *deviaţia standard* calculată ca media abaterii fiecărui punct de la media calculată a eşantionului. O altă măsură a dispersiei datelor este şi *varianţa*, care reprezintă pătratul deviaţiei standard.

Aşa după cum am văzut în ***Subcapitolul 5.5.1***, subpunctul 1 „***Estimarea parametrică punctuală clasică***”, atât media cât şi varianţa distribuţiei empirice sunt parametri de selecţie şi ei sunt adesea utilizaţi ca estimaţii ale parametrilor statistici corespunzători, din populaţie.

După cum vom vedea în cele ce urmează, atât media cât şi varianţa reprezintă, în fapt, cele mai utilizate două măsuri cantitative asociate unei v.a. dintr-o, practic, infinitate de măsuri definite teoretic şi care sunt cunoscute generic sub numele de **momente**.

[47] Se cunosc mai multe măsuri cantitative ce descriu tendinţa centrală. Dintre acestea amintim aici şi *modul*, *media geometrică*, *media armonică* etc.

1. Momentele în analiza statistică. Funcţii caracteristice

Momentele unei distribuţii reprezintă cantităţi de interes în studiul variabilelor şi vectorilor aleatori. Mai jos avem definite **momentele pentru o v.a**. x:

Momentele: $m_n = E\{x^n\} = \int_{-\infty}^{+\infty} a^n f_x(a)da$

Momentele centrate: $\mu_n = E\{(x-m_x)^n\} = \int_{-\infty}^{+\infty} (a-m_x)^n f_x(a)da$

Momentele absolute: $E\{|x|^n\}$, $E\{|x-m_x|^n\}$

Momentele generalizate: $E\{(x-\alpha)^n\}$, $E\{|x-\alpha|^n\}$

În practică, cele mai utilizate sunt momentele şi momentele centrate de ordin 1 şi, respectiv, 2: $m_1 = \boldsymbol{m_x}$, $\mu_2 = \boldsymbol{\sigma_x^2}$.

Trebuie să reţinem faptul că momentele unei v.a. nu sunt numere arbitrare ci ele satisfac numeroase inegalităţi cum ar fi, de exemplu: $\sigma^2 = m_2 - m_1^2 \geq 0$ (vezi ***Problema 5.26*** rezolvată în anexă).

Pe lângă momentele prezentate mai sus există definite şi momentele centrate normalizate (standardizate); acestea sunt date de momentul centrat de ordin n corespunzător, divizat prin σ^n, respectiv,

$$\textit{moment centrat de ordin n normalizat} = \mu_n/\sigma^n \qquad (5.169)$$

Aceste momente normalizate sunt cantităţi fără unitate de măsură ce reprezintă distribuţia într-un mod independent de orice modificare liniară a scalei.

În mod corespunzător, **momentele comune pentru două variabile aleatoare**, x şi y, se definesc şi ele astfel:

Momentele: $m_{kl} = E\{x^k y^l\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} a^k b^l f_{xy}(a,b)dadb$

Media produsului $x^k y^l$ reprezintă momentul comun de ordin $n = k + l$ al variabilelor aleatoare x şi y.

Momentele centrate: $\mu_{kl} = E\{(x-m_x)^k (y-m_y)^l\} =$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (a-m_x)^k (b-m_y)^l f_{xy}(a,b)dadb$$

Momentele absolute: $E\{|x|^k \cdot |y|^l\}$, $E\{|x-m_x|^k \cdot |y-m_y|^l\}$

Momentele generalizate: $E\{(x-\alpha)^k (y-\beta)^l\}$, $E\{|x-\alpha|^k \cdot |y-\beta|^l\}$

Din definiţiile de mai sus rezultă că:

- *momentele comune de ordin unu* sunt: $m_{10} = m_x$ şi $m_{01} = m_y$;
- *momentele comune de ordin doi* sunt: $m_{20} = E\{x^2\}$, $m_{11} = E\{xy\}$ şi $m_{02} = E\{y^2\}$;
- momentele *centrate comune de ordin unu* sunt: $\mu_{10} = \mu_{01} = 0$;
- momentele *centrate comune de ordin doi* sunt: $\mu_{21} = \sigma_x^2$, $\mu_{11} = Cov(x,y)$[48] şi $\mu_{02} = \sigma_y^2$.

Observaţia 5.38: Definiţiile momentelor (respectiv, a momentelor comune) pentru vectori aleatori sunt similare celor prezentate, mai sus, pentru variabile aleatoare, cu particularitatea, bineînţeles, a folosirii operatorului de transpunere în cazul produsului scalar al vectorilor.

Utilitatea momentelor este una multiplă:

- astfel, am văzut până acum rolul jucat de acestea în *estimarea parametrică* (statistica inferenţială) – acolo unde anumite momente coincid chiar cu parametrii distribuţiei teoretice ce se presupune a modela funcţia de densitate necunoscută a unei populaţii; în acest caz, doar calcularea respectivelor momente la nivel de eşantion conduc direct la obţinerea unui estimat al densităţii de probabilitate dorite;
- o altă utilitate a momentelor definite mai sus, întâlnită de această dată în *statistica descriptivă*, rezidă în capacitatea lor de a furniza o serie de informaţii descriptive privind datele analizate; în cele ce urmează vom discuta această ultimă direcţie de utilizare a momentelor.

Un rezultat important legat de momentele unei v.a. (respectiv, vector aleator) x este cel al unei teoreme, numită ***teorema momentului***, care afirmă că: *dacă m_n este cunoscut pentru fiecare n, atunci, în anumite condiţii* (vezi mai jos), *funcţia densitate de probabilitate a lui x este* ***determinată în mod unic***. O prezentare pe scurt a acestei teoreme, precum şi a noţiunilor noi pe care le utilizează aceasta, o redăm în tabelul de mai jos:

Funcţii ale unei v.a. x	**Definiţii/Enunţ**
Funcţia caracteristică	Notată $\boldsymbol{\Phi_x(\omega)}$, este prin definiţie dată de integrala, $\Phi_x(\omega) = \int_{-\infty}^{+\infty} f_x(a) e^{j\omega a} da = E\{e^{j\omega x}\}$ (5.170)

[48] Coeficientul de covarianţă al variabilelor aleatoare x şi y.

Funcţia (generatoare) de momente	Notată, $\boldsymbol{M_x(s)}$, se obţine din funcţia caracteristică astfel: pentru $M_x(j\omega) = \Phi_x(\omega)$ şi $s = j\omega$, $$M_x(s) = \int_{-\infty}^{+\infty} f_x(a) e^{sa}\, da = E\{e^{sx}\} \quad (5.171)$$
Teorema momentului	Derivând (5.235) de n ori obţinem, $$M^{(n)}(s) = E\{x^n e^{sx}\} \quad (5.172)$$ şi, corespunzător, $$M^{(n)}(0) = E\{x^n\} = m_n \quad (5.173)$$ Dacă toate momentele m_n sunt finite şi dacă, în plus, dezvoltarea în serie a lui $M_n(s)$ în apropierea originii, $$M_x(s) = \sum_{n=0}^{\infty} \frac{m_n}{n!} s^n \quad (5.174)$$ este o serie convergentă, atunci *densitatea* $f_x(a)$ – ce poate fi scrisă în termenii lui $M_n(s)$ (vezi şi relaţia (5.239)) – este, atunci când se cunosc toţi m_n, *unic determinată.*

Notă:

(a) Întrucât funcţia caracteristică, $\boldsymbol{\Phi_x(\omega)}$, este strâns legată de transformata Fourier – mai exact, $\Phi_x(\omega)$ a lui x este complex-conjugatul transformatei Fourier continue a funcţiei de densitate a lui x (vezi **Anexa: Transformata Fourier**), atunci proprietăţile lui $\boldsymbol{\Phi_x(\omega)}$ sunt, în esenţă, aceleaşi cu proprietăţile transformatelor Fourier. În acest caz, $f_x(a)$ poate fi exprimat în termenii lui $\Phi_x(\omega)$, astfel:

$$f_x(a) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi(\omega) e^{-j\omega a}\, d\omega \quad (5.175)$$

(b) Dacă $x = [x_1, ..., x_d]^T$ este un vector aleator d-dimensional atunci argumentul scalar ω devine vectorul linie, d-dimensional, $\Omega = [\omega_1, ..., \omega_d]$ iar produsul ωx devine produsul scalar Ωx:

$$\Phi_x(\Omega) = E\{e^{j\Omega x}\} = E\{e^{j(\omega_1 x_1 + \cdots + \omega_d x_d)}\} = M_x(j\Omega) \quad (5.176)$$

(c) Funcţia caracteristică a unei v.a. există întotdeauna ceea ce nu este şi cazul funcţiei generatoare de moment. În cazul în care aceasta din urmă nu există funcţia caracteristică este cea utilizată pentru a determina momentele.

În concluzie, *orice distribuţie de probabilitate* (pe R sau pe R^n) *are o funcţie caracteristică* şi, reciproc, *pentru orice funcţie caracteristică există exact o singură distribuţie de probabilitate*. Altfel spus, funcţia caracteristică a unei v.a. *defineşte complet şi în mod unic distribuţia de probabilitate* a respectivei variabile.

O concluzie imediată ce se poate trage de aici este şi faptul că, momentele de ordin unu si doi – de departe cele mai utilizate în aplicaţiile practice – nu dau decât o caracterizare limitată a funcţiei de densitate necunoscută a populaţiei. În acest context, cunoaşterea şi a altor momente poate fi utilă mai ales în faza de alegere între două densităţi ce prezintă aceleaşi momente de ordin unu şi doi şi care sunt candidate la calitatea de estimat al unei funcţii densitate de probabilitate necunoscute; în consecinţă, momentele ne pot asista în mod efectiv în procesul de selecţie a acelei distribuţii teoretice care se apropie cel mai mult de alura (dezvăluită prin indicatori sintetici adecvaţi) a funcţie densitate reale.

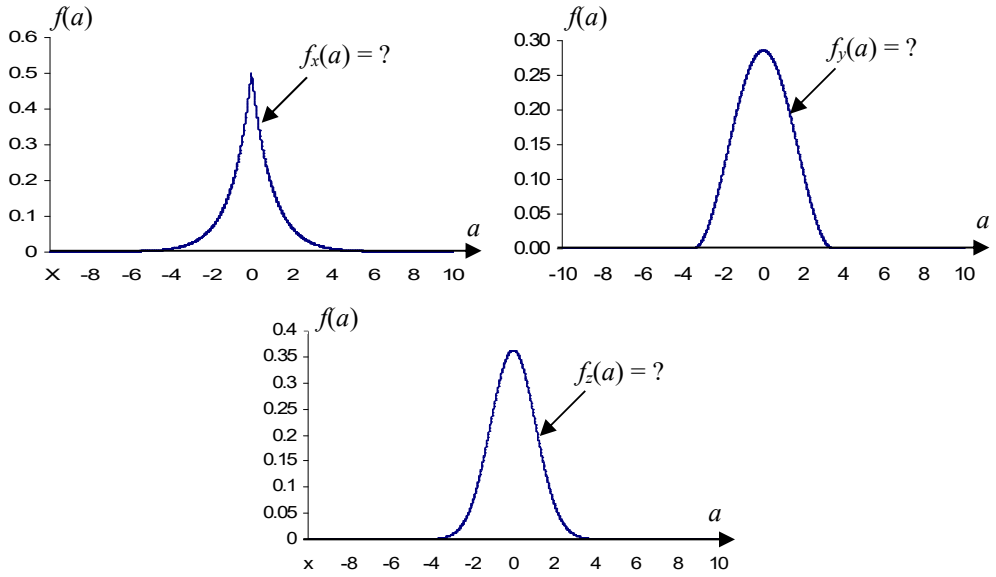


Figura 5.21. Funcţii densitate de probabilitate de aceeaşi medie şi varianţă

Exemplu 5.11: În **Figura 5.21** sunt reprezentate trei densităţi empirice unidimensionale (obţinute cu ajutorul histogramei) – densităţi care, prezintă aceeaşi medie, m, şi aceeaşi varianţă, σ^2. Sunt aceste densităţi din familia de distribuţii gaussiene?

În urma unei prime inspecţii vizuale am fi tentaţi să spunem că din cele trei densităţi numai ultimele două ar corespunde descrierii aşa numitului „clopot al lui Gauss” (vezi ***Subcapitolul 5.6.2***).

O suprapunere a celor trei grafice (vezi **Figura 5.22**) ne dezvăluie însă o situaţie destul de contradictorie şi anume, suntem în situaţia în care pentru aceiaşi doi parametri, *medie* şi, respectiv, *varianţă* –

parametri ce ştim că definesc complet o distribuţie gaussiană –, obţinem densităţi *Gauss*-iene distincte.

