

B. C. H.

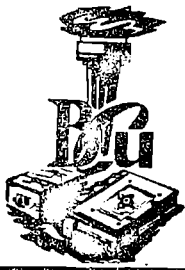
III 667/66

LUCIA MUTIHAC

VASILE DAVID

# CHEMOMETRIE

EDITURA UNIVERSITĂȚII DIN BUCUREȘTI  
- 1997 -



BIBLIOTECA CENTRALĂ  
UNIVERSITARĂ  
București

Cota III 467/166  
Inventar C798/97

**LUCIA MUTIHAC**

**VASILE DAVID**

# **CHEMOMETRIE**

**EDITURA UNIVERSITĂȚII DIN BUCUREȘTI**  
**1997**

<https://biblioteca-digitala.ro> / <https://unibuc.ro>

Referenți științifici: Prof. dr. C-TIN PĂTROESCU  
Conf. dr. ION TĂNASE

BIBLIOTECA CENTRALĂ UNIVERSITARĂ

BUCUREȘTI

COTA III 1167/166

Contribuția autorilor la elaborarea lucrării:

LUCIA MUTIAC

Capitolele I–III, VI–VIII și IX

VASILE DAVID

Capitolele IV, V și IX

**B.C.U. București**



C 00798 97

Toate drepturile sunt rezervate Editurii Universității București.  
Orice reproducere sau traducere, fie și parțială, precum și  
contrafacerile de orice tip intră sub incidența legii penale.

# CUPRINS

	INTRODUCERE	
Cap. I	CHEMOMETRIA ȘI CHIMIA ANALITICĂ	9
	I.1. Definierea chemometriei	9
	I.2. Metode chemometrice	10
	Bibliografie	12
Cap. II	CRITERII UTILIZATE ÎN EVALUAREA METODELOR ANALITICE	13
	II.1. Tipuri de erori în analiza chimică	13
	II.2. Precizia și corectitudinea în chimia analitică	14
	II.3. Selectivitatea și rezoluția	16
	II.4. Sensibilitatea și limita de detecție	19
	II.5. Informația în chimia analitică	21
	II.6. Amplitudinea semnalului și raportul semnal/zgomot	21
	Bibliografie	22
Cap. III	UTILIZAREA STATISTICII MATEMATICE ÎN CHIMIA ANALITICĂ	24
	III.1. Generalități	24
	III.2. Distribuții de probabilitate	25
	III.2.1. Distribuția variabilelor aleatoare	25
	III.2.2. Funcția de distribuție normală	27
	III.2.3. Funcția de distribuție $\chi^2$	29
	III.2.4. Funcția de distribuție t (Student)	30
	III.2.5. Funcția de distribuție F	32
	III.2.6. Distribuția binomială	32
	III.3. Estimarea parametrilor	32
	III.4. Intervale de încredere	33
	III.4.1. Intervalul de încredere pentru media unei distribuții	34
	III.4.2. Intervalul de încredere pentru dispersia de selecție	35
	III.4.3. Intervalul de încredere pentru parametrii distribuției binomiale	36
	III.5. Testarea ipotezelor	36
	III.6. Analiza de dispersie în chimia analitică	39
	Bibliografie	40

Anexa 3.1	Valorile funcției Laplace	41
Anexa 3.2	Valorile variabilei $\chi^2$	42
Anexa 3.3	Valorile variabilei t (Student)	43
Anexa 3.4	Valorile variabilei F	44
Cap. IV	ANALIZA DE REGRESIE	45
	IV.1 Generalități	45
	IV.2 Tipuri de regresii	45
	IV.3 Etapele analizei de regresie	49
	IV.3.1 Definierea modelului	49
	IV.3.2 Estimarea parametrilor modelului	52
	IV.3.3 Validarea modelului	58
	IV.3.4 Calculul intervalelor de încredere	62
	IV.4 Concluzii	63
	Bibliografie	63
Cap. V	NOȚIUNI DE PRELUCRARE A SEMNALELOR ANALITICE	65
	V.1. Generalități	65
	V.2. Tipuri de semnale	65
	V.3. Eșantionarea semnalelor	68
	V.4. Prelucrarea semnalelor analitice	71
	V.4.1. Aspecte fundamentale ale prelucrării semnalelor	71
	V.4.1.1. Caracterul temporal - frecvențial al semnalelor	72
	V.4.1.2. Contaminarea cu zgomot a semnalelor	82
	V.4.2. Tipuri de zgomot conținute în semnalele analitice	87
	V.4.3. Tehnici de prelucrare a semnalelor analitice	89
	V.4.3.1. Determinarea constantei de timp a instrumentelor analitice	89
	V.4.3.2. Determinarea unor parametri fundamentali ai semnalelor	94
	V.4.3.3. Tehnici de îmbunătățire a raportului semnal/zgomot	98
	V.4.3.3.1. Tehnici instrumentale	98
	V.4.3.3.1.1. Scurt istoric al utilizării tehnicilor de transformare în chimie	99

	V.4.3.3.1.2. Sisteme analitice cu transformată Hadamard	100
	V.4.3.3.2. Tehnici aplicate după achiziționarea semnalelor	103
	V.5. Concluzii	105
	Bibliografie	105
<b>Anexa 5.1</b>	<b>Analiza spectrală a semnalelor deterministe în domeniu de timp continuu</b>	<b>107</b>
	A5.1.1. Funcții ortogonale	107
	A5.1.2. Serii Fourier	108
	A5.1.3. Transformata Fourier continuă (CFT)	111
<b>Anexa 5.2</b>	<b>Noțiuni de simulare a proceselor aleatoare</b>	<b>113</b>
<b>Anexa 5.3</b>	<b>Ansamblu simulat reprezentat prin 16 funcții de probă</b>	<b>117</b>
<b>Anexa 5.4</b>	<b>Planuri experimentale de cântărire</b>	<b>119</b>
	A5.4.1. Cazul clasic (al cântăririlor secvențiale)	119
	A5.4.2. Cazul cântăririlor în grup punând obiectele de cântărit pe ambele talere	120
	A5.4.3. Cazul cântăririlor în grup punând obiectele de cântărit pe un singur taler	121
	A.5.4.4. Generalizare	122
<b>Cap VI</b>	<b>METODOLOGIA SUPRAFETEI RĂSPUNSULUI</b>	<b>123</b>
	VI.1. Suprafața de răspuns	123
	VI.2. Experimente factoriale	126
	VI.2.1. Experiment factorial complet la două nivele (Plan experimental 2 <sup>n</sup> )	127
	VI.2.2. Experimente factoriale fracționate	132
	VI.2.3. Experiment factorial la trei nivele	132
	Bibliografie	133
<b>Cap. VII</b>	<b>OPTIMIZAREA ÎN CHIMIA ANALITICĂ</b>	<b>134</b>
	VII.1. Generalități	134
	VII.2. Reprezentarea matematică a sistemului de optimizat	134
	VII.3. Funcția de răspuns și restricțiile	135
	VII.4. Strategia optimizării	136
	Bibliografie	139

88	VIII	METODE DE EVALUARE A OPTIMULUI UNEI FUNCȚII OBIECTIV	130
		VIII.1. Metode analitice	140
		VIII.2. Metode numerice	140
		VIII.3. Metode mixte	140
		Bibliografie	141
	IX	ELEMENTE DE TEORIA INFORMAȚIEI UTILIZATE ÎN CHIMIA ANALITICĂ	142
		IX.1. Generalități	142
		IX.2. Procesele analitice ca sisteme de transmisie a informației	142
		IX.3. Entropie și informație	143
		IX.4. Aplicații ale teoriei informației în chimia analitică	146
		IX.4.1. Cuantificarea informației obținute în experimentele de identificare	146
		IX.4.2. Cuantificarea informației obținute în experimentele de determinare	147
		IX.4.3. Selectarea metodelor analitice pe baza informației furnizate	148
		IX.5. Alte mărimi informaționale utilizate în caracterizarea proceselor analitice	151
		IX.5.1. Redundanța materialului supus cercetării analitice	151
		IX.5.2. Puterea informațională a unui procedeu analitic	152
		IX.5.3. Debitul și costul informației	153
		Bibliografie	153
a)	X	REȚELE NEURONALE	154
		X.1. Introducere	154
		X.2. Elemente de teoria modelării	155
		X.3. Modele de rețele neuronale	158
		X.3.1. Rețele neuronale formale cu instruire supervizată (memorii asociative)	158
		X.3.2. Rețele neuronale formale cu instruire nesupervizată (clasificatori vectoriali)	161
		X.4. Modele matematice, reguli de propagare și algoritmi de instruire a rețelelor neuronale formale	163



X.4.1. Modele matematice generale	163
X.4.2. Algoritmi de instruire generalizați	164
X.4.2.1. Algoritmul de instruire exhaustivă	164
X.4.2.2. Algoritmul gradientului	167
X.4.2.3. Algoritmul lui Newton	168
X.5. Aplicații ale rețelelor neuronale	169
X.5.1. Concepte fundamentale	169
X.5.2. Clase de aplicații specifice ale rețelelor neuronale	171
X.5.3. Abordarea modelării neuronale	175
X.5.4. Aplicații ale rețelelor neuronale în chimie	176
X.5.5. Concluzii finale privind aplicațiile rețelelor neuronale în chimie	186
Bibliografie	187

## INTRODUCERE

Chemometria ca disciplină are ca obiectiv introducerea și dezvoltarea metodelor matematice moderne de analiză, sinteză și prelucrare a datelor chimice. Aplicarea tehnicilor specifice tehnologiei informației în optimizarea proceselor analitice și evidențierea direcțiilor fundamentale de evoluție a metodelor computaționale constituie alt aspect al obiectivelor chemometriei.

Chimia analitică este bogată în informații și este logic ca dezvoltarea sa să beneficieze de metodele informaticii. Această problemă a adaptării metodelor informaticii la cerințele necesare analiștilor constituie de asemenea un obiectiv principal al chemometriei. Chimia analitică înglobează analiza chimică precum și ansamblul metodelor de interpretare al rezultatelor, este deci o metodologie care se bazează pe un ansamblu de metode chimice, fizice și fizico-chimice, numite metode de analiză chimică și metode informatice, statistice și matematice adaptate cerințelor specifice chimiștilor numite metode chemometrice. Obiectivul lor constă în rezolvarea problemelor prin intermediul procesării informației chimice.

Lucrarea de față tratează o parte dintre numeroasele posibilități de aplicare a metodelor chemometrice în chimia analitică.

În primul capitol se face o introducere generală privind implicarea chemometriei în chimia analitică. În capitolul al doilea sunt prezentate principalele criterii utilizate în evaluarea metodelor analitice și anume : precizia, corectitudinea, selectivitatea, rezoluția, sensibilitatea, limita de detecție, informația precum și amplitudinea semnalului și raportul semnal - zgomot. Capitolele III-VI cuprind într-un mod sintetic câteva dintre cele mai importante metode statistico-matematice utilizate în chimia analitică. Metodele statistico-matematice sunt prezente atât în perfecționarea unor metode de analiză existente, cât și la elaborarea de noi metode de analiză. Pentru selecționarea unei metode anume de analiză din mai multe metode se utilizează testele statistice privind compararea caracteristicilor de performanță ale metodelor. Planificarea experimentelor, analiza de dispersie și analiza funcțională sunt metode de statistică matematică speciale utilizate în chimia analitică pentru obținerea informației chimice. În continuare sunt prezentate cele mai importante aspecte ale procesului de optimizare în cazul unor sisteme analitice. Lucrarea de față prezintă succint domeniile chemometriei cunoscute de mai mult timp și aplicate astăzi frecvent la interpretarea datelor analitice și la optimizarea metodelor de analiză, și acordă mai multă atenție unor domenii mai puțin cunoscute, cum sunt prelucrarea semnalelor analitice (cap.V) și utilizarea rețelelor neuronale în chimie (cap.X). Aplicațiile rețelelor neuronale în chimie, oferă noi metode de cercetare în modelarea, prelucrarea și interpretarea datelor experimentale.

Lucrarea se adresează studenților și chimiștilor analiști preocupați de introducerea metodelor matematice moderne în prelucrarea datelor chimice.

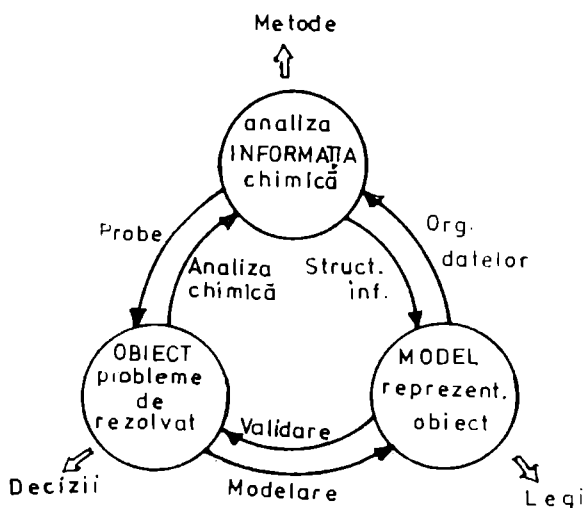
Lucia Mutihac

# I. CHEMOMETRIA ȘI CHIMIA ANALITICĂ

## I.1. Definirea chemometriei

Chimia analitică, ca mod de investigare a unei științe experimentale, este considerată a aparține unei cercetări aplicate în care scopul este acela de a găsi o nouă metodă de analiză, de informare, de achiziționare sau de îmbunătățire a celor existente. În majoritatea cazurilor chimia analitică nu poate descoperi singură o lege ci ea trebuie să dispună de alte informații ca de exemplu fizice sau biologice.

Specificitatea chimiei analitice se poate înțelege dacă aceasta este abordată ca o știință experimentală. Acest aspect este schematizat în jurul a 3 poli prezentați în figura 1 care constituie trinitatea analitică: obiectul studiat (o problemă), reprezentarea sa (un model) și informația cercetată.



*Fig. 1.1. Trinitatea analitică*

Deci plecând de la o problemă, prima etapă constă în reprezentarea sa sub formă de model, în general matematic. Acest model servește la definirea naturii informației cercetate și organizează banca de date pe un plan teoretic. Originalitatea sa constă în informația cu caracterul chimic, ceea ce reprezintă o informație asupra materiei care cere o cunoaștere a concepțiilor de bază ale chimiei și care se obțin cu ajutorul metodelor de măsurare chimică, fizică și

fizico-chimică puse în lucru de către analiza chimică. Rezultatele acestor analize vor fi structurate în acord cu modelul prestabilit și evident se va lua decizia potrivită.

În concluzie, chimia analitică înglobează analiza chimică precum și ansamblul metodelor de interpretare a rezultatelor în urma cărora se iau decizii cu scopul de a rezolva probleme propuse ca subiecte de studiu. Astfel, evaluarea sau controlul poluării unui râu în conținutul de Pb este o problemă pasă chimiei analitice, și în mod distinct, dozajul de Pb într-o probă dată este relevat de analiza chimică.

Chimia analitică este deci o metodologie care se bazează pe un ansamblu de metode chimice, fizice și fizico-chimice, numite *metode de analiză chimică* și metode informatice, statistice și matematice adaptate nevoilor specifice chimiștilor numite *metode chemometrice*. Ele au ca obiectiv rezolvarea problemelor prin intermediul procesării informației chimice. De asemenea, rolul lor este de a optimiza și îmbunătăți colecția de date utile, să se întrebe asupra conținutului lor, asupra validității lor și asupra semnificației lor.

## 1.2. Metode chemometrice

Începând cu anii 1970, atenția mai multor grupuri de cercetători [1-3] a fost îndreptată spre elaborarea unor ansambluri de metode al căror obiectiv încă de la apariție, se poate defini astfel:

- permit o mai bună prelucrare a rezultatelor obținute de analiști în scopul extragerii informației celei mai pertinente posibil;
- creșterea calității măsurătorilor chimice prin optimizarea condițiilor de colectare ale acestora.

S-au numit metode chemometrice. Chemometria este deci o disciplină încă tânără dar care a reușit să se impună serios în chimie. În 1975 s-a creat Chemometrics Society iar în fiecare an se ține regulat un congres intitulat "Chemometrics in Analytical Chemistry". Există, de asemenea două reviste "Journal of Chemometrics" și "Journal of Intelligent Laboratory Systems" unde sistematic sunt prezentate informațiile importante privind acest fenomen.

Literatura de specialitate [4,5] arată deci, că metodele chemometrice se referă la metode matematice, statistice și informatice cunoscute dar adaptate la specificul informației chimice. Se disting 4 domenii semnificative de investigație în ceea ce privește posibilitățile de aplicare a metodelor chemometrice și anume:

- a) achiziționarea semnalelor la ieșirea din aparatele de măsură, stocarea lor și aranjarea lor într-o formă corespunzătoare;
- b) transformarea acestor semnale, expresia lor sub forma unor rezultate prin etalonare și validarea lor în termeni de precizie;
- c) interpretarea rezultatelor în vederea rezolvării problemelor puse analiștilor;
- d) gestionarea informațiilor din laboratoarele de analiză.

Ca istoric, apariția chemometriei și dezvoltarea sa face parte integrantă din evoluția chimiei. La momentul oportun, chemometria a propus soluții noi bazate pe un fundament matematic, la problemele puse chimiștilor și în principal, analiștilor.

Dezvoltarea chemometriei în măsura în care aceasta afectează chiar concepția de instrument de măsură dovedește o tentativă de a defini specificitatea chimiei analitice ce influențează în consecință utilizarea aparaturii, natura probelor precum și modul de funcționare al laboratorului. Într-adevăr, calculatorul rămâne util cercetării și dezvoltării, iar un obiectiv permanent îl constituie integrarea aparatelor la noile tehnici chemometrice.

Tehnologia informatică a început să se implanteze în chimia analitică în jurul anilor 1960, în particular în tot ceea ce înseamnă instrumente de laborator. În consecință, majoritatea aparatelor de măsură disponibile astăzi conțin microprocesoare. Acestea vin să înlocuiască componentele electronice cablate ale aparatelor. Acestea cresc fiabilitatea instrumentelor și îmbunătățesc funcțiile de filtrare ale semnalului, de asemenea și calculul statistic sau reprezentarea grafică.

Dacă astăzi se vorbește de un instrument inteligent, această inteligență provine de la metodele informatice instalate în aparate (software specializat) și nu de la microprocesorul însuși.

Trăim într-o societate a informației, chimia analitică este bogată în informații și este logic ca dezvoltarea sa să beneficieze de metodele informaticii. Această problemă a adaptării metodelor informaticii la cerințele necesare analiștilor constituie unul dintre obiectivele principale ale chemometriei.

## ***Bibliografie***

1. D. L. Massart, A. Dijkstra și L. Kaufman, **Evaluation and Optimization of Laboratory Methods and Analytical Procedures**, Elsevier, Amsterdam, 1978.
2. G. Kateman și P. Pipers, **Quality Control in Analytical Chemistry**, John Wiley, New York, 1981.
3. M. A. Sharaf, D. L. Illman și B. R. Kowalski, **Chemometrics**, John Wiley, New York, 1986.
4. D. L. Massart și colab., **Chemometrics**, John Wiley, New York, 1988.
5. M. Feinberg, ***L'Actualité Chimique***, **3**, 191, 1991.

## II. CRITERII UTILIZATE ÎN EVALUAREA METODELOR ANALITICE

### II.1. Tipuri de erori în analiza chimică

Scopul determinărilor în chimia analitică îl constituie estimarea valorii adevărate. Dar datorită faptului că un proces analitic, în general, este supus unor surse de zgomot (perturbații) inerente, rezultatele măsurătorilor și apoi evident rezultatele de analiză vor fi însoțite de erori. Cu cât concentrația componentului de determinat este mai mică, cu atât probabilitatea apariției erorilor în determinare este mai mare.

În concluzie, pentru ridicarea performanțelor unei determinări analitice este necesară îndepărtarea surselor de erori, lucru posibil evident prin cunoașterea naturii acestora.

Eroarea de determinare analitică este dată de diferența dintre rezultatul determinării și valoarea adevărată a determinării. Dacă se notează cu  $\hat{x}_o$  valoarea adevărată a conținutului componentului de determinat, iar cu  $x$  rezultatul analizei, eroarea de determinare se poate scrie:

$$\Delta x = x - \hat{x}_o \quad (2.1)$$

Eroarea relativă de determinare are următoarea expresie:

$$(\Delta x)_r = (x - \hat{x}_o) / \hat{x}_o \quad (2.2)$$

iar relativ procentual în raport cu valoarea adevărată a conținutului, eroarea de determinare se va scrie astfel:

$$(\Delta x)\% = \frac{(x - \hat{x}_o)}{\hat{x}_o} \cdot 100 \quad (2.3)$$

Clasificarea erorilor de determinare în mod frecvent cuprinde [1]: erori întâmplătoare, sistematice și grosolane. Între erorile întâmplătoare și cele sistematice nu există o diferențiere rigidă și anume în funcție de modul de examinare a erorilor, erorile sistematice pot trece în erori întâmplătoare și invers. Erorile întâmplătoare se pot evidenția atunci când la repetarea unui experiment de determinare se obține de fiecare dată alt rezultat. În concluzie, eroarea întâmplătoare este dată de diferența dintre rezultatele obținute și

valoarea medie a acestora la efectuarea repetată a unui experiment. Erorile întâmplătoare nu pot fi prevăzute individual, dar mulțimea lor poate fi descrisă prin funcții de repartiție bine determinate.

**Erorile sistematice** sunt acele erori de analiză care prezintă un caracter constant, adică distorsionează rezultatele determinărilor într-un singur sens cu o cantitate constantă față de valoarea adevărată. De asemenea este necesar, în funcție de stabilitatea sistemului analitic în timp, să fie luat în considerare și parametrul timp la clasificarea erorilor sistematice și anume: erori sistematice variabile în timp și erori sistematice constante în timp. Erorile sistematice pot avea diverse origini, ca de exemplu folosirea incorectă a instrumentelor de măsură datorită unei funcții de calibrare incorecte, folosirea unor reactivi de puritate necorespunzătoare, fenomene de interferență, etc.

Într-un alt mod, erorile sistematice se pot clasifica în dependente și independente de conținutul componentului de determinat.

**Eroarea grosolană** este definită ca fiind eroarea întâmplătoare intermitentă, care provoacă o abatere mare a rezultatelor de la valoarea adevărată. Ele se identifică ușor, deosebindu-se semnificativ de restul rezultatelor.

## II.2. Precizia și corectitudinea în chimia analitică

Precizia și corectitudinea reprezintă unele dintre cele mai importante criterii de caracterizare a sistemelor analitice precum și a rezultatelor lor.

1. **Precizia** (fidelitatea) este dată de împrăștierea rezultatelor unei mulțimi de determinări în jurul mediei lor și este evaluată de abaterea standard  $s_x$ . Dacă  $x_1, x_2, \dots, x_n$  este mulțimea rezultatelor,  $n$  este numărul rezultatelor și  $\bar{x}$  este media rezultatelor, atunci dispersia de selecție  $s_x^2$  este:

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad (2.4)$$

În cazul în care seria are valori care se repetă, expresia 2.4 va avea forma de mai jos:

$$s^2 = \frac{\sum_{i=1}^n n_i (x_i - \bar{x})^2}{N} \quad (2.5)$$



unde  $n_i$  este numărul valorilor  $x_i$ ; iar  $N = \sum_{i=1}^n n_i$ .

Abaterea standard sau abaterea medie pătratică de selecție  $s_x$  este rădăcina pătrată a dispersiei de selecție  $s_x^2$ :

$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}, \quad (2.6)$$

respectiv:

$$s_x = \sqrt{\frac{\sum_{i=1}^n n_i (x_i - \bar{x})^2}{N}}. \quad (2.7)$$

Rădăcina pătrată a dispersiei de selecție modificată este abaterea standard de selecție modificată:

$$s = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (2.8)$$

Aceasta este abaterea standard folosită în practica statistică pentru estimarea abaterii standard a întregii populații [2].

În baza definiției abaterii standard rezultă că, *abaterea standard de estimare* pentru rezultatele medii a  $n$  determinări este:

$$s_x = \frac{s_x}{\sqrt{n}}, \quad (2.9)$$

*abaterea standard relativă de estimare* este  $\frac{s_x}{\bar{x}}$ ; *abaterea standard relativă*

*procentuală de estimare (coeficientul de variație)* este  $\frac{s_x}{\bar{x}} \cdot 100$ , iar *deviația*

*medie absolută de estimare* este  $\frac{\sum_{i=1}^n |\bar{x} - x_i|}{n}$ .

Pentru a elimina orice eroare de interpretare, și de asemenea, pentru ca abaterea standard să aibă o semnificație tehnică, este necesară menționarea condițiilor în care a fost obținută mulțimea de date  $x_1, x_2, \dots, x_n$ .

În literatura de specialitate, pentru noțiunea de precizie se mai utilizează și următoarea nomenclatură: reproductibilitate, repetabilitate, împrăștiere, variabilitate, imprecizie [3-6].

Deci, evaluarea preciziei se face prin intermediul abaterii standard,  $s_x$ , cu specificarea obligatorie a condițiilor și modului de obținere a mulțimii de date  $x_1, x_2, \dots, x_n$ .

Precizia este un criteriu de măsurare a reproductibilității măsurătorii. Dacă diferența între măsurătorile rezultate este mare, precizia este mică. O precizie bună nu înseamnă că și măsurătoarea este exactă deoarece este posibil să se comită sistematic aceeași greșeală.

**2. Corectitudinea unei metode de analiză** este un alt criteriu de performanță utilizat pentru alegerea optimă a unei metode de analiză. Corectitudinea este dată de deviația rezultatelor simple sau a rezultatelor medii de la valoarea adevărată. Pentru evaluarea corectitudinii este necesar ca probele standard să aibă compoziție similară cu cea a probei de analizat, deoarece în majoritatea cazurilor corectitudinea este influențată de compoziția probei [7-9].

### II.3. Selectivitatea și rezoluția

Problema definirii și evaluării selectivității în chimia analitică a fost abordată în literatura de specialitate de mulți autori [10-14].

Selectivitatea este dată de posibilitatea unei metode de a detecta sau determina unul sau mai mulți componenți din probă în prezența altor componenți. Ea poate fi evaluată cu ajutorul unei funcții de răspuns, adică al funcției de dependență dintre răspuns ( $y$ ) și compoziția probei de analizat. În cazul unei funcții de răspuns liniare în raport cu conținuturile componenților se poate scrie următoarea relație:

$$y = a_1c_1 + a_2c_2 + \dots + a_nc_n \quad (2.10)$$

unde  $c_1$  este conținutul componentului de analizat iar  $c_2, c_3, \dots, c_n$  reprezintă conținuturile celorlalți componenți. Metoda este selectivă dacă  $a_1$  este singurul coeficient diferit de zero.

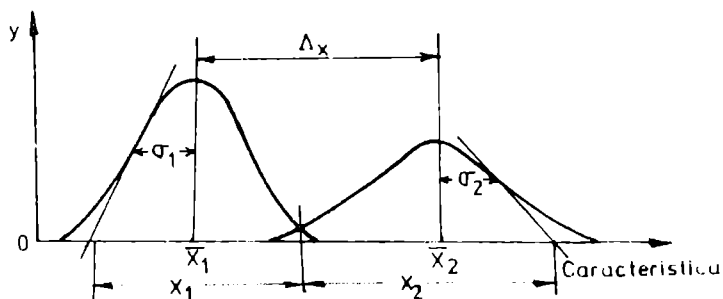
Dacă se consideră cazul unui sistem cu  $m$  componenți și funcția de răspuns liniară în raport cu conținuturile componenților, sistemul de  $m$  ecuații se poate scrie:



Măsura gradului de suprapunere a două semnale, de orice formă, este evaluată prin integrala:

$$A = \int_{-\infty}^{+\infty} y_1(x)y_2(x)dx \quad (2.15)$$

în care  $y_1(x)$  și  $y_2(x)$  reprezintă expresiile funcțiilor celor două semnale. Când cele două semnale sunt separate,  $A = 0$ , iar când se suprapun  $A > 0$ , cu cât gradul de suprapunere este mai mare, cu atât parametrul  $A$  are o valoare mai mare.



*Fig.2.1. Reprezentarea schematică a rezoluției*

Fenomenul de *interinfluență* este dat de modificarea amplitudinii semnalului în funcție de valorile conținuturilor altor componente din probă. Acest fenomen se întâlnește adesea în cazul spectroscopiei atomice de emisie și absorbție.

În chimia analitică, îmbunătățirea selectivității sistemelor analitice se poate realiza fie prin perfecționarea instrumentației analitice, fie prin prelucrarea semnalului.

## II.4. Sensibilitatea și limita de detecție

Adesea se spune despre o metodă că este foarte sensibilă atunci când limita de detecție are o valoare foarte mică. Sensibilitatea este definită ca fiind derivata funcției de răspuns  $f(\hat{x})$  în raport cu conținutul de determinat,

și anume  $\frac{\partial f(\hat{x})}{\partial \hat{x}}$ .

În cazul unei funcții de răspuns liniare, sensibilitatea este egală cu panta funcției, iar pentru o funcție de răspuns neliniară, sensibilitatea se modifică în raport cu valoarea conținutului de determinat.

Caracterizarea capacității de detecție a unei metode de analiză constituie una dintre cele mai importante probleme ale chimiei analitice. Bazele acestei probleme au fost puse de Kaiser [15] și Wilson [16], Kaiser fiind acela care introduce noțiunea de *limită de detecție*.

Contribuții însemnate în rezolvarea problemei sensibilității și a limitei de detecție aduc Ehrlich și Gerbatsch [17], Nalimova și colaboratorii [18], Liteanu și Rîcă [19] care fundamentează definirea limitei de detecție pe baze statistice de detecție a semnalelor luând în considerare dependența dintre semnalul analitic și conținutul componentului de determinat.

Semnalul analitic corespunzător limitei de detecție în sensul unui conținut minim sigur detectabil pentru un raport semnal/zgomot  $r_d = 6$ , atunci când detecția se face cu un rezultat simplu, are următoarea expresie:

$$\hat{y}_d = \hat{y}_0 + 6\sigma_y \quad (2.16)$$

unde  $\sigma_y$  este *abaterea standard*, iar pentru media a  $n$  determinări simple avem expresia de mai jos:

$$(\hat{y}_d)_n = \hat{y}_0 + 6 \frac{\sigma_y}{\sqrt{n}} \quad (2.17)$$

În cazul unei dependențe liniare între semnal și conținut, este satisfăcută următoarea relație:

$$\hat{y} = \hat{y}_0 + b\hat{c}_d \quad (2.18)$$

$$(\hat{y}_d)_n = \hat{y}_0 + r_d \frac{\sigma_y}{\sqrt{n}} \quad (2.19)$$

În concluzie, între limita de detecție corespunzătoare rezultatelor medii a  $n$  rezultate simple și limita de detecție pentru rezultate simple, este relația următoare:

$$(\hat{c}_d)_n = \frac{\hat{c}_d}{\sqrt{n}} \quad (2.20)$$

În figura 2.2 este prezentată semnificația limitei de detecție în sensul unui conținut minim sigur detectabil.

Liteanu și Florea [20] au elaborat o metodă frecventometrică cu ajutorul căreia se poate evalua riguros limita de detecție în sensul unui conținut minim detectabil. Long și Winefordner [21] tratează metodele de determinare a limitei de detecție în funcție de erorile de măsură ale determinărilor analitice.

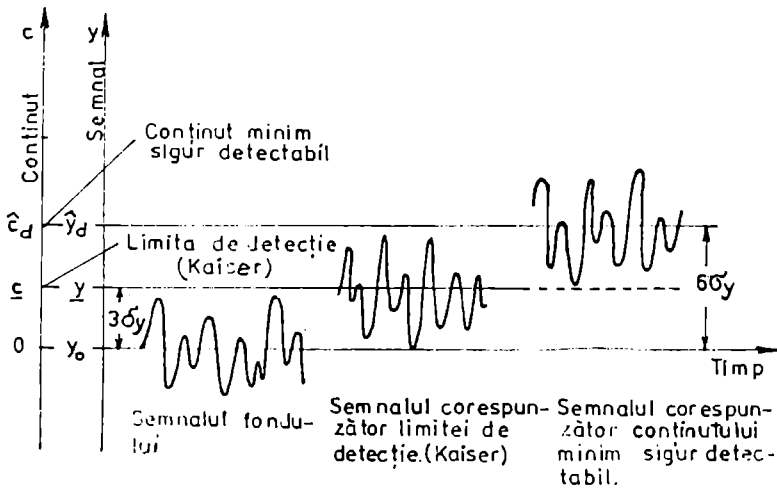


Fig. 2.2. Limita de detecție a unui conținut minim sigur detectabil

## II.5. Informația în chimia analitică

Utilizarea teoriei informației în chimia analitică prezintă un real interes, mai ales în cazul proceselor analitice de măsură, constituind una dintre direcțiile cele mai moderne și eficiente. Aplicarea conceptelor de teoria informației în studiul proceselor de măsură analitice este prezentată detaliat în capitolul IX.

## II.6. Amplitudinea semnalului și raportul semnal/zgomot

În chimia analitică, rezultatul unei măsurători poate fi considerat suma următoarelor semnale: semnalul fondului, semnalul dat de componentul de determinat și zgomotul ce însoțește orice proces de măsură. Dacă se reprezintă ipotetic o funcție de răspuns pentru un sistem analitic, aceasta are forma  $\hat{y} = \hat{y}_0 + f(c)$ , în care  $\hat{y}_0$  reprezintă semnalul fondului pentru o valoare  $\hat{c}_0$  a conținutului componentului de determinat care produce un semnal egal cu semnalul fondului. În acest caz, valoarea adevărată pentru semnalul  $f(\hat{c}_i)$  dat de conținutul  $\hat{c}_i$  pentru componentul de determinat este data de expresia:

$$f(\hat{c}_i) = \Delta\hat{y}_i = \hat{y}_i - \hat{y}_0 \quad (2.21)$$

În cazul unei distribuții normale a semnalului  $\Delta y_i$ , raportul semnal/zgomot corespunzător unei concentrații  $c_i$  are următoarea expresie:

$$\frac{S}{N} = \frac{\hat{y}_i - \hat{y}_0}{\sigma_{\Delta y_i}} \quad (2.22)$$

în care  $\sigma_{\Delta y_i}$  este abaterea standard. Raportul semnal/zgomot este considerat caracteristică de performanță în domeniul conținuturilor mici; în domeniul conținuturilor mari, semnificativ este inversul raportului semnal/zgomot, adică:

$$\frac{N}{S} = \frac{\sigma_{\Delta y_i}}{\hat{y}_i - \hat{y}_0} \quad (2.23)$$

care reprezintă abaterea standard relativă pentru rezultatul măsurării semnalului conținutului componentului de determinat.

### ***Bibliografie***

1. C. Liteanu și I. Rică, **Optimizarea proceselor analitice**, Ed. Academiei Române, București, 1985.
2. D. Ceaușescu, **Utilizarea statisticii matematice în chimia analitică**, Ed. Tehnică, București, 1982.
3. C. Liteanu și I. Rică, **Teoria și metodologia statistică a analizei urmelor**, Ed. Scrisul Românesc, Craiova, 1979.
4. D. L. Massart, A. Dijkstra și L. Kaufman, **Evaluation and Optimization of Laboratory Methods and Analytical Procedures**, Elsevier, Amsterdam, 1978.
5. A. L. Wilson, *Talanta*, **17**, 31, 1970.
6. W. J. Younden și E. H. Steiner, **Statistical Manual of the Association of Official Analytical Chemists**, The Association of Official Analytical Chemists, Washington DC, 1975.
7. E. F. McFarren, R. J. Lishka și J. H. Parker, *Anal. Chem.*, **42**, 358 1970
8. K. Eckschlager, *Anal. Chem.*, **44**, 878, 1972.
9. K. Eckschlager, *Coll. Czech. Chem. Comm.*, **39**, 1426, 1974.
10. R. Belcher, *Talanta*, **12**, 129, 1965; *Talanta*, **23**, 883, 1976.
11. R. Belcher și D. Betteridge, *Talanta*, **13**, 535, 1966.
12. H. Kaiser, *Z. Analyt. Chem.*, **260**, 252, 1972.
13. A. L. Wilson, *Talanta*, **12**, 701, 1965; *Talanta*, **21**, 1109, 1974.
14. I. Pszonicki, *Talanta*, **24**, 613, 1977; *Chem. Anal.*, **25**, 439, 1980.



15. H. Kaiser, *Optik*, **21**, 309, 1964; *Spectrochim. Acta*, **3**, 40, 1957; *Z. Analyt. Chem.*, **209**, 1, 1965; *Pure Appl. Chem.*, **34**, 35, 1972.
16. A. L. Wilson, *Analyst*, **86**, 72, 1961.
17. G. Ehrlich și R. Gherbatsch, *Z. Analyt. Chem.*, **220**, 260, 1966.
18. V. V. Nalimon, V. V. Nadler și N. P. Menshova, *Zavodskaya Lab.*, **27**, 861, 1961.
19. C. Liteanu și I. Rîcă, *Microchim. Acta*, **745**, 1973; *Pure Appl. Chem.*, **44**, 535, 1975; *Microchim. Acta*, 311, 1975.
20. C. Liteanu și I. Florea, *Microchim. Acta*, 983, 1966.
21. G. L. Long și J. D. Winefordner, *Anal. Chem.*, **55**, 712 A, 1983.

# III. UTILIZAREA STATISTICII MATEMATICE ÎN CHIMIA ANALITICĂ

## III.1. Generalități

Sistemele analitice, în sensul definiției conceptului de sistem formulată sintetic de Ludwing van Bertalanffy (1956) considerat fondatorul teoriei sistemelor, și anume "un complex de elemente aflate în interacțiune", sunt sisteme informaționale, rezultatele funcționării lor fiind informațiile obținute asupra compoziției calitative și cantitative a materialului cercetat.

În general, analiza chimică a unei probe are ca scop imediat obținerea unei informații chimice, care poate fi cantitativă, ca de exemplu concentrația unui constituent, sau calitativă, specificând prezența sau absența unui radical într-o moleculă. Numeroase lucrări de specialitate descriu etapele unei analize chimice. Datorită chemometriei se constată că folosirea unui ansamblu de tehnici simple, permite obținerea unui răspuns corect prin interpretarea rezultatelor cu ajutorul modelelor statistice.

În chimia analitică, metodele statisticii matematice sunt utilizate în următoarele direcții: a) obținerea, prelucrarea și prezentarea rezultatelor determinărilor analitice; b) evaluarea caracteristicilor de performanță ale instrumentelor și metodelor de analiză; c) elaborarea de noi metode de analiză; d) optimizarea proceselor analitice.

În cazul rezultatelor unor măsurători, ca de exemplu: spectrele, cromatogramele, polarogramele de diverse tipuri, pentru prelucrarea lor se utilizează metode speciale ale statisticii matematice, și anume: netezirea semnalelor, derivarea semnalelor, autocorelația și intercorelația semnalelor, descompunerea în componente a semnalelor, filtrarea semnalelor. De asemenea, pentru calibrarea instrumentelor analitice sunt necesare metode matematice de estimare a funcțiilor de calibrare liniare sau neliniare, de una sau mai multe variabile. Planificarea experimentelor, analiza de dispersie și analiza funcțională sunt metode de statistică matematică speciale utilizate de asemenea în chimia analitică pentru obținerea informației chimice.

Pentru selecționarea unei metode anume de analiză din mai multe metode posibile, se utilizează testele statistice privind compararea caracteristicilor de performanță ale metodelor. Metodele statistico-matematice sunt prezente atât în perfecționarea unor metode de analiză existente, cât și la elaborarea de noi metode de analiză.

În continuare vom prezenta câteva dintre cele mai importante metode statistico-matematice utilizate în chimia analitică.

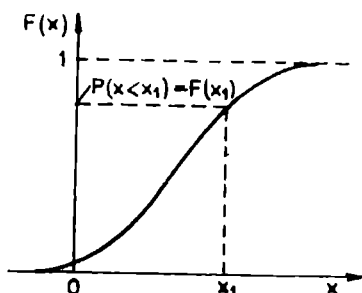
## III.2. Distribuții de probabilitate

Distribuțiile de probabilitate sunt adesea utilizate în prelucrarea rezultatelor experimentale în scopul obținerii unei cantități maxime de informație. Utilizarea distribuțiilor de probabilitate la prelucrarea datelor analitice se face prezentă în:

- estimări de parametri,
- evaluări de intervale de încredere în legătură cu parametrii estimați,
- verificări de ipoteze în legătură cu parametrii estimați.

### III.2.1. Distribuția variabilelor aleatoare

Dacă se repetă un experiment analitic, rezultatele acestuia alcătuiesc o mulțime ce este descrisă de către o funcție de distribuție  $F$ . Această funcție de distribuție măsoară probabilitatea ca variabila aleatoare  $x$  să fie mai mică decât o anumită valoare.



*Fig. 3.1. Funcția de distribuție a variabilelor întâmplătoare*

Probabilitatea ca variabila  $x$  să fie mai mică decât un nivel  $x_1$  este:

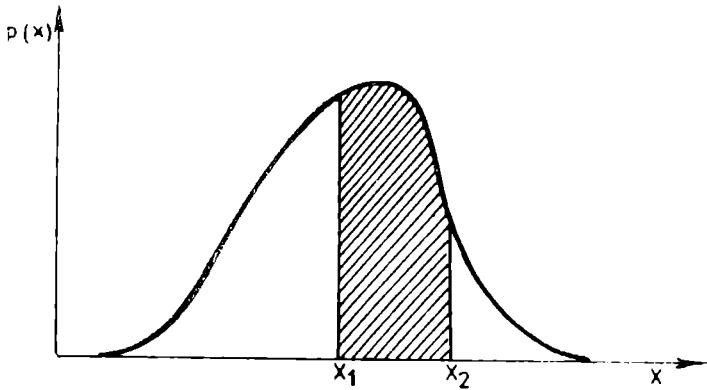
$$P(x < x_1) = F(x_1) \quad (3.1)$$

Funcția de distribuție are o valoare cuprinsă în domeniul  $[0, 1]$

$$0 \leq F(x) \leq 1 \quad (3.2)$$

Prin derivarea funcției de distribuție  $F$ , se obține funcția de probabilitate corespunzătoare  $p$  (fig. 3.2.):

$$p(x) = \frac{dF(x)}{dx} \quad (3.3)$$



*Fig. 3.2. Funcția densității de probabilitate pentru variabile aleatoare*

Probabilitatea ca variabila aleatoare să ia o valoare situată în domeniul  $(x_1, x_2)$  este dată de integrala:

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} p(x) dx \quad (3.4)$$

Funcția de distribuție este caracterizată de valoarea medie  $m$  și dispersia  $\sigma^2$ , adică:

$$m = \int_{-\infty}^{+\infty} xp(x) dx = M(x) \quad (3.5)$$

$$\sigma^2 = \int_{-\infty}^{+\infty} [x - m(x)]^2 p(x) dx \quad (3.6)$$

Dispersia variabilei aleatoare  $x$ , este momentul centrat de ordinul 2

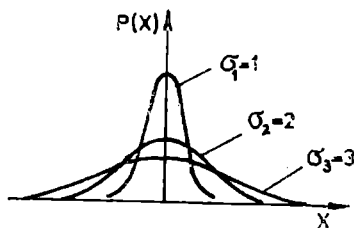
### III. 2.2. Funcția de distribuție normală

Funcția de distribuție normală în concepția lui Gauss poate fi considerat drept model pentru studiul probabilistic al proceselor de măsură. Funcția de distribuție normală, fundamentală în statistica matematică și în teoria erorilor, are densitatea de probabilitate conform expresiei de mai jos:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right], \quad -\infty < x < \infty \quad (3.7)$$

în care  $x$  este variabila aleatoare;  $m$  este media distribuției (momentul de ordinul unu), iar  $\sigma^2$  este dispersia (momentul centrat de ordinul doi).

Funcția de distribuție normală se notează prin  $N(m, \sigma)$  (fig. 3.3.).



*Fig. 3.3. Curbele unor funcții de densitate de probabilitate normală cu aceeași medie și dispersii diferite.*

Funcția densitate de probabilitate are următoarele proprietăți mai importante:

- admite un maxim pentru  $x = m$ ,
  - este simetrică în raport cu  $m$ ,
  - modificarea parametrului  $m$  determină deplasarea curbei normale de-a lungul absciselor fără să-i afecteze forma,
  - modificarea lui  $\sigma$  conduce la lățirea sau îngustarea curbei fără ca media ( $m$ ) să fie afectată,
  - curba are două puncte de inflexiune corespunzătoare valorilor  $x = \pm \sigma$
- Se disting două teoreme a căror utilitate este importantă pentru multe din aplicațiile distribuției normale:

1. Media a  $n$  valori de selecție aparținând unei funcții de distribuție normală  $N(m, \sigma)$  are la rândul ei o funcție de distribuție normală  $N(m; \sigma/\sqrt{n})$ .

2. Media a  $n$  valori dintr-o selecție, aparținând unei distribuții nenormale, pentru  $n$  număr mare are o funcție de densitate de probabilitate normală sau apropiată de aceasta.

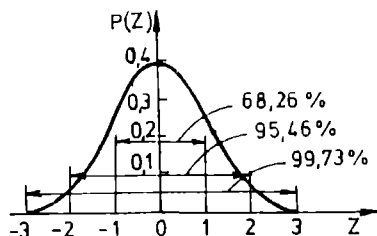
Dacă  $x$  este o variabilă aleatoare cu funcția de distribuție  $N(m, \sigma)$ , atunci variabila aleatoare  $z$  (abaterea normată):

$$z = (x - m) / \sigma \quad (3.8)$$

are o funcție de distribuție  $N(0, 1)$  cu o funcție de densitate de probabilitate normală standard:

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]; \quad -\infty < z < \infty \quad (3.9)$$

cu  $m = 0$  și dispersia  $\sigma^2 = 1$ . Aceasta funcție este prezentată în fig. 3.4.



*Fig. 3.4. Funcția de densitate de probabilitate normală standard*

Funcția de probabilitate a distribuției normale standard este:

$$P(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-z^2/2} dz \quad (3.10)$$

Această funcție este tabelată în anexa31.

Funcția de densitate de probabilitate normală este simetrică în raport cu  $z = 0$ , deci:

$$p(-z) = 1 - p(z) \quad (3.11)$$

Funcția de distribuție normală a fost acceptată drept model probabilistic pentru evaluarea, prelucrarea și cercetarea fluctuațiilor determinărilor analitice. Totuși, în condițiile unui sistem analitic bine conceput și cu o execuție corectă a analizelor, rezultatele experimentale trebuie să aibă o distribuție normală sau apropiată de aceasta. Abaterile pronunțate ale distribuției rezultatelor analitice de la o distribuție normală pune în evidență nereguli ale sistemului analitic studiat.

### III. 2.3. Funcția de distribuție $\chi^2$

Dacă  $x_1, x_2, \dots, x_n$  sunt variabile aleatoare independente cu o funcție de distribuție normală  $N(0, 1)$ , variabila aleatoare  $\chi^2$  prezentată mai jos:

$$\chi^2 = \sum_{i=1}^n x_i^2 ; \quad 0 < \chi^2 < +\infty \quad (3.12)$$

are o distribuție  $\chi^2$  cu  $\nu$  grade de libertate și o funcție de densitate de probabilitate (fig. 3.5.):

$$p(\chi^2) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} (\chi^2)^{(\nu/2)-1} e^{-\chi^2/2} ; \quad \nu > 0; \quad 0 \leq \chi^2 \leq \infty \quad (3.13)$$

Media acestei distribuții este egală cu  $\nu$  iar dispersia egală cu  $2\nu$ .

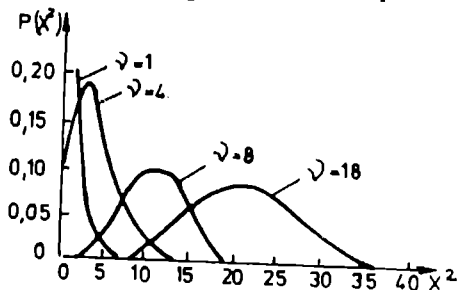


Fig. 3.5. Funcția de densitate de probabilitate a distribuției  $\chi^2$

Funcția de probabilitate a distribuției  $\chi^2$  este:

$$P(\chi^2) = \int_0^{\chi^2} \frac{1}{2^{v/2} \Gamma(v/2)} (\chi^2)^{(v/2)-1} e^{-\chi^2/2} d(\chi^2) \quad (3.14)$$

Valorile pentru variabila  $\chi^2$  precum și valorile pentru  $P(\chi^2)$  sunt prezentate în anexa 3.2.

Funcția de distribuție  $\chi^2$  poate fi utilizată pentru descrierea unei funcții densitate de probabilitate pentru dispersia de selecție  $s^2$ . Acest aspect permite evaluarea unui interval de încredere pentru dispersie, respectiv abaterea standard. De asemenea, cu ajutorul distribuției  $\chi^2$  se poate construi un test de semnificație pentru compararea a două dispersii de selecție.

Pentru utilizarea distribuției  $\chi^2$  la verificarea ipotezelor referitoare la funcțiile de distribuție ale variabilelor aleatoare se iau în considerare următoarele aspecte. Fiind dată o variabilă aleatoare discretă  $x$  cu o distribuție

$$x = \left| \begin{array}{c} x_1, x_2, \dots, x_m \\ p_1, p_2, \dots, p_m \end{array} \right| \quad (3.15)$$

asupra căreia se fac  $n$  experimente independente, variabila  $x$  luând de  $n_1$  ori valoarea  $x_1$ , de  $n_2$  ori valoarea  $x_2, \dots$ , de  $n_m$  ori valoarea  $x_m$ , astfel încât  $n_1 + n_2 + \dots + n_m = n$ , se poate arăta că variabila:

$$\frac{(n_1 - n p_1)^2}{n p_1} + \frac{(n_2 - n p_2)^2}{n p_2} + \dots + \frac{(n_m - n p_m)^2}{n p_m} \quad (3.16)$$

tinde către o repartiție  $\chi^2$ , cu parametri  $v = n - 1$  grade de libertate când  $n \rightarrow \infty$ .

### III.2.4. Funcția de distribuție t (Student)

Variabila aleatoare  $t$  a funcției de distribuție Student reprezintă raportul a două variabile aleatoare, și anume  $z$ , variabila normală standard cu funcție de distribuție normală  $N(0,1)$  și variabila aleatoare  $\sqrt{\chi^2/v}$ , independentă de prima:



$$t = \frac{z}{\sqrt{\chi^2/v}} = \frac{z}{s/\sigma} = \frac{\bar{x} - m}{\sigma} \frac{1}{s/\sigma} = \frac{\bar{x} - m}{s} \quad (3.17)$$

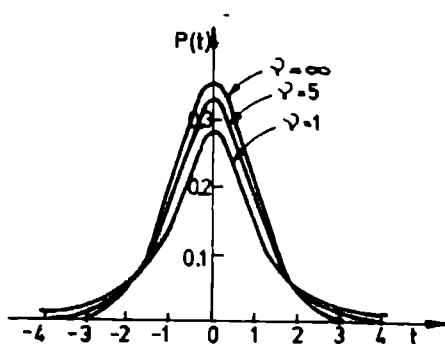
Funcția densitate de probabilitate pentru variabila  $t$  are expresia următoare:

$$p(t) = \frac{1}{\sqrt{\pi v}} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}}, \quad -\infty < t < \infty \quad (3.18)$$

unde  $v$  reprezintă numărul gradelor de libertate asociat cu  $s^2$ . Funcția de probabilitate corespunzătoare funcției densitate de probabilitate pentru variabila  $t$  are expresia:

$$P(t) = \int_{-\infty}^t p(t) dt \quad (3.19)$$

În figura 3.6. sunt prezentate câteva densități de probabilitate  $p$  pentru variabila  $t$ . Când numărul gradelor de libertate  $v \rightarrow \infty$ , distribuția  $t$  tinde către o distribuție normală standard.



**Fig. 3.6.** Densitatea de probabilitate pentru variabila  $t$ .

În anexa 3.3. sunt tabelate valorile variabilei  $t$ .

În general funcția de distribuție  $t$  se folosește pentru evaluarea intervalului de încredere pentru media unei distribuții normale și de asemenea

pentru verificarea ipotezelor statistice pe care le implică analiza chimică pentru selecții mici ( $n < 30$ ).

### III.2.5. Funcția de distribuție F

O variabilă aleatoare F are o funcție de distribuție F cu  $\nu_1, \nu_2$  grade de libertate și o funcție de densitate de probabilitate cu următoarea expresie:

$$P(F) = \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} F^{\left(\frac{\nu_1}{2}-1\right)} \left(1 + \frac{\nu_1}{\nu_2} F\right)^{-\frac{\nu_1 + \nu_2}{2}}, \quad 0 < F < \infty \quad (3.20)$$

Funcția de distribuție F se utilizează la verificarea ipotezelor de semnificație a două dispersii de selecție. Dacă se consideră selecțiile  $x_{11}, x_{12}, \dots, x_{1n}$  și  $x_{21}, x_{22}, \dots, x_{2n}$  efectuate asupra aceleiași distribuții normale sau a doua distribuții normale cu aceeași dispersie, variabila:

$$F = s_1^2 / s_2^2 = (\chi_1^2 / \nu_1) / (\chi_2^2 / \nu_2) \quad (3.21)$$

are o distribuție F cu  $\nu_1 = n_1 - 1$  și  $\nu_2 = n_2 - 1$  grade de libertate. Funcția de distribuție F este prezentată în anexele 3.4 și 3.5.

### III.2.6. Distribuția binomială

Se poate defini funcția de distribuție binomială a unui eveniment cu o probabilitate de apariție p pentru probabilitatea ca evenimentul să apară de x ori în urma a n experimente independente:

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \text{unde } x = 0, 1, \dots, n \quad (3.22)$$

Pentru  $0,3 \leq p \leq 0,70$  și  $n \geq 30$ , distribuția binomială este exprimată bine de către o distribuție normală cu  $m = n p$  și dispersia  $\sigma^2 = n p (1 - p)$ .

### III.3. Estimarea parametrilor

Estimatorul este o regulă de estimare a unor parametri în baza unei selecții, iar estimat, valoarea care se obține într-un anumit caz. Un exemplu

tipic de estimare îl constituie estimarea mediei și dispersiei distribuției unei variabile aleatoare  $x$  cu ajutorul unei selecții  $x_1, x_2, \dots, x_n$  conform relațiilor de mai jos:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (3.23)$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.24)$$

În cazul proceselor analitice sunt numeroase exemple de estimare, cum ar fi: estimarea funcțiilor de calibrare; estimarea parametrilor distribuțiilor rezultatelor și erorilor de analiză, estimarea performanțelor proceselor analitice; estimarea funcției de răspuns a procesului analitic, etc.

Generalizat, problema estimării parametrilor poate fi prezentată astfel: se consideră situația în care între variabilele de interes  $X = |x_1, x_2, \dots, x_m|$  și răspunsul  $y$  există o dependență de tip liniar:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_m x_m + e \quad (3.25)$$

Pentru estimarea parametrilor  $B = |b_1, b_2, \dots, b_m|$  se utilizează o mulțime de  $n$  perechi de date  $X(i) = |x_1(i), x_2(i), \dots, x_m(i)|$ ;  $y(i)$ , generate în acord cu relația:

$$y(i) = BX(i) + e(i), \quad i = 1, 2, \dots, n \quad (3.26)$$

unde  $e(i)$  este zgomotul cu o valoare de așteptare egală cu zero.

Se notează cu  $\hat{B}$  valorile de estimare pentru parametri iar regula de estimare prin:

$$\hat{B} = G(x; y) \quad (3.27)$$

Printre cei mai importanți estimatori [1-4] se pot enumera: estimatori Bayes, estimatorul verosimilității maxime, estimatorul Markov și estimatorul celor mai mici pătrate.

### III.4. Intervale de încredere

Intervalul statistic de încredere ca noțiune a fost utilizat pentru prima dată de către Laplace pentru a mări gradul de semnificație al estimării parametrilor distribuției binomiale. Apoi Neymann a fost cel care a dezvoltat

teoria și terminologia intervalului de încredere. Noțiunea de estimare implică noțiunea de interval de încredere. Deci problema intervalului de încredere apare ori de câte ori este pusă o problemă de estimare.

Dacă se consideră o variabilă aleatoare cu o funcție de densitate de probabilitate ipotetică  $p(x; b)$ , parametrul  $b$  va fi estimat în baza unei selecții:  $x(1), x(2), \dots, x(n)$  cu ajutorul unui estimator:

$$\hat{b} = G[x(1), x(2), \dots, x(n)] \quad (3.28)$$

Valoarea de estimatie  $\hat{b}$  este la rândul ei o variabilă aleatoare a cărei funcție de densitate de probabilitate o notăm cu  $p(\hat{b}; b)$ . Probabilitatea  $P$  ca variabila  $\hat{b}$  să se găsească într-un interval  $b_1 < \hat{b} < b_2$  este dată de integrala:

$$P = P(b_1 < \hat{b} < b_2) = \int_{b_1}^{b_2} p(\hat{b}, b) d\hat{b} \quad (3.29)$$

Mărimea  $P$  este numită și coeficient de încredere. În general, în statistică se pune problema invers, adică: dacă se dă o probabilitate  $P$ , să se stabilească un interval de încredere pentru parametrul existent.