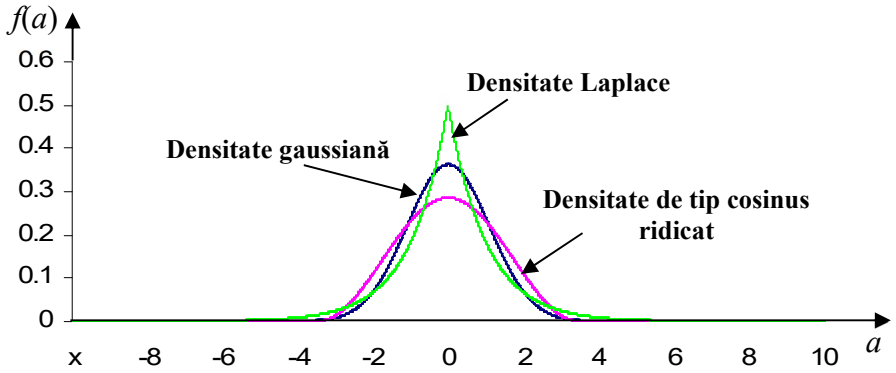


Figura 5.22. Forme funcţionale diferite pentru trei densităţi cu aceeaşi medie şi aceeaşi varianţă, însă cu coeficient de boltire diferit.

Răspunsul la această dilemă îl vom afla, însă, abia în momentul în care vom calcula valoarea unor **indicatori sintetici** suplimentari (momente centrate normalizate, vezi (5.169)), indicatori care ne oferă informaţii privitoare de această dată la **alura funcţiei** de densitate empirică.

În principal, vorbim de două categorii de astfel de indicatori sintetici, respectiv:

(a) indicatori de asimetrie, respectiv,
(b) indicatori de boltire (sau ai excesului).

Indicatorii de asimetrie dau informaţii asupra modului de repartizare a frecvenţelor de o parte sau alta a valorii centrale a unei serii aleatoare. În literatura de specialitate s-au propus mai mulţi indicatori pentru măsurarea gradului de asimetrie al distribuţiilor. Cel folosit de noi în ***Exemplul 5.11*** a fost calculat cu formula:

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \tag{5.177}$$

Valori ale lui γ_1:

- egale cu zero → indică o distribuţie simetrică;
- strict negative → indică o distribuţie mai împrăştiată spre stânga, faţă de valoarea medie, cu o coadă a distribuţiei mai mare spre stânga (*asimetrie de stânga*);

- strict pozitive → indică o distribuţie mai împrăştiată spre dreapta, faţă de valoarea medie, cu o coadă a distribuţiei mai mare spre dreapta (*asimetrie de dreapta*).

Indicatorii de boltire dau informaţii asupra gradului cât sunt de ascuţite, respectiv, de plate, densităţile empirice faţă de o densitate de probabilitate normală. Dintre indicatorii propuşi, prezentăm doar *coeficientul de boltire al lui Fisher*, respectiv:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} \tag{5.178}$$

Valori ale lui γ_2:

- egale cu 3 → indică o distribuţie normală (*distribuţie mezokurtică*);
- mai mari decât 3 → indică o distribuţie cu un vârf aproape de medie mai ascuţit decât cel al unei distribuţii normale (*distribuţie leptokurtică*);
- mai mici decât 3 → indică o distribuţie cu un vârf aproape de medie mai aplatizat faţă de cel al unei distribuţii normale (*distribuţie platikurtică*).

Observaţia 5.39: Pentru o corectă interpretare a valorilor calculate, în momentul implementării sau a folosirii pentru indicatorii de mai sus a unor funcţii gata predefinite în anumite medii de lucru (de exemplu, Matlab, Microsoft Office Excel etc.) trebuie avut grijă la următoarele aspecte: (a) dacă în relaţia (5.178) este sau nu substrasă valoarea 3 (în prima situaţie coeficientul de boltire calculat pentru o distribuţie normală va avea valoarea zero în loc de trei) şi (b) întrucât aceşti indicatori sunt calculaţi pe un eşantion iar nu pe întreaga populaţie, valorile obţinute vor fi deplasate (vor tinde să difere de valorile lor reale din populaţie cu o cantitate sistematică ce depinde de mărimea eşantionului); în acest caz este indicată realizarea corespunzătoare a unei corecţii a valorilor obţinute.

Revenind la ***Exemplul 5.11***, calculul coeficientului de asimetrie, respectiv, al coeficientului de boltire pentru cele trei densităţi prezentate ne conduce la următoarele rezultate – rezultate ce concordă cu caracteristicile (în particular, alura) funcţiilor teoretice folosite la generarea densităţilor empirice respective:

Densitate empirică	**Momente normalizate**	**Distribuţii teoretice folosite**
$f_x(a)$	$\gamma_1 = 0$, $\gamma_2 = 6$	Distribuţia Laplace: $$f_x(a) = \frac{1}{2\beta}\exp\left(-\frac{\mid a - m_x \mid}{\beta}\right) \quad (5.179)$$ Calculată pentru β (factorul de scalare) = 1.
$f_y(a)$	$\gamma_1 = 0$, $\gamma_2 = 2.7$	Distribuţia normală, $N(0, 1.1)$.
$f_z(a)$	$\gamma_1 = 0$, $\gamma_2 = 1$	Distribuţia cosinus ridicat: $$f_x(a) = \begin{cases} \frac{1}{2s}\left[1 + \cos\left(\frac{a - m_x}{s}\pi\right)\right] & pentru \mid a \mid \leq s \\ 0 & în\ rest. \end{cases} \quad (5.180)$$ calculată pentru s (ce defineşte limitele intervalului pe care funcţia ia valori diferite de zero) =3.5.

Notă: Implementarea coeficienţilor s-a făcut în Microsoft Excel™ iar dimensiunea eşantionului de date a fost $n = 2000$.

În concluzie, complexitatea realităţii impune folosirea unui sistem de indicatori. Altfel spus, atunci când alegem o distribuţie teoretică sau alta pentru a modela densitatea necunoscută din populaţie, pe lângă informaţiile furnizate de momentele de ordin unu şi doi, de mare utilitate ne sunt şi informaţiile furnizate de momentele de ordin superior, ce ne dau o imagine mai corectă privind forma densităţii empirice. Din acest punct de vedere şi histograma este o tehnică grafică eficientă ea punând în evidenţă această caracteristică a datelor.

O altă utilitate a momentelor, ce se încadrează tot în statistica descriptivă, este dată de informaţiile descriptive pe care ni le furnizează, în principal, momentele comune de ordin unu şi doi a două variabile aleatoare, respectiv, a doi vectori aleatori. Aceste informaţii utile (ca de exemplu, independenţa sau gradul de corelare a două au mai multe variabile aleatoare) rezidă în însăşi proprietăţile acestor momente care se definesc aşa cum s-a prezentat mai sus.

În continuarea acestui subcapitol vom defini momentele de ordin I şi II pentru un vector aleator, vom prezenta proprietăţile acestora precum şi câteva proceduri simple de estimare a acestor momente din setul de date. Reamintim că prin estimare parametrică se înţelege o *evaluare aproximativă* a parametrilor de interes.

2. Momentul de ordin întâi

Fie x un vector aleator, discret sau continuu. Atunci, *valoarea sperată* sau ***media*** (momentul statistic de ordinul I, notat cu m_x) al acestui vector aleator este, prin ***definiţie,*** dată de relaţia:

$$E\{x\} \overset{def}{=} \int_{-\infty}^{+\infty} a\, f_x(a)\, da \tag{5.181}$$

$$E\{x\} \overset{def}{=} \sum_{-\infty}^{+\infty} a\ p_x(a) \tag{5.182}$$

Relaţia (5.181) este pentru un vector aleator caracterizat de variabile aleatoare continue, în timp ce relaţia (5.182) este pentru un vector aleator caracterizat de variabile aleatoare discrete.

Dacă integrala sau suma date de relaţiile anterioare, pentru cel puţin una dintre componentele vectorului aleator x, nu converge către un număr finit vom spune că estimatorul vectorului aleator x nu există.

Media unui vector aleator este un ***operator liniar***, el fiind caracterizat de următoarele proprietăţi.

Proprietăţi:

1. Pentru orice vectori aleatori x şi y avem:

$$E\{x+y\} = E\{x\} + E\{y\} \tag{5.183}$$

2. Pentru orice vector aleator x şi pentru orice constantă scalară α şi vector constant a^0 avem:

$$E\{\alpha x + a^0\} = \alpha E\{x\} + a^0 \tag{5.184}$$

3. Fie x un vector aleator şi fie $g(x)$ o cantitate derivată din el. De exemplu, această cantitate, $g(x)$, poate fi scalară, vectorială sau matricială, precum: $g(x) = x\, x^{*T}$ sau $g(x) = (x - m_x)(x - m_x)^{*T}$. Media (momentul statistic de ordinul I) a lui $g(\cdot)$ este notată $E\{g(x)\}$ şi ea este definită de relaţiile matematice:

$$E\{g(x)\} = \int_{-\infty}^{+\infty} g(a) f_x(a)\, da \tag{5.185}$$

$$E\{g(x)\} = \sum_{-\infty}^{+\infty} g(a) p_x(a) \tag{5.186}$$

Relaţia (5.185) este pentru un vector aleator caracterizat de variabile aleatoare continue în timp ce relaţia (5.186) este pentru un vector aleator ce conţine variabile aleatoare discrete.

Dacă $g(a)$ este un vector sau o matrice, în relaţia (5.185), se aplică estimarea statistică pentru fiecare element în parte iar rezultatul estimării statistice va fi un alt vector sau o altă matrice cu aceleaşi dimensiuni.

Exemplu 5.12: Pentru o posibilă aplicaţie a semnalelor aleatoare în analiza convertoarelor AD (analog – digitale) consultaţi **Anexa: Analiza raportul semnal/zgomot pentru un convertor ideal**.

Media condiţionată a unui vector aleator x, presupunând evenimentul B, este dată de integrala (5.181) în care fdp $f_x(a)$ este înlocuită de fdp condiţionată $f_x(a|B)$:

$$E\{x \mid B\} = \int_{-\infty}^{+\infty} a f_x(a \mid B)\, da \tag{5.187}$$

Cele mai comune momente sunt cele de ordin unu şi doi.
Momentul de ordin întâi sau media vectorului aleator este definit de relaţia:

$$m_x = E\{x\} \overset{def.}{=} \int_{-\infty}^{+\infty} a\, f_x(a) da \tag{5.188}$$

Deoarece x este un vector d dimensional, m_x este şi el, la rândul lui, un vector d – dimensional de forma:

$$m_x = \begin{bmatrix} m_1 \\ m_2 \\ \dots \\ m_d \end{bmatrix} \tag{5.189}$$

Componenta k a vectorului m_x este dată de:

$$m_k = E\{x_k\} = \int_{-\infty}^{+\infty} a_k f_x(a) da = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} a_k f_x(a)\, da_N \dots da_1 \tag{5.190}$$

După integrarea relaţiei (5.190) pe toate componentele vectorului se obţine:

$$m_k = \int_{-\infty}^{+\infty} a_k f_{xk}(a_k) da_k \tag{5.191}$$

Dacă integrala (5.191) sau suma (în cazul exprimării relaţiei (5.191) pentru o variabilă aleatoare discretă – relaţie similară (5.182)) nu converge către un număr real spunem că estimatorul statistic de ordinul întâi nu există.

3. Momente de ordin doi

În acest subcapitol am reunit generic, sub numele de momente de ordin doi (a se vedea şi S**ubcapitolul 5.5.2.1.**):

- momentele de ordin doi ale unui vector aleator x,
  - → *matricea de corelaţie*, R_x,
  - → *matricea de covarianţă*, C_x;
- momentul centrat comun de ordin doi a două variabile aleatoare,
  - → *covarianţa între două variabile aleatoare*;
- momentele comune de ordin doi pentru doi vectori aleatori, x şi y:
  - → *matricea de cros-corelaţie*, R_{xy},
  - → *matricea de cros-covarianţă*, C_{xy}.