### III.4.1. Intervalul de încredere pentru media unei distribuții

Se pot diferenția două situații:

a) Dispersia distribuției normale este cunoscută. În această situație, pentru estimarea mediei  $\bar{x}$  se efectuează o selecție  $x(1), x(2), \dots, x(n)$  referitoare la o variabilă  $x$  distribuită  $N(m, \sigma)$ . La rândul ei, media va avea o distribuție  $N(m, s/\sqrt{n})$ . Variabila normală standard  $z = (\bar{x} - m)\sqrt{n}/s$  are o distribuție normală  $N(0, 1)$ . Pentru o probabilitate de încredere  $P$ , mediei  $m$  a distribuției îi corespunde un interval de încredere simetric conform expresiei  $\bar{x} \pm z_{(1-\alpha/2)}s/\sqrt{n}$  în care,  $\alpha=1 - P$  iar  $z_{(1-\alpha/2)}$  se deduce din condiția:

$$P\{z(1-\alpha/2)\} = (1+p)/2 \quad (3.30)$$

**Exemplu.** Să se evalueze un interval de încredere simetric pentru medie, ce corespunde unei probabilități de încredere  $P \geq 0,95$ .

Deoarece  $\bar{x} = 0,41$ ,  $\sigma^2 = 0,000625$ ,  $n = 9$  determinări analitice iar din anexa I  $z_{0,975} = 1,96$ , intervalul de încredere al mediei este

$$0,41 - 0,016 < m < 0,41 + 0,016$$

b) Dispersia distribuției normale este necunoscută. În acest caz, pentru evaluarea intervalului de încredere simetric, evident corespunzător unei probabilități de încredere  $P$ , în cazul unei distribuții  $N(\mathbf{m}, \sigma)$  la care nu se cunoaște nici media, nici dispersia, se ia în considerare variabila  $t = (\bar{x} - \mathbf{m}) / (s / \sqrt{n})$  cu o distribuție  $t$  cu  $v = n - 1$  grade de libertate. În consecință, expresia intervalului de încredere este:

$$\bar{x} \pm t_{(1-\alpha/2; v)} s / \sqrt{n} \quad (3.31)$$

unde  $\alpha = 1 - P$  iar  $t_{(1-\alpha/2; v)}$  se deduce din ecuația  $P[t_{(1-\alpha/2; v)}] = (1+P)/2$ .

Acest caz al evaluării intervalului de încredere pentru media unei distribuții normale cu dispersie necunoscută este mult utilizat în chimia analitică.

În același mod se poate calcula intervalul de încredere pentru diferența a două medii ( $\mathbf{m}_1 - \mathbf{m}_2$ ) corespunzătoare variabilelor aleatoare  $x$  și  $y$  cu distribuțiile  $N(\mathbf{m}_1, \sigma)$  și  $N(\mathbf{m}_2, \sigma)$ .

Ca aplicație în chimia analitică, acest caz se utilizează în special la evaluarea intervalului de încredere pentru diferența dintre mediile rezultatelor obținute în două laboratoare diferite, prin aceeași metodă.

### III.4.2. Intervalul de încredere pentru dispersia de selecție

Pentru estimarea dispersiei unei variabile aleatoare cu o distribuție  $N(\mathbf{m}, \sigma)$  se consideră o selecție  $x(1), x(2), \dots, x(n)$ , din care se evaluează dispersia de selecție  $s^2$ . Se cunoaște faptul că variabila:  $\chi^2 = s^2 (n - 1) / \sigma^2$  are o distribuție cu  $v = n - 1$  grade de libertate. Intervalul de încredere pentru dispersie, pentru o probabilitate  $P$ , are expresia următoare:

$$\frac{(n-1)s^2}{c_{(1-\alpha/2; v)}^2} < t^2 < \frac{(n-1)s^2}{c_{(1-\alpha/2; v)}^2} \quad (3.32)$$

dacă se utilizează volume de selecție mari ( $n \geq 30$ ) se demonstrează că abaterea standard de selecție  $s$  are o distribuție  $N(s; s / \sqrt{2n})$ . Deci, se poate admite că variabila:

$$z = \frac{s - s}{s / \sqrt{2n}} \quad (3.33)$$

are o distribuție  $N(0,1)$ . În concluzie, pentru o probabilitate de încredere corespunzătoare  $P$ , se poate evalua un interval de încredere pentru abaterea

standard a unei distribuții normale, pentru volume de selecție mari, cu expresia:

$$s \pm z_{(1-\alpha/2)} s / \sqrt{2v} \quad (3.31)$$

### III.4.3. Intervalul de încredere pentru parametrii distribuției binomiale

În urma unor determinări repetate de  $n$  ori, fenomenul care ne interesează s-a manifestat de  $m$  ori. Probabilitatea  $p$  de manifestare a evenimentului este estimată de raportul  $\hat{p} = m/n$ . Ținând seama de distribuția binomială, variabila:

$$z = \frac{\hat{p} - p}{[\hat{p}(1 - \hat{p}) / n]^{1/2}} \quad (3.35)$$

are aproximativ distribuția  $N(0, 1)$  și deci intervalul de încredere pentru parametrul  $p$ , corespunzător unei probabilități de încredere  $P$ , are următoarea expresie:

$$\hat{p} \pm z_{(1-\alpha/2)} [\hat{p}(1 - \hat{p}) / n]^{1/2} \quad (3.36)$$

### III.5. Testarea ipotezelor

În general, testările de ipoteze se aplică la rezolvarea problemelor de evaluare și optimizare a proceselor analitice. Orice decizie luată în privința optimizării unui proces analitic trebuie elaborată în baza unor teste statistice. Teoria testelor de verificare a ipotezelor a fost elaborată de Neymann și Pearson și se aplică mult în analiza chimică.

Cazul general în chimia analitică este prezentat prin testarea unei ipoteze  $H_0$  în raport cu alte ipoteze alternative:  $H_1, H_2, \dots$ . În cazul unei metode de analiză se pune problema testării ipotezei  $H_0$  - media erorilor de analiză este egală cu zero (lipsește erorile sistematice) în raport cu ipotezele alternative:  $H_1$  - media erorilor de analiză  $< 0$  și  $H_2$  - media erorilor de analiză  $> 0$ . În altă situație analitică, se pune problema testării ipotezei  $H_0$  - introducerea sau modificarea unui factor în procesul de analiză duce la îmbunătățirea performanțelor procesului de analiză.

Pentru a studia ipoteza  $H_0$  se consideră că parametrul  $b$  al procesului luat în studiu are valoarea  $b = b_0$  în raport cu ipoteza  $H_1$  pentru care parametrul  $b$  are valoarea  $b \neq b_0$ . Evident, în continuare pentru rezolvarea acestei probleme se efectuează o selecție în baza căreia se obține o valoare  $\hat{b}$

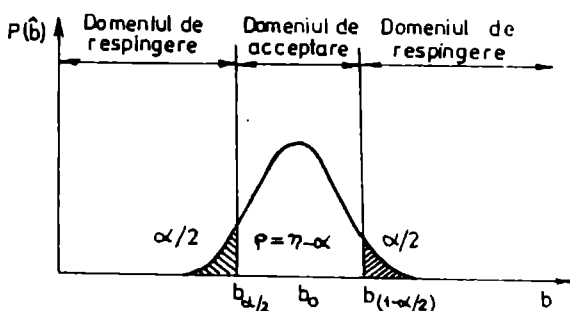
pentru parametrul  $b$ . Variabilei  $\hat{b}$  îi corespunde conform teoriei statistice densitatea de probabilitate  $p(\hat{b})$ . Întrebarea este cât de mare trebuie să fie diferența  $\hat{b} - b_0$  astfel ca ipoteza  $H_0$  să fie respinsă? În figura 3.7. este prezentată densitatea de probabilitate  $p(\hat{b})$  pentru cazul când ipoteza  $H_0$  este adevărată. Probabilitatea ca estimatorul  $\hat{b}$  să ia o valoare mai mică decât un prag  $b_{\alpha/2}$  este:

$$P(b \leq b_{\alpha/2}) = \int_{-1}^{b_{\alpha/2}} p(\hat{b}) d\hat{b} = \alpha / 2 \quad (3.37)$$

iar probabilitatea ca estimatorul  $\hat{b}$  să ia o valoare mai mare decât un prag  $b_{(1-\alpha/2)}$  este:

$$P(\hat{b} \geq b_{(1-\alpha/2)}) = \int_{b_{(1-\alpha/2)}}^{\infty} p(\hat{b}) d\hat{b} = \alpha / 2 \quad (3.38)$$

În concluzie, pentru formularea deciziilor se selectează o valoare denumită nivelul de semnificație al testului așa încât probabilitatea de a depăși pragurile  $b_{\alpha/2}$  și  $b_{(1-\alpha/2)}$  să aibă o valoare prestabilită egală cu **0,10**, **0,05** sau **0,01**. Dacă valoarea de estimare  $\hat{b}$  este mai mică decât  $b_{\alpha/2}$  sau mai mare decât  $b_{(1-\alpha/2)}$ , ipoteza  $H_0$  este respinsă, iar dacă se situează în regiunea  $b_{\alpha/2} < \hat{b} < b_{(1-\alpha/2)}$ , ipoteza este acceptată.



**Fig.3.7.** Reprezentarea schematică a procesului de testare a ipotezelor

Evident, pe baza rezultatelor testului, ipoteza propusă este acceptată sau respinsă. Respingerea ipotezei este o consecință a faptului că rezultatele diferă semnificativ. Esențial este faptul că testele nu sunt folosite pentru demonstrarea unei ipoteze, ci pentru îndrumarea acțiunilor ca și când ipoteza ar fi adevărată. În testarea ipotezelor pot apare două tipuri de erori: eroare de tip I, când se respinge o ipoteza adevărată și eroare de tip II, când se admite o ipoteză falsă.

Pentru verificarea ipotezelor în lucrările de statistică (testare) sunt prezentate mai multe criterii (teste). În funcție de problemele abordate în procesul analitic, se selectează testul potrivit. De exemplu, pentru eliminarea valorilor extreme dintr-un șir de rezultate, eliminare care nu trebuie făcută arbitrar, ci în baza unui criteriu statistic, pentru a fi veridic se folosesc următoarele criterii: criteriul t (Student), criteriul R (Grubbs), criteriul Q (Dixon), criteriul R. Liteanu C. prezintă detaliat aceste criterii în lucrarea **Teoria și metodologia statistică a analizei urmelor** (1979).

Pentru verificări de ipoteze în legătură cu media unei distribuții sau mediile a doua distribuții trebuie acceptată ca model distribuția normală. De asemenea, aceste teste utilizează selecții de valori bine determinate iar în cazul în care volumul selecției este o variabilă, se utilizează teste secvențiale precum și teste neparametrice care nu sunt elaborate pentru un anumit tip de distribuție. Pentru testarea a mai mult de două medii se utilizează metoda analizei de dispersie.

Să luam un exemplu de utilizare al testului t la verificarea ipotezei în legatură cu media unei distribuții normale. Pentru verificarea ipotezei absenței erorilor sistematice într-un proces analitic s-a analizat de 10 ori o probă standard, având un conținut al probei de determinat  $m_0 = 0,45$ . Din totalul mulțimii de rezultate, se calculează media și abaterea standard, acestea având următoarele valori:  $\bar{x} = 0,46$  și  $s = 0,008$ . Cu ajutorul mediei și abaterii standard calculate mai sus, se estimează variabila t cu ajutorul expresiei:

$$t_{\text{exp}} = \frac{\bar{x} - m_0}{s / n^{1/2}} \quad (3.39)$$

și se obține  $t_{\text{exp}} = 0,95$ .

Dacă ipoteza  $H_0: m = m_0$  ( $m =$  media distribuției), este adevărată, variabila  $t_{\text{exp}}$  aparține unei distribuții t cu  $v = n - 1$  grade de libertate. Drept umare, pentru un nivel de semnificatie  $\alpha$ , ipoteza  $H_0$  este respinsă dacă:

$$|t_{\text{exp}}| \geq t_{(1-\alpha/2;v)} \quad (3.40)$$



Din anexa 3 pentru distribuția  $t$  se obține  $t_{(0,95;9)} = 2,26$ . Deoarece  $|t_{\text{exp}}| > t_{(0,95;9)}$  rezultă că ipoteza absenței erorilor sistematice pentru acest nivel de semnificație  $\alpha = 0,10$  trebuie respinsă.

### III.6. Analiza de dispersie în chimia analitică

Analiza de dispersie a fost introdusă de Fisher. Cu toate că este o metodă eficientă de evaluare și optimizare a proceselor analitice cu aplicații în biologie și chimie, studiile referitoare la acest subiect sunt relativ puține. Analiza de dispersie este o tehnică statistică utilizată la evaluarea unor factori suspecți de a distorsiona veridicitatea rezultatelor. În cazul chimiei analitice, analiza de dispersie este utilizată la identificarea și evaluarea celor mai semnificative surse de erori. Din punct de vedere al metodologiei se poate menționa analiza de dispersie monofactorială, bifactorială, trifactorială.

Studiile menționate în literatura de specialitate a analizei de dispersie [5-10] prezintă exemple de utilizare a analizei de dispersie monofactorială și bifactorială privind descompunerea dispersiei rezultatelor de analiză în două componente (între zile și în cursul aceleași zile de lucru) sau trei componente. Analiza de dispersie este utilizată în studiile efectuate între laboratoare.

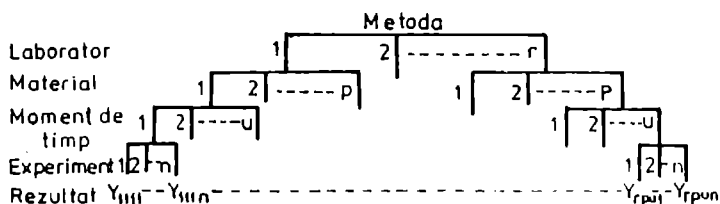


Fig.3.8. Plan experimental pentru analiza de dispersie trifactorială

Prezentăm în continuare un model de utilizare a analizei de dispersie trifactoriale pentru evaluarea unei metode de analiză conform unui plan experimental. Pentru testarea metodei de analiză se iau în considerare 3 factori: laboratorul, compoziția materialului de analizat și timpul. Avem deci  $r$  laboratoare,  $p$  materiale cu un conținut constant în componentul de determinat și diferit de restul componentelor și  $u$  zile de funcționare, deci timp; la fiecare moment fiecare material este analizat de  $n$  ori. În cazul acestui experiment dispersia totală este descompusă în componente pe factori: laboratorul, compoziția materialului și timpul.

În analiza de dispersie trifactorială observațiile aparțin unei rețele tridimensionale formată din  $r$ ,  $p$ ,  $u$  celule, în fiecare celulă având  $n$  observații. Notațiile  $r$ ,  $p$  și  $n$  reprezintă numărul de nivele la care se situează cei trei factori studiați și  $n$ , numărul de observații la un anumit nivel.

### *Bibliografie*

1. P. Eykhoff, "Process-parameter estimation" in **Progress in Control Engineering**, vol. 2, R. H. Macmillan Heywood, London, 1964.
2. R. Deusch, **Estimation Theory**, Prentice Hall, Englewood Cliffs, New Jersey, 1965.
3. N. E. Nahi, **Estimation Theory and Applications**, J. Wiley, New York, 1969
4. P. Eykhoff, **Identificarea sistemelor, estimarea parametrilor și stărilor pentru sisteme tehnice, economice, biologice**, Ed. Tehnica, București, 1977 (traducere din limba engleză).
5. R. F. Hirsch, *Analyt. Chem.*, **49**, 691A, 1977.
6. J. A. H. Gooszen, *Clinica Chim. Acta*, **5**, 431, 1960.
7. J. S. A. Amenta, *Amer. J. Clin. Pathol.*, **4**, 842, 1968.
8. J. H. Riddick, R. Flora, Q. L. Van Meter, *Clin. Chem.*, **18**, 250, 1972
9. G. Gootschalk, Z., *Analyt. Chem.*, **276**, 81, 1975.
10. G. Wernimont, *Mat. Res. Stds.*, **9-10**, 8, 1969.

Anexa 3.1

Valorile funcției Laplace  $P(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt$

z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3354	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4962	0,4963	0,4964	0,4965
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4983	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993

Anexa J.2

Valorile variabilei  $\chi^2$  pentru  $P = P(\chi^2) = \int_0^{\chi^2} p(\chi^2; v) d\chi^2$

$P$	0,01	0,025	0,05	0,10	0,90	0,95	0,975	0,99	0,995
1	0,00018	0,00098	0,0039	0,0156	2,71	3,84	5,02	6,63	7,88
2	0,0201	0,0508	0,1026	0,2107	4,61	5,99	7,38	9,21	10,60
3	0,115	0,218	0,353	0,584	6,25	7,81	9,35	11,34	12,84
4	0,297	0,484	0,711	1,064	7,78	9,49	11,14	13,28	14,86
5	0,554	0,831	1,15	1,61	9,24	11,07	12,83	15,09	16,75
6	0,872	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96
9	2,09	2,70	3,33	4,17	14,68	16,92	18,92	21,67	23,59
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73	26,76
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,68	5,63	6,57	7,79	21,08	23,68	26,12	29,14	31,32
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27
18	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
24	10,86	12,40	13,85	15,66	32,20	36,42	39,36	42,98	45,56
30	14,95	16,79	18,49	20,60	40,28	43,77	46,98	50,89	53,67
40	22,16	24,43	26,51	29,05	51,81	53,78	59,34	63,69	66,77
60	37,48	40,48	43,19	48,40	74,40	79,08	83,30	88,38	91,95
120	86,02	91,58	95,70	100,62	140,23	146,57	152,21	158,95	163,64

Valorile varianțelor (Student) pentru  $P = P(t) = \int_{-\infty}^t p(t; v) dt$

$\frac{P}{v}$	0,70	0,80	0,90	0,95	0,99	0,999
1	1,083	3,078	6,314	12,708	63,657	639,010
2	1,386	1,888	2,920	4,303	9,925	31,598
3	1,250	1,638	2,353	3,182	5,841	12,941
4	1,190	1,533	2,132	2,776	4,604	8,010
5	1,156	1,478	2,015	2,571	4,032	6,859
6	1,134	1,440	1,943	2,447	3,707	5,959
7	1,119	1,415	1,895	2,365	3,499	5,405
8	1,108	1,397	1,860	2,308	3,355	5,041
9	1,100	1,383	1,833	2,262	3,250	4,781
10	1,093	1,372	1,812	2,228	3,169	4,587
11	1,088	1,363	1,796	2,201	3,106	4,437
12	1,083	1,356	1,782	2,179	3,055	4,318
13	1,079	1,350	1,771	2,160	3,012	4,221
14	1,076	1,345	1,761	2,145	2,977	4,140
15	1,074	1,341	1,753	2,131	2,947	4,073
16	1,071	1,337	1,746	2,120	2,921	4,015
17	1,069	1,333	1,740	2,110	2,898	3,965
18	1,067	1,330	1,734	2,101	2,878	3,922
19	1,076	1,328	1,729	2,093	2,861	3,883
20	1,064	1,325	1,725	2,086	2,845	3,850
21	1,063	1,323	1,721	2,080	2,831	3,819
22	1,061	1,321	1,717	2,074	2,819	3,792
23	1,060	1,319	1,714	2,069	2,807	3,767
24	1,059	1,318	1,711	2,064	2,797	3,745
25	1,058	1,316	1,708	2,060	2,787	3,725
26	1,058	1,315	1,706	2,056	2,779	3,707
27	1,057	1,314	1,703	2,052	2,771	3,690
28	1,056	1,313	1,701	2,048	2,763	3,674
29	1,055	1,311	1,699	2,045	2,756	3,659
30	1,055	1,310	1,697	2,042	2,750	3,646
40	1,050	1,303	1,684	2,021	2,704	3,551
60	1,046	1,296	1,671	2,000	2,660	3,460
120	1,041	1,289	1,659	1,980	2,617	3,373
	1,036	1,282	1,645	1,960	2,576	3,291

Anexa 3.4

Valorile variabilei  $F$  pentru  $P = P(F) = \int_0^F p(F; v_1, v_2) dF = 0,95$

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	161	200	216	225	230	234	237	239	241	242	243	244	245	245	246	246	247
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,73	8,71	8,70	8,69	8,68
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,89	5,87	5,86	5,84	5,83
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,60	4,59
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,92	3,91
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,53	3,51	3,49	3,48
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22	3,20	3,19
9	5,12	4,26	4,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01	2,99	2,97
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85	2,83	2,81
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,78	2,74	2,72	2,70	2,69
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62	2,60	2,58
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53	2,51	2,50
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46	2,44	2,43
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40	2,38	2,37
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35	2,33	2,32
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,32	2,31	2,29	2,27
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,28	2,27	2,25	2,23
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,21	2,20
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20	2,18	2,17
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,22	2,20	2,18	2,16	2,14
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,20	2,18	2,15	2,13	2,09
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,23	2,20	2,18	2,15	2,13	2,11	2,09
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,21	2,18	2,15	2,13	2,11	2,09	1,07
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,11	2,09	2,07	2,05
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,12	2,09	2,07	2,05	2,03
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13	2,10	2,08	2,06	2,04	2,02
28	4,20	3,34	2,95	2,72	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,09	2,06	2,04	2,02	2,00
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,08	2,05	2,03	2,01	1,99

## IV. ANALIZA DE REGRESIE

### IV.1. Generalități

Un instrument matematic întrebuițat în mod frecvent în aplicațiile științifice îl reprezintă analiza de regresie care permite modelarea datelor afectate de erori experimentale. Pentru atenționarea studenților români care consultă literatură științifică străină, menționăm că în limba engleză, pentru denumirea analizei de regresie se folosesc, cu valoare sinonimă, mai multe sintagme lingvistice cum sunt: *empirical modeling*, *curve fitting*, *regression analysis* sau *forecasting*.

În literatura românească dedicată chimiei analitice, termenul adoptat pentru denumirea acestui capitol a fost statuat de către Liteanu și Râcă [1] care au arătat că în analiza chimică cele mai numeroase aplicații ale analizei de regresie se întâlnesc la evaluarea funcțiilor de etalonare a sistemelor analitice, la evidențierea și corectarea fenomenelor de interferență, la stabilirea condițiilor optime de lucru pentru sistemele analitice de măsură și la evaluarea performanțelor calitative ale sistemelor analitice. În general însă, analiza de regresie se poate întrebuița ori de câte ori chimistul analist se confruntă cu nevoia de a găsi un model matematic care să descrie cât mai bine setul de date de care dispune.

### IV.2. Tipuri de regresii

Răspunsul sistemelor analitice (potențial de electrod, absorbantă, aria picurilor cromatografice, rezultatul unei metode de analiză etc), notat prescurtat  $y$ , este influențat, în general, de un număr foarte mare de factori experimentali (concentrații de analiți și de interferenți, condiții de operare și de mediu ș.a.m.d.). De obicei se dorește modelarea răspunsului în funcție de un număr limitat de factori în condițiile în care toți ceilalți sunt menținuți constanți. Pentru aceasta, factorii experimentali considerați la elaborarea modelului (denumiți variabile independente) sunt variați sistematic și se determină valoarea răspunsului (variabila dependentă). Între perechile de date astfel obținute există o relație de forma:

$$y_i = f(x_{1i}, x_{2i}, \dots, x_{mi}; b_0, b_1, \dots, b_m) + e_i; \quad i = 1, n \quad (4.1)$$

unde  $y_i$  reprezintă valoarea răspunsului obținută în experimentul numărul  $i$ ;  $x_{ji}$  - valorile din experimentul numărul  $i$  a factorilor  $x_j$  ( $j = 1, m$ );

$b_k$  - constante denumite parametrii modelului ( $k = 0, m$ );  $f(x_{ji}; b_k)$  - componenta deterministă a valorii răspunsului și  $e_i$  - componenta stocastică sau aleatoare. Această ultimă componentă modelează variabilitatea datelor analitice care se manifestă ca urmare a fluctuațiilor condițiilor experimentale (temperatură, umiditate, tensiune de alimentare a instrumentelor de măsură etc) pe parcursul determinărilor.

Analiza de regresie constă în modelarea componentei deterministe a semnalelor în prezența componentei stocastice [2] iar tipul de regresie care trebuie utilizat într-un anumit caz depinde de problema de rezolvat. Acesta se referă la numărul variabilelor independente din expresia matematică a modelului precum și la natura relației de dependență dintre variabila independentă și parametrii  $b_k$ . O sistematizare a principalelor tipuri de regresii întâlnite în analiza chimică este prezentată în Tabelul 4.1.

Regresia de o singură variabilă (monovariabilă) presupune existența unei singure variabile independente, indiferent care este relația de dependență dintre aceasta și variabila dependentă, în timp ce în regresia de mai multe variabile (multivariabilă) se presupune că variabila dependentă este o funcție de mai multe variabile independente. În continuare, peste tot se va nota cu  $y$  variabila dependentă în timp ce pentru variabilele independente vom folosi notațiile  $x$  - în regresia monovariabilă, și, respectiv,  $x_1, x_2, x_3, \dots, x_m$  - în regresia multivariabilă. În cazul repetării experimentelor, valorile concrete luate de către variabilele independentă și dependente vor fi reprezentate printr-un indice adițional notat, în caz general, cu litera  $l$ . De asemenea, acolo unde este necesar, cu  $z$  vom reprezenta variabila dependentă transformată iar cu  $u$  - variabila independentă transformată.

După forma matematică a modelului utilizat, putem face distincție între metode de regresie liniare și, respectiv, neliniare. În cazul regresiei liniare, variabila dependentă poate fi reprezentată printr-o combinație liniară de operatori care depind (nu neapărat liniar) de variabila independentă (sau de variabilele independente). În cazul regresiei monovariabile, astfel de operatori pot fi  $x^c$  ( $c = \text{const}$ ),  $\log x$ ,  $1/x$ ,  $\sin x$  etc.

O clasă importantă de modele sunt cele întrebuințate în regresia polinomială monovariabilă care se aplică frecvent pentru a aproxima, pe anumite porțiuni, funcții neliniare. În acest scop, menționăm că se mai pot utiliza și curbele spline de ordinul trei [3].

În cazul în care modelul matematic polinomial se rezumă la primul doi termeni se obține tipul de regresie liniară propriu zisă, caz în care datele experimentale se înscriu pe o dreaptă. Acest ultim tip de regresie este foarte mult utilizat în studii de calibrare ale metodelor de



analiză, când variabila independentă semnifică concentrația componentului analizat din etaloane iar  $b_0$  și  $b_1$  - ordonata la origine a dreptei de calibrare și, respectiv, panta acesteea, ultima reprezentând o caracteristică foarte importantă a unei metode de analiză legată direct de sensibilitatea metodei.

Tabelul 4.1  
Principalele tipuri de regresii utilizate în analiza chimică

Nr. crt.	Tipul de regresie	Caracteristici	Forma modelului
1.	Monovariabilă liniară	Nr. de variabile independente = 1 Nr. de parametri = 1 sau 2	$y = b_0 + b_1 x$ Uneori $b_0 = 0$
2.	Monovariabilă polinomială	Nr. de variabile independente = 1 Nr. de parametri = 2 sau mai mulți În general $m + 1$	$y = b_0 + \sum_{k=1}^m b_k x^k$
3.	Monovariabilă neliniară	Nr. de variabile independente = 1 Nr. de parametri = 2 sau mai mulți	$y = f(x)$ f - orice funcție neliniară
4.	Multivariabilă liniară	Nr. de variabile independente = 2 sau mai multe (m) Nr. de parametri = 2 sau mai mulți	$y = b_0 + \sum_{i=1}^m b_i x_i$ Frecvent $b_0 = 0$

Regresia multivariabilă se aplică frecvent în cazul metodelor spectrometrice de determinare simultană a mai multor componente din soluții (determinări multicomponent) când se presupune de regulă că  $b_0 = 0$  [2, 4, 5].

Metodele neliniare de regresie utilizează pentru exprimarea relației de dependență dintre variabila dependentă și variabila independentă ecuații care nu sunt liniare în toți parametri. Ele pot fi totuși, în anumite condiții, liniarizate prin dezvoltări în serie. Astfel, o funcție oarecare depinzând de mai mulți parametri,  $b_1, b_2, \dots, b_m$ , și având forma:

$$y = f(x; b_1, \dots, b_m) \quad (4.2)$$

poate fi foarte bine aproximată, în jurul unui punct din spațiul parametrilor,  $\mathbf{b}_0$ , definit prin vectorul linie:  $\mathbf{b}_0^T = (b_{10}, b_{20}, \dots, b_{m0})$ , printr-o dezvoltare în serie Taylor din care se rețin termenii de ordinul 0 și 1, conform ecuației:

$$y = f(x; b_{10}, \dots, b_{m0}) + \left( \frac{\partial f}{\partial b_1} \right)_{\mathbf{b}=\mathbf{b}_0} \cdot \Delta b_1 + \left( \frac{\partial f}{\partial b_2} \right)_{\mathbf{b}=\mathbf{b}_0} \cdot \Delta b_2 + \dots + \left( \frac{\partial f}{\partial b_m} \right)_{\mathbf{b}=\mathbf{b}_0} \cdot \Delta b_m \quad (4.3)$$

cu condiția ca  $\Delta b_k$  ( $k=1, m$ ) să fie suficient de mici. Utilizând această proprietate, regresia neliniară poate fi efectuată în mod similar celei liniare în cazul în care inițial se cunoaște sau se poate determina o estimată suficient de bună a valorilor parametrilor modelului.

Ca exemple de modele monoparametrice neliniare, aplicate frecvent în chimie, putem menționa:

- funcții exponențiale:

$$y = b_1 \exp(b_2 x) \quad (4.4)$$

care descriu funcțiile de răspuns ale instrumentelor analitice după aplicarea unui semnal în formă de treaptă;

- funcții de tip Gauss:

$$y = b_1 \exp \left[ -b_2 (x - b_3)^2 \right] \quad (4.5)$$

care descriu forma picurilor cromatografice;

- funcții de tip Lorenz:

$$y = \frac{b_1}{1 + b_2 (x - b_3)^2} \quad (4.6)$$

care descriu forma picurilor de rezonanță magnetică nucleară;

- funcții putere:

$$y = b_1 x^{b_2} \quad (4.7)$$

care descriu intensitatea liniilor spectrale în spectrografia de emisie;

- funcții logaritmice:

$$y = b_1 + b_2 \log x \quad (4.8)$$

care descriu răspunsul electrozilor ion-selectivi.

Unele ecuații neliniare pot fi aduse ușor, prin transformări de variabile convenabile, la o formă liniară. Astfel, în cazul potențio-metriei, unde relația de dependență potențial - concentrație, dată de ecuația lui Nernst, este logaritmică, regresia liniară se poate aplica după ce se efectuează următoarea transformare a variabilei independente:

$$x = \log a \quad (4.9)$$

### IV.3. Etapele analizei de regresie

Având deja determinat un set de date, analiza de regresie presupune parcurgerea următoarelor etape: definirea modelului, estimarea parametrilor modelului (regresia propriu zisă), validarea modelului și calculul intervalelor de încredere a parametrilor.

#### IV.3.1. Definirea modelului

Această etapă este foarte importantă deoarece stabilește tipul de regresie care va fi utilizat la estimarea parametrilor. Ea presupune explicitarea relației de dependență dintre variabilele dependentă și, respectiv, independente (estimarea funcției  $f$  din relația 4.1) și, implicit, stabilirea numărului de parametri ai modelului (constantele  $b_k$ ).

Modelul matematic adoptat este uneori sugerat de anumite legi fizice care se respectă în cazul investigat (de exemplu, în cazul calibrărilor, legea Lambert-Beer în spectrometrie, ecuația lui Nernst în potențiometrie etc). Alteori modelul cel mai probabil se impune de la sine cu necesitate. Astfel, de exemplu, dacă se dorește identificarea tipurilor de interferențe care afectează o anumită metodă de analiză se compară rezultatele obținute cu metoda studiată pentru un set de probe,  $y_i$ , cu rezultatele obținute, pentru aceleași probe, cu o metodă de referință considerată exactă,  $x_i$ . Un model liniar de forma:

$$y = b_0 + b_1 x \quad (4.10)$$

este foarte probabil să fie respectat în astfel de cazuri. Dacă modelul este validat, valorile estimate ale parametrilor furnizează următoarele concluzii:

- a) Dacă  $b_0 = 0$  și  $b_1 = 1$ , metoda este exactă;
- b) Dacă  $b_0 \neq 0$  și  $b_1 = 1$ , metoda de analiză este afectată de o eroare sistematică de tip aditiv și pentru obținerea rezultatului corect,  $b_0$  trebuie scăzut din rezultatul calculat;
- c) Dacă  $b_0 = 0$  și  $b_1 \neq 1$ , metoda este afectată de o eroare de tip multiplicativ și, pentru obținerea rezultatului corect, se poate aplica metoda adășiilor standard;
- d) Dacă  $b_0 \neq 0$  și  $b_1 \neq 1$ , metoda este afectată atât de erori de tip aditiv, cât și multiplicativ.

Există și situații, întâlnite mai ales la optimizarea parametrilor de lucru ai metodelor de analiză, când relația dintre răspuns și factorii de optimizat nu este cunoscută a priori și când se recurge la modelare empirică. Inspectarea vizuală a graficelor reprezentând variația răspunsului

în funcție de factorii considerați sugerează funcțiile matematice care vor fi utilizate la definirea modelului în astfel de cazuri.

La aplicarea regresiei monovariabile liniare se presupune că între perechile de date experimentale,  $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$ , măsurate în condiții reale și, prin urmare, afectate de zgomot, există următoarele relații:

$$\begin{aligned} y_1 &= b_0 + b_1 x_1 + e_1 \\ y_2 &= b_0 + b_1 x_2 + e_2 \\ &\dots \\ y_i &= b_0 + b_1 x_i + e_i \\ &\dots \\ y_n &= b_0 + b_1 x_n + e_n \end{aligned} \quad (4.11)$$

Ecuatiile de mai sus pot fi scrise în formă condensată utilizând notații matriciale. Astfel, notând cu  $\mathbf{y}$ ,  $\mathbf{b}$ ,  $\mathbf{e}$ , și  $\mathbf{X}$  vectorii coloană ai măsurătorilor, ai parametrilor, ai erorilor (sau a reziduurilor) și, respectiv, matricea variabilelor independente, definiți prin ecuațiile următoare:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_i \\ \dots \\ y_n \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}; \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_i \\ \dots \\ e_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_i \\ \dots & \dots \\ 1 & x_n \end{pmatrix} \quad (4.12)$$

sistemul de ecuații 4.11 este echivalent cu:

$$\mathbf{y} = \mathbf{X} \mathbf{b} + \mathbf{e} \quad (4.13)$$

Peste tot în acest capitol se adoptă convențiile uzuale de scriere a mărimilor vectoriale și matriceale [4]. Astfel, literele mici scrise întărit reprezintă vectori coloană iar majusculele - mărimi matriceale. Transpusa vectorilor sau matricelor se reprezintă adăugând litera T în dreapta sus la simbolul inițial. De exemplu, dacă  $\mathbf{y}$  reprezintă vectorul coloană cu  $n$  elemente definit prin prima dintre ecuațiile 4.12, atunci  $\mathbf{y}^T$  reprezintă următorul vector linie:

$$\mathbf{y}^T = (y_1 \quad y_2 \quad \dots \quad y_i \quad \dots \quad y_n) \quad (4.14)$$

De asemenea:

$$\mathbf{X}^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \quad (4.15)$$

Estimațiile oricăror mărimi, indiferent de natura acestora, se vor reprezenta prin adăugarea unui accent circumflex ( $\wedge$ ) deasupra simbolului mării respective. Astfel, estimațiile obținute pe baza modelului

de regresie adoptat pentru vectorii măsurătorilor și al parametrilor se vor nota cu  $\hat{y}$  și, respectiv, cu  $\hat{b}$ .

Tabelul 4.2  
Forma vectorului parametrilor și a matricei măsurătorilor în trei tipuri diferite de regresii

Nr. crt.	Tipul de regresie	Forma vectorului $b$	Forma matricei $X$
1.	Monovariabilă polinomială	$\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \\ \vdots \\ b_m \end{pmatrix}$	$\begin{pmatrix} 1 & x_1 & \dots & x_1^j & x_1^m \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & x_i & \dots & x_i^j & x_i^m \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & x_n & \dots & x_n^j & x_n^m \end{pmatrix}$
2.	Monovariabilă neliniară	$\begin{pmatrix} 1 \\ \Delta b_1 \\ \vdots \\ \Delta b_k \\ \vdots \\ \Delta b_m \end{pmatrix}$	$\begin{pmatrix} f_1 & \frac{\delta f_1}{\delta b_1} & \dots & \frac{\delta f_1}{\delta b_k} & \frac{\delta f_1}{\delta b_m} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ f_i & \frac{\delta f_i}{\delta b_1} & \dots & \frac{\delta f_i}{\delta b_k} & \frac{\delta f_i}{\delta b_m} \\ \vdots & \vdots & \dots & \vdots & \vdots \\ f_n & \frac{\delta f_n}{\delta b_1} & \dots & \frac{\delta f_n}{\delta b_k} & \frac{\delta f_n}{\delta b_m} \end{pmatrix}$
3.	Multivariabilă liniară	$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_m \end{pmatrix}$	$\begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{m1} \\ 1 & x_{12} & x_{22} & \dots & x_{m2} \\ 1 & x_{13} & x_{23} & \dots & x_{m3} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{mn} \end{pmatrix}$

Și pentru celelalte tipuri de regresii prezentate în tabelul 4.1 se pot scrie ecuații de dependență între variabilele dependente măsurate în prezența zgomotului și cele independente asemănătoare ca formă cu ecuațiile 4.11 din regresia monovariabilă liniară. Singura deosebire constă în modificarea componentelor deterministe care derivă din forma modelelor de regresie utilizate.

Pentru ca ecuațiile astfel obținute să depindă liniar de toți parametrii și în cazul regresiei neliniare, expresia modelului utilizat în

acest caz se poate înlocui cu aproximarea sa dată de ecuația 4.3 dacă sunt îndeplinite condițiile de valabilitate a acesteia menționate anterior.

Deși relațiile dintre valorile măsurătorilor efectuate și a factorilor experimentali considerați diferă de la un tip de regresie la altul, totuși, ecuația 4.13 este generală. Ea se aplică în toate cazurile de regresie liniară și, în condițiile restrictive menționate, la regresia neliniară cu condiția să se utilizeze expresii adecvate pentru vectorul parametrilor și pentru matricea măsurătorilor. Forma matematică a acestora este prezentată în tabelul 4.2 pentru tipurile de regresie prezentate în tabelul 4.1 (altele decât regresia monovariabilă liniară deja discutată) presupunând că modelul matematic de regresie utilizat este cel din ultima coloană a tabelului 4.1. În cazul regresiei monovariabile neliniare, toate elementele matricii  $X$  ( $f_i$  și  $\frac{\delta f_i}{\delta b_k}$ ) reprezintă modelul de regresie,  $f$ , sau derivatele de ordinul unu ale acestuia în raport cu diferiți parametri și se calculează la valorile  $x_i$  ale variabilei independente. De asemenea, toate acestea se calculează în punctul  $b_0$ , acesta reprezentând vectorul estimatie inițială a parametrilor modelului.

Utilizarea algebrei matriceale [6] în analiza de regresie prezintă avantajul că ecuațiile și soluțiile se scriu într-o formă clară, compactă, generală (aplicabilă oricărei probleme) și ușor de transpus în codul calculatoarelor.

#### IV. 3.2. Estimarea parametrilor modelului

În această etapă se calculează valorile necunoscute ale parametrilor prin suprapunerea cât mai bună a modelului peste perechile de date experimentale ceea ce presupune minimizarea unui criteriu matematic a cărui expresie depinde de natura distribuției erorilor în setul de date.

În cazul cel mai frecvent se consideră că valorile factorilor ( $x_{ji}$ ) sunt cunoscute fără eroare, acestea necontribuind la diferențele dintre valorile măsurate ale răspunsului ( $y_i$ ) și cele estimate pe baza modelului ( $\hat{y}_i$ ), diferențe denumite frecvent deviații sau reziduuri. În majoritatea aplicațiilor din analiza chimică, unde  $x_j$  reprezintă concentrații de analiți, această presupunere este adevărată deoarece de regulă precizia răspunsului analitic este considerabil mai redusă decât cea obținută în prepararea soluțiilor standard.

Dacă modelul adoptat aproximează suficient de bine dependența exprimată în relația 4.1 prin funcția  $f$  pe tot domeniul de variație al factorilor, deviațiile ( $y_i - \hat{y}_i$ ) reprezintă o estimatie a erorilor expe-

imentale ( $e_i$ ). Criteriul matematic,  $R$ , prin minimizarea cărui se estimează corect valorile parametrilor modelului îl reprezintă suma pătratelor deviațiilor:

$$R = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4.16)$$

cu condiția ca erorile experimentale de determinare a răspunsului să reprezinte variabile aleatoare care să îndeplinească următoarele condiții:

- a) să fie distribuite normal cu media zero;
- b) să fie independente;
- c) să aibă aceeași varianță,  $\sigma^2$ .

Suma pătratelor deviațiilor,  $R$ , definită prin relația 4.16, reprezintă norma vectorului reziduurilor, notată  $|\mathbf{e}|$  și calculată, în notații matriceale, cu expresia:

$$|\mathbf{e}| = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) \quad (4.17)$$

Aceasta este o funcție dependentă de parametrii modelului,  $b_k$ . Valorile cele mai probabile ale acestora, reprezentate prin următorul vector linie:

$\hat{\mathbf{b}}^T = (\hat{b}_1, \hat{b}_2, \dots, \hat{b}_m)$ , se estimează rezolvând următorul sistem de ecuații normale liniare:

$$\frac{\partial R}{\partial b_k} = 0; \quad k = 0, m \quad (4.18)$$

care reprezintă condiția de minim a criteriului  $R$ .

Numărul de ecuații din sistemul 4.18 este egal cu numărul de parametri ai modelului,  $m+1$ . În toate cazurile de regresie liniară, soluția generală a acestor ecuații este dată de expresia:

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (4.19)$$

în care  $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$  reprezintă inversa generalizată a matricii  $\mathbf{X}$  de tip Moore-Penrose. Este preferabil ca această matrice să aibă o formă cât mai simplă, ceea ce se poate realiza adoptând un plan experimental corespunzător. În cazul calibrărilor unor determinări multicomponent, această presupune prepararea unor standarde conținând, rând pe rând, câte un singur component.

Indiferent de numărul liniilor și coloanelor matricii  $\mathbf{X}$ , care nu trebuie să fie neapărat egale, matricea  $\mathbf{X}^T \mathbf{X}$  are o formă pătrată. Soluția reprezentată prin ecuațiile 4.19 există doar dacă matricea  $\mathbf{X}^T \mathbf{X}$  este inversabilă.

În cazul regresiei neliniare, estimarea parametrilor urmează o cale întrucâtva diferită de cea prezentată anterior. Dacă în ecuația 4.17 se utilizează vectorul  $\hat{\mathbf{y}}$  exprimat prin modelul de regresie neliniar adoptat, sistemul de ecuații 4.18 nu mai are o formă liniară ușor de rezolvat. Presupunând însă că se poate determina, indiferent prin ce mijloace, o estimatie inițială suficient de bună a parametrilor modelului neliniar, notată  $\mathbf{b}_0$ ,  $\hat{\mathbf{y}}$  poate fi înlocuit în ecuația 4.17 prin  $\mathbf{X}\mathbf{b}$ , unde  $\mathbf{X}$  este matricea măsurătorilor cu expresia dată în tabelul 4.2 pentru regresia neliniară iar  $\mathbf{b}$  este definit prin relația:

$$\mathbf{b}^T = (1, b_{10}, b_{20}, \dots, b_{m0}) \quad (4.20)$$

În aceste condiții, se verifică ușor că ecuația 4.17 devine:

$$\begin{aligned} |\mathbf{e}| &= (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) = (\mathbf{y} - \hat{\mathbf{y}} - \mathbf{J}\Delta\mathbf{b})^T (\mathbf{y} - \hat{\mathbf{y}} - \mathbf{J}\Delta\mathbf{b}) = \\ &= (\Delta\mathbf{y} - \mathbf{J}\Delta\mathbf{b})^T (\Delta\mathbf{y} - \mathbf{J}\Delta\mathbf{b}) \end{aligned} \quad (4.21)$$

unde:  $\Delta\mathbf{y}$  reprezintă vectorul abaterilor valorilor măsurate ale răspunsului,  $\mathbf{y}$ , față de valorile estimate pe baza modelului neliniar adoptat,  $\hat{\mathbf{y}}$ , în care s-au utilizat valorile estimate inițial ale parametrilor;  $\Delta\mathbf{b}$  - vectorul diferențelor dintre valorile adevărate ale parametrilor și cele estimate inițial iar  $\mathbf{J}$  este matricea lui Jacobi definită prin ecuația:

$$\mathbf{J} = \begin{pmatrix} \frac{\delta f_1}{\delta b_1} & \frac{\delta f_1}{\delta b_k} & \frac{\delta f_1}{\delta b_m} \\ \frac{\delta f_i}{\delta b_1} & \frac{\delta f_i}{\delta b_k} & \frac{\delta f_i}{\delta b_m} \\ \frac{\delta f_n}{\delta b_1} & \frac{\delta f_n}{\delta b_k} & \frac{\delta f_n}{\delta b_m} \end{pmatrix} \quad (4.22)$$

Elementele acestei matrici au aceleași semnificații ca și elementele corespunzătoare din matricea  $\mathbf{X}$  (vezi paragraful precedent).

Cu aceste observații, sistemul de ecuații normale 4.18, în care criteriul  $R$  se înlocuiește cu ecuația 4.21, se aplică și în regresia neliniară, în același mod ca și în cea liniară. Rezolvarea sa conduce la vectorul  $\Delta\hat{\mathbf{b}}$ . Prin similitudine cu ecuația 4.19, expresia sa este dată de:

$$\Delta\hat{\mathbf{b}} = (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \Delta\mathbf{y} \quad (4.23)$$

Vectorul  $\hat{\mathbf{b}} = \mathbf{b}_0 + \Delta\hat{\mathbf{b}}$  reprezintă de obicei o estimatie a parametrilor modelului neliniar mai bună decât cea inițială,  $\mathbf{b}_0$ , dacă aceasta din urmă este suficient de exactă. Pentru a obține o estimatie și mai bună a parametrilor modelului, algoritmul de regresie neliniară poate fi reluat prin înlocuirea estimatiei inițiale a parametrilor,  $\mathbf{b}_0$ , cu ultima dată obținută,  $\hat{\mathbf{b}}$ . Un număr mic de iterații (3 - 5) conduce, de



regulă, la valorile adevărate ale parametrilor modelului. Aceste valori se recunosc ușor deoarece două iterații succesive conduc practic la același rezultat. Totuși, în cazul în care estimarea inițială a parametrilor,  $b_0$ , diferă mult de valorile reale, algoritmul de regresie menționat poate să eșueze, fapt care se recunoaște prin lipsa de convergență a valorilor estimate ale parametrilor obținute prin iterații succesive.

Datorită formei criteriului de minimizat,  $R$ , metodele de estimare a parametrilor prezentate până în prezent se numesc **regresii prin metoda celor mai mici pătrate**.

În timp ce proprietățile a) și b) ale erorilor experimentale menționate anterior sunt îndeplinite în practică în cvasitotalitatea cazurilor, ultima proprietate, denumită **homoscedasticitate**, nu este de obicei respectată într-un domeniu larg de variație al factorilor. De aceea, se recomandă testarea statistică a acestora ori de câte ori există dubii asupra valabilității sale. Pentru aceasta, pentru fiecare pereche de valori a factorilor,  $\mathbf{x}_i^T = (x_{1i}, x_{2i}, \dots, x_{mi})$ , se repetă experimentul analitic de mai multe ori măsurându-se valorile răspunsului,  $y_{ij}$ , de fiecare dată, apoi se calculează dispersiile și, în final, se testează egalitatea acestora (de exemplu, cu ajutorul testului lui Bartlett [1]).

În cazul multor tehnici analitice, incluzând măsurătorile radiometrice și spectrometrice, varianța erorilor experimentale este dependentă de valoarea răspunsului,  $y_i$ , proprietatea aceasta numindu-se **heteroscedasticitate** [2]. Aceasta se întâmplă datorită tipurilor specifice de zgomot care afectează măsurătorile. Astfel, în determinările radiometrice, bazate pe numărarea particulelor care ajung la detector, predomină un tip de zgomot supus distribuției lui Poisson. În astfel de cazuri, dispersia erorilor experimentale este proporțională cu rădăcina pătrată a valorilor semnalelor chiar în prezența altor surse de erori (în afara celor de numărare datorate detectorului) cum ar fi: erori de pipetare, variații ale timpilor de reacție sau a temperaturii etc. În tehnicile spectrometrice, cum ar fi absorbția atomică, zgomotul datorat fluctuațiilor intensității sursei de lumină este proporțional cu intensitatea semnalelor. În unele tehnici analitice poate să predomine zgomotul datorat componentelor electronice ale instrumentului de măsură care poate avea varianță constantă sau, uneori, dependentă de valoarea semnalelor.

În cazul heteroscedasticității nu se mai poate aplica metoda celor mai mici pătrate și, pentru efectuarea regresiei, se apelează la una din următoarele alternative: aplicarea metodei celor mai mici pătrate ponderate sau transformarea variabilelor pentru obținerea homoscedasticității.

La aplicarea metodei celor mai mici pătrate ponderate trebuie cunoscute dispersiile,  $s_i$ , pentru valorile răspunsului, notat  $y_i$ , în fiecare punct,  $x_i$ , din domeniul factorilor. Pentru aceasta este nevoie de un număr, notat  $n_i$ , suficient de mare de valori,  $y_{i1}$ , ale răspunsului, măsurate prin repetarea experimentului la fiecare set de valori ale factorilor, acesta constituind un important neajuns al metodei.

Trebuie menționat că dacă numerele de repetări,  $n_i$ , sunt prea mici, la testarea egalității mediilor aplicând testul lui Bartlett, probabilitatea respingerii ipotezei heteroscedasticității, în cazul în care aceasta este adevărată, este destul de mare.

Totuși, dacă dispersiile,  $s_i$ , s-au estimat suficient de corect, criteriul de optimizat, R, are, în acest caz, următoarea expresie:

$$R = \sum_{i=1}^n (w_i e_i)^2 = \sum_{i=1}^n [w_i (\bar{y}_i - \hat{y}_i)]^2 \quad (4.25)$$

unde  $w_i$  reprezintă factorii de pondere ai reziduurilor, dați de relația:

$$w_i = \frac{1}{s_i} \quad (4.26)$$

$\bar{y}_i$  - media măsurătorilor individuale ale variabilei dependente obținute la aceleași valori ale variabilelor independente:

$$\bar{y}_i = \sum_{l=1}^{n_i} y_{il} \quad (4.27)$$

iar restul mărimilor au semnificațiile cunoscute.

Cu următoarele notații:

$$W = \begin{pmatrix} w_1 & & \\ & w_2 & \\ & & \ddots \\ & & & w_n \end{pmatrix} \quad (4.28)$$

$$U = W^T W = \begin{pmatrix} w_1^2 & & \\ & w_2^2 & \\ & & \ddots \\ & & & w_n^2 \end{pmatrix} \quad (4.29)$$

se verifică ușor că soluția sistemului de ecuații normale 4.18, în care s-a înlocuit criteriul R prin expresia 4.25, are forma:

$$b = (X^T U X)^{-1} (X^T U y) \quad (4.30)$$

Ideea de bază a regresiei ponderate prin metoda celor mai mici pătrate este, asadar, de a atribui cea mai mare importanță datelor

măsurate cu cea mai mare precizie. Cea mai dificilă problemă la aplicarea acestui tip de regresii constă în estimarea corectă a factorilor de pondere deoarece este nevoie de un număr mare de repetări ale experimentelor, ceea ce uneori poate fi foarte costisitor și/sau poate necesita un volum mare de timp.

În absența unui volum de date suficient de mare, se poate presupune, luând în considerare și experiența anterioară desprinsă din tipuri similare de experiențe, o relație funcțională între varianță și răspuns. Astfel, se cunoaște că în radioimunoanaliză  $s_i$  depinde aproape liniar de  $\sqrt{y_i}$ . Un alt caz foarte important este cel în care  $s_i$  depinde liniar de  $y_i$  și în care deviația relativă standard este constantă, așa cum se întâmplă uneori, la concentrații ridicate, în analiza prin ICP.

În astfel de cazuri se poate aplica o transformare convenabilă a variabilei dependente în vederea obținerii homoscedasticității. Dacă  $s_i \propto \sqrt{y_i}$ , homoscedasticitatea se obține prin transformarea  $z = \sqrt{y}$ , iar dacă  $s_i \propto y_i$ , se utilizează în același scop transformarea  $z = \log y$ .

Trebuie menționat că transformarea variabilei dependente modifică foarte mult forma modelului inițial. Uneori este convenabil să se aplice o transformare și variabilei independente,  $x$ . Astfel, o transformare dublu logaritmică ( $z = \log y$ ,  $u = \log x$ ) în cazul regresiei monovariabile, transformă o dreaptă care trece prin origine (așa cum este cazul în majoritatea calibrărilor) într-o altă dreaptă de forma:

$$z = a + u \quad (4.31)$$

După aplicarea unei transformări asupra variabilei dependente care conduce la homoscedasticitate, parametrii noului model obținut se estimează prin regresia celor mai mici pătrate aplicate noii variabile,  $z$ .

Pentru ilustrarea proprietăților de homo- și heteroscedasticitate, în figura 4.1 se prezintă trei cazuri diferite de dependență a dispersiei de valoarea răspunsului în cazul regresiei liniare monovariabile.

Revenind la exemplul din paragraful IV.3.1. în care se discută identificarea tipurilor de erori sistematice care afectează metodele noi de analiză, se constată ușor că nu sunt îndeplinite condițiile de aplicare a metodelor de regresie prezentate anterior, respectiv presupunerea că variabila independentă nu este afectată de erori este falsă. Într-adevăr, chiar dacă metoda de referință este exactă, ceea ce implică lipsa erorilor sistematice, rezultatele obținute cu ajutorul său,  $x_i$ , sunt afectate de erori întâmplătoare care nu pot fi neglijate. În cazul în care metoda nouă și cea de referință prezintă precizii foarte asemănătoare, estimarea parametrilor se va realiza prin minimizarea sumei pătratelor distanțelor

$|M_i O_i|$  dintre punctele experimentale și model calculate în direcție perpendiculară pe dreapta de regresie așa cum se sugerează în figura 4.2. Acest caz este mult mai dificil de tratat matematic comparativ cu cele discutate anterior și nu va fi aprofundat în acest capitol.

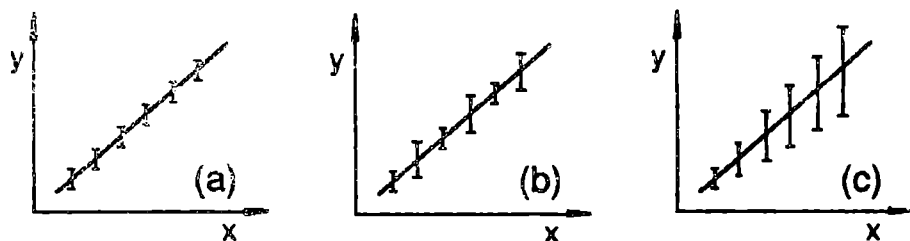


Fig. 4.1. Ilustrarea proprietăților de homo- și heteroscedasticitate. a) Homoscedasticitate; b) Heteroscedasticitate; c) Heteroscedasticitate cu deviație relativă standard constantă.

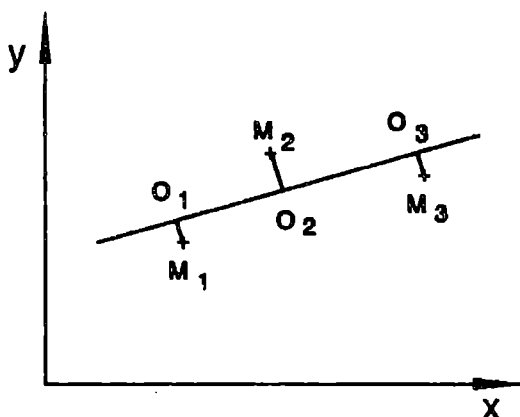


Fig. 4.2. Modul de efectuare a regresiei când variabila independentă,  $x$ , este afectată de erori comparabile cu ale variabilei dependente,  $y$ .

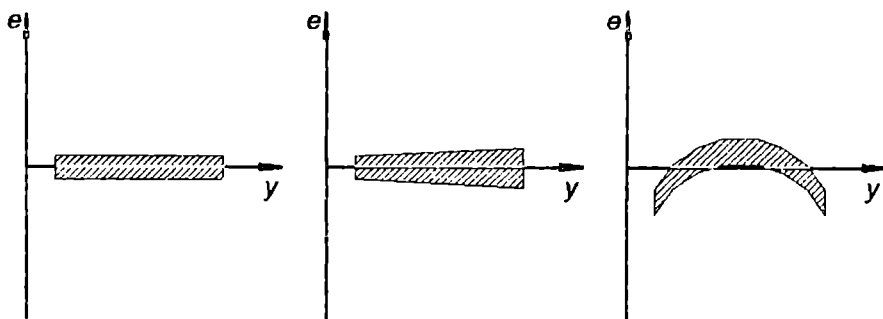
### IV.3.3. Validarea modelului

Această etapă are ca scop testarea concordanței dintre modelul matematic obținut în etapa anterioară și perechile de date experimentale. Dacă nu se identifică o lipsă de concordanță între acestea, presupunerile referitoare la model sunt corecte și acesta descrie bine relația deterministă existentă între răspuns și factorii experimentali

considerații. Doar astfel de modele vor fi utilizate ulterior pentru calculul concentrațiilor de analiz în probe noi sau pentru stabilirea optimului de funcționare al unor sisteme analitice, în cazul în care modelul a fost elaborat într-un experiment de calibrare sau, respectiv, de optimizare.