Matricea de corelaţie şi de covarianţă

Matricea de corelaţie reprezintă mulţimea completă de momente de ordin doi ale unui vector aleator, x, şi ea este definită în modul următor:

$$R_x \overset{def}{=} E\{x\, x^{*T}\} \tag{5.192}$$

Pentru un vector aleator real obţinem:

$$R_x = \begin{bmatrix} E\{(x_1)^2\} & E\{x_1 x_2\} & \cdots & E\{x_1 x_d\} \\ E\{x_2 x_1\} & E\{(x_2)^2\} & \cdots & E\{x_2 x_d\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{x_d x_1\} & E\{x_d x_2\} & \cdots & E\{(x_d)^2\} \end{bmatrix} \tag{5.193}$$

Pentru un vector aleator complex matricea are forma:

$$R_x = \begin{bmatrix} E\{|\, x_1\,|^2\} & E\{x_1 x_2^*\} & \cdots & E\{x_1 x_d^*\} \\ E\{x_2 x_1^*\} & E\{|\, x_2\,|^2\} & \cdots & E\{x_2 x_d^*\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{x_d x_1^*\} & E\{x_d x_2^*\} & \cdots & E\{|\, x_d\,|^2\} \end{bmatrix} \tag{5.194}$$

Atenţie! În cazul determinării matricei de corelaţie (prin intermediul relaţiilor (5.192) sau (5.194)) sau/şi a matricei de covarianţă (în modul în care va fi prezentată ulterior în cadrul acestui capitol) pentru cazul vectorilor aleatori complecşi sau variabile aleatoare complexe este **obligatorie**, pe lângă utilizarea transpunerii, şi utilizarea complex conjugatului în definirea tuturor acestor relaţii. În cazul neutilizării transpunerii complex conjugate, ci doar a

transpunerii simple în definirea matricei de corelaţie (relaţiile (5.192) şi (5.194)) vom obţinem pentru orice vector aleator x:

$$E\{x\, x^{\mathrm{T}}\} = [0] \tag{5.195}$$

Evident că rezultatul dat de relaţia (5.195) ($R_x = [0]$), relaţie valabilă pentru orice vector aleator x complex, este unul care nu ne furnizează nici o informaţie relevantă despre vectorul x, această relaţie fiind astfel total inutilă în orice gen de analiză.

Problema 5.25: Demonstraţi validitatea relaţiei (5.195)[49].

Matricea de covarianţă este, la rândul ei, definită drept o mulţime de momente de ordinul doi centrate în jurul valorii medii. Această matrice este dată de relaţia:

$$C_x \overset{def}{=} E\{ (x - m_x) (x - m_x)^{*\mathrm{T}} \} \tag{5.196}$$

Forma matricială pentru C_x este similară cu cea prezentată în relaţiile (5.193) şi (5.194), cu menţiunea că, de această dată, elementele matricei sunt:

$$c_{kl} = \mathrm{Cov}\,(x_k, x_l) = E\{ (x_k - m_k) (x_l - m_l)^* \} \tag{5.197}$$

cu

$$c_{kk} = E\{ |x_k - m_k|^2 \} = \sigma_{x_k}^2 \tag{5.198}$$

Dacă matricea de covarianţă se calculează pentru un vector de trăsături, atunci:

1. elementul c_{kl} **este coeficientul de covarianţă** între trăsătura k şi trăsătura l – notat în literatura de specialitate cu Cov (x_k, x_l). Un coeficient de covarianţă pozitiv ne arată că deviaţiile de la valorile medii ale trăsăturilor x_k şi x_l au, în medie, acelaşi semn. Dacă, creşterea (scăderea),

---

[49] Analiza vectorilor aleatori complecşi, $x = x_r + j \cdot x_i$ cu x_r şi x_i vectori aleatori reali, (**numai** pentru cazul prelucrării digitale a semnalelor şi pentru problemele de clasificare) poate fi redusă numai la acei vectori pentru care există următoarele relaţii de simetrie [Therrien, 1992]:

$$\begin{aligned} &E\{x_r x_r^T\} = E\{x_i x_i^T\} \quad (a) \\ &E\{x_i x_r^T\} = -E\{x_r x_i^T\} \quad (b) \end{aligned} \tag{a}$$

Aceste două relaţii ne furnizează informaţia că părţile reale şi imaginare ale vectorului satisfac condiţiile:

$$E\{x_{rk}\, x_{rl}\} = E\{x_{ik}\, x_{il}\} \tag{b}$$

şi

$$E\{x_{ik}\, x_{rl}\} = -E\{x_{rk}\, x_{il}\} \tag{c}$$

în medie, a valorii unei trăsături faţă de media ei va determina, în medie, şi creşterea (scăderea) valorii celei de a doua trăsături faţă de media proprie, atunci spunem că trăsăturile sunt ***corelate pozitiv***. O valoare negativă a acestui coeficient ne indică că, în medie, deviaţiile acestor două trăsături au sensuri opuse faţă de mediile lor şi, în această situaţie trăsăturile spunem că sunt ***corelate negativ***. Dacă acest coeficient este zero spunem că variabilele aleatoare x_k şi x_l sunt ***necorelate***. În situaţia în care x_k şi x_l sunt ***statistic independente*** atunci $\boldsymbol{c_{kl} = 0}$.

2. ***c_{kk}* este varianţa trăsăturii *k*** – acest parametru furnizează o informaţie despre „întinderea", dispersia spaţială, a valorilor trăsăturii *k*.

În continuare exemplificăm forma matricii de covarianță pentru cazurile unor distribuții unidimensionale (1D), bidimensionale (2D) și tridimensionale (3D). Pentru astfel de vectori:

$$x_{(1D)} = [x_1] = x_1 \qquad x_{(2D)} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad x_{(3D)} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{5.199}$$

matricile de covarianță (pentru cele trei situații prezentate anterior) iau următoarele forme:

$$\begin{aligned} C_{x\,(1D)} &= \left[E\left\{\left(x_1 - m_{x_1}\right)\left(x_1 - m_{x_1}\right)^*\right\}\right] = \\ &= \left[E\left\{\left|x_1 - m_{x_1}\right|^2\right\}\right] = \left[\sigma_{x_1}^2\right] = \sigma_{x_1}^2 \end{aligned} \tag{5.200}$$

$$\begin{aligned} C_{x\,(2D)} &= \begin{bmatrix} E\{(x_1 - m_{x_1})(x_1 - m_{x_1})^*\} & E\{(x_1 - m_{x_1})(x_2 - m_{x_2})^*\} \\ E\{(x_2 - m_{x_2})(x_1 - m_{x_1})^*\} & E\{(x_2 - m_{x_2})(x_2 - m_{x_2})^*\} \end{bmatrix} = \\ &= \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} \sigma_{x_1}^2 & Cov(x_1, x_2) \\ Cov(x_2, x_1) & \sigma_{x_2}^2 \end{bmatrix} \end{aligned} \tag{5.201}$$

$$\begin{aligned} C_{x\,(2D)} &= \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \\ &= \begin{bmatrix} \sigma_{x_1}^2 & Cov(x_1, x_2) & Cov(x_1, x_3) \\ Cov(x_2, x_1) & \sigma_{x_2}^2 & Cov(x_2, x_3) \\ Cov(x_3, x_1) & Cov(x_3, x_2) & \sigma_{x_3}^2 \end{bmatrix} \end{aligned} \tag{5.202}$$

Generalizând relațiile (5.200), (5.201) și (5.202) constatăm că pe diagonala principală vom avea întotdeauna varianța trăsăurilor în timp ce restul termenilor sunt dați de coeficienții de covarianță – având semnificația

varianței relative, a varianței unei variabile (de exemplu, x sau x_1) relativă la varianța unei alte variabile aleatoare (de exemplu, y sau x_3). Din acest motiv numele corect al acestei matrici este: **matricea de varianță-covarianță** (a unui anumit vector aleator). Cu toate acestea numele care s-a impus este acela de **matrice de covarianță** (potențial datorită numărului superior de termeni de covarianță, față de cei de varianță, începând de la matrici cu dimensionalitate minimă de 3 x 3).

Problema 5.26: Știind că media unei variabile aleatoare reale este $m_x = E\{x\}$ în timp ce varianţa aceleiaşi variabile este dată de $\sigma_x^2 = E\{(x - m_x)^2\}$ determinaţi $E\{x^2\}$ în funcţie numai de m_x şi σ_x^2.

Din relaţiile (5.192) şi (5.123) se poate observa că ambele matrici, atât cea de corelaţie cât şi cea de covarianţă sunt **matrici Hermetiene simetrice**, adică au proprietatea:

$$R_x = R_x^{*T} \tag{5.203}$$

și

$$C_x = C_x^{*T} \tag{5.204}$$

Problema 5.27: Demonstraţi validitatea relaţiilor (5.203) şi (5.204).

Mai mult decât atât, ambele matrici sunt şi **pozitiv semidefinite**, deci satisfac relaţiile:

$$a^{*T} R_x a \geq 0 \tag{5.205}$$

și

$$a^{*T} R_x a \geq 0 \tag{5.206}$$

pentru orice vector complex a. Relaţiile (5.205) şi (5.206) se pot demonstra foarte uşor, ţinând cont că[50]:

$$a^{*T} R_x a = a^{*T} E\{x x^{*T}\} a = E\{|x^{*T} a|^2\} \tag{5.207}$$

este întotdeauna mai mare sau egală cu zero.

Matricea de covarianţă şi cea de corelaţie sunt legate între ele printr-o relaţie matematică. Observând că E este un operator liniar iar m_x este un vector având elemente constante putem scrie:

$$\begin{aligned} E\{(x - m_x)(x - m_x)^{*T}\} &= E\{x x^{*T} - x m_x^{*T} - m_x x^{*T} + m_x m_x^{*T}\} = \\ &= E\{x x^{*T}\} - E\{x\} m_x^{*T} - m_x E\{x^{*T}\} + m_x m_x^{*T} \\ &= E\{x x^{*T}\} - m_x m_x^{*T} \end{aligned} \tag{5.208}$$

Rezultând **o relaţie foarte importantă**:

[50] Pentru demonstrarea relaţiilor (5.205) şi (5.206) ţineţi cont şi de relaţia: $(A B^T)^T = B A^T$

$$R_x = C_x + m_x m_x^{*\mathrm{T}} \tag{5.209}$$

Covarianţa între două variabile aleatoare

Aşa după cum am văzut anterior, un element al matricei de covarianţă, care nu este poziţionat pe diagonala principală, este caracterizat de relaţia:

$$c_{kl} = E\{ (x_k - m_k)(x_l - m_l)^* \} \tag{5.210}$$

Această valoare ***măsoară tendinţa trăsăturilor*** x_k şi x_l ***de a varia împreună***.

După cum se va prezenta ulterior, dacă a^1, a^2, …, a^K sunt K realizări particulare ale vectorului aleator de trăsături x, atunci relaţia de estimare a covarianţei între trăsăturile k şi l ale vectorului x (caracterizate de variabilele aleatoare x_k şi x_l) se calculează cu relaţia (vezi ***Observaţia 5.27***):

$$\hat{c}_{kl} = \frac{1}{K}\left[(a_k^1 - \hat{m}_k)(a_l^1 - \hat{m}_l)^* + (a_k^2 - \hat{m}_k)(a_l^2 - \hat{m}_l)^* + \cdots + (a_k^K - \hat{m}_k)(a_l^K - \hat{m}_l)^*\right] \tag{5.211}$$

În relaţia anterioară $\hat{m}_z$, cu z egal cu k sau l, este estimatorul mediei trăsăturii k sau l. Dacă realizările particulare ale vectorului aleator x sunt vectori reali, complex conjugatul nu-şi mai are rolul în relaţia anterioară.

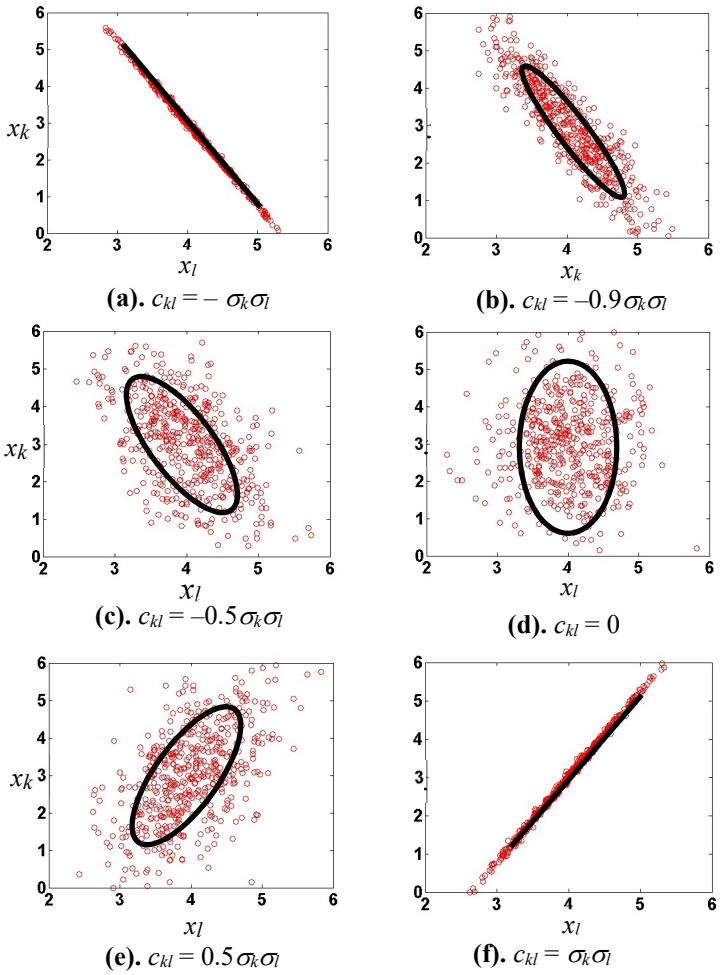
Observaţia 5.40: Dacă, în contextul de mai sus, dorim să determinăm un coeficient de covarianţă oarecare, c_{kl}, în alt mod decât cel prezentat mai sus, atunci, putem estima mai întâi matricea de covarianţă a setului de date după care, ulterior preluăm din aceasta spre analiză numai valoarea elementului care ne interesează.

Coeficientul de covarianţă între două variabile aleatoare oarecare – fie acestea componentele de ordin k şi l ale unui vector aleator de trăsături x – are, aşa cum am mai amintit, următoarele ***proprietăţi fundamentale***:

- dacă trăsătura k şi trăsătura l au tendinţa, în medie, de a varia în acelaşi sens, atunci $c_{kl} > 0$;
- dacă, în medie, valoarea trăsăturii k creşte iar cea a trăsăturii l scade (sau invers) faţă de mediile lor corespunzătoare, atunci $c_{kl} < 0$;
- în cazul în care trăsăturile sunt independente avem $c_{kl} = 0$;
- $|c_{kl}| \le \hat{\sigma}_k \cdot \hat{\sigma}_l$, unde $\hat{\sigma}_k$ şi $\hat{\sigma}_l$ sunt deviaţiile standard estimate pentru cele două trăsături, k şi, respectiv, l;
- $c_{kk} = \hat{\sigma}_k^2$ este varianţa estimată a trăsăturii k.

Proprietăţile enumerate anterior pentru coeficientul de covarianţă sunt adevărate numai dacă anumite condiţii sunt satisfăcute. Astfel, în cazul

trăsăturilor necorelate, c_{kl} este egal cu zero numai dacă există un număr infinit de realizări particulare ale variabilei aleatoare x; în caz contrar (număr mare dar finit de exemplare) această valoare nu va fi zero ci va fi o valoare foarte mică, apropiată de zero. Punerea în evidenţă a influenţei mărimii setului de date asupra diferiţilor parametri estimaţi ai matricei de covarianţă va fi evidenţiată în ***Capitolul 7***.



(a). $c_{kl} = -\sigma_k\sigma_l$ **(b).** $c_{kl} = -0.9\sigma_k\sigma_l$

(c). $c_{kl} = -0.5\sigma_k\sigma_l$ **(d).** $c_{kl} = 0$

(e). $c_{kl} = 0.5\sigma_k\sigma_l$ **(f).** $c_{kl} = \sigma_k\sigma_l$

Figura 5.23. Relaţia ce există între diferite valori ale coeficientului de covarianţă şi distribuţia reală a elementelor în spaţiul trăsăturilor

Pentru o problemă reală, caracterizată de un număr suficient de mare de eşantioane, valoarea lui c_{kl}, pentru situaţia trăsăturilor dependente, va fi cu mult mai mică decât valoarea maximă pe care o poate lua acest coeficient, respectiv valoare egală cu produsul deviaţiilor standard ale trăsăturilor k şi l.

Covarianţa a două variabile aleatoare, x_k şi x_l, ce descriu două trăsături ale unei populaţii poate lua valori în intervalul $[-\sigma_k \cdot \sigma_l, +\sigma_k \cdot \sigma_l]$ şi ea măsoară dependenţa/independenţa între cele două variabile.

Corespondenţa între valoarea coeficientului de covarianţă între două trăsături şi o posibilă distribuţie a elementelor unei clase funcţie de aceste două trăsături este dată în **Figura 5.23**. Elipsele prezentate în **Figura 5.23** sunt contururile funcţiei densitate de probabilitate (intersecţia dintre funcţia densitate de probabilitate şi diferite planuri paralele cu planul trăsăturilor, vezi **Figura 5.7**) proiectate în planul trăsăturilor (x_k versus x_l). Aceste elipse poartă numele de *elipse de concentrare*. În **Figura 5.24** se prezintă legăturile care există între diferite reprezentări ale elipselor de concentrare şi posibilele matrici de covarianţă generatoare ale acestor reprezentări.

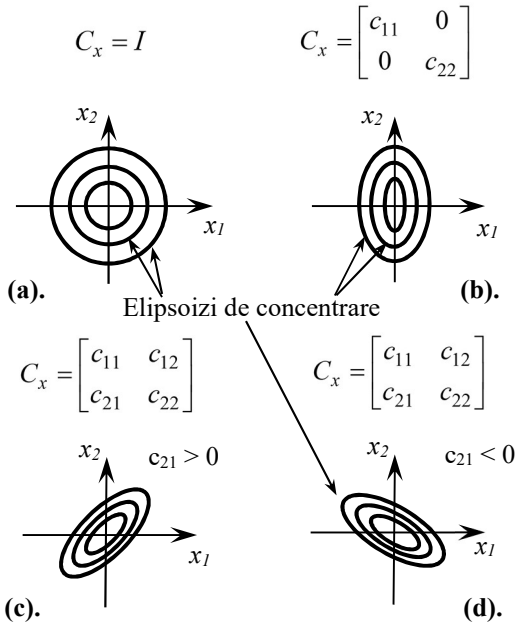


Figura 5.24. Reprezentarea diferitelor distribuţii *Gauss*-iene bidimensionale prin intermediul proiecţiei cotelor de egală probabilitate – coeficienţii c_{kl} sunt daţi de relaţiile (5.197) şi (5.198)

Aplicaţia 5.2: Utilizând programul din directorul ***TrasareDensitateGauss*** asociat acestui capitol:

1. Introduceţi diferite valori pentru coeficienţii matricei C_x astfel încât să obţineţi în mod calitativ toate cazurile prezentate în **Figura 5.24**.

2. Introduceţi diferite numere întregi în *EditBox*-ul „*Levels*” pentru a obţine diferite contururi de echiprobabilitate ale funcţiei densitate de probabilitate *Gauss*-iană. Modificaţi numărul de puncte pe care faceţi analiza şi rotiţi reprezentările pentru îmbunătăţirea vizibilităţii. Pentru a observa simultan atât funcţia densitate de probabilitate cât şi repartiţia punctelor în plan modificaţi transparenţa funcţiei densitate de probabilitate gauss-iene reprezentată grafic prin intermediul *EditBox*-ului „*Transparency*”.

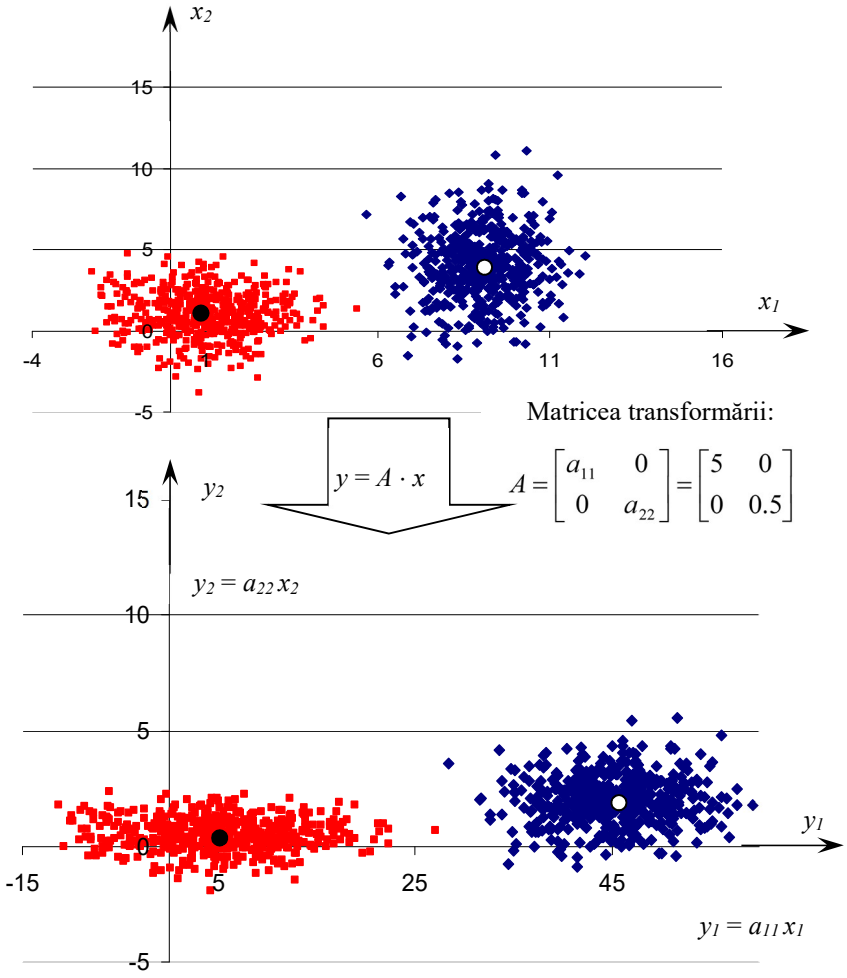


Figura 5.25. Efectul scalării asupra distribuţiilor celor două clase

Întrebări:

1. Cum trebuie să fie coeficienţii c_{11} şi c_{22} astfel încât o elipsă de concentrare a distribuţiei *Gauss*-iene să aibă axa mare paralelă cu axa x_1 a sistemului de coordonate (vezi comparativ şi **Figura 5.24 (b)**)?
2. Coeficienţii c_{21} şi c_{12} pot lua maxim valoarea $\hat{\sigma}_1 \cdot \hat{\sigma}_2$. Atribuiţi această valoare, cu semnul „-” sau „+”, acestor coeficienţi. Explicaţi rezultatele obţinute şi vizualizate.

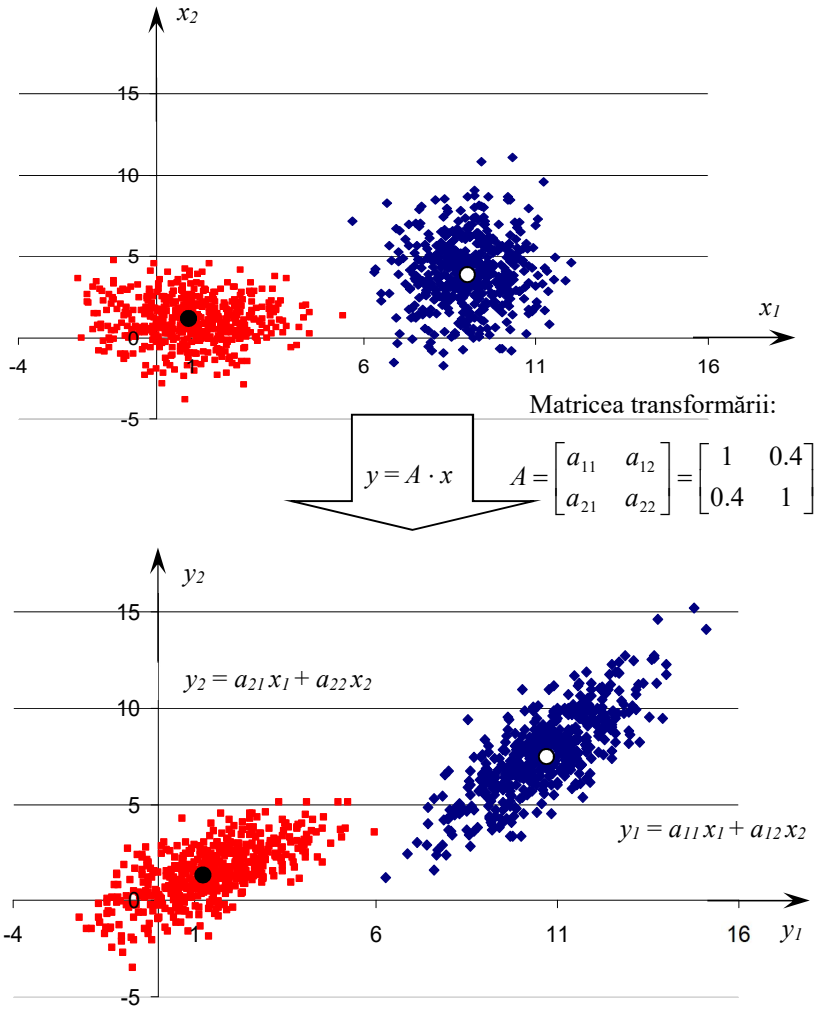


Figura 5.26. Un posibil efect al unei transformări generale asupra distribuţiilor a două clase

Observaţia 5.41: Diferenţa conceptuală care există între **Figura 5.24** şi

Figura 5.23 este dată în principal de faptul că în ultimul set de figuri este tratat cazul unui vector de trăsături bidimensional şi al corelaţiei ce există între cele două trăsături ale lui, în timp ce în **Figura 5.23** este prezentat un caz mai general şi posibilele situaţii existente între oricare două trăsături (de exemplu între trăsăturile k şi l) ale aceluiaşi vector care de această dată poate să fie unul d dimensional.