Pentru validarea modelului se pot utiliza metode grafice sau statistice. Acestea pornesc de la premiza că, dacă modelul este corect, deviațiile dintre datele experimentale și cele calculate pe baza modelului sunt variabile aleatoare cu caracteristici apropiate de a erorilor experimentale care afectează măsurarea valorilor răspunsului. Evident, metodele statistice sunt de preferat, acestea fiind mai obiective. Dar indiferent ce metode se aplică în această etapă, trebuie știut că acestea sunt foarte insensibile când numărul perechilor de date este redus.

După calcularea reziduurilor se poate testa natura distribuției acestora, fie grafic, prin reprezentarea frecvențelor de apariție a acestora, fie statistic, aplicând testul  $\chi^2$  [1]. În cazul în care ipoteza distribuției normale a reziduurilor nu poate fi respinsă, modelul poate fi validat.



**Fig. 4.3.** Exemple de forme de distribuție a reziduurilor.

Reziduurile calculate mai pot fi reprezentate grafic în funcție de valorile estimate, sau măsurate experimental, ale răspunsului ( $y_i$  sau  $\hat{y}_i$ ), sau de valorile factorilor (de exemplu,  $x_i$  - în cazul regresiei monovariabile) și se analizează forma benzii în care acestea sunt dispersate. Următoarele situații, dintre care primele trei sunt sugerate în figura 4.3, sunt mai probabile:

a) Reziduurile se distribuie aleator într-o bandă cu media zero și cu lățime constantă, situație care poate apărea în caz de homoscedasticitate și dacă modelul este corect;

b) Reziduurile se distribuie aleator într-o bandă cu media zero și cu lățime variabilă, această situație putând descrie un caz de heteroscedasticitate, dacă modelul este corect;

c) Reziduurile se distribuie aleator într-o bandă a cărei medie, în unele porțiuni, se depătează în mod sistematic de la valoarea zero, caz în care modelul va fi invalidat;

d) Majoritatea reziduurilor sunt distribuite într-o bandă de forma celor din figura 4.3 (a) și (b), cu excepția unora care se situează mult în afara limitelor benzii. Perechile de date,  $(y_i, x_{ij})$ , care conduc la valori foarte mari ale reziduurilor în comparație cu majoritatea celorlalte, pot fi considerate valori aberante. Acestea se pot elimina din setul de date inițial și apoi se recalculază, mai exact, valorile parametrilor. De asemenea, intervalele de încredere atașate parametrilor modelului se restrâng mult dacă valorile aberante nu sunt considerate în calcule.

Cel mai riguros mod de validare sau invalidare a unui model îl reprezintă aplicarea analizei de varianță (ANOVA) [2, 7], aceasta constituind o tehnică statistică de analiză a măsurătorilor depinzând de mai mulți factori care acționează simultan în vederea estimării importanței și efectelor fiecăruia. Analiza de varianță permite, cu alte cuvinte, să se determine ce parte a variabilității unei populații este cauzată de diverși factori sistematici și, respectiv, de întâmplare.

În continuare se prezintă metoda generală de validare a modelelor de regresie prin analiza de varianță în cazul în care proprietatea de homoscedasticitate este îndeplinită și presupunând că, în etapă de achiziționare a datelor, se repetă de  $n_i$  ori măsurarea răspunsului la fiecare combinație de valori a factorilor  $(x_{ij})$ , pentru  $j = 1, m$ . Notăm cu  $\bar{y}_i$  - media răspunsului la fiecare combinație de valori a factorilor, dată de ecuația 4.27 și cu  $\bar{y}$  - media globală a valorilor determinate ale răspunsului în toate experimentele:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^n \sum_{l=1}^{n_i} y_{il} \quad (4.32)$$

unde  $N$  este numărul total de experimente ( $N = \sum_{i=1}^n n_i$ ).

Variația valorilor lui  $y$  față de media globală,  $\bar{y}$ , este dată de suma de pătrate totală:

$$S_{T^2} = \sum_{i=1}^n \sum_{l=1}^{n_i} (y_{il} - \bar{y})^2 \quad (4.33)$$

Această sumă poate fi descompusă în trei componente, deoarece se poate scrie ecuația următoare:

$$y_{il} - \bar{y} = (y_{il} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \quad (4.34)$$

care introdusă în relația 4.33 conduce, după efectuarea calculului și ținând cont de relațiile existente între variabile, la:

$$SP_T = \sum_{i=1}^n \sum_{l=1}^{n_i} (y_{il} - \bar{y}_i)^2 + \sum_{i=1}^n n_i (\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^n n_i (\hat{y}_i - \bar{y})^2 \quad (4.35)$$

În tabelul 4.3 se prezintă semnificația termenilor sumei totale de pătrate împreună cu modul de efectuare al testului de validare a modelului matematic de regresie prin analiza de varianță.

Tabelul 4.3

Descompunerea sumei de pătrate și modul de validare al modelului de regresie prin analiza de varianță (ANOVA)

Sursa de variație	Suma de pătrate (SP)	Grade libertate	Media sumei de pătrate (MSP)*	Testul statistic F
Datorat regresiei	$(SP_{REG})$ $\sum_{i=1}^n n_i (\hat{y}_i - \bar{y})^2$	1	$MSP_{REG}$	$F = \frac{MSP_{LC}}{MSP_{EE}}$
Variația mediei grupurilor față de modelul de regresie	$(SP_{LC})$ $\sum_{i=1}^n n_i (\bar{y}_i - \hat{y}_i)^2$	$n - 2$	$MSP_{LC}$	
Variația măsurătorilor repetate	$(SP_{FF})$ $\sum_{i=1}^n \sum_{l=1}^{n_i} (y_{il} - \bar{y}_i)^2$	$N - n$	$MSP_{FF}$	
Variația totală	$(SP_T)$ $\sum_{i=1}^n \sum_{l=1}^{n_i} (y_{il} - \bar{y})^2$	$N - 1$		

\* Mediile sumelor de pătrate se obțin făcând câtul dintre valorile din coloanele a doua și a treia.

Ultimul termen al sumei,  $SP_{REG}$ , este de obicei mare comparativ cu ceilalți doi termeni și exprimă variabilitatea valorilor estimate pentru

variabila independentă,  $y$ , pe baza modelului de regresie față de media globală,  $\bar{y}$ .

Primul termen al sumei,  $SP_{EE}$ , exprimă variabilitatea măsurătorilor repetate și este afectată de  $(N - n)$  grade de libertate. Media sumei de pătrate,  $MSP_{EE}$ , definită conform relației din tabel, reprezintă o estimare pură a erorilor experimentale,  $\sigma^2$ .

Termenul al doilea al sumei de pătrate, notat  $SP_{LC}$  și denumit sumă de pătrate datorată lipsei de concordanță, exprimă variabilitatea mediilor valorilor obținute prin repetarea experimentului față de modelul matematic de regresie și este afectat de  $n - 2$  grade de libertate. Media acestui termen,  $MSP_{LC}$ , reprezintă și ea o estimare a varianței erorilor experimentale,  $\sigma^2$ , cu condiția ca modelul teoretic pentru efectuarea regresiei să fie adecvat. Așadar, validarea sau invalidarea modelului de regresie se poate face prin compararea statistică a egalității celor două medii a sumei de pătrate,  $MSP_{LC}$  și  $MSP_{EE}$ , utilizând testul statistic  $F$  [1], așa după cum rezultă din tabelul ANOVA.

Dacă  $F < F_{n, N-n}^{\alpha}$ , unde  $n$  și  $N - n$  reprezintă gradele de libertate corespunzătoare iar  $\alpha$  este nivelul de semnificație al testului, atunci se poate admite, cu probabilitatea  $P = 1 - \alpha$ , că modelul de regresie utilizat exprimă corect relația deterministă dintre variabilele dependentă și independente.

#### IV.3.4. Calculul intervalelor de încredere

Informația necesară determinării intervalelor de încredere a parametrilor este conținută în matricea de varianță - covarianță:

$$V(\mathbf{b}) = \begin{pmatrix} v(b_0) & \text{cov}(b_0, b_1) & \text{cov}(b_0, b_m) \\ \text{cov}(b_1, b_0) & v(b_1) & \text{cov}(b_1, b_m) \\ \text{cov}(b_m, b_0) & & v(b_m) \end{pmatrix} \quad (4.36)$$

unde  $v(b_i)$  reprezintă varianța parametrilor,  $b_i$ , considerând că aceștia reprezintă variabile aleatoare iar  $\text{cov}(b_i, b_j)$  - covarianța parametrilor  $b_i$  și  $b_j$ . Se poate arăta că aceasta are expresia:

$$V(\mathbf{b}) = s_{cc}^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (4.37)$$

în care  $s_{cc}^2$  reprezintă dispersia erorilor experimentale. În cazul în care nu este cunoscută varianța erorilor experimentale, aceasta poate fi estimată prin metoda celor mici pătrate și este cunoscută iar modelul de regresie a fost validat, atunci varianța estimată poate fi exprimată în termenii următorii:



$$s_{cc}^2 = \frac{SP_e}{n - m - 1} \quad (4.38)$$

în care  $SP_e$  reprezintă suma pătratelor tuturor reziduurilor iar  $n$  și  $m$  au semnificațiile anterioare. În acest caz, expresia intervalului de încredere pentru oricare din parametri  $b_j$ , ( $j = 0, m$ ), este:

$$b_j \pm t_{n-m-1}^{1-\frac{\alpha}{2}} \sqrt{v(b_j)} \quad (4.39)$$

unde  $t_{n-m-1}^{1-\frac{\alpha}{2}}$  este valoarea testului Student pentru  $n - m - 1$  grade de libertate și pentru un nivel de semnificație  $\alpha$ , iar  $v(b_j)$  - valoarea elementului de pe linia  $j + 1$  a matricii pătrate de varianță - covarianță,  $V(\mathbf{b})$ .

Trebuie subliniat faptul că ecuația 4.39, deși perfect valabilă pentru oricare parametru individual, nu redă corect expresia domeniului de încredere simultan pentru doi sau mai mulți parametri decât în cazul în care covarianțele acestora, considerați rînd pe rînd doi câte doi, sunt egale cu zero. Modul de calcul general al domeniilor de încredere pentru doi sau mai mulți parametri considerați simultan poate fi găsit în referința [8].

#### IV.4. Concluzii

În acest capitol s-au prezentat algoritmi generali de efectuare a celor mai uzuale tipuri de regresii folosite în analiza chimică și doar s-au sugerat câteva domenii de aplicare a acestora deoarece există numeroase lucrări prezentând aplicații care ilustrează suportul teoretic discutat. Dintre acestea cităm doar câteva [1-3, 7, 9, 10].

Deși algoritmi de regresie prezentați pot părea complicați la prima vedere, în realitate ei se implementează foarte ușor pe calculatoarele personale, mai ales dacă se utilizează pentru aceasta programe profesionale de calcule matematice cum sunt: **Mathematica** al firmei Wolfram Research [11], **Matlab** (MathWorks), **Mathcad** (MathSoft) sau **Slide** (Advanced Graphics Software).

#### Bibliografie

1. C. Litanu și I. Rîcă, **Teoria și metodologia statistică a analizei urmelor**, Editura Scrisul Românesc, Craiova, 1979.

2. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte și L. Kaufman, **Chemometrics: A Textbook**, Elsevier, Amsterdam, 1988.
3. R. Shiavi, **Introduction to Applied Statistical Analysis**, Aksen Publishers, Boston, 1991.
4. S. D. Brown și R. S. Bear, Jr., *Crit. Rev. Anal. Chem.*, **24**, 99, 1993.
5. E. V. Thomas, *Anal. Chem.*, **66**, 795A, 1994.
6. I. Gh. Șabac; P. Cocârlan; O. Stănășilă și A. Topală, **Matematici Speciale**, vol. II, E. D. P., București, 1983.
7. C. Liteanu și I. Rică, **Optimizarea proceselor analitice**, Editura Academiei Române, București, 1985.
8. N. R. Draper și H. Smith, **Applied Regression Analysis**, Wiley, New York, 1981.
9. D. Ceaușescu, **Utilizarea statisticii matematice în chimia analitică**, Editura Tehnică, București, 1982.
10. K. Ebert; H. Ederer și T. L. Isenhour, **Computer Applications in Chemistry. An Introduction for PC Users**, VCH, 1989.
11. S. Wolfram, **Mathematica™. A System for Doing Mathematics by Computer**, Addison-Wesley, New York, 1988.
12. R. Jain și J. S. Schultz, *Anal. Chem.*, **56**, 141, 1984.
13. J. F. Rusling, *Crit. Rev. Anal. Chem.*, **21**, 49, 1989.

# V. NOȚIUNI DE PRELUCRARE A SEMNALELOR ANALITICE

## V.1. Generalități

Prelucrarea semnalelor analitice reprezintă o parte fundamentală a chemometriei care are ca scop extragerea informației conținute în datele chimice obținute experimental prin investigarea sistemelor materiale de interes.

Întrucât, din punct de vedere istoric, prin semnal se înțelege orice set de semne sau simboluri capabile de a transmite informații sau mesaje [1], sub denumirea de "semnale analitice" pot fi grupate toate tipurilor de date obținute în practica analitică. Cel mai frecvent însă, și în sens puțin mai restrâns, literatura publicată sub denumirea generică de "*Signal Processing*" include doar tehnicile matematice de analiză a datelor înregistrate în mod continuu ca funcții care depind de una sau mai multe variabile continue. În acord cu aceasta, prin **semnal** vom înțelege, în acest capitol, doar mulțimi de date cum sunt, de exemplu, cromatogramele, spectrele, înregistrările obținute cu ajutorul analizoarelor continue sau cu instrumente care posedă detectori care operează continuu etc.

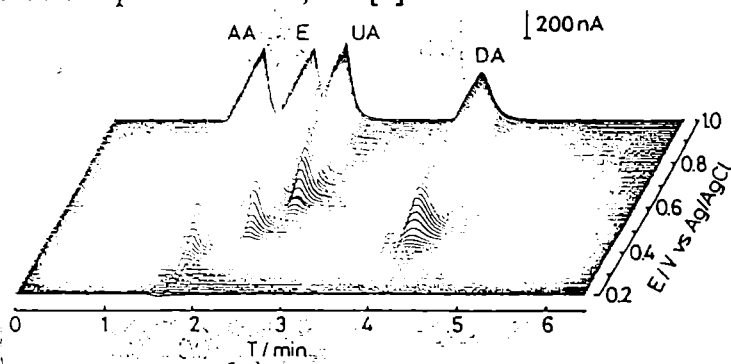
## V.2. Tipuri de semnale

Majoritatea semnalelor de interes științific sunt cuantificabile. Prin **domeniul unui semnal** se înțeleg variabilele independente în funcție de care se face analiza fenomenului investigat. Astfel, ca exemple frecvent utilizate în analiza chimică menționăm: domeniul timpului (întâlnit la achiziționarea semnalelor de către analizoarele continue și cele cromatografice), domeniul spațial în două sau trei dimensiuni (cu aplicații numeroase în analiza imaginilor), domeniul lungimilor și al numerelor de undă (cu aplicabilitate în spectrometrie) etc.

Instrumentele analitice au fost clasificate după numărul variabilelor de care depind semnalele produse de către acestea [2]. Astfel, analizoarele cromatografice și spectrometrele de diferite tipuri au fost denumite instrumente de ordinul întâi deoarece semnalele obținute cu ajutorul lor depind de o singură variabilă (fie timpul, fie lungimea de undă, fie raportul masă/sarcină etc). Spre deosebire de acestea, instrumentele de ordinul doi,

rezultate prin cuplarea în tandem a două instrumente de ordinul întâi, furnizează semnale care admit o reprezentare tridimensională, acestea fiind funcții de două variabile. Cele mai reprezentative tehnici analitice cuplate sunt GC/MS, MS/MS, 2D NMR și GC/FT-IR. De asemenea, tot din această categorie fac parte seturile de senzori (multidiodice, electrozi amperometrici în serie polarizați diferit etc) care au început să fie utilizate în ultimii ani ca multidetectori în sistemele continue de analiză.

Un exemplu de semnal tridimensional este prezentat în figura 5.1. Acesta reprezintă cronovoltamograma componentelor eluați dintr-o coloană cromatografică înregistrată la 80 de valori diferite ale potențialului cu un set de 16 microelectrozi de aur, acestora fiindu-le aplicată o scară de potențial în cinci trepte la fiecare 0,26 s [3].



**Fig. 5.1.** Cronovoltamograma obținută prin HPLC pentru un amestec de acid ascorbic (AA), acid uric (UA), epinefrină (E) și dopamină (DA), toate la concentrația 0,3 mM.

Până în prezent, instrumentele de ordin trei, care furnizează semnale mai greu de vizualizat, sunt relativ puține și se întrebuințează mai rar. Exemple de acest tip sunt măsurătorile de luminiscentă în timp în care se înregistrează intensitatea radiației de luminiscentă în funcție de timp și de lungimile de undă ale acesteea și ale radiației excitatoare ("Time-resolved Luminescence") [4].

În ciuda faptului că interpretarea statistică a semnalelor multidimensionale nu este încă complet elaborată, s-au conștientizat deja avantajele unice ale analizelor efectuate cu instrumente de ordin trei [2], acestea constând în posibilitatea de determinare exactă a componentelor analizate din probe conținând interferenți care nu au fost prezenți în etaloanele utilizate la calibrare.

În acest capitol vor fi prezentate doar tehnicile de prelucrare a semnalelor monovariabile obținute cu instrumente de ordinul întâi dar facem precizarea că acestea se aplică, într-o manieră corespunzătoare, și semnalelor multidimensionale. De asemenea, menționăm că, deși în scopul studierii semnalelor de o singură variabilă se utilizează, de obicei, domeniul timpului, teoriile generale obținute astfel se aplică și pentru semnale înregistrate în alte domenii.

La rândul lor, semnalele monovariabile pot fi clasificate după mai multe criterii. Astfel, după valorile luate de către variabila dependentă, putem menționa: semnale continue (temperatura etc), discontinue (numărul vânzărilor dintr-un produs) și fără amplitudine. Acestea din urmă dau informații despre timpul sau locul de desfășurare al unui fenomen și se numesc procese de punct. Ele au aplicații în codificarea neuronală a informației. De asemenea, după valorile luate de către variabila independentă, se pot distinge: semnale în domeniu continuu, notate  $x(t)$ , sau în domeniu discret, notate  $x(n)$  (vezi paragraful următor).

După formă putem deosebi semnale periodice, care au proprietatea  $x(t) = x(t + P)$ , unde  $P$  este durata unei perioade, și aperiodice iar după durată - infinite sau finite. Dintre acestea, cel mai ușor de descris matematic sunt semnalele periodice cu durată infinită care descriu procese ciclice relativ puțin numeroase. Ca urmare a acestui avantaj, semnalele cu durată finită sunt considerate, în mod frecvent, ca reprezentând o singură perioadă a unor semnale periodice infinite.

După natura proceselor descrise, se disting semnale deterministe și, respectiv, stochastice. Acestea din urmă descriu procesele aleatoare care determină apartența zgomotului de fond conținut întotdeauna în semnalele reale în timp ce semnalele deterministe constituie componenta repetabilă a înregistrărilor obținute la desfășurarea în condiții identice a unui proces.

O altă clasificare a semnalelor poate fi efectuată după valoarea energiei acestora,  $E$ . Pentru semnale cu durată infinită în domeniu de timp continuu, energia se definește prin relația:

$$E = \int_{-\infty}^{+\infty} x^2(t) dt \quad (5.1)$$

iar pentru cele din domeniul discret:

$$E = T \sum_{n=-\infty}^{+\infty} x^2(nT) \quad (5.2)$$

în care  $T$  reprezintă intervalul de prelevare.

Semnalele energetice sunt semnale cu energie finită. Astfel de semnale sunt cele tranziente care prezintă valori diferite de zero într-un interval de timp finit. Alte tipuri de semnale, denumite semnale de putere, au energie infinită. Pentru acestea se definește **puterea**,  $P$ , prin:

$$P = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x^2(t) dt \quad (5.3)$$

în cazul semnalelor cu durată infinită desfășurate în domeniu continuu, și respectiv:

$$P = \lim_{N \rightarrow \infty} \frac{1}{2N + 1} \sum_{n=-N}^{+N} x^2(nT) \quad (5.4)$$

pentru cele în domeniu discret. În ecuația 5.3,  $T$  are o semnificație diferită de cea din ecuațiile 5.2 și 5.4, reprezentând jumătate din perioada totală a semnalului continuu.

### V.3. Eșantionarea semnalelor

Semnalele analitice se desfășoară, de obicei, în domeniul de timp continuu iar pentru prelucrarea lor cu ajutorul calculatoarelor, valorile se achiziționează direct de către acestea. Calculatoarele însă nu admit decât o reprezentare în domeniu de timp discret a semnalelor fiind mai puțin apte de a descrie mulțimi continue de valori. Procesul de transformare a semnalului continuu într-unul discret se realizează printr-o operație de conversie analogic/digitală (A/D) și se numește **eșantionare**. Aceasta este reprezentată schematic în figura 5.2. Semnalul  $f(t)$  se măsoară continuu de către un sensor prevăzut cu un traductor. Traductorul transformă semnalul, de obicei în mod proporțional, într-o mărime electrică,  $x(t)$ , care este apoi aplicată eșantionatorului reprezentat printr-un comutator cu perioada  $T$  și având durata de trecere a semnalului  $\tau$ . Eșantionatorul convertește semnalul electric continuu într-o secvență de timp discret,  $(x(T), x(2T), \dots, x(nT), \dots, x(NT))$ , denumită **serie de timp**. Aceasta conține valorile semnalului la anumite valori ale timpului care sunt separate între ele prin aceeași durată,  $T$ , mai numită și **interval de prelevare**, și se reprezintă simplu astfel:  $x(n)$ ,  $0 \leq n \leq N - 1$ , prin omiterea intervalului de prelevare care se presupune subînțeles.

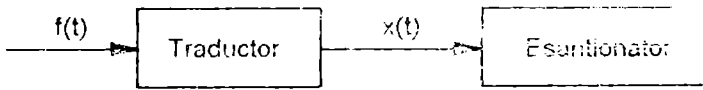


Fig. 5.2. *Reprezentarea schematică a procesului de eșantionare*

După modul de efectuare a conversiei A/D, în literatură [5] sunt menționate: dispozitivul de eșantionare mediator care calculează valoarea medie a semnalului  $x(t)$  pe durata  $\tau$  a impulsului în acord cu relația:

$$x(n) = \frac{1}{\tau} \int_{nT - \tau}^{nT} x(t) dt \quad (5.5)$$

și dispozitivul de eșantionare cu autoblocare care reține valoarea instantanee a semnalului la momentul  $nT$  pe durata,  $\tau$ , necesară efectuării conversiei. Principiul eșantionării cu aceste dispozitive este prezentat în figura 5.3.

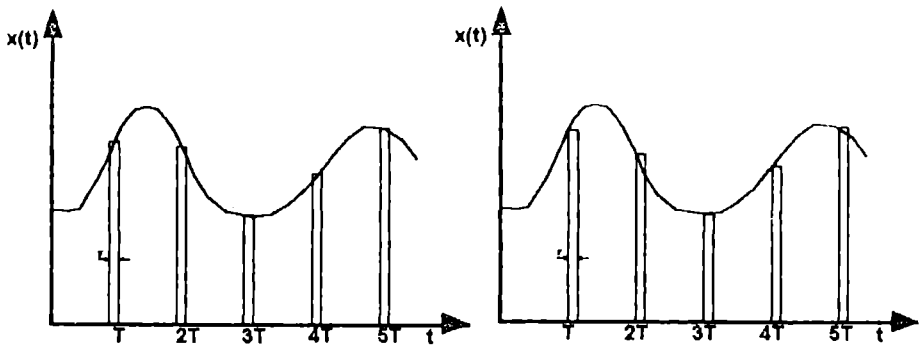


Fig. 5.3. *Eșantionare: a) cu mediere și b) cu autoblocare.*

Operația de eșantionare trebuie să conserve spectrul semnalului. Teorema eșantionării a lui Shannon arată că un semnal poate fi reconstituit din eșantioanele sale cu condiția ca **frecvența de eșantionare**,  $f_e$ , definită ca numărul de valori prelevate într-o perioadă de timp, să fie egală sau mai mare decât dublul celei mai înalte frecvențe,  $f_{\max}$ , conținută în spectrul semnal, adică:

$$f_e \geq 2 f_{\max} \quad (5.6)$$

Convertoarele A/D tipice lucrează la frecvențe de eșantionare de până la 100.000 de prelevări/secundă.

Trebuie menționat că operația de eșantionare produce nu numai discretizarea domeniului variabilei independente, timpul, ci și a domeniului semnalului ca urmare a faptului că memoria calculatorului nu poate reprezenta o mulțime continuă și, prin urmare, infinită de valori. Într-adevăr, memoria fiind compusă din cuvinte în cod binar cu o lungime finită, notată

nu pot fi reprezentate în mod distinct decât  $2^L$  valori. De aceea, domeniul semnalului,  $[x_{\min}, x_{\max}]$ , este divizat, de către convertorul A/D, în intervale de clasă cu lățimea  $\Delta x$  de forma  $(x_{\min} + (i-1)\Delta x, x_{\min} + i\Delta x]$ , unde  $i$  sunt numere întregi variind de la 1 la  $2^L$ . Toate valorile semnalului dintr-un anumit interval sunt approximate prin mijlocul intervalului respectiv fiind reprezentate în memorie prin același cuvânt. În consecință, fiecare valoare a semnalului,  $x(n)$ , conține în sine o eroare de rotunjire,  $\eta(n)$ . Aceste erori sunt distribuite uniform în intervalul  $\left[-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right]$  și au varianța,  $\sigma^2$ , dată de relația:

$$\sigma^2 = \frac{(\Delta x)^2}{12} \quad (5.7)$$

după cum se poate simplu verifica.

Așadar, pe lângă alte tipuri de zgomot conținute în semnal, operația de eșantionare introduce un zgomot suplimentar, denumit **zgomot de cuantificare** care însă poate fi minimizat prin reprezentarea în memorie a valorilor semnalului prin cuvinte cu lungime mai mare deoarece se micșorează astfel erorile de rotunjire prin reducerea lățimii intervalelor de clasă care este dată de relația:

$$\Delta x = \frac{x_{\max} - x_{\min}}{2^L} \quad (5.8)$$

În practică, pentru majoritatea aplicațiilor, o lungime de 12 biți a cuvintelor care stochează valorile semnalelor este suficientă pentru neglijarea zgomotului de cuantificare. Această lungime variază, de obicei, între 4 și 32 de biți. Rareori se utilizează convertoare cu lungime a cuvintelor mai mare, doar în cazurile în care valorile semnalelor se găsesc într-un domeniu foarte larg și este necesară o precizie deosebit de bună. Totuși, în coprocesoarele matematice, care efectuează prelucrări complexe, lungimea cuvintelor poate ajunge până la 80 de biți.



## **V.4. Prelucrarea semnalelor analitice**

Tehnicile matematice utilizate la prelucrarea semnalelor analitice sunt foarte diverse și, dintre acestea, menționăm aici doar câteva: medierea ansamblului, derivarea, transformarea, filtrarea și modularea semnalelor, testarea statistică de ipoteze, modelarea statistică a componentei stochastice a semnalelor, analiza spectrală, convoluții și deconvoluții, interpolări și extrapolări, precum și tehnicile de regresie ale căror principii au fost deja prezentate în capitolul precedent. Ele pot fi aplicate "on-line" de către instrumente specializate sau pot fi implementate, utilizând produse software adecvate, pentru prelucrarea semnalelor cu ajutorul calculatoarelor care, de regulă, sunt cuplate cu instrumentele analitice. Aceste prelucrări pot fi efectuate în timp real sau "off-line".

Cu scopul de a ușura înțelegerea modalităților de aplicare a acestora în analiza chimică, vor fi prezentate mai întâi, într-o manieră succintă și generală, instrumentele matematice de bază utilizate cel mai frecvent la prelucrarea seriilor de timp.