Există o serie largă de mecanisme care pot determina corelarea trăsăturilor unui vector aleator. De exemplu, aplicarea unei **transformări liniare** asupra setului de date poate avea un asemenea efect (pentru o tratare mai în detaliu a transformatelor liniare vezi ***Subcapitolele 6.1*** și ***6.2***).

Operaţia de scalare este un exemplu particular al ***transformărilor liniare***. Din punct de vedere geometric operaţia de scalare, „dilată" sau „contractă" pe una sau pe mai multe direcţii pe care se aplică aceasta (care corespund cu una sau mai multe axe ale sistemului de coordonate) forma aglomerărilor de realizări particulare ale unei variabilei aleatoare. Astfel, de exemplu, clusteri de vectori care iniţial erau circulari devin de formă elipsoidală, cu axele principale ale elipsoizilor de concetrare ce cracterizează aceste distribuţii orientate paralel cu axele sistemului de coordonate.

O transformare liniară mai generală introduce în plus şi o rotire în planul trăsăturilor. Astfel, de exemplu, clusterii care iniţial erau circulari devin acum elipsoidali cu axele principale ce fac un anumit unghi faţă de axele sistemului de coordonate, **Figura 5.26**. În acest mod se introduce o covarianţă intre componentele vectorului de trăsături.

Exemplu 5.13: Dacă considerăm situaţia achiziţionării simultane, pe N canale, a semnalului EEG (în multe din situaţiile practice N poate lua şi valori mai mari de 20), atunci acestea pot fi grupate sub forma unui ***vector aleator***, $x = [x_1, ..., x_N]^T \in R^N$, ce reprezintă o secvenţă de N variabile aleatoare. Deoarece conductorii pe care se transmite semnalul EEG sunt foarte apropiaţi, fiecare canal induce în toate celelalte un anumit semnal perturbator prin intermediul cuplajului capacitiv (prin intermediul câmpului electric) şi prin intermediul cuplajului inductiv (prin intermediul câmpului magnetic)[51]. Astfel, la intrarea sistemului de achiziţie a unui canal vom avea informaţia de la nivelul electrodului ce corespunde acelui canal plus o combinaţie de semnale (considerate aici perturbatoare) de la celelalte 19 canale:

$$y_i = a_{i1} x_1 + a_{i2} x_2 + \ldots + a_{iN} x_N \qquad (5.212)$$

[51] Acest fenomen poartă denumirea de **diafonie**.

Global pentru toate cele N canale de achiziţie putem scrie:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \tag{5.213}$$

Se observă că *transformarea liniară* prezentată în (5.213) este obţinută prin simpla transmisie a informaţiilor de la electrozii de achiziţie la intrările sistemului de conversie analog-numerică. Această simplă transmisie a datelor determină o corelare mai mică sau mai mare a setului de date de intrare – achiziţionat de la senzorii poziționați pe scapul subiectului.

Aplicaţia 5.3: Utilizând programul din directorul ***Corelare seturi de date 2D*** asociat acestui capitol precum şi setul de date *„Date 2 clase necorelate B.txt”*, parcurgeţi şi răspundeţi cerinţelor de la următoarele puncte:

1. Aplicaţi transformările prezentate în **Figura 5.25** şi **Figura 5.26** şi obţineţi transformări similare ale setului de date, conforme cu figurile anterioare.
2. Determinaţi o transformată A (vezi **Figura 5.25** şi **Figura 5.26**) astfel încât să obţineţi o corelare „negativă” a setului de date – caracterizată de o distribuţie a elementelor conformă cu **Figura 5.23 (b)** sau **(c)**.

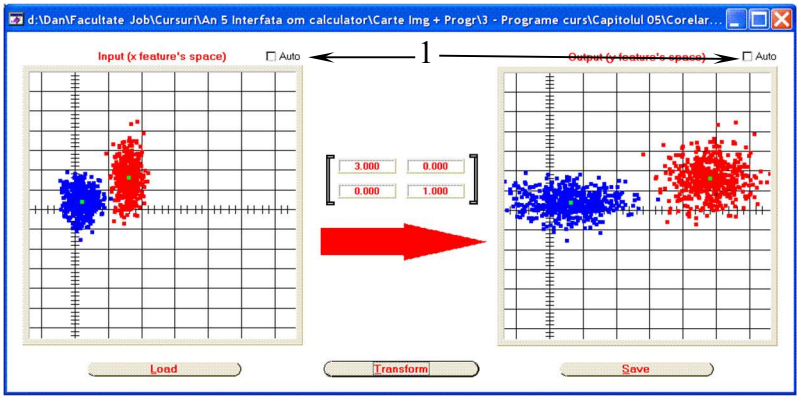


Figura 5.27. Interfaţa grafică a programului

3. Intuitiv, ţinând constante valorile elementelor a_{11} şi a_{22} din matricea A şi considerând egale valorile elementelor a_{12} şi a_{21}, în care din cele două

situaţii valorile elementului a_{12} va fi mai mic dacă se doreşte obţinerea unor distribuţii similare cu cele din **Figura 5.23(b)** sau **(c)**?

Programul de mai sus este capabil să reprezinte atât spaţiul de intrare cat şi cel de ieşire, păstrând aceleaşi unităţi de măsură şi intervale pentru ambele axe astfel încât să se poată face o comparaţie corectă a distribuţiei elementelor în cele două spaţii. În momentul în care se doreşte o scalare a seturilor de date pe cele două clase se ţine cont şi de valorile maxime ale elementelor se utilizează check-boxurile 1, vezi **Figura 5.27**. Încărcarea setului de date de intrare, salvarea rezultatelor şi aplicarea transformării liniare, A, se iniţiază din butoanele poziţionate în partea de jos a aplicaţiei, **Figura 5.27**.

Matricea de cros-corelaţie şi cros-covarianţă

Când o cantitate $g(\cdot)$ depinde de doi vectori aleatori x şi y (cum este cazul, de exemplu, al momentelor comune pentru doi vectori aleatori), media acesteia poate fi scrisă:

$$E\{\, g(x,y)\,\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(a,b) f_{xy}(a,b)\,da\,db \tag{5.214}$$

Momentele statistice comune cele mai folosite pentru doi vectori aleatori oarecare sunt ***matricea de cros-corelaţie***:

$$R_{xy} = E\{\, x\, y^{*T}\,\} \tag{5.215}$$

şi ***matricea de cros-covarianţă***:

$$C_{xy} = E\{\, (x - m_x)\,(y - m_y)^{*T}\,\} \tag{5.216}$$

În cazul cel mai general nici una dintre matricile R_{xy} sau C_{xy} nu este pătratică. Acestea pot fi însă pătratice numai în situaţia în care x şi y au acelaşi număr de elemente. Printr-o demonstraţie similară cu cea dată în momentul obţinerii relaţiei (5.209) se poate arăta că:

$$R_{xy} = C_{xy} + m_x m_y^{*T} \tag{5.217}$$

Deşi ambele matrici R_{xy}, C_{xy} nu sunt, în general, simetrice se poate arăta că:

$$R_{xy} = R_{yx}^{*T} \tag{5.218}$$

și

$$C_{xy} = C_{yx}^{*T} \tag{5.219}$$

Problema 5.28: Demonstraţi validitatea relaţiilor (5.218) şi (5.219).

Proprietăţi ale momentelor statistice de ordin doi

Pentru anumite valori particulare ale momentelor statistice de ordin doi acestea caracterizează, în mod specific, vectorii aleatori asociaţi. Astfel:

(a). Vectorii aleatori x şi y sunt **necorelaţi** dacă:

$$R_{xy} = E\{ x\, y^{*T} \} = E\{ x \}\, E\{ y^{*T} \} = m_x m_y^{*T} \qquad (5.220)$$

Datorită relaţiei (5.217), putem afirma, în mod echivalent, că vectorii aleatori sunt **necorelaţi** dacă:

$$C_{xy} = E\{ (x - m_x) (y - m_y)^{*T} \} = [\, 0 \,] \qquad (5.221)$$

deci, dacă **matricea de cros-covarianţă este zero** – aceasta este, de altfel, şi cea mai uzuală definiţie.

(b). Doi vectori sunt **ortogonali**, dacă:

$$R_{xy} = E\{ x\, y^{*T} \} = [\, 0 \,] \qquad (5.222)$$

deci, dacă **matricea de cros-corelaţie este zero**.

Din relaţia (5.217) putem vedea că în momentul în care media ambilor vectori este zero atunci ambele matrici, de cros-corelaţie şi, respectiv, de cros-covarianţă sunt similare; numai în această situaţie vectorii aleatori ortogonali sunt şi necorelaţi şi invers.

(c). Din relaţia (5.220) este uşor de demonstrat că **atunci când vectorii aleatori *x* şi *y* sunt independenţi[52] atunci sunt şi necorelaţi**. Inversa acestei afirmaţii este în general falsă. Putem spune că această condiţie de independenţă este o condiţie mai „dură", mai restrictivă decât cea de necorelare deoarece independeţa implică automat şi necorelarea.

Problema 5.29: Încercaţi să demonstraţi inversa afirmaţiei precedente, atunci când cei doi vectori sunt caracterizaţi de o distribuţie *Gauss*-iană de probabilitate.

Problema 5.30: Demonstraţi că atunci când vectorii aleatori x şi y sunt independenţi atunci sunt şi necorelaţi.

Aceste concepte de ortogonalitate şi necorelaţie sunt definite în mod similar şi pentru componentele x_k şi x_l ale unui vector aleator x. Astfel:

[52] Doi vectori aleatori x şi y sunt statistic independenţi dacă: $f_{xy}(x,y) = f_x(x) \cdot f_y(y)$

(d). **Componentele *k* şi *l*** ale vectorului aleator x vor fi **ortogonale** dacă:

$$E\{ x_k x_l^{*T} \} = 0; \quad k \neq l \tag{5.223}$$

Ca o consecinţă directă, matricea de corelaţie, R_x, a vectorului aleator x este o matrice diagonală atunci când toate componentele vectorului sunt ortogonale două câte două.

(e). Aceleaşi **componente vor fi necorelate** dacă:

$$E\{ (x_k - m_k) (x_l - m_l)^{*T} \} = 0; \quad k \neq l \tag{5.224}$$

În mod echivalent putem spune că **matricea de covarinţă, C_x, a vectorului aleator *x* este o matrice diagonală atunci când componentele vectorului aleator sunt necorelate** (două câte două).

Atenţie! Deosebirea majoră între proprietăţile **(a)**, **(b)** şi proprietăţile **(d)** şi **(e)** este dată de faptul că primele două proprietăţi **se referă la doi vectori aleatori diferiţi** în timp ce proprietăţile **(d)** şi **(e) se referă la două componente (variabile aleatoare) ale aceluiaşi vector aleator**.

Observaţia 5.42: Este util de ştiut faptul că, dându-se un vector aleator x cu orice tip de distribuţie, se poate găsi o transformare liniară din spaţiul iniţial în alt spaţiu astfel încât componentele vectorului, în noul spaţiu, să fie ortogonale şi/sau necorelate. Astfel de tehnici vor fi prezentate în capitolele următoare. Folosindu-ne de o astfel de transformare liniară putem depăşi cu succes neajunsul prezentat în **Subcapitolul 4.3.2**, respectiv prezenţa corelaţiei între diferitele componente ale vectorului de trăsături; în acest context, în cazul seturilor de date caracterizate de trăsături puternic corelate inclusiv clasificatorul de tip minimă distanţă poate furniza performanţe superioare cu condiţia, bineînţeles, a unei pre-procesări prealabile constând în decorelarea datelor.

5.5.3. Estimarea parametrilor statistici din setul de date

În acest subcapitol se prezintă modul în care se pot estima parametrii statistici (medie, matrice de covarianţă etc.) a unei populaţii sau proces atunci când funcţia densitate de probabilitate este necunoscută, dar se cunosc, în schimb, o serie de realizări particulare ale vectorului aleator x: a^1, a^2, …, a^K.

Dacă ar fi să determinăm media, matricea de corelaţie sau cea de covarianţă a unui vector aleator x atunci, conform modului cum au fost definite aceste momente (vezi ***Subcapitolul 5.5.2***, subpunctul 1 „***Momentele în analiza statistică. Funcţii caracteristice***”), avem nevoie să cunoaştem, în

prealabil, funcţia densitate de probabilitate a lui *x*. Reluând mai jos relaţiile prin care sunt definite aceste cantităţi, avem că:

$$m_x = E\{x\} \overset{def.}{=} \int_{-\infty}^{+\infty} a\, f_x(a)\, da \tag{5.225}$$

$$R_x = E\{x\, x^{*T}\} = \int_{-\infty}^{+\infty} a\, a^{*T} f_x(a)\, da \tag{5.226}$$

$$C_x = E\left\{(x - m_x)(x - m_x)^{*T}\right\} = \int_{-\infty}^{+\infty} (a - m_x)(a - m_x)^{*T} f_x(a)\, da \tag{5.227}$$

În continuare prezentăm şi demonstrăm modul în care se pot estima, din setul de date, diferiţi parametri statistici fără a cunoaşte efectiv funcţia densitate de probabilitate. Această prezentare va fi făcută pentru un vector aleator real monodimensional – deci, vom lua în considerare cazul unei variabile aleatoare simple. Generalizarea acestei demonstraţii (de la situaţia monodimensională la o alta, în mai multe dimensiuni) este una directă.

1. Estimarea momentelor statistice pentru un vector aleator

Fie $g(x)$ o funcţie ce are ca argument vectorul aleator x. Dacă ambele mărimi $f_x(a)$ şi $g(a)$ nu variază substanţial pe toate intervalele definite, atunci, folosind relaţia (5.166), putem realiza aproximarea:

$$E\{g(x)\} = \int_{-\infty}^{+\infty} g(a) f_x(a) da \cong \sum_{toti\ a^0} g(a^0) f_x(a^0) \Delta a \cong \frac{1}{K} \sum_{toti\ a^0} g(a^0) K_{a^0} \tag{5.228}$$

unde K reprezintă numărul total de valori ale lui x dintr-un eşantion iar K_{a^0} este numărul de valori din eşantion aflate în intervalul $(a^0, a^0 + \Delta a]$; deci, pe un singur interval de lungime Δa care porneşte din punctul a^0 avem o mulţime formată din K_{a^0} elemente, respectiv:

$$\left\{a^i, a^{i+1}, ..., a^{i+K_{a^0}-1}\right\} \tag{5.229}$$

Pentru toţi vectorii a^i aparţinând aceluiaşi interval, în ipoteza $g(a^i) \approx g(a^0)$ (ipoteză necesară aproximării integralei din relaţia (5.228)), cantitatea $g(a^0) K_{a^0}$ poate fi scrisă sub forma:

$$K_{a^0}\cdot g\left(a^0\right)=\underbrace{\overbrace{g\left(a^0\right)}^{g\left(a^i\right)}+\overbrace{g\left(a^0\right)}^{g\left(a^{i+1}\right)}+\ldots+\overbrace{g\left(a^0\right)}^{g\left(a^{i+K_{a^0}-1}\right)}}_{de\ K_{a^0}\ ori}=\sum_{j=i}^{i+K_{a^0}-1}g\left(a^j\right) \qquad (5.230)$$

Relaţia (5.228) poate fi rescrisă prin sumarea tuturor cantităţilor de tipul (5.230). Dacă această procedură este urmată pentru toate intervalele, relaţia (5.228) devine:

$$E\{g(x)\}\cong\frac{1}{K}\sum_{k=1}^{K}g(a^k) \qquad (5.231)$$

În acest mod am obţinut un *estimat empiric* (valoare aproximată) pentru media vectorului/matricei aleator(oare) $g(x)$ a lui x; în plus, se poate observa, din relaţia de mai sus, că **estimatorul** este dat chiar de ***media aritmetică a eşantionului***, calculată pentru $g(x)$.

Particularizând definiţia funcţiei $g(x)$, respectiv, $g(x) = x$, $g(x) = x\,x^{*T}$ şi $g(x) = (x - m_x)\,(x - m_x)^{*T}$, vom obţine următoarele estimaţii pentru media vectorului aleator x, matricea de corelaţie şi, respectiv, matricea de covarianţă a lui x:

$$\hat{m}_x=E\{g(x)\}=E\{x\}=\frac{1}{K}\sum_{k=1}^{K}a^k \qquad (5.232)$$

$$\hat{R}_x=E\{g(x)\}=E\left\{x\cdot x^{*T}\right\}=\frac{1}{K}\sum_{k=1}^{K}a^k\left(a^k\right)^{*T} \qquad (5.233)$$

$$\hat{C}_x=E\{g(x)\}=E\left\{(x-m_x)\cdot(x-m_x)^{*T}\right\}=\frac{1}{K}\sum_{k=1}^{K}\left(a^k-m_x\right)\left(a^k-m_x\right)^{*T} \qquad (5.234)$$

2. Estimarea momentelor statistice comune pentru doi vectori aleatori

Un rezultat similar se obţine şi în cazul estimării cantităţilor $g(x,y)$, dependente de doi vectori aleatori, x şi y – vectori ce sunt caracterizaţi prin realizările lor particulare: a^1, a^2, …, a^k şi, respectiv, b^1, b^2, …, b^k. În această situaţie un estimat pentru parametrul *media* vectorului/matricei aleator(oare) $g(x,y)$ este dat de relaţia:

$$E\{g(x,y)\} \cong \frac{1}{K}\sum_{k=1}^{K} g(a^k, b^k) \tag{5.235}$$

unde estimatorul este, de asemenea, *media aritmetică a eşantionului*, calculată însă, de această dată, pentru $g(x,y)$.

Şi de această dată, particularizând definiţia funcţiei $g(x,y)$, respectiv, $g(x,y) = xy^{*T}$ şi $g(x,y) = (x - m_x)(y - m_y)^{*T}$, obţinem următoarele estimaţii pentru matricile de cros-corelaţie, respectiv, cros-covarianţă ale vectorilor aleatori x şi y:

$$\hat{R}_{xy} = E\{g(x,y)\} = E\{x \cdot y^{*T}\} = \frac{1}{K}\sum_{k=1}^{K} a^k \left(b^k\right)^{*T} \tag{5.236}$$

$$\hat{C}_{xy} = E\{g(x,y)\} = E\{(x - m_x)\cdot(y - m_y)^{*T}\} = \frac{1}{K}\sum_{k=1}^{K}\left(a^k - m_x\right)\left(b^k - m_y\right)^{*T} \tag{5.237}$$

3. Estimarea momentelor statistice din matricea setului de date

Pentru estimarea, de exemplu, a matricelor de corelaţie şi cros-corelaţie a unui vector aleator x, respectiv, a doi vectori aleatori, x şi y, se pot folosi, aşa cum am arătat mai sus, în mod direct, relaţiile (5.231) şi, corespunzător, (5.235). La acelaşi rezultat se poate ajunge, însă, şi printr-o altă metodă echivalentă, metodă pe care o prezentăm în cele ce urmează.

Astfel, într-un prim pas se defineşte matricea vectorilor de date sub forma:

$$X = \begin{bmatrix} - & \left(a^1\right)^{*T} & - \\ - & \left(a^2\right)^{*T} & - \\ & \vdots & \\ - & \left(a^K\right)^{*T} & - \end{bmatrix} \tag{5.238}$$

Matricea X mai poartă numele de *matricea eşantionului* sau *matricea setului de date*.

Folosind următorul produs:

$$X^{*T}X = \sum_{k=1}^{K} a^k \left(a^k\right)^{*T} \tag{5.239}$$

putem estima, de exemplu, matricea de corelaţie, *în mod direct*, prin relaţia:

$$\hat{R}_x = \frac{1}{K} X^{*T} X \tag{5.240}$$

Formulele de calcul pentru estimarea matricei de corelaţie, date de relaţiile (5.233) şi (5.240) sunt, deci, echivalente, cu menţiunea că ultima dintre acestea este, însă, una mult mai directă.

Estimarea matricei de covarianţă se poate face într-un mod similar celui prezentat mai sus. Singura deosebire este, în acest caz, dată de maniera în care definim, pentru început, matricea de calcul. Astfel, în locul matricei setului de date, X, construim matricea X_0, dată de elementele eşantionului din care s-a extras, mai întâi, vectorul medie m_x:

$$a_0^i = a^i - m_x, \quad pentru\ i = \overline{1,K} \tag{5.241}$$

$$X_0 = \begin{bmatrix} - & \left(a_0^1\right)^{*T} & - \\ - & \left(a_0^2\right)^{*T} & - \\ & \vdots & \\ - & \left(a_0^K\right)^{*T} & - \end{bmatrix} \tag{5.242}$$

În continuare, pentru determinarea matricei de covarianţă aplicăm relaţia:

$$\hat{C}_x = \frac{1}{K} X_0^{*T} X_0 \tag{5.243}$$

Modalitatea de estimare a matricii de covarianță, dată de relația (5.243), implică calcularea matricii X_0 prin eliminarea din fiecare vector de trăsături în parte a valorii medii (conform relației (5.241)) și apoi se va compune matricea X_0 din acești vectori. Această abordare este una ușor de abordat pentru un operator uman. Un sistem informatic (embedded – DSP, microcontroler etc.) va lucra în mod direct cu matrici, pornind de la matricea setului de date, relația (5.238), și prin utilizarea unor calcule matriciale este de dorit a obține direct matricea X_0. O astfel de abordare se prezintă mai departe:

$$X_0 = X - \frac{1}{K} \cdot [1] \cdot X \tag{5.244}$$

În relația (5.244) matricea [1] este o matrice pătratică, având o dimensionalitate K x K (K este numărul de vectori de trăsături utilizați în estimarea matricii de covarianță), cu toate elementele egale cu 1.

Problema 5.31: Demonstrați corectitudinea relației (5.244).

Problema 5.32: Pentru următoarele realizări particulare ale unui vector aleator *x* bidimensional, $a^1=\begin{bmatrix}1\\2\end{bmatrix}$, $a^2=\begin{bmatrix}2\\1\end{bmatrix}$, $a^3=\begin{bmatrix}3\\3\end{bmatrix}$, $a^4=\begin{bmatrix}2\\2\end{bmatrix}$ estimaţi:

1. Media vectorului aleator, ce caracterizează clasa generată de realizările particulare anterior prezentate.
2. Matricea de covarianţă a clasei (estimarea matricei de covarianţă se va realiza prin două metode diferite).

Rezolvare:

1. Pentru estimarea vectorului mediu al clasei ne folosim de relaţia (5.232) particularizată după cum urmează:

$$\hat{m}_x=\frac{1}{4}\sum_{i=1}^{4}a^i=\frac{1}{4}\left\{\begin{bmatrix}1\\2\end{bmatrix}+\begin{bmatrix}2\\1\end{bmatrix}+\begin{bmatrix}3\\3\end{bmatrix}+\begin{bmatrix}2\\2\end{bmatrix}\right\}=\begin{bmatrix}2\\2\end{bmatrix}.$$

2. Pentru estimarea matricei de covarianţă vom extrage, într-un prim pas, vectorul mediu din toate realizările particulare ale vectorului aleator *x*. Astfel obţinem:

$$a_0^1=\begin{bmatrix}-1\\0\end{bmatrix},\quad a_0^2=\begin{bmatrix}0\\-1\end{bmatrix},\quad a_0^3=\begin{bmatrix}1\\1\end{bmatrix},\quad a_0^4=\begin{bmatrix}0\\0\end{bmatrix}$$

   a. Prima metodă de estimare a matricei de covarianţă se bazează pe utilizarea relaţiei (5.234).

$$\hat{C}_x=E\left\{(x-m_x)(x-m_x)^{*T}\right\}\cong\frac{1}{4}\sum_{i=1}^{4}\left(a^i-\hat{m}_x\right)\left(a^i-\hat{m}_x\right)^{*T}=$$

$$=\frac{1}{4}\left\{\begin{bmatrix}-1\\0\end{bmatrix}\begin{bmatrix}-1&0\end{bmatrix}+\begin{bmatrix}0\\-1\end{bmatrix}\begin{bmatrix}0&-1\end{bmatrix}+\begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1&1\end{bmatrix}+\begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}0&0\end{bmatrix}\right\}=\begin{bmatrix}1/2&1/4\\1/4&1/2\end{bmatrix}$$

$$\hat{C}_x=\begin{bmatrix}1/2&1/4\\1/4&1/2\end{bmatrix}$$

   b. În cea de a doua metodă de estimare, bazată pe matricea setului de date, construim, mai întâi, matricea X_0 dată de elementele eşantionului din care am extras, anterior, vectorul mediu. Astfel:

$$X_0 = \begin{bmatrix} - & \left(a_0^1\right)^{*T} & - \\ - & \left(a_0^2\right)^{*T} & - \\ & \vdots & \\ - & \left(a_0^4\right)^{*T} & - \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$$

Matricea de covarianţă estimată va fi egală cu:

$$\hat{C}_x = \frac{1}{K} X_0^{*T} X_0 = \frac{1}{4} \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

5.5.4. Coeficientul de corelaţie

Coeficientul de corelaţie este o valoare ce **caracterizează gradul de dependenţă liniară** dintre două variabile aleatoare şi el ia întotdeauna valori în intervalul [-1, 1]. Cu cât coeficientul de corelaţie este mai apropiat de capetele intervalului de variaţie (respectiv, de valorile -1 sau 1), cu atât cele două variabile aleatoare sunt mai puternic dependente liniar una faţă de cealaltă.