### **V.4.1. Aspecte fundamentale ale prelucrării semnalelor**

Spre deosebire de prelucrarea altor tipuri de date (cum sunt, de exemplu, valorile absorbanțelor unor etaloane utilizate pentru a calcula ecuații de calibrare), în analiza matematică a seriilor de timp trebuie acordată o importanță deosebită ordinei în care sunt colectate valorile semnalelor, deoarece, în majoritatea cazurilor, acestea sunt dependente de valorile anterioare.

Astfel de situații se modelează matematic cu ajutorul lanțurilor stochastice de tip Markov [6] și apar la înregistrarea oricăror semnale deoarece instrumentele de măsură prezintă o inerție de răspuns care se traduce printr-o constantă de timp diferită de zero.

În plus, cu privire la prelucrarea semnalelor, trebuie avute în vedere următoarele două aspecte fundamentale: caracterul dublu, temporal și frecvențial, al semnalelor și contaminarea cu zgomot a semnalelor experimentale.

### V.4.1.1. Caracterul temporal - frecvențial al semnalelor

Fiecare semnal achiziționat în domeniul timpului poate fi considerat ca fiind compus dintr-o sumă de unde sinusoidale sau cosinusoidale denumite **componente spectrale** iar determinarea parametrilor acestora (amplitudine - mai denumită uneori și magnitudine, frecvență și fază) se numește **analiză spectrală**.

În mod paradoxal, deși majoritatea semnalelor sunt obținute în domeniul timpului de către instrumentele analitice, deseori acestea sunt mai sugestive în domeniul frecvențelor. De asemenea, chiar și în natură, percepția senzorială (auzul, văzul) se face în domeniul frecvențelor.

Reprezentările semnalelor în domeniile timpului și al frecvențelor sunt complementare. Astfel, în timp ce domeniul timpului asigură informații despre amplitudinea semnalului la diverse momente de timp, domeniul frecvențelor conține informații despre frecvențele prezente în semnal și magnitudinile lor.

Modul de efectuare al analizei spectrale depinde de tipul semnalelor. Pentru semnale deterministe desfășurate în domeniu de timp continuu se aplică în acest scop fie dezvoltarea în serii Fourier (dacă semnalul este periodic), fie transformata Fourier continuă (pentru semnalele aperiodice). Acestea sunt prezentate în anexa 5.1.

Întrucât calculatoarele electronice nu operează cu funcții continue, în practică se utilizează, pentru analiza spectrală a seriilor de timp, varianta discretă a transformatei Fourier, notată DFT, în care atât domeniul timpului cât și cel al frecvențelor au valori discrete. Pentru aceasta este însă necesar ca numărul valorilor din seria de timp,  $N$ , să fie finit.

DFT este un caz particular al următoarei ecuații generale de transformare a seriilor de timp:

$$X(m) = \sum_{n=0}^{N-1} x(n) K(m, n) \quad (5.9)$$

în care  $K(m, n)$  reprezintă factorul de transformare aplicat seriei de timp. Pentru DFT, acesta are expresia:

$$K(m, n) = \exp(-j2\pi mn / N) \quad (5.10)$$

Considerăm că un semnal continuu,  $x(t)$ , cu durata  $P$ , ( $P = NT$ ), a fost eșantionat cu un interval de prelevare  $T$ , obținându-se seria de timp  $x(n)$ ,  $0 \leq n \leq N - 1$ . În vederea calculării transformatei Fourier discrete a

unui astfel de semnal, domeniul frecvențelor se discretizează în  $N$  valori,  $(f_m)$ ,  $m = 0, N - 1$ , fiecare dintre acestea reprezentând un multiplu întreg al frecvenței de spațiere,  $f_s$ :

$$f_m = m f_s = m \frac{1}{P} \quad (5.11)$$

Una din definițiile perechilor de transformate Fourier discrete [1, 7] este reprezentată prin ecuațiile:

$$X(m) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \exp\left(-\frac{j2\pi mn}{N}\right), \quad 0 < m < N-1 \quad (5.12)$$

și

$$x(n) = \sum_{m=0}^{N-1} X(m) \exp\left(\frac{j2\pi mn}{N}\right), \quad 0 < n < N-1 \quad (5.13)$$

În primele din aceste ecuații, prin  $X(m)$  s-a notat valoarea transformatei Fourier directe în punctul  $f = f_m$  iar ecuațiile 5.13, care permit calculul semnalului în domeniul de timp discret cunoscând caracteristicile de frecvență ale acestuia, exprimă transformata Fourier inversă.

Se observă că efectul discretizării domeniului frecvențelor constă în transformarea semnalului inițial finit într-unul infinit și periodic cu perioada  $P$ , ceea ce este echivalent cu:

$$x(n) = x(n + kN), \text{ pentru } k \in \mathbb{Z} \quad (5.14)$$

Calculul transformatei Fourier discrete utilizând ecuațiile 5.12 și 5.13 este foarte laborios deoarece fiecare set de ecuații implică un număr de  $N^2$  operații de multiplicare a unor numere complexe, ceea ce, pentru  $N$  mare, poate necesita un timp de calcul foarte lung chiar dacă acesta se efectuează cu calculatoare performante. Algoritmul Cooley-Tukey (1965) de calcul rapid al transformatei Fourier, a cărui descriere poate fi găsită în lucrarea [8], reduce semnificativ timpul de calcul deoarece numărul de operații necesare este mult mai mic, fiind egal cu  $N \log_2(N)$ . Astfel, aplicarea sa pentru  $N = 1024$  este de cca 100 de ori mai rapidă. Singura condiție necesară în algoritmul Cooley-Tukey este ca numărul valorilor din seria de timp să fie o putere naturală a lui 2 ( $N = 2^l$ ).

În ecuațiile 5.12 și 5.13, numărul valorilor din semnal nu prezintă restricții. Cu toate acestea, considerarea unor valori pare pentru  $N$  prezintă unele avantaje în interpretarea DFT după cum va rezulta în continuare. De aceea, mai departe, această condiție o presupunem îndeplinită.

Cunoașterea detaliată a proprietăților DFT este necesară pentru a minimiza numeroasele tipuri de erori care pot să apară la calcularea sa și

deoarece multe tehnici de prelucrare a semnalelor analitice sunt bazate pe ele. Aceste proprietăți pot fi rezumate prin următoarele teoreme ale căror demonstrații sunt imediate:

**Liniaritatea.** DFT a unei combinații liniare a două semnale,  $x(n)$  și  $y(n)$ , prelevate la aceleași valori ale timpului, este egală cu aceeași combinație liniară a transformatei Fourier corespunzătoare sau, în formă simbolică, dacă  $x(n) \Leftrightarrow X(m)$  și  $y(n) \Leftrightarrow Y(m)$ , atunci  $a x(n) + b y(n) \Leftrightarrow a X(m) + b Y(m)$ , pentru orice valori complexe ale lui  $a$  și  $b$ .

**Periodicitatea.** DFT a oricărei serii de timp care se produce într-un domeniu de timp finit este o funcție periodică cu perioada egală cu  $N f_s$ , ceea ce se exprimă printr-o ecuație similară cu 5.14, și anume:

$$X(m) = X(m + k N), \text{ pentru } k \in Z \quad (5.15)$$

**Simetria conjugată.** În domeniul frecvențelor, componentele simetrice ale transformatei Fourier discrete ale oricărui semnal,  $x(n)$ , reprezintă valori complex conjugate, adică pentru orice  $m$  ( $1 < m < N/2 - 1$ ):

$$X(N - m) = X^*(m) \quad (5.16)$$

**Teorema de deplasare în domeniul timpului.** Dacă  $x(n)$  și  $X(m)$  sunt două perechi de transformate Fourier discrete, atunci transformata Fourier directă a unui semnal  $y(n) = x(n - k)$  se calculează cu relația:

$$Y(m) = X(m) \exp\left(-j \frac{2\pi m k}{N}\right) \quad (5.17)$$

**Convoluția în domeniul frecvenței.** Fiind date semnalele  $x(n)$  și  $y(n)$  prelevate la aceleași valori ale timpului și transformatele lor Fourier,  $X(m)$  și  $Y(m)$ , transformata Fourier discretă a produsului celor două semnale se calculează cu relația:

$$\text{DFT}[x(n) y(n)] = \sum_{k=0}^{N-1} Y(k) X(m - k) \quad (5.18)$$

și se notează  $X(m) * Y(m)$ .

**Convoluția în domeniul timpului.** Fiind date semnalele frecvențiale  $X(m)$  și  $Y(m)$  cunoscute la aceleași frecvențe și transformatele lor Fourier,  $x(n)$  și  $y(n)$ , transformata Fourier discretă a produsului semnalelor frecvențiale se calculează cu relația următoare:

$$\text{DFT}[X(m) Y(m)] = \frac{1}{N} \sum_{k=0}^{N-1} x(k) y(n - k) \quad (5.19)$$

și se notează  $x(n) * y(n)$ .

Ultimele două teoreme pot fi reprezentate simbolice astfel:  $DFT[x(n) y(n)] \Leftrightarrow X(m) * Y(m)$ , și  $DFT[X(m) Y(m)] \Leftrightarrow x(n) * y(n)$ . Sensul lor poate fi rezumat spunând că procesul de convoluție a două semnale într-un domeniu echivalează cu produsul scalar al semnalelor în domeniul transformat. Relațiile 5.18 și 5.19 se numesc ecuații de convoluție în domeniul frecvențelor și, respectiv, al timpului.

Teoremele de mai sus prezintă câteva consecințe importante. Astfel, teorema de simetrie conjugată arată că, dintre cele  $N$  componente ale  $DFT[x(n)]$ , doar  $N/2 + 1$  sunt independente, și anume  $X(m)$  pentru  $0 < m < N/2$ , motiv pentru care unele programe de calcul furnizează doar valorile acestora. Componentele pentru  $N/2 + 1 < m < N - 1$ , exprimă o informație redundantă, deoarece, după cum arată ecuațiile 5.16, acestea se confundă cu valorile  $X(m)$  pentru  $1 < m < N/2 - 1$ . În capitolul prezent, efectul acesta, ("*folding*" în limba engleză), îl vom denumi **redundanță**.

Componentele  $X(N/2)$  și  $X(0)$  au întotdeauna valori reale iar ultima dintre acestea, termenul DC, este egală cu media valorilor semnalului din domeniul timpului.

Datorită redundanței, frecvențele  $f_m$  pentru  $N/2 + 1 < m < N - 1$  nu se deosebesc de cele pentru  $1 < m < N/2 - 1$ . Pentru a dovedi acest lucru, observăm că, datorită periodicității și simetriei conjugate, au loc relațiile:

$$X(N - m) = X(-m) = X^*(m) \quad (5.20)$$

pentru fiecare  $m$ ,  $0 < m < N/2 - 1$ . Ultima parte a ecuațiilor precedente este similară cu ecuația A1.31 și cu relația dintre coeficienții seriilor Fourier. Ecuațiile 5.13 pot fi rescrise astfel:

$$x(n) = \sum_{m=-N/2+1}^{N/2} X(m) \exp\left(\frac{j2\pi mn}{N}\right), \quad 0 < n < N-1 \quad (5.21)$$

și, mai departe, prin analogie cu seriile Fourier (vezi anexa 5.1):

$$x(n) = C_0 + \sum_{m=1}^{N/2} C_m \cos(2\pi f_m nT + \theta_m), \quad 0 < n < N-1 \quad (5.22)$$

unde  $C_0 = X(0)$ ;  $C_m = |X(m)|$ ;  $\theta_m = -\arctan \frac{\text{Im}[X(m)]}{\text{Re}[X(m)]}$ , iar restul mărimilor au semnificațiile cunoscute. Ecuațiile 5.22 dovedesc prezența în spectrul semnalului obținut prin aplicarea DFT a numai  $N/2$  frecvențe discrete, în afara termenului DC.

Cea mai mare frecvență neredundantă din semnal apare pentru  $m = N/2$ . Ea este jumătate din frecvența de prelevare și se numește **frecvență Nyquist**. DFT este definită în mod unic doar pentru semnale conținând frecvențe mai mici decât această frecvență și, deci, având un spectru de bandă limitat, fapt concordant cu teorema de eșantionare a lui Shannon.

Spectrul din regiunea frecvențelor negative este complexul conjugat al celui din regiunea pozitivă. Aceasta poate fi observat și în figura 5.4 în care sunt prezentate atât valorile reale și imaginare ale componentelor DFT, cât și spectrele de magnitudine și de fază corespunzătoare unui semnal gaussian care simulează un pic cromatografic. Se remarcă că semnalul analitic prezentat, care simulează componenta deterministă a unor semnale înregistrate în mod real, are cea mai mare parte a energiei repartizată pe un număr redus de frecvențe situate în vecinătatea frecvenței zero (vezi fig. 5.4.d).

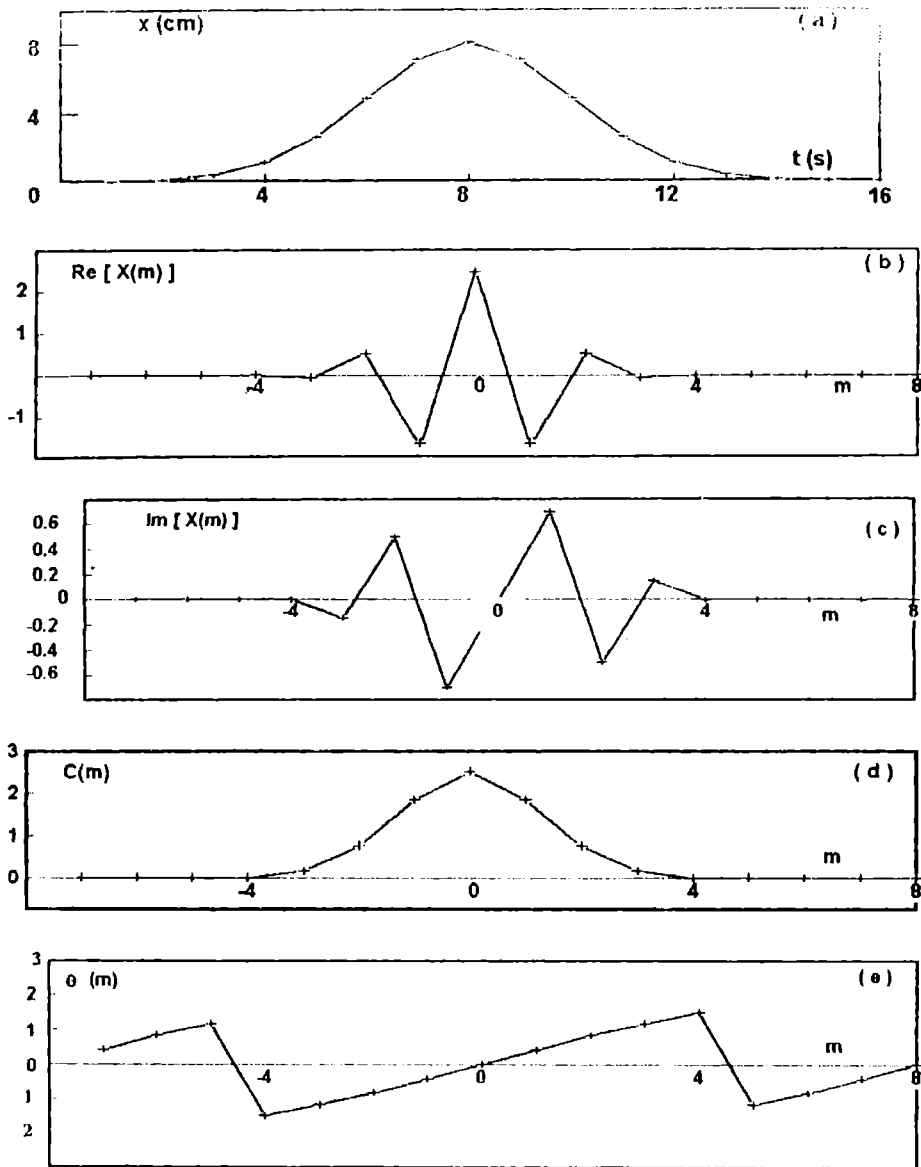
O caracteristică a DFT constă în aceea că ea asigură valori ale lui  $X(f)$  doar la anumite frecvențe. Aceasta nu are consecințe grave în cazul unui spectru continuu. Totuși, dacă o valoare importantă din spectru, cum ar fi, de exemplu, maximul unui pic foarte ascuțit, există la o frecvență  $f \neq m f_s$ , aceasta nu va putea fi pusă în evidență. Această eroare ("*picket fence effect*") se numește **efect de grilă**. Ea se poate înlătura fie prin mărirea frecvenței de achiziție a semnalului în aceeași durată de timp, fapt nu întotdeauna posibil de realizat, fie prin completarea semnalului cu un anumit număr de zerouri. Cea de a doua metodă constituie, de fapt, o metodă de interpolare bazată pe extinderea arbitrară a duratei semnalului.

Dacă la sfârșitul unui semnal cu durata  $NT$  se adaugă  $M$  zerouri, se obține un semnal nou cu durata  $LT = (M + N)T$ . La calculul DFT al noului semnal, frecvența de spațiere,  $f'_s$ , dată de relația:

$$f'_s = \frac{1}{LT} = \frac{1}{(M + N)T} \quad (5.23)$$

se va micșora, ceea ce va permite evidențierea unui număr mai mare de frecvențe.

Completarea cu zerouri nu modifică forma de ansamblu a spectrului. Ea se aplică și la utilizarea algoritmilor de calcul rapid al DFT în cazurile în care numărul valorilor din semnale diferă de o putere naturală a lui 2.



**Fig. 5.4.** Semnal discret (a) obținut prin eșantionarea ( $N = 16$ ,  $T = 1$  s) unui pic cromatografic simulat având următorii parametri: înălțimea maximă - 8 cm; poziția maximului - 8 s; varianța -  $4 \text{ s}^2$ ; Valorile reale (b) și imaginare (c) ale spectrului semnalului; Spectrele de magnitudine (d) și de fază (e) ale semnalului.

O sursă majoră de erori la calcularea DFT apare ca o consecință a trunchierii semnalelor care este întotdeauna prezentă la prelevarea unui semnal cu durată finită dintr-un semnal cu o durată teoretic infinită.

Trunchierea poate fi descrisă matematic considerând că semnalul real,  $x(n)$ , se multiplică cu o fereastră de date dreptunghiulară,  $d_R(n)$ , pentru a obține semnalul prelevat,  $y(n)$ :

$$y(n) = x(n) d_R(n) \quad (5.24)$$

$$\text{unde: } d_R(n) = \begin{cases} 1 & \text{pentru } 0 < n < N-1 \\ 0 & \text{pentru alți } n \end{cases} \quad (5.25)$$

DFT[ $y(n)$ ] diferă de DFT[ $x(n)$ ] datorită teoremei de convoluție:

$$Y(m) = X(m) * D_R(m) \quad (5.26)$$

unde:  $y(n) \Leftrightarrow Y(m)$ ;  $x(n) \Leftrightarrow X(m)$  și  $d_R(n) \Leftrightarrow D_R(m)$ . Genul acesta de eroare ("leakage error") se numește eroare de dispersare.

Pentru a vedea care sunt consecințele erorii de dispersare, considerăm două semnale în domeniul continuu:  $[x_i(t)]$ , ( $i = 1, 2$ ), definite prin:

$$x_i(t) = \cos(2\pi f_{i0}t) \quad (5.27)$$

cu  $f_{10} = 1/8$  și  $f_{20} = 1/9$ . Spectrele Fourier ale acestora, obținute prin aplicarea CFT, sunt reprezentate în figura 5.5 și au expresia:

$$X_i(f) = 0,5 \delta(f - f_{i0}) + 0,5 \delta(f + f_{i0}), \quad i = 1, 2 \quad (5.28)$$

în care  $\delta$  semnifică funcția lui Dirac.

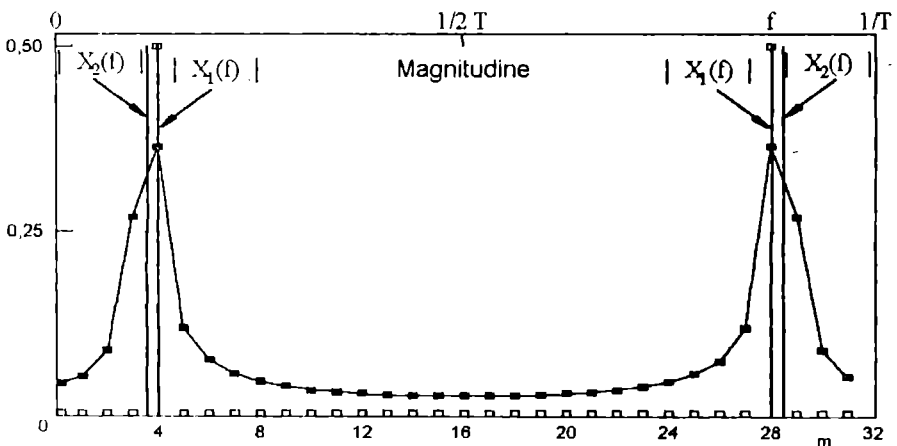


Fig. 5.5. Spectrele Fourier de magnitudine continue,  $|X_i(f)|$ , și discrete,  $|X_i(m)|$ , pentru cele două semnale sinusoidale definite în text.  $\square$  pentru  $|X_1(m)|$  și  $\blacksquare$  pentru  $|X_2(m)|$ .



Considerăm că semnalele  $x_i(t)$  sunt discretizate, începând cu momentul  $t = 0$ , în  $N = 32$  de valori cu o frecvență de eșantionare  $f_c = 1/NT = 1/32$ , obținându-se semnalele  $x_i(n)$ ,  $0 \leq n \leq 31$ . Transformatele Fourier discrete ale acestor două semnale,  $X_i(m)$ , sunt de asemenea prezentate în figura 5.5. Se observă că  $X_1(m)$  prezintă, pentru  $m = 4$  și  $m = 28$ , valorile 0,5, ceea ce este corect, semnificând prezența în semnal a frecvenței  $f = 4 f_c = 1/8$  cu magnitudinea 1 ( $2 \times 0.5$ ). Pentru al doilea semnal însă, deoarece  $f_{20} \neq m f_c$ , apare o distorsionare importantă a spectrului magnitudinilor calculat aplicând DFT, ceea ce este o consecință directă a trunchierii semnalului.

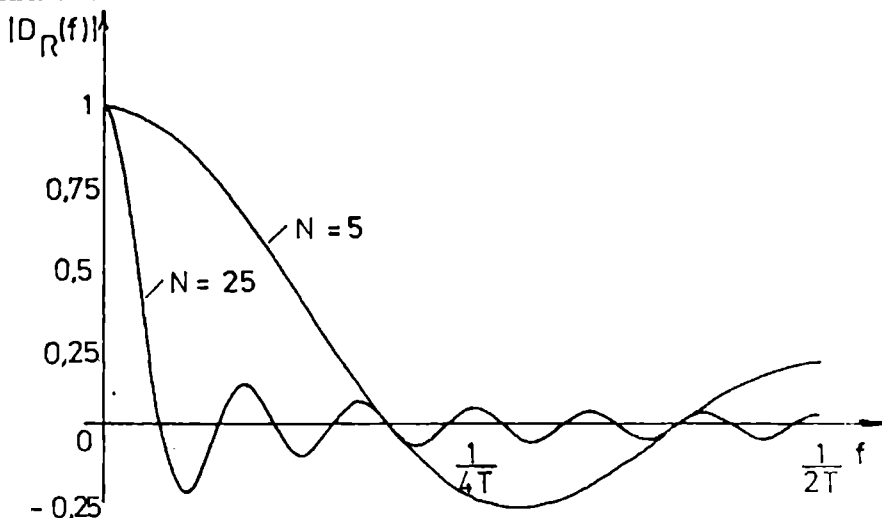


Fig. 5.6. Spectrul de magnitudine al ferestrei de date dreptunghiulare pentru două valori diferite ale numărului valorilor din semnal.

Eroarea de dispersare este inacceptabilă. Minimizarea sa poate fi înțeleasă analizând forma factorului de convoluție,  $D_R(m)$ , din ecuația (5.26) mai denumit **fereastră spectrală dreptunghiulară**. În domeniul de frecvențe continuu, ecuația sa este:

$$D_R(f) = \frac{\sin(\pi fNT)}{\sin(\pi fT)} \exp(-j\pi fNT) \quad (5.29)$$

iar spectrul său Fourier, care este o funcție pară, este prezentat în figura 5.6 pentru două valori diferite ale lui  $N$  doar în domeniul frecvențelor pozitive. Se observă că există multe intersecții ale axei frecvențelor iar porțiunile dintre două intersecții consecutive se numesc **lobi**. Dintre

acestia, lobul centrat în jurul frecvenței zero reprezintă lobul principal iar ceilalți - lobi secundari.

Operația de convoluție dispersează efectul de ponderare al lobilor secundari asupra întregului domeniu de frecvențe. Se poate dovedi că din ecuația (5.26) rezultă  $X(m) = Y(m)$  doar dacă fereastra spectrală de convoluție este identică cu funcția lui Dirac care descrie un pic frecvențial infinit de îngust. Din figura 5.6 se observă că pentru  $N$  mai mare,  $D_R(f)$  prezintă un lob principal mai apropiat de forma funcției Dirac fiind mai îngust. Așadar, o modalitate de micșorare a erorii de dispersare constă în mărirea lui  $N$ , ceea ce însă nu este întotdeauna posibil. De aceea, în practică, se recurge la o altă metodă, și anume la modificarea formei ferestrei de date,  $d(n)$ , cu scopul de a obține ferestre spectrale,  $D(m)$ , ale căror lobi secundari să conțină mult mai puțină energie decât lobul principal.

În literatură au fost propuse peste 20 de ferestre de date diferite [9]. Dintre acestea, figura 5.7 prezintă doar fereastra dreptunghiulară deja discutată și fereastra triunghiulară a lui Bartlett împreună cu ferestrele spectrale corespunzătoare pentru o aceeași valoare a lui  $N$ .

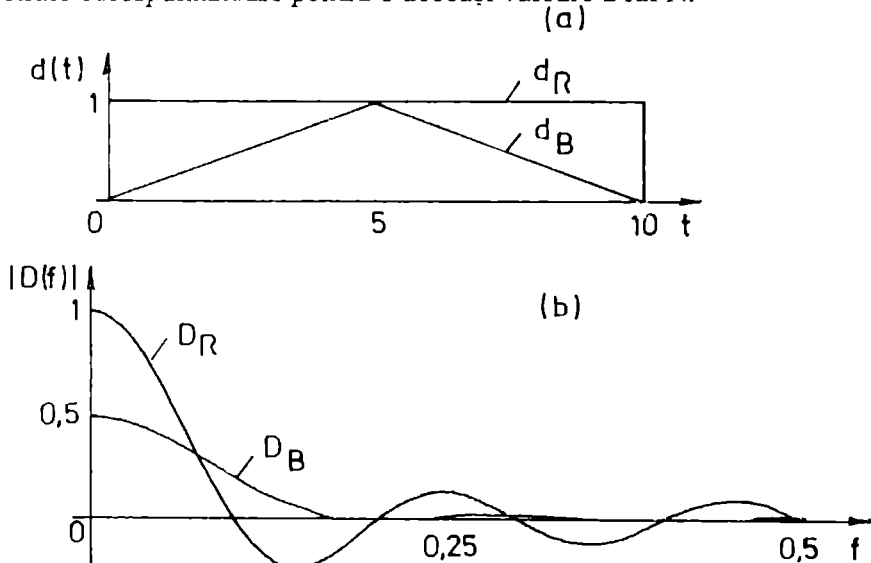


Fig. 5.7. Forma ferestrelor de date dreptunghiulară și triunghiulară (a) și a ferestrelor spectrale corespunzătoare (b).

Caracteristicile principale care diferențiază ferestrele spectrale sunt nivelul lobului principal, definit ca raportul dintre magnitudinile lobului principal și al celui mai mare lob secundar, și lățimea lobului principal

care reprezintă lățimea domeniului de frecvențe al acestuia. Aceste caracteristici, împreună cu ecuațiile ferestrelor de date și spectrale corespunzătoare, sunt prezentate în tabelul 5.1.

Tabelul 5.1  
Două ferestre de date și caracteristicile  
ferestrelor spectrale corespunzătoare

Fereastra de date $d(n)$	Fereastra spectrală $D(f)$	Lățimea lobului principal	Nivelul lobului principal
Dreptunghiulară $d_R(n) = 1, 0 \leq n \leq N-1$	$D_R(f) = T \frac{\sin(\pi fNT)}{\sin(\pi fT)} \times \exp[-j\pi fT(N-1)]$	$\frac{2}{NT}$	22,4 %
Triunghiulară-Bartlett $d_B(n) = \begin{cases} \frac{2n}{N-1}, & 0 \leq n \leq (N-1)/2 \\ 2 - \frac{2n}{N-1}, & (N-1)/2 \leq n \leq N-1 \end{cases}$	$D_B(f) = \frac{2T}{N} \times \left( \frac{\sin(\pi fNT/2)}{\sin(\pi fT)} \right)^2$	$\frac{4}{NT}$	4,5 %

Atât din tabelul 5.1, cât și din figura 5.7 rezultă că micșorarea nivelului lobului principal este însoțită de creșterea lățimii acestuia, ceea ce are ca efect scăderea rezoluției. Într-adevăr, simulările pe calculator [1] ale DFT pentru un semnal de forma:

$$x(n) = \cos(2\pi f_1 nT) + \cos[2\pi(f_1 + \Delta f)nT] \quad (5.30)$$

utilizând diferite ferestre de date, arată că cele două frecvențe conținute în semnal,  $f_1$  și  $f_1 + \Delta f$ , sunt evidențiate în mod distinct în spectrul Fourier doar dacă  $\Delta f$  este mai mică decât jumătate din lățimea lobului principal al ferestrei spectrale corespunzătoare ferestrei de date utilizate.

Alegerea ferestrei de date optime depinde de scopul urmărit. În general, utilizarea unor ferestre asemănătoare cu cea triunghiulară face ca spectrul Fourier al semnalului prelevat să fie mult mai apropiat, ca formă, de cel al semnalului real. Totuși, ca urmare a diminuării amplitudinilor valorilor semnalului din domeniul timpului, în astfel de cazuri se pierde o parte din energia semnalului, fenomenul acesta fiind denumit pierdere de procesare. El poate fi corectat prin multiplicarea energiei calculate cu un factor care depinde de fereastra de date utilizată.

#### V.4.1.2. Contaminarea cu zgomot a semnalelor

Interpretarea semnalelor analitice este mult îngreunată de faptul că înregistrările obținute la monitorizarea unor procese fizico-chimice reale se compun întotdeauna dintr-un semnal stohastic denumit **zgomot** suprapus peste semnalul determinist de interes. Tehnicile prin care se înlătură parțial consecințele negative ale contaminării cu zgomot a semnalelor reale asupra estimării corecte a parametrilor fizico-chimici de interes sunt fundamentate pe metode matematice care asimilează zgomotul cu procese aleatoare tratate într-o manieră specifică. Noțiuni privind modul de reprezentare și de caracterizare al proceselor aleatoare sunt prezentate succint în continuare cu scopul de a facilita înțelegerea proprietăților zgomotului din sistemele analitice de măsură și a tehnicilor de combatere ale acestuia.

Un proces aleator reprezintă o variabilă aleatoare dependentă de timp și poate fi descris printr-un semnal stohastic pe care îl vom nota  $x(t)$  în domeniu de timp continuu și, respectiv,  $x(n)$  în domeniu discret. Deși aceste notații sunt identice cu cele utilizate în mod uzual și pentru semnale deterministe, proprietățile celor două tipuri de semnale nu trebuie confundate. Diferențele majore provin din faptul că semnalele deterministe sunt repetabile pe când cele stohastice nu. De aceea, caracterizarea acestora din urmă nu este posibilă decât în termeni probabilistici pe baza unui **ansamblu** format dintr-un număr mare de înregistrări a procesului aleator, fiecare dintre acestea fiind denumită **funcție de probă**.

Ansamblul poate fi descris cu o funcție bivariabilă,  $x(t, \zeta_n)$ , în care argumentul cu valori discrete și naturale,  $\zeta_n$ , reprezintă numărul înregistrării. În anexa 5.2 sunt prezentate noțiuni de simulare a proceselor aleatoare necesare la testarea metodelor de prelucrare a semnalelor stohastice și la generarea de semnale cu proprietăți cunoscute iar în anexa 5.3 este simulat un ansamblu format din 16 funcții de probă reprezentând un pic de formă gaussiană contaminat cu zgomot.

La fiecare valoare a timpului,  $t_0$ , valorile funcțiilor de probă,  $x(t_0, 1)$ ,  $x(t_0, 2)$ ,  $x(t_0, 3)$ , etc.), reprezintă o variabilă aleatoare. Notând cu  $F_{x(t_0)}(\alpha)$  funcția de distribuție a acesteea, care reprezintă probabilitatea ca la momentul  $t_0$  semnalul să aibă o valoare mai mică decât  $\alpha$ , distribuția variabilei aleatoare  $x(t_0, \zeta_n)$  se caracterizează cu o funcție de

probabilitate de ordinul unu,  $f_{x(t_0)}(\alpha)$ , definită similar cu cea pentru variabilele aleatoare obișnuite:

$$f_{x(t_0)}(\alpha) = \frac{dF_{x(t_0)}(\alpha)}{d\alpha} \quad (5.31)$$

În cazul considerării a două momente de timp,  $t_1$  și  $t_2$ , se definește funcția de probabilitate de ordinul doi,  $f_{x(t_1)x(t_2)}(\alpha_1, \alpha_2)$ , iar în cazul mai multor valori ale timpului - funcții de probabilitate de ordin superior.

Utilizând funcțiile de probabilitate, cu relațiile următoare se definesc funcțiile de moment ale ansamblului:

$$m_{x(t_0)} = E[x(t_0)] = \int_{-\infty}^{\infty} \alpha f_{x(t_0)}(\alpha) d\alpha \quad (5.32)$$

$$\sigma_{x(t_0)}^2 = E[(x(t_0) - m_{x(t_0)})^2] = \int_{-\infty}^{\infty} (\alpha - m_{x(t_0)})^2 f_{x(t_0)}(\alpha) d\alpha \quad (5.33)$$

$$\begin{aligned} \varphi_{x(t_1, t_2)} &= E[x(t_1) x(t_2)] = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \alpha_1 \alpha_2 f_{x(t_1)x(t_2)}(\alpha_1, \alpha_2) d\alpha_1 d\alpha_2 \end{aligned} \quad (5.34)$$

$$\begin{aligned} \gamma_{x(t_1, t_2)} &= E[(x(t_1) - m_{x(t_1)})(x(t_2) - m_{x(t_2)})] = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\alpha_1 - m_{x(t_1)})(\alpha_2 - m_{x(t_2)}) f_{x(t_1)x(t_2)}(\alpha_1, \alpha_2) d\alpha_1 d\alpha_2 \end{aligned} \quad (5.35)$$

$$\xi_{x(t_1, t_2)} = \frac{\varphi_{x(t_1, t_2)} - m_{x(t_1)} m_{x(t_2)}}{\sigma_{x(t_1)} \sigma_{x(t_2)}} = \frac{\gamma_{x(t_1, t_2)}}{\sigma_{x(t_1)} \sigma_{x(t_2)}} \quad (5.36)$$

Media,  $m_{x(t_0)}$ , și varianța,  $\sigma_{x(t_0)}^2$ , ansamblului reprezintă valorile de așteptare (sau cele mai probabile) ale semnalului și, respectiv, ale pătratului deviației semnalului de la medie la  $t = t_0$ . Ele sunt funcții de un singur moment de timp și se numesc, de aceea, momente de ordinul întâi spre deosebire de funcțiile de autocorelație (ACF), de autocovarianță (ACVF) și de autocovarianță normalizată (NACVF), notate  $\varphi_{x(t_1, t_2)}$ ,  $\gamma_{x(t_1, t_2)}$  și, respectiv,  $\xi_{x(t_1, t_2)}$ , care depind de două momente de timp,  $t_1$  și  $t_2$ , și sunt numite, de aceea, momente de ordinul doi. ACF este valoarea de așteptare a produsului semnalelor la momentele  $t_1$  și  $t_2$ , ACVF semnifică valoarea de așteptare a produsului deviațiilor de la medie a

semnalului la momentele considerate iar NACVF reprezintă ACVF normalizată în raport cu produsul deviațiilor standard ale semnalului.

Funcțiile de moment reprezintă caracteristici fundamentale cu ajutorul cărora se descriu procesele aleatoare în domeniul timpului. Analiza semnalelor are ca scop determinarea acestora și a funcțiilor de probabilitate. Utilizarea în acest scop a unui ansamblu format dintr-un număr mare de funcții de probă este practic nefezabilă. Este preferabil ca aceste determinări să se efectueze pe baza unei singure funcții de probă discrete. Condițiile în care poate fi făcut acest lucru sunt rezumate în teoremele de ergodicitate. Câteva concluzii ale acestora, de importanță fundamentală în prelucrarea semnalelor analitice, sunt prezentate sumar în continuare. Fiind vorba despre o singură funcție de probă din cadrul ansamblului, argumentul  $\zeta_n$  va fi omis.

Un proces aleator prezintă staționaritate de ordinul întâi dacă funcția de probabilitate de ordinul unu nu depinde de timp ceea ce presupune respectarea relației:

$$f_{x(t_1)}(\alpha) = f_{x(t_2)}(\alpha) \quad (5.37)$$

pentru oricare două momente de timp,  $t_1$  și  $t_2$ , arbitrare iar staționaritatea de ordinul doi are loc dacă:

$$f_{x(t_1+\theta)x(t_2+\theta)}(\alpha_1, \alpha_2) = f_{x(t_1)x(t_2)}(\alpha_1, \alpha_2) \quad (5.38)$$

pentru oricare  $t_1$  și  $t_2$  și pentru orice interval de timp,  $\theta$ .

Procesele aleatoare reale prezintă de obicei staționaritate în sens larg, aceasta implicând staționaritatea momentelor de ordinul unu și doi și, deci, satisfacerea ecuațiilor 5.37 și 5.38. Doar acest tip de procese vor fi considerate în continuare. Pentru acestea, mediă și varianța nu depind de timp acestea notându-se de aceea  $m_x$  și  $\sigma_x^2$ . Totodată, în aceste condiții, funcțiile de moment de ordinul doi depind doar de diferența dintre momentele de timp,  $t_1 - t_2$ , denumită timp de lag și notată cu  $\tau$ . Dintre acestea din urmă, funcțiile de autocovarianță, care indică orice dependență liniară între valorile procesului aleator separate printr-un anumit timp de lag, au proprietăți similare. Pentru ACVF acestea sunt următoarele:

- 1)  $\gamma_x(0) = \sigma_x^2$ ;

- 2) ACVF este o funcție pară:  $\gamma(\tau) = \gamma(-\tau)$ ;

- 3) Dacă  $\gamma_x(\tau) = \gamma_x(\tau + P)$  pentru toți  $\tau$ , atunci  $x(t)$  este o funcție periodică cu perioada  $P$ ;

$$\text{și } 4) |\gamma_{x(\tau)}| \leq \gamma_{x(0)}.$$

Pentru NACVF, prima și ultima proprietate devin:  $\xi_{x(0)} = 1$  și, respectiv,  $|\xi_{x(\tau)}| \leq 1$ , această funcție având, pentru un timp de lag stabil, semnificația unui coeficient de corelație a zgomotului care afectează semnalul la momente de timp separate prin timpul de lag considerat.

În tabelul 5.2 se arată expresiile momentelor de ordinul unu și doi pentru un proces aleator infinit, înregistrat în domeniu de timp continuu,  $x(t)$ , și, respectiv, discret,  $x(n)$ , precum și ale unor estimatori ai acestora obținuți pe baza unei înregistrări finite a semnalului,  $x(n)$ ,  $n=1, N$ . Indicele  $x$  din dreapta jos a momentelor a fost omis, semnalul fiind subînțeles.

Tabelul 5.2

Expresiile momentelor de ordinul unu și doi pentru un proces aleator staționar și pentru unii estimatori ai acestora

Momentul	Expresii pentru momentele unui semnal		Estimatori
	Continuu	Discret	
Media	$m = \lim_{P \rightarrow \infty} \frac{1}{2P} \int_{-P}^P x(t) dt$	$m = \lim_{n \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x(n)$	$\hat{m}_N = \frac{1}{N} \sum_{n=1}^N x(n)$
Varianța	$\sigma^2 = \lim_{P \rightarrow \infty} \frac{1}{2P} \int_{-P}^P (x(t)-m)^2 dt$	$\sigma^2 = \lim_{n \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N (x(n)-m)^2$	$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{n=1}^N (x(n)-\hat{m}_N)^2$
ACF	$\rho(\tau) = \lim_{P \rightarrow \infty} \frac{1}{2P} \int_{-P}^P x(t)x(t+\tau) dt$	$R(k) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x(n)x(n+k)$	$\hat{R}(k) = \frac{1}{N-k} \sum_{n=1}^{N-k} x(n)x(n+k)$ $\hat{R}'(k) = \frac{1}{N-k} \sum_{n=1}^{N-k} x(n)x(n+k)$
ACVF	$\gamma(\tau) = \rho(\tau) - m^2$	$C(k) = R(k) - m^2$	
NACVF	$\xi(\tau) = \gamma(\tau)/\gamma(0)$	$\rho(k) = C(k)/C(0)$	

Deoarece, după cum s-a menționat, tehnicile de studiu ale semnalelor stochastice prezentate necesită staționaritatea în sens larg a proceselor aleatoare investigate, această proprietate trebuie testată statistic. În cazul semnalelor reale, modificarea unor proprietăți complexe ca cele ale ACF sau ale funcțiilor densității de probabilitate sunt însoțite de modificări ale mediei și/sau varianței. Prin urmare, o metodă adecvată de probare a staționarității în sens larg o constituie testarea schimbărilor în timp ale acestor momente. Pentru aceasta, semnalul se divide în mai multe segmente și se compară statistic între ele mediile și varianțele acestora.

Pentru testarea egalității mediilor,  $\hat{m}_1$  și  $\hat{m}_2$ , a două segmente cu lungimile  $N_1$  și  $N_2$ , se utilizează statistica  $T$  definită prin:

$$T = \frac{\hat{m}_1 - \hat{m}_2}{\left[ \left( \frac{1}{N_1} + \frac{1}{N_2} \right) \left( \frac{N_1 \hat{\sigma}_1^2 + N_2 \hat{\sigma}_2^2}{N_1 + N_2 - 2} \right) \right]^{1/2}} \quad (5.39)$$

care, în cazul staționarității semnalului, prezintă o distribuție Student cu  $v = N_1 + N_2 - 2$  grade de libertate, iar în cazul varianțelor  $\hat{\sigma}_1^2$  și  $\hat{\sigma}_2^2$  - statistica  $F$ :

$$F = \frac{(N_1 / v_1) \sigma_1^2}{(N_2 / v_2) \sigma_2^2} \quad (5.40)$$

care urmează o distribuție  $F$  cu  $v_1 = N_1 - 1$  și  $v_2 = N_2 - 1$  grade de libertate.

Segmentele trebuie să fie suficient de lungi pentru a reflecta corect proprietățile semnalului. Astfel, de exemplu, dacă semnalul prezintă componente oscilante, în fiecare segment trebuie incluse mai multe cicluri de oscilare.

Tehnicile moderne de filtrare a zgomotului conținut în semnalele analitice se bazează pe compoziția spectrală diferită a zgomotului față de cea a componentei deterministe. Analiza spectrală a oricărui semnal stohastic care prezintă staționaritate în sens larg prin utilizarea transformatei Fourier discrete,  $X(f)$ , a unei funcții de probă,  $x(n)$ ,  $n=1, N$ , nu este posibilă deoarece  $X(f)$  reprezintă o variabilă aleatoare care nici măcar nu poate fi definită deoarece funcția de probă continuă,  $x(t)$ ,  $-\infty < t < \infty$ , din care s-a prelevat seria de timp nu îndeplinește condițiile lui Dirichlet, respectiv energia semnalelor stochastice este infinită. De aceea, pentru astfel de semnale, în scopul caracterizării spectrale se utilizează, în locul funcției densitate de energie, funcția densitate de putere,  $S(f)$ , care, prin analogie cu relația (A1.36), se definește prin:

$$P_{\text{tot}} = \int_{-\infty}^{\infty} |S(f)|^2 df \quad (5.41)$$

unde  $P_{\text{tot}}$  reprezintă puterea zgomotului dată de relațiile 5.3 sau 5.4. Pe baza unei serii de timp,  $x(n)$ , cuprinzând un număr foarte mare de valori achiziționate cu un timp de prelevare  $T$ ,  $S(f)$  se poate defini prin cu relația:



$$S(f) = \lim_{N \rightarrow \infty} E \left( \frac{X_P(f) X_P^*(f)}{(2N+1)T} \right) df$$

în care  $X_P(f)$  reprezintă transformata Fourier a unui semnal,  $x_P(n)$  obținut prin trunchierea semnalului inițial cu relația:

$$x_P(n) = \begin{cases} x(n), & |n| \leq N \\ 0, & |n| > N \end{cases} \quad (5.43)$$

Funcția densitate de putere este reală ( $S(f) = S^*(f)$ ), pară ( $S(f) = S(-f)$ ) și pozitivă ( $S(f) \geq 0$ ). Toată informația referitoare la faza componentelor spectrale ale procesului aleator se pierde în funcția densitate de putere datorită primei dintre aceste proprietăți.

Estimarea clasică a spectrului de putere al zgomotului se realizează pe baza teoremelor lui Wiener - Hincin și a lui Wold [10, 11] care, pentru procese în domeniu de timp continuu și, respectiv, discret, afirmă că funcțiile densitate de putere și de autocorelație a semnalului aleator reprezintă perechi de transformate Fourier.

Informații detaliate despre seriile de timp ale proceselor aleatoare, inclusiv spectrul de putere, pot fi obținute prin modelarea acestora folosind conceptele de sisteme de timp discret, stabile și cauzale. În acest cadru, tratat în cursuri avansate de prelucrare a semnalelor [1], se arată că toate procesele aleatoare corespunzând unor sisteme tehnice reale pot fi reprezentate prin modele mixte (vezi Anexa 5.2 - relația A2.14) având componente autoregresive și/sau de medie alunecătoare, modele care, odată deduse prin tehnici statistice laborioase, servesc la estimarea modernă a spectrului de putere al zgomotului.

#### V.4.2. Tipuri de zgomot conținute în semnalele analitice

Zgomotul sistematic care afectează măsurătorile analitice poate fi, de regulă, diminuat prin adoptarea unui plan experimental potrivit. În schimb zgomotul aleator este inerent semnalului și/sau măsurării. Pentru evaluarea gradului de contaminare a unui semnal cu zgomot aleator se utilizează raportul semnal/zgomot dat de relația:

$$S/N = \frac{\text{media semnalului}}{\text{deviația standard a semnalului}} \quad (5.44)$$

Sursele de zgomot sunt diverse: din mediul înconjurător (vibrațiile, instabilitatea rețelei electrice de alimentare, fluctuațiile de temperatură, de

lumină etc.), care pot fi reduse prin măsuri tehnice adecvate, surse fundamentale, datorate caracterului discret al materiei și energiei, precum și fluctuațiilor termice, și zgomotul de cuantificare, care apare în procesul de măsurare în urma cuantificării valorilor semnalului de evaluat.

Sursele fundamentale de zgomot pot fi clasificate [8] după dependența zgomotului de puterea semnalului astfel:

a) zgomot independent de puterea semnalului ( $N_A = \text{const.}$ ), sau zgomot limitat de detector. Predomină la detecția radiației cu energia fotonilor ( $h\nu$ ) mai mică decât energia termică,  $k_B T$ , unde  $T$  reprezintă temperatura absolută a detectorului. Sursa acestui tip de zgomot o constituie detectorul. Zgomotul de acest tip nu depinde de puterea sursei și apare la detecția radiației IR, din domeniul microundelor și al radiofrecvențelor.

b) zgomot proporțional cu rădăcina pătrată a puterii semnalului ( $N_B \propto (\text{semnal})^{1/2}$ ), sau zgomot limitat de sursă. El predomină când energia fotonilor este mai mare decât  $k_B T$  și când detecția (sau numărarea) fotonilor sau electronilor individuali, supusă statisticii Poisson, se face cu eficiență aproximativ unitară. Depinde de radicalul puterii semnalului și apare în spectrometria UV-Vis și în spectrometria cu particule încărcate (cu fotoelectroni, cu electroni de impact, ESCA etc.)

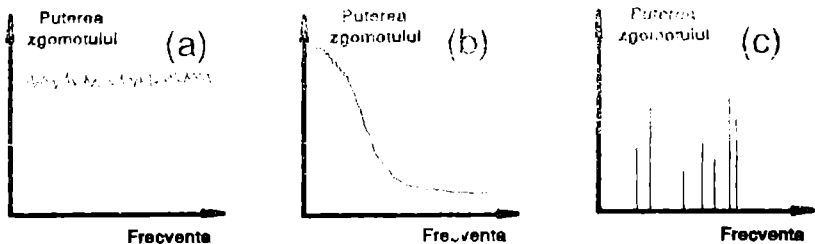
c) Zgomot proporțional cu puterea semnalului ( $N_C \propto \text{semnal}$ ), sau de "fluctuație", de "modulare", de "scintilație". Apare la detecția semnalelor foarte mici. Tipul corespunzător de zgomot din circuitele electronice se numește zgomot  $1/\nu$ , deoarece mărimea sa variază invers proporțional cu frecvența.

Zgomotul experimental total,  $N_{\text{total}}$ , însumează cele trei tipuri de zgomot care, de regulă, sunt independente din punct de vedere statistic. Deci:

$$N_{\text{total}} = (N_A^2 + N_B^2 + N_C^2)^{1/2} \quad (5.45)$$

unde prin  $N$  s-au notat varianțele zgomotului. Totuși, deși prezente simultan, în anumite tipuri de măsurători predomină un anumit tip de zgomot.

Zgomotul poate fi clasificat și în funcție de forma spectrului de putere. Trei tipuri diferite de zgomot (zgomot alb, zgomot de tip  $1/\nu$  și zgomot de interferență) sunt prezentate în figura 5.8.



**Fig. 5.8.** Spectrele de putere ale principalelor tipuri de zgomot. a) zgomot alb; b) zgomot de tip  $1/v$ ; c) zgomot de interferență.

### V.4.3. Tehnici de prelucrare a semnalelor analitice

Scopurile cele mai frecvente ale prelucrării semnalelor analitice [5, 8] sunt determinarea unor caracteristici fundamentale ale semnalelor (cromatograme, spectre și polarograme), prin descrierea lor cu un număr mic de parametri, și îmbunătățirea caracteristicilor metodelor de analiză: a sensibilității, aplicând tehnici de diminuare a zgomotului, și a selectivității, utilizând tehnici de restaurare a semnalelor cum ar fi cele de deconvoluție. La acestea se mai adaugă și estimarea inerției de răspuns a instrumentelor analitice prin determinarea constantelor de timp ale acestora.

#### V.4.3.1. Determinarea constantei de timp a instrumentelor analitice

Constanta de timp reprezintă o caracteristică a analizoarelor continue legată de posibilitatea acestora de a urmări cu fidelitate fluctuațiile de concentrație. Determinarea sa se realizează analizând forma semnalelor obținute după aplicarea unui semnal cu formă cunoscută, cel mai adesea sub formă de impuls cu amplitudine constantă un timp cât mai scurt, de treaptă unitară sau cu aspect periodic sinusoidal (vezi figura 5.9).

Răspunsul analizorului la un impuls infinit îngust ("impulse response function") se notează de obicei  $h(t)$ . Cunoașterea sa permite caracterizarea deplină a răspunsului analizorului deoarece acest răspuns se obține [1] prin operația de convoluție a funcției de concentrație de la intrarea în analizor cu funcția  $h(t)$ .

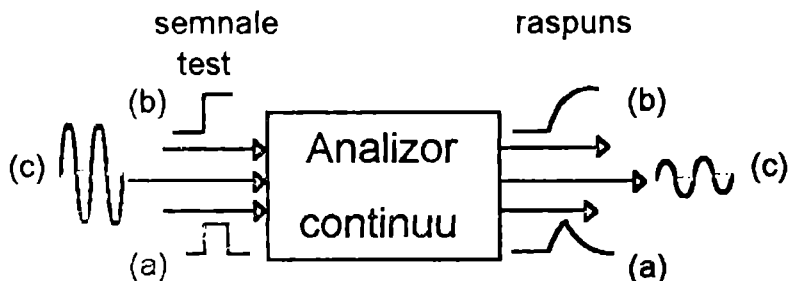


Fig. 5.9. Semnale test pentru determinarea caracteristicilor dinamice ale unui analizor continuu. (a) Impuls infinit îngust; (b) Treaptă de concentrație; (c) Funcție periodică.

Pentru a ușura înțelegerea sensului fizic al constantei de timp a analizatoarelor continue se va discuta în continuare un caz particular deosebit de relevant. Astfel, presupunem că un analizor în flux complet continuu care monitorizează absorbanta unui compus colorat este prevăzut cu o celulă în flux cu volumul  $V$  care are proprietățile unui reactor cu amestecare perfectă [12]. Dacă inițial prin analizor circulă o soluție incoloră înregistrarea obținută va reprezenta linia de bază. Aplicarea unui semnal în formă de treaptă în acest caz echivalază cu introducerea continuă, cu debitul  $q$ , prin sistemul de analiză, începând cu un anumit moment dat, a unui flux cu o concentrație constantă de specie colorată trasoare, notată  $C$ . Începând din momentul  $t=0$  din care specia trasoare intră în celulă în flux, concentrația acesteia în celulă se modifică dar, datorită proprietăților de amestecare ale celulei, ea nu crește brusc de la zero la  $C$ . Modul de variație în timp al concentrației de traser din celulă,  $c(t)$ , se obține efectuând bilanțul de masă pentru un interval infinitesimal de timp,  $dt$ :

$$V dc(t) = q c(t) dt - q C dt \quad (5.46)$$

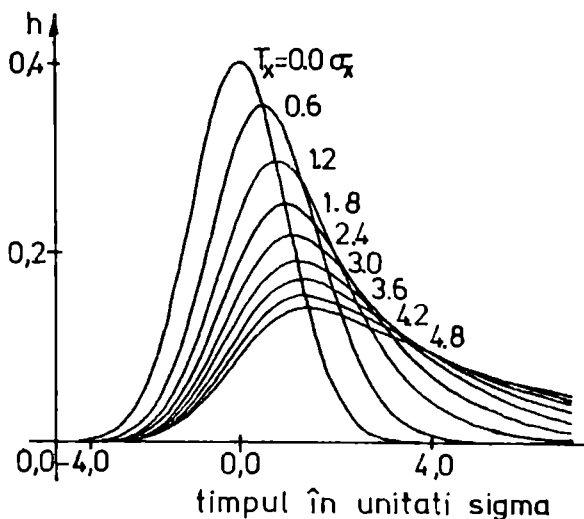
Soluția acestei ecuații, ținând cont de condițiile la limită impuse, este:  $c(t) = C [1 - \exp(-qt/V)]$  (5.47)

Se observă că valoarea  $C$  a concentrației de traser din celulă nu se atinge decât după un timp îndelungat, profilul de variație exponențial al  $c(t)$  fiind controlat de raportul  $V/q$  care, având dimensiuni de timp, se notează  $T_x$  și definește constanta de timp a analizorului. Cu cât aceasta are o valoare mai mare, cu atât este necesar un timp mai îndelungat pentru a se ajunge la o concentrație de traser în celulă egală cu  $C$ .

Dacă se poate înregistra răspunsul analizorului după aplicarea unui semnal treaptă de concentrație, așa cum este cazul pentru electrozii ion selectivi unde timpul de răspuns definit de către IUPAC [13] este indisolubil legat de  $T_x$ , atunci constanta de timp se determină prin aceleași metode prin care se determină constanta de viteză a reacțiilor de ordinul întâi.

Consecința practică imediată a unei constante de timp diferite de zero în cazul sistemelor analitice de măsură o reprezintă deformarea semnalelor inițiale. În general, cu cât un sistem de măsură prezintă o constantă de timp mai mare, cu atât semnalele înregistrate cu el reflectă mai puțin fidel semnalele de intrare fiind mai largi și mai puțin înalte. Totuși, ariile de sub curbele de variație ale semnalelor înregistrate în sisteme cu constante de timp diferite pentru un același semnal de intrare sunt egale între ele și, de asemenea, egale cu suprafața integrată a semnalului de intrare.

Modificarea formei semnalului de intrare de către sisteme reale de măsură având constante de timp diferite este ilustrată și în figura 5.10.



**Fig. 5.10.** Forma semnalelor obținute cu un analizor continuu când profilul semnalului de intrare este gaussian pentru diferite valori ale raportului dintre constanta de timp a analizorului și deviația standard a semnalului de intrare.