Spunem că avem o relaţie de *dependenţă liniară totală*[53] între, să zicem, variabilele aleatoare x_k şi x_l (reprezentând componentele k şi l ale unui vector aleator de trăsături), atunci când coeficientul de corelaţie dintre acestea va avea valoarea 1 sau -1. În această situaţie particulară, reprezentarea grafică, în spaţiul trăsăturilor (x_k, x_l), a vectorilor de trăsături, va conduce la un set de puncte poziţionate pe o dreaptă cu pantă pozitivă (vezi **Figura 5.23.(f))** pentru o valoare egală cu +1 a coeficientului de corelaţie. În cazul în care coeficientul de corelaţie ia valoarea -1 avem de-a face, din nou, cu o relaţie de liniaritate perfectă între cele două variabile, iar vectorii de trăsături se vor poziţiona pe o dreaptă care va avea de această dată, însă, o pantă negativă.

Corelaţia dintre două variabile aleatoare va avea o **valoare pozitivă** dacă, în medie, creşterea valorii unei variabile aleatoare faţă de media ei va determina creşterea valorii celeilate variabile faţă de propria medie, iar scăderea valorii primei variable aleatoare faţă de medie va determina, de cele

53 *Dependenţa liniară totală* poate fi tradusă, de exemplu, prin aceea că **<u>întotdeauna</u>** creşterea/scăderea valorii unei trăsături faţă de media ei va determina creşterea/scăderea, cu aceeaşi cantitate echivalentă, şi a valorii celei de a doua trăsături faţă de media corespunzătoare ei.

mai multe ori, scăderea valorii celei de a doua variabile aleatoare faţă de propria ei medie. Spunem despre aceste variabile aleatoare că sunt ***corelate pozitiv***. Exemple de variabile corelate pozitiv: educaţia şi venitul obţinut, consumul grăsimilor saturate şi nivelul colesterolului, nivelul colesterolului şi probabilitatea unui atac de cord etc.

Pentru variabilele ***corelate negativ*** mecanismul de variabilitate presupune următoarele: de cele mai multe ori (în medie) creşterea valorii unei variabile aleatoare faţă de media ei va determina scăderea valorii celeilate variabile faţă de propria medie, iar scăderea valorii primei variabile aleatoare faţă de medie va determina, de cele mai multe ori (în medie), creşterea valorii celei de a doua variabile aleatoare faţă de propria ei medie. Exemple de variabile corelate negativ: numărul de ţigări fumate de mamă şi greutatea copilului la naştere, masa maşinii şi consumul de carburant pe 100 de Km, numărul de ţigări fumate şi durata de viaţă etc.

Dacă coeficientul de corelaţie ia valoarea zero atunci spunem că ***nu există nici o relaţie linară*** între cele două variabile aleatoare.

În prezent se cunosc mai multe metode de calcul al coeficientului de corelaţie. Unele dintre aceste metode sunt mai potrivite comparativ cu altele, funcţie de tipul variabilelor aleatoare analizate. Astfel, spre exemplu, există *coeficientul de corelaţie Pearson*, *Spearman*, *coeficientul de corelaţie intraclasă* (*intraclass correlation coefficient*, ICC) etc.

În cadrul acestei cărţi vom prezenta doar ***coeficientul de corelaţie Pearson*** care este definit de relaţia:

$$\rho \overset{def}{=} \frac{c_{kl}}{\sqrt{c_{kk} \cdot c_{ll}}} = \frac{Cov(x_k, x_l)}{\sqrt{Cov(x_k)Cov(x_k)}} \tag{5.245}$$

sau, echivalent,

$$\rho = \frac{Cov(x_k, x_l)}{\sqrt{\sigma_{x_k}^2 \cdot \sigma_{x_l}^2}} = \frac{Cov(x_k, x_l)}{\sigma_{x_k} \cdot \sigma_{x_l}} = \frac{E\left\{(x_k - m_k)(x_l - m_l)^*\right\}}{\sqrt{E\left\{\left|(x_k - m_k)\right|^2\right\} \cdot E\left\{\left|(x_l - m_l)\right|^2\right\}}} \tag{5.246}$$

Problema 5.33: Pentru următoarele două matrici de covarianţă:

$$C_x = \begin{bmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{bmatrix} \text{ şi } C_y = \begin{bmatrix} 4 & -2 \\ -2 & 9 \end{bmatrix}$$

ce caracterizează doi vectori aleatori bidimensionali, *x* şi *y*, determinaţi coeficientul de corelaţie între trăsăturile acestor vectori.

Aplicaţia 5.4: Cu ajutorul programului ce se află în directorul ***CoefCorr-Pearson***, asociat acestui capitol, şi care determină coeficientul de corelaţie Pearson între două semnale reprezentate vectorial prin doi vectori aleatori *x* şi *y* (indiferent de dimensiunile acestora), se cere să:

1. Determinaţi relaţiile matematice ce ar fi necesare în implementarea programului.
2. Generaţi cu ajutorul programului ce se află în directorul ***Generare distributie bivariata Gauss-iana*** (vezi ***Capitolul 6*** al acestei cărţi) diferite distribuţii *Gauss*-iene bidimensionale, al căror coeficient de corelaţie între variabilele aleatoare ce formează vectorul aleator îl cunoaşteţi. Convertiţi seriile de timp astfel încât să poată fi încărcate în programul ce determină coeficientul de corelaţie – folosiţi-vă, de exemplu, de programul de calcul tabelar Microsoft Office Excel. Verificaţi corectitudinea coeficientului de corelaţie determinat de program faţă de cel introdus de dumneavoastră.
3. În directorul seturilor de date asociat acestui capitol se găsesc două serii de timp ***„x - distrib parabolica.txt”*** şi ***„y - distrib parabolica.txt”***. În primul fişier avem o variabilă aleatoare uniform distribuită în intervalul [-5, 5]. În cel de al doilea fişier este un alt şir de valori obţinut din primul cu ajutorul relaţiei $y = x^2$; deci, variabila aleatoare *y* este complet determinată de variabila aleatoare *x*. Dacă le reprezentăm grafic, prima versus cea de a doua, obţinem reprezentarea grafică din **Figura 5.28**.

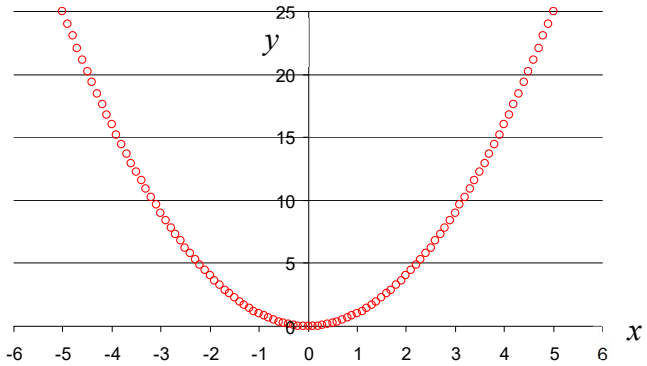


Figura 5.28. Reprezentarea grafică a seriilor de timp ***„x - distrib parabolica.txt”*** şi ***„y - distrib parabolica.txt”***

Calculând cu ajutorul programului de mai sus coeficientul de corelaţie vom obţine pentru aceste două serii de timp o valoare egală

cu zero, valoare identică cu cea obţinută pentru o distribuţie de tipul celei prezentate în **Figura 5.23(d)**, în care variabilele aleatoare sunt necorelate. Prin urmare, variabilele aleatoare *x* şi *y* sunt dependente (cunoaştem chiar relaţia funcţională ce le leagă) dar corelaţia lor este zero. Din ce cauză am obţinut această valoare?

Rezolvare:

1. Relaţia (5.246) se generalizează foarte uşor pentru cazul unor variabile aleatoare *x* şi *y* reale sub forma:

$$\rho = \frac{Cov(x,y)}{\sqrt{\sigma_y^2 \cdot \sigma_y^2}} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{E\{(x-m_x)(y-m_y)\}}{\sigma_x \cdot \sigma_y} \tag{5.247}$$

Ţinând cont, în plus, că:

$$m_x = E\{x\} \tag{5.248}$$

$$\sigma_x^2 = E\{x^2\} - E^2\{x\} \tag{5.249}$$

putem foarte uşor să deducem relaţia:

$$\rho = \frac{E\{x \cdot y\} - E\{x\}E\{y\}}{\sqrt{E\{x^2\} - E^2\{x\}}\sqrt{E\{y^2\} - E^2\{y\}}} \tag{5.250}$$

sau

$$\rho = \frac{E\{x \cdot y\} - m_x m_y}{\sqrt{E\{x^2\} - m_x^2}\sqrt{E\{y^2\} - m_y^2}} \tag{5.251}$$

Estimând termenii din relaţia (5.251) conform relaţiei (5.231) şi ştiind că dispunem de un set de date de lungime *N* elemente, rezultă imediat că:

$$\rho = \frac{\frac{1}{N} \cdot \sum_{i=1}^{N}\left(x^i \cdot y^i\right) - \frac{1}{N}\sum_{i=1}^{N} x^i \cdot \frac{1}{N}\sum_{j=1}^{N} y^j}{\sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N}\left(x^i\right)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x^i\right)^2}\sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N}\left(y^i\right)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} y^i\right)^2}} \tag{5.252}$$

sau

$$\rho = \frac{N \cdot \sum_{i=1}^{N} \left(x^i \cdot y^i \right) - \sum_{i=1}^{N} x^i \cdot \sum_{j=1}^{N} y^j}{\sqrt{N \cdot \sum_{i=1}^{N} \left(x^i \right)^2 - \left(\sum_{i=1}^{N} x^i \right)^2} \sqrt{N \cdot \sum_{i=1}^{N} \left(y^i \right)^2 - \left(\sum_{i=1}^{N} y^i \right)^2}} \qquad (5.253)$$

2. În mod intuitiv pentru distribuţia din **Figura 5.28** obţinem un coeficient de corelaţie egal cu zero deoarece valorile mari ale variabilei aleatoare *y* sunt asociate atât cu valori mari (pozitive) dar şi cu cele mici (negative) ale variabilei aleatoare *x*.

   Existenţa unei relaţii neliniare între variabilele aleatoare pentru care calculăm coeficientul de corelaţie *este o sursă de erori* a coeficientului de corelaţie Pearson. Acest fapt apare deoarece **coeficientul de corelaţie Pearson măsoară doar existenţa sau nu a unei relaţii liniare** între variabilele aleatoare implicate în calcul.

Anterior am făcut afirmaţia că dacă ***doi vectori aleatori sau variabile aleatoare, x şi y, sunt independente atunci ele sunt şi necorelate***; în schimb, s-a afirmat că ***reciproca nu este întotdeauna adevărată***. Inversa afirmaţiei precedente nu este adevărată întotdeauna deoarece corelaţia pune în evidenţă doar relaţiile liniare ce există între variabile. Dacă parcurgem punctul trei al aplicaţiei precedente observăm existenţa a două variabile aleatoare necorelate (coeficientul de corelaţie este zero) dar dependente între ele – variabila aleatoare *y* este complet determinată de variabila aleatoare *x*, prin intermediul relaţiei $y = x^2$.

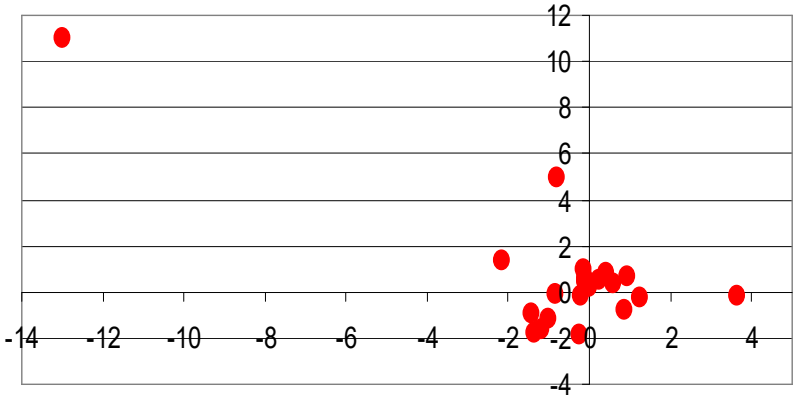


Figura 5.29. Influenţa elementelor aberante asupra valorii coeficientului de corelaţie

Atunci când în setul de date apar înregistrări aberante ("*outlier*" în limba engleză), acestea influenţează în mod negativ valoarea coeficientului de corelaţie. Astfel, spre exemplu, dacă analizăm numai setul de date poziţionat în centrul sistemului de coordonate, vezi **Figura 5.29,** obţinem o valoare a coeficientului de corelaţie de 0.002853; ţinând cont că este un set de date mic, de doar 20 de vectori de trăsături, putem trage concluzia că aceste trăsături, care formează vectorul de trăsături, sunt necorelate. Dacă în schimb, în analiză luăm în calcul şi elementul "aberant" (cel poziţionat în partea stânga sus), atunci obţinem un coeficient de corelaţie al setului de date de -0.7847, valoare pe baza căreia suntem tentaţi să tragem concluzia că variabilele aleatoare ce caracterizează această distribuţie sunt corelate negativ, ceea ce, din nou, nu este adevărat. Raportarea acestui coeficient de corelaţie, de -0.7847, este falsă deoarece coeficientul astfel calculat nu este reprezentativ pentru cea mai mare parte a setului de date.

Aplicaţia 5.5: Cu ajutorul programului ce se află în directorul ***CoefCorr-Pearson*** încărcaţi seriile de timp *„outliers - x.txt"* şi *„outliers - y.txt"*. Vizualizaţi rezultatele în două situaţii distincte, respectiv cu şi fără primul element din ambele serii. Determinaţi coeficientul de corelaţie şi reprezentaţi grafic rezultatele pentru ambele cazuri prezentate precedent, utilizând, de această dată, programul de calcul tabelar Microsoft Office Excel.

Elementele aberante pot să apară, de exemplu, în practică atunci când punctul zecimal nu este poziţionat corect (setul de date lucrează cu caracterul „ **.** " drept punct zecimal iar programul ce realizează analiza utilizează drept punct zecimal caracterul „ **,** ") sau atunci când unul din electrozii cu ajutorul căruia s-a făcut achiziţia unui semnal biomedical s-a desprins ş.a.m.d.

Datorită erorilor pe care le pot genera aceste elemente aberante, anterior oricărei analize, fiecare variabilă aleatoare este preprocesată şi sunt menţinuţi spre analiză şi procesare, de exemplu, numai acele valori ale vectorului aleator de trăsături pentru care valorile elementelor componente (realizări particulare ale variabilelor aleatoare componente) se găsesc, fiecare în parte, în intervalul $[m - 3\cdot\sigma, m + 3\cdot\sigma]$ (prin m am notat valoarea medie a variabilei aleatoare iar σ este deviaţia standard a aceleiaşi variabile aleatoare); pentru o distribuţie normală în intervalul astfel specificat ar trebui să se regăsească, teoretic cel puţin, cca. 99.72% dintre valori.

O altă problemă ce poate să apară şi să influenţeze în mod negativ calcularea coeficientului de corelaţie este dată de existenţa, în setul de date, a unui număr de clase distincte care, însă, sunt considerate în cadrul analizei,

drept o singură clasă. Astfel, dacă analizăm distribuţia prezentată în **Figura 5.30,** ambele clase sunt caracterizate de vectori de trăsături necorelate. Coeficientul de corelaţie pentru prima clasă este de 0.0028 iar pentru cea de a doua clasă este de -0.0099. Dacă, în schimb, nu am fi dispus de nici o informaţie *a priorică* privind structura setului de date şi, în consecinţă, l-am fi privit ca pe un întreg, ca pe un set de date ce caracterizează o singură clasă, atunci am fi obţinut un coeficient de corelaţie global de -0.773, ceea ce nu ar fi reflectat situaţia reală existentă.

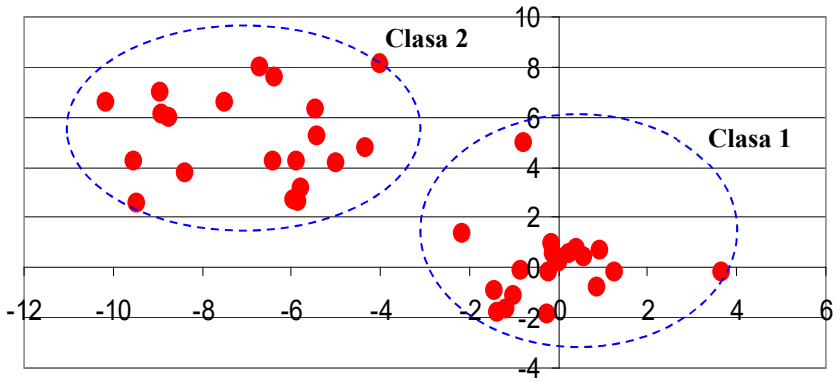


Figura 5.30. Influenţa existenţei a două clase distincte în setul de date asupra coeficientului de corelaţie

Aplicaţia 5.6: Cu ajutorul programului ce se află în directorul ***CoefCorr-Pearson*** încărcaţi seriile de timp *„2 clase - x.txt”* şi *„2 clase - y.txt”*. Vizualizaţi şi analizaţi rezultatele în următoarele două situaţii distincte: **(a)** ambele clase considerate drept un întreg (drept o singură clasă) şi **(b)** cazul în care clasele sunt considerate şi analizate separat. Determinaţi coeficientul de corelaţie şi reprezentaţi grafic rezultatele pentru toate situaţiile prezentate anterior, utilizând pentru aceasta programul de calcul tabelar, Microsoft Office Excel.

Din ultimele trei exemple s-a putut observa că există un număr de situaţii (dependenţă neliniară a datelor, valori aberante în setul de date, mai multe clase distincte în setul de date etc.) când valorile coeficientului de corelaţie nu reflectă situaţia reală şi mecanismele de cuplare ce au generat datele respective. Din acest motiv, şi nu numai, apare ca fiind **imperios necesar să generăm, într-o primă etapă, o reprezentare vizuală a seturilor de date** şi abia ulterior să determinăm valoarea coeficientului de corelaţie, în mod specific, pentru fiecare set de date.

O aplicaţie practică, directă, a coeficientului de corelaţie ţine de *reducerea dimensionalităţii spaţiului de trăsături*. Dacă, de exemplu, două trăsături ale unui vector de trăsături, *d*-dimensional, sunt caracterizate de un coeficient de corelaţie apropiat de valorile 1 sau -1, aceste trăsături pot fi combinate într-o singură trăsătură sau, eventual, una dintre ele poate fi eliminată complet. În acest mod, dimensionalitatea vectorului de trăsături se reduce cu o unitate şi, simultan, informaţia redundantă existentă în setul de date este eliminată.

5.5.5. Funcţia de coerenţă spectrală

Dependenţa dintre două semnale staţionare (vezi **Anexa: Noţiunea de semnal. Semnal staţionar**) şi mărginite poate fi caracterizată de anumiţi parametri care evaluează şi evidenţiază corelaţia dintre acestea.

Pentru două procese $x(t)$ şi $y(t)$ ***funcţia de coerenţă spectrală*** este dată de relaţia:

$$\left|R_{xy}(\omega)\right|^2 = \lim_{T\to\infty}\left| corr\left\{F_x^T(\omega)\cdot F_y^T(\omega)\right\}\right| \qquad (5.254)$$

Această funcţie furnizează o măsură a gradului de corelaţie între procesele $x(t)$ şi $y(t)$ pentru fiecare componentă spectrală în parte. Relaţia (5.254) poate fi văzută ca pătratul amplitudinii funcţiei de corelaţie între transformatele Fourier ale secvenţelor $x(t)$ şi $y(t)$ [Amjad, 1997].

În relaţia (5.254), $F_x^T(\omega)$ reprezintă transformata Fourier a unui segment de semnal de lungime T, aparţinând procesului $x(t)$.

Deoarece rezultatele obţinute în mod practic reprezintă doar nişte estimări ale funcţiei de coerenţă spectrală, se creează şi, în acest caz, premizele calculului estimatorilor iar, în final, prezentăm inclusiv gradul corespunzător de încredere al estimării.

Pentru a calcula un estimat al funcţiei de coerenţă, împărţim, mai întâi, seriile de timp în secvenţe de o anumită durată/lungime (fie T lungimea unor astfel de secvenţe). În consecinţă, pentru fiecare serie de timp înregistrată se obţin, să zicem, L astfel de segmente, fiecare de lungime T. Transformata Fourier a unui astfel de segment, $F_x^T(\omega,l)$, $l = 1 \ldots L$, pentru pulsaţia ω este dată de relaţia:

$$F_x^T(\omega,l) = \int_{(l-1)T\Delta t_s}^{lT\Delta t_s} x(t)\cdot e^{-i\omega t}dt \quad \approx \sum_{t=(l-1)T}^{lT} e^{-i\omega t}x(t) \qquad (5.255)$$

Definiţia funcţiei de corelaţie între transformatele Fourier a celor două procese în funcţie de varianţa şi covarianţa acestora pe un anumit segment este dată de:

$$corr\left\{F_x^T(\omega)F_y^T(\omega)\right\} = \frac{\mathrm{cov}\left\{F_x^T(\omega)F_y^T(\omega)\right\}}{\sqrt{\mathrm{var}\left\{F_x^T(\omega)\right\}\cdot\mathrm{var}\left\{F_y^T(\omega)\right\}}} \qquad (5.256)$$

Folosindu-ne de relaţia (5.256) putem obţine o definiţie alternativă a funcţiei de coerenţă spectrală între două procesele $x(t)$ şi $y(t)$:

$$\left|R_{xy}(\omega)\right|^2 = \frac{\left|f_{xy}(\omega)\right|^2}{f_{xx}(\omega) f_{yy}(\omega)} \tag{5.257}$$

unde $f_{xy}(\omega)$ este **densitatea cros-spectrală** a celor două procese $x(t)$ şi $y(t)$ iar $f_{xx}(\omega)$ şi $f_{yy}(\omega)$ sunt **densităţile auto-spectrale**, pentru pulsaţia ω, ale fiecărui proces în parte.

Relaţia (5.257) indică gradul de corelaţie liniară în domeniul spectral între două semnale, pe o scală de la 0 (semnificând independenţa liniară) la unu (completa dependenţă liniară a unui proces de celălalt), acest parametru putând fi interpretat şi ca o măsură a predictibilităţii liniare.

Un avantaj al acestei metode este dat de faptul că, indiferent de unităţile de măsură ale seriilor $x(t)$ şi $y(t)$, valoarea coerenţei spectrale pentru o anumită frecvenţă nu are nici o unitate de măsură, furnizând doar o interpretare a relaţiei în care se găsesc cele două semnale, la frecvenţa respectivă, conform cu specificaţiile anterioare.

Auto-spectrul şi cros-spectrul necesare în relaţia (5.257) pot fi estimate prin medierea segmentelor de semnal disjuncte, anterior obţinute. Pentru calcularea (estimarea) cros-spectrului relaţia de calcul este dată de:

$$\hat{f}_{xy}(\omega) = \frac{1}{2\pi LT} \sum_{l=1}^{L} F_x^T(\omega, l) \cdot \overline{F_y^T(\omega, l)} \tag{5.258}$$

În calcularea auto-spectrului secvenţei x, $f_{xx}(\omega)$, se foloseşte relaţia (5.258) în care y se înlocuieşte cu x. În relaţia (5.258) bara orizontală de deasupra termenului $F_y^T(\omega, l)$ simbolizează complex conjugatul acestuia.

În final, pentru estimarea funcţiei de coerenţă spectrală din relaţia (5.257) am utilizat estimatorii calculaţi cu relaţia (5.258). Astfel:

$$\left|\hat{R}_{xy}(\omega)\right|^2 = \frac{\left|\hat{f}_{xy}(\omega)\right|^2}{\hat{f}_{xx}(\omega) \hat{f}_{yy}(\omega)} \tag{5.259}$$