În general, sistemele reale monitorizate (de ex., se poate urmări variația de concentrație a unui component într-un proces tehnologic sau într-un râu) prezintă și ele o inerție de răspuns la fluctuațiile de concentrație și, de aceea, pot fi și ele caracterizate [7] printr-o constantă de timp,  $(T_x)_p$ . În astfel de cazuri, între varianțele fluctuațiilor reale de concentrație,  $s_p^2$ , și, respectiv, a semnalelor înregistrate cu un sistem analitic având constanta de timp  $T_x$ ,  $\sigma_p^2$ , există relația:

$$\frac{s_p^2}{\sigma_p^2} = \frac{(T_x)_p}{(T_x)_p + T_x}$$

(5.48)

Se consideră că sistemul analitic de măsură introduce o deformare neglijabilă a semnalelor reale doar dacă constanta de timp a acestuia este de cel puțin 10 ori mai mică decât cea a procesului monitorizat. Aceasta conduce la concluzia că selectarea sistemelor de măsură pe baza valorii constantei de timp trebuie făcută în corelație cu procesul fizico-chimic de investigat, fapt bine cunoscut spectrometriștilor care lucrează în IR în care viteza de baleiere la înregistrarea spectrului se calculează [7] în funcție de rezoluția dorită la înregistrarea benzilor spectrale.

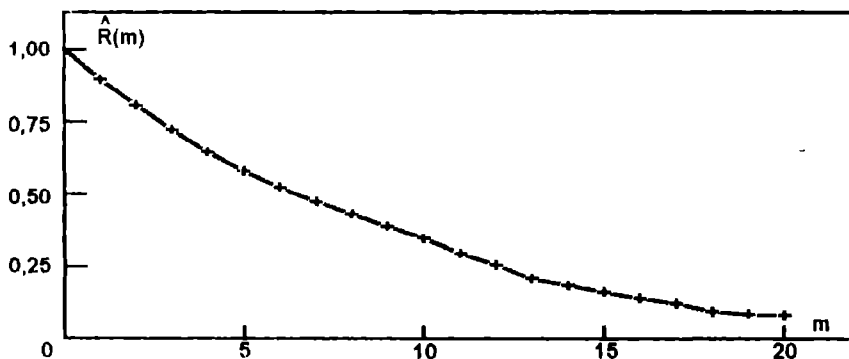


Fig. 5.11. Funcția de autocorelație a procesului aleator simulat în Anexa 5.3 care stă la baza determinării constantei de timp.

Nu în toate cazurile practice (cum ar fi cazul unui fluviu în care concentrația în diferiți componenți nu poate fi programată) este posibil să

se determine constanta de timp pe baza răspunsului obținut la aplicarea unui semnal în formă de treaptă sau de altă natură. În astfel de cazuri, deoarece constanta de timp a sistemului afectează și modificările aleatoare ale concentrației, la baza determinării valorii  $T_x$  poate sta funcția de autocovarianță normalizată a zgomotului în același mod ca și în cazul utilizării răspunsului la un semnal în formă de treaptă de concentrație, ca având, de asemenea, pentru sisteme afectate de zgomot cu structură autoregresivă (zgomot echivalent cu un proces Markov), o formă exponențială după cum rezultă și din figura 5.11. pentru zgomotul simulat în Anexa 5.3.

În majoritatea cazurilor practice din chimia analitică, zgomotul are structură autoregresivă și, prin urmare, sistemele respective pot fi caracterizate integral prin valoarea constantei de timp. Totuși, în unele cazuri răspunsul sistemelor analitice la semnale test nu poate fi explicat numai pe această bază. De exemplu, sunt menționați în literatură [7] electrozi ion selectivi cu funcția de răspuns la un semnal în formă de treaptă de forma:

$$E(t) = E_0 + E_1 \exp[1 - (t/T_1)] + E_2 \exp[1 - (t/T_2)] \quad (5.49)$$

sugerând faptul că răspunsul lor este determinat de două procese paralele independente care se desfășoară la suprafața electrodică.

În general, pentru orice tip de zgomot afectând un proces tehnic de măsură, fenomenul de degradare a semnalelor poate fi pe deplin explicat pe baza funcției de răspuns a sistemului la un semnal înfinit de îngust ("impulse response function"),  $h(t)$ , știut fiind [1] faptul că răspunsul de timp înregistrat,  $r(t)$ , este convoluția funcției  $h(t)$  cu semnalul de intrare real,  $s(t)$ :

$$r(t) = s(t) * h(t) \quad (5.50)$$

Funcția  $h(t)$  se determină, de obicei, înregistrând răspunsul sistemului la un semnal dreptunghiular foarte îngust și a cărui varianță este de cel puțin 10 ori mai mică decât a semnalului înregistrat fiind echivalentă cu acest ultim semnal.

### V.4.3.2. Determinarea unor parametri fundamentali ai semnalelor

Există numeroase situații în care componenta deterministă a semnalelor analitice poate fi descrisă printr-un model matematic cu formă cunoscută în care trebuie estimați unul sau mai mulți parametri. Un caz de acest tip a fost deja întâlnit în secțiunea anterioară când s-a arătat că răspunsul analizoarelor continue la o modificare în treaptă a concentrației prezintă o formă exponențială complet caracterizată prin constanta de timp a analizorului.

În caz general, presupunând că semnalul a fost eșantionat cu un interval de prelevare  $T$  obținându-se seria de timp  $x(n)$ , ( $n = 1, N$ ) și că acesta se aproximează cu o funcție continuă de forma:

$$\hat{x}(t) = f(t; \alpha_1, \alpha_2, \dots, \alpha_M), \quad 0 \leq t \leq NT \quad (5.51)$$

, estimarea parametrilor modelului,  $\alpha_1, \alpha_2, \dots, \alpha_M$ , se poate realiza prin minimizarea erorii pătratice a aproximării,  $E_M$ , care are expresia:

$$E_M = \sum_{n=1}^N [x(n) - \hat{x}(nT)]^2 \quad (5.52)$$

Pentru aceasta se utilizează frecvent analiza de regresie neliniară prezentată în capitolul IV, caz în care tehnica este denumită, în limba engleză, "*curve fitting method*".

Uneori înregistrările analitice pot fi caracterizate printr-un număr redus de parametri, caz în care componenta deterministă a semnalelor înregistrate poate fi scrisă ca suma unui număr de funcții având aceeași formă depinzând de un anumit număr de parametri. Astfel de situații se întâlnesc în cromatografie, când forma funcției este gaussiană, în spectrometrie, când forma este lorentziană, și în polarografie. În toate aceste cazuri semnalul  $x(t)$  poate fi scris sub forma:

$$\hat{x}(t) = \sum_{i=1}^I f(\alpha_{1i}, \alpha_{2i}, \alpha_{3i}) \quad (5.53)$$

În situațiile descrise mai sus trebuie aplicată regresia neliniară față de un număr de  $3 \times I$  parametri care poate ușor să eșueze dacă nu se cunosc estimații inițiale suficient de exacte pentru toți parametrii. De aceea, în astfel de cazuri, minimizarea criteriului dat de relația 5.52 se realizează prin metode de regresie cuplate cu algoritmi convenabili de



optimizare. Un astfel de algoritm este cel al lui Rosenbrock [14] a cărei aplicare computerizată este ilustrată în lucrarea [5].

Un alt caz de estimare foarte important întâlnit în practică este cel în care fiecare valoare înregistrată din seria de timp,  $x(i)$ , depinde liniar de concentrațiile ( $c_j$ ) mai multor specii chimice (în număr de  $N$ ):

$$x(i) = \sum_{j=1}^N h_{ij}c_j + e_i, \text{ pentru } i = 1, n \quad (5.54)$$

așa cum este cazul în tehnicile multicomponent de analiză prin spectrometrie UV-Vis. Deși regresia multivariabilă este aplicabilă în astfel de cazuri, o metodă alternativă de estimare, echivalentă cu aceasta și care prezintă numeroase avantaje, este algoritmul recursiv al lui Kalman ("Kalman filter"), propus pentru prima dată în domeniul controlului de proces [15] și care găsește largi aplicații în analiza chimică [16]. Noțiunile cu care operează tehnica Kalman [17] sunt următoarele: vectorul concentrațiilor, considerat o funcție de timp:

$$c(i)^T = (c_1(i), \dots, c_j(i), \dots, c_N(i)) \quad (5.55)$$

; funcția de modificare în timp a vectorului concentrațiilor,  $F(i)$ , definită, pentru fiecare valoare a timpului printr-o matrice pătrată ( $N \times N$ ) și care, în cazul sistemelor dinamice (cu reacție chimică desfășurată pe parcursul achiziționării datelor), face legătura între concentrațiile speciilor la două momente succesive de timp; vectorul măsurătorilor:

$$x_i^T = (x_i(1), \dots, x_i(j), \dots, x_i(N)) \quad (5.56)$$

; funcția de legătură dintre răspunsul măsurat și concentrația componentelor, văzută ca o mărime dependentă de timp:

$$h_i^T = (h_{1i}, \dots, h_{ji}, \dots, h_{Ni}) \quad (5.57)$$

;  $R$  - varianța măsurătorilor experimentale definită printr-un scalar sau printr-un vector având componente dependente de timp;  $P(i)$  - matricea de varianță - covarianță ( $N \times N$ ) dintre concentrațiile diversilor componente, care se calculează, în doi pași diferiți, pentru fiecare valoare de timp asimilată valorilor lui  $i$ ;  $Q(i)$  - varianța modelului de stare reprezentată, la fiecare valoare  $i$ , printr-o matrice pătrată ( $N \times N$ ) și care ia în considerare abaterile posibile de la desfășurarea procesului chimic conform modelului dinamic elaborat; și, în sfârșit,  $g(i)$  - performanța filtrului Kalman ("filter

gain”) definită printr-un vector cu N componente egal cu diferența dintre estimățiile concentrațiilor obținute în doi pași consecutivi.

Fiind o metodă recursivă de estimare, filtrul Kalman se desfășoară după un program stabilit. Mai întâi se inițializează vectorul concentrație,  $c(0)$ , printr-un vector nul și matricea  $P(0)$  punând toți  $P(0; i, i) = 10 \times R \times R$  și  $P(0; i, j) = 0$ , unde  $R$  este eroarea maximă estimată a fi conținută în datele experimentale. Apoi, pentru fiecare valoare a lui  $i$ , se efectuează cinci etape de calcul rezumate astfel:

1. Extrapolarea vectorului concentrație conform modelului dinamic stabilit:

$$c(i) = F(i-1) c(i-1) \quad (5.58)$$

2. Extrapolarea matricii de varianță - covarianță:

$$P(i) = P(i-1) + Q(i-1) \quad (5.59)$$

3. Actualizarea performanței filtrului Kalman:

$$g(i) = P(i) h_i [h_i^T P(i) h_i + R] \quad (5.60)$$

4. Actualizarea concentrațiilor:

$$c(i) = c(i) + g(i) [x_i - h_i^T c(i)] \quad (5.61)$$

5. Actualizarea matricii de varianță covarianță:

$$P(i) = P(i) - g(i) h_i^T P(i) \quad (5.62)$$

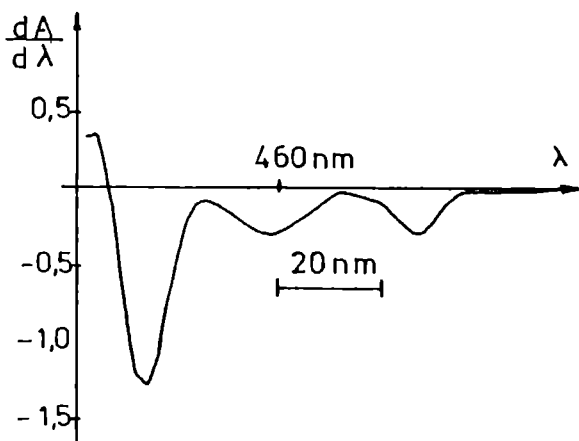
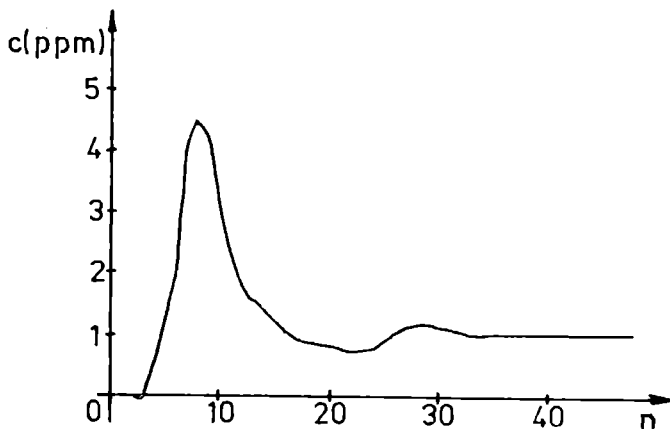


Fig. 5.12. Spectrul derivat al unui preparat farmaceutic tricomponent conținând piridoxină (1 ppm), acid ascorbic și tirozină.

După derularea pașilor filtrului, vectorul concentrație conține esența informației inițiale din amestec pentru toate componentele și matricea  $P$  - informația referitoare la varianțele acestora și la intercorelațiile dintre valorile estimate pentru toate componentele.

Filtrul Kalman prezintă o flexibilitate deosebită deoarece ține cont de un număr mare de factori, în aceasta constând marea sa avantaj comparativ cu tehnicile clasice de estimare, și furnizează un volum mare de informații. Memorarea valorilor calculate în pașii filtrului pentru diverșii săi parametri și reprezentarea lor grafică după încheierea calculelor permite obținerea de concluzii referitoare la modul de operare al sistemului de măsură sau la prezența în amestec a unor compuși noi față de cei din probele cu care s-a efectuat calibrarea. Astfel de fenomene pot fi luate în considerație când filtrul nu se stabilizează, respectiv când valorile estimate succesiv pentru concentrațiile componentelor nu converg spre o limită bine definită.

Modul normal de funcționare al filtrului Kalman, utilizând datele spectrului derivat din figura 5.12, îl ilustrăm [18] cu variația concentrației estimate de piridoxină pe parcursul derulării filtrului (vezi figura 5.13). Se observă stabilizarea rapidă a filtrului și convergența valorilor estimate spre valoarea reală a concentrației de piridoxină.



**Fig. 5.13.** Variația concentrației de piridoxină dintr-un preparat farmaceutic pe parcursul derulării filtrului Kalman.

### V.4.3.3. Tehnici de îmbunătățire a raportului semnal/zgomot

Pentru îmbunătățirea raportului semnal/zgomot se utilizează atât tehnici instrumentale care acționează chiar în momentul achiziționării datelor, cât și tehnici chemometrice "off-line" efectuate cu computerizat utilizând produse software adecvate.

#### V.4.3.3.1. Tehnici instrumentale

Principalele tehnici instrumentale de îmbunătățire a raportului semnal/zgomot sunt modularea semnalelor [5], aplicată în special în spectrometria de absorbție atomică, medierea ansamblului, tehnicile bazate pe transformarea semnalelor (tehnici de multiplexare optică) care, datorită importanței lor în istoricul dezvoltării analizei instrumentale vor fi tratate distinct, și, respectiv, pe intercorelația semnalelor.

Medierea ansamblului se aplică în special la înregistrarea unor semnale de intensitate foarte slabă și când repetarea experimentului de măsurare nu prezintă inconveniente, așa cum este cazul în spectrometria RMN [4]. Pentru ilustrarea performanțelor acestea, figura 5.14 reprezintă medierea ansamblului funcțiilor de probă simulate în Anexa 5.3. Se observă că medierea conduce la netezirea apreciabilă a picurilor cu intensitate situată în apropierea limitei de detecție. În general, prin această tehnică raportul semnal/zgomot se îmbunătățește de  $\sqrt{N}$  ori, unde  $N$  reprezintă numărul funcțiilor de probă mediate.

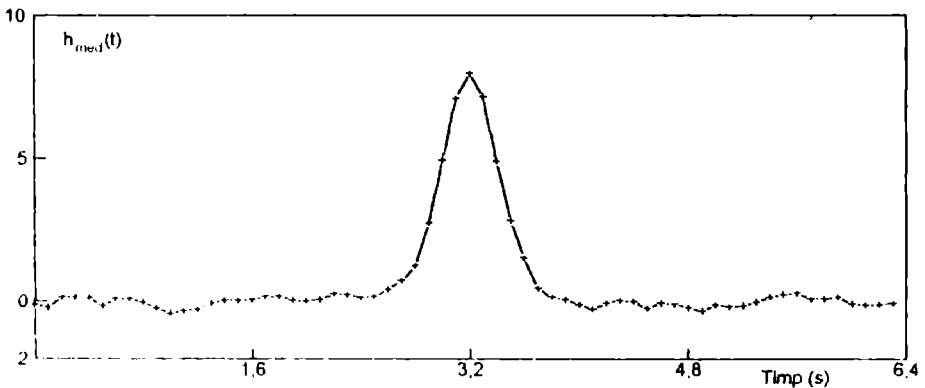
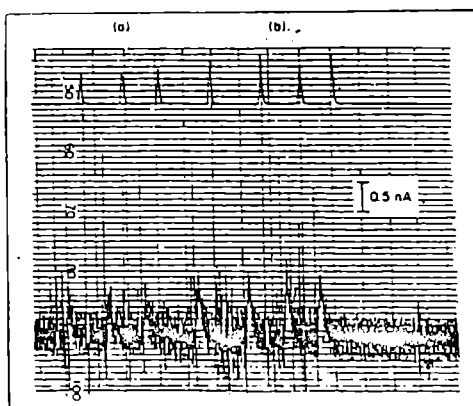


Fig. 5.14. Medierea unui ansamblu simulat de funcții de probă.

Tehnicile de intercorelație sunt foarte eficiente în coborârea limitelor de detecție ale metodelor analitice de determinare. Ele se bazează pe calcularea funcției de intercorelație dintre două semnale (definiție similară cu funcția de autocorelație pentru un semnal) care prezintă componente deterministe similare dar care sunt afectate de zgomot aleatoriu distribuit necoerent.

Pentru a ilustra performanțele metodei de intercorelare, figura 5.15 prezintă picuri înregistrate la determinarea dopaminei (jos) prin FIA cu detecție amperometrică [19] comparativ cu cele obținute după prelucrarea primelor prin tehnica intercorelației (sus). Pe această bază, limita de detecție a dopaminei a putut fi îmbunătățită de cca. 100 de ori.



**Fig. 5.15.** Înregistrarea curentului afectat de zgomot și semnalul intercorelat pentru injectarea a (a) 7,4 nM și (b) 15 nM dopamină în soluție tampon acetat 0,1 M ( $pH = 5,4$ ).

#### **V.4.3.3.1.1. Scurt istoric al utilizării tehnicilor de transformare în chimie**

La sfârșitul secolului trecut, Michelson a proiectat interferometrul care astăzi îi poartă numele. Încercările de determinare cu ajutorul său a profilelor liniilor spectrale a condus la obținerea interferogramelor. Lordul Rayleigh a arătat că nu se poate obține distribuția spectrală reală decât cu ajutorul transformatei Fourier. Michelson a încercat să suplinească lipsa

posibilităților tehnice de calculare a transformatei Fourier din timpul său prin construirea unui sintetizator armonic. Invenția sa nu a influențat însă decisiv dezvoltarea ulterioară a științei.

Rubens și Wood (1911) publică prima interferogramă veritabilă, obținută în domeniul IR îndepărtat cu un interferometru diferit de al lui Michelson și au recunoscut avantajul multiplexării. Totuși, utilizarea pe scară largă a interferometriei nu se produce decât după ce Jacquinet (1948 - Franța) și Fellget (1951 - Anglia) au evidențiat avantajele acesteia. De altfel Fellget (1951) a și publicat prima interferogramă transformată numeric și a accentuat avantajele multiplexării informației spectrale.

Primul spectrometru în IR cu transformată Fourier a fost comercializat în 1964 de către Research and Industrial Instruments Corporation iar elaborarea, la scurt timp după aceasta, a algoritmului de transformare Fourier rapidă de către Cooley și Tukey (1965) a condus la reducerea considerabilă a timpului de calcul al spectrelor. Alte contribuții importante la dezvoltarea spectrometriei în IR cu transformată Fourier au mai fost aduse de către P. R. Griffiths care a și editat prima monografie dedicată tehnicilor de transformare în chimie [8].

#### **V.4.3.3.1.2. Sisteme analitice cu transformată Hadamard**

Spectrul de absorbție al unei substanțe poate fi obținut în două moduri distincte:

- a) Se masoară secvențial (rând pe rând) intensitatea radiației la fiecare lungime de undă (calea clasică);
- b) Se utilizează o tehnică de modulare a componentelor spectrale și se analizează simultan de un număr mare de ori.

Scopul principal al modulării este de a mări cantitatea de radiație care ajunge la un moment dat la detector deoarece, în cazul în care zgomotul detectorului nu depinde de intensitatea radiației incidente, se îmbunătățește raportul semnal/zgomot (S/N).

**Multiplexarea reprezintă modularea independentă și simultană a unui mare număr de componente spectrale.**

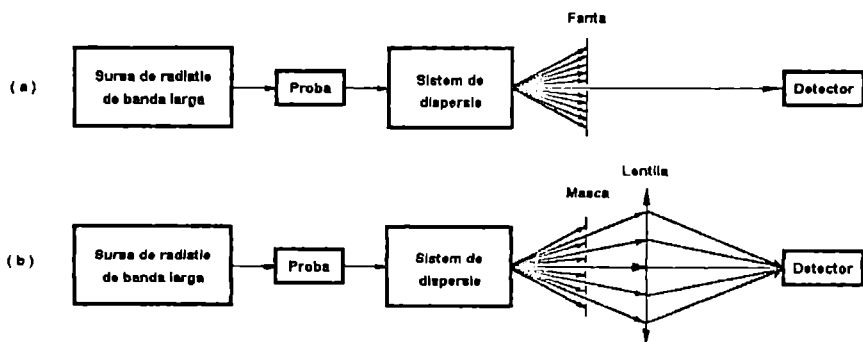
Se cunosc două tehnici distincte de multiplexare:

1. Ține cont de caracterul de undă al radiației, utilizează interferometria pentru achiziționarea semnalelor în domeniul timpului și face apel la transformata Fourier pentru calcularea spectrului;

2. Utilizează spectrometrele cu sistem de dispersie obișnuite din care lipsește fanta de intrare și în care fanta de ieșire se înlocuiește cu o mască de codificare, care are unele zone transparente (acestea permițând trecerea radiației) și alte zone opace (reținând radiația).

Teoria matematică pe care se bazează această ultimă tehnică este transformata Walsh - Hadamard, aplicată în domeniul spectrometriei pentru prima dată de către Decker și Harwit (1973) [20].

Figura 5.16 prezintă, comparativ, schemele de principiu a unui spectrometru clasic (a) și a unui cu transformată Hadamard (b).



**Fig. 5.16.** Schemele de principiu ale unui spectrometru clasic (a) și a unui spectrometru cu transformată Hadamard (b).

În spectrometrele cu transformată Hadamard, spre deosebire de cele clasice, se interpune o mască între sistemul de detecție și detector. Masca se construiește astfel încât cele mai mici deschideri ale sale să aibă aceeași lățime ca și fanta din spectrometrele clasice cu scanarea lungimii de undă, dar cu aproximativ jumătate din pozițiile posibile deschise. Alternarea zonelor închise și deschise din calea radiației este pseudoaleatoare.

Notăm cu  $x_1, x_2, \dots, x_N$  intensitățile transmise de probă. Când masca se află în poziția din figura 5.16.b, intensitatea,  $y$ , măsurată de către detector va fi:

$$y = w_1 x_1 + w_2 x_2 + \dots + w_N x_N, \quad (5.63)$$

unde  $w_n = 0$  sau  $1$ ;  $n = 1, N$ , ( $0$  - dacă fanta este deschisă,  $1$  - dacă este închisă).

Cu alte cuvinte, detectorul furnizează o singură observabilă ( $y$ ) care depinde de  $N$  necunoscute (intensitățile spectrale individuale,  $x_1, x_2, \dots, x_N$ ), conform codului din ecuația 5.63. Situația este asemănătoare cu cazul A5.3.3 discutat în anexa 5.3. Pentru a putea calcula cele  $N$  valori dorite ( $x_1, x_2, \dots, x_N$ ) și a regăsi astfel tot spectrul, ar trebui să utilizăm  $N$  măști diferite, astfel încât fiecare să aibă un aranjament al fantelor deschise și liniar independente de a celorlalte, problemă care ar fi foarte complicată.

Teoria matematică elaborată de Hadamard simplifică foarte mult această problemă. Astfel, se arată că, dacă  $N$  este un număr impar, atunci se poate construi o singură mască cu  $2N - 1$  fante (închise sau deschise) cu ajutorul cărora pot fi obținute, prin translatarea măștii cu o poziție la fiecare pas în fața ferestrei spectrale a detectorului,  $N$  măști diferite și liniar independente între ele, fiecare derivând una din cealaltă prin permutare ciclică și având cca. jumătate din fante deschise la fiecare măsurare.

Pentru  $N = 7$ , masca este: (0 1 0 1 1 1 0 0 1 0 1 1 1) iar matricea ciclică corespunzătoare va fi:

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

În figura 5.17. se reprezintă schematic această situație.

Dacă în sistemul de ecuații 5.63 intervine o matrice ciclică de tip Hadamard ( $W$ ), atunci rezolvarea sistemului este banală. De exemplu, din sistemul:



$$\begin{aligned} y_1 &= x_1 + x_2 & e_1 \\ y_2 &= x_1 & e_2 \\ y_3 &= x_2 + x_3 & e_3 \end{aligned}$$

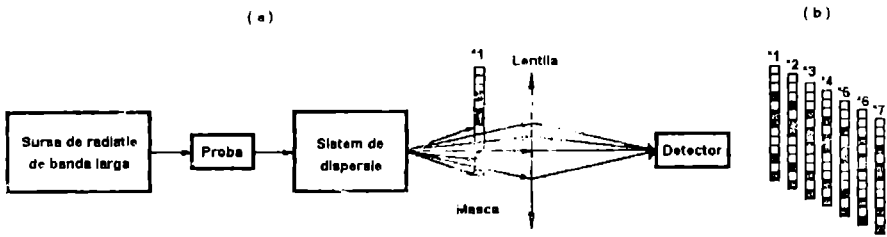
rezultă:

$$\hat{x}_1 = \frac{1}{2} (y_1 + y_2 - y_3) + \frac{1}{2} (e_1 + e_2 - e_3)$$

$$\hat{x}_2 = \frac{1}{2} (y_1 - y_2 + y_3) + \frac{1}{2} (e_1 - e_2 + e_3) \quad (5.55)$$

$$\hat{x}_3 = \frac{1}{2} (-y_1 + y_2 + y_3) + \frac{1}{2} (-e_1 + e_2 + e_3)$$

sau  $\hat{x} = W^{-1}(y - e).$  (5.66)



**Fig. 5.17.** Ilustrarea principiului multiplexării optice în spectro-metria cu transformată Hadamard.

#### V.4.3.3.2. Tehnici aplicate după achiziționarea semnalelor

Ca modalități de tratare a semnalelor pentru îmbunătățirea raportului semnal/zgomot după achiziționarea lor integrală se cunosc tehnicile de filtrare (care mai pot fi efectuate și cu dispozitive electronice) și, respectiv, cele de netezire.

Operația de filtrare reprezintă un subiect complex constând, în general, în modificarea profilului de frecvențe al unui semnal de intrare. Aceasta presupune, de regulă, transmiterea unor domenii de frecvențe și îndepărtarea altora. În funcție de aceasta, filtrele pot fi clasificate [1, 5] în filtre trece sus, trece jos, trece bandă și oprește bandă.

În ultimul timp, filtrele digitale tind să le înlocuiască pe cele analogice datorită avantajelor astfel obținute. Astfel, filtrele digitale sunt programabile, stabile și predictibile, nu depind de parametrii experimentali

de operare, ele eliminând driftul cu temperatura și umiditatea al filtrelor analogice, și au raportul performanță/cost mai favorabil [21].

Există două modalități diferite de filtrare a datelor după achiziționarea lor. Cea mai importantă dintre acestea este filtrarea în domeniul frecvențelor. Ea implică calculul transformatei Fourier discrete a semnalului, multiplicarea spectrului de frecvențe astfel obținut cu un filtru dorit și retransformarea spectrului modificat în domeniul timpului aplicând transformata Fourier inversă. Cea de a doua cale, efectuată în domeniul timpului, constă în operația de convoluție a semnalului cu funcția de răspuns la un impuls a unui filtru potrivit.

Un exemplu de semnal filtrat prin prima din metodele amintite este prezentat în figura 5.18. Semnalul de pornire l-a constituit un pic gaussian contaminat cu zgomot având structură autoregresivă eșantionat la 64 de momente diferite de timp. Deoarece, în general, zgomotul prezintă benzi spectrale intense la frecvențe înalte, în timp ce, așa după cum rezultă și din figura 5.4, semnalele deterministe au energia distribuită preponderent pe frecvențe joase, din semnalul frecvențial obținut aplicând DFT semnalului inițial s-au reținut doar 12 frecvențe înaintea aplicării transformatei Fourier inverse.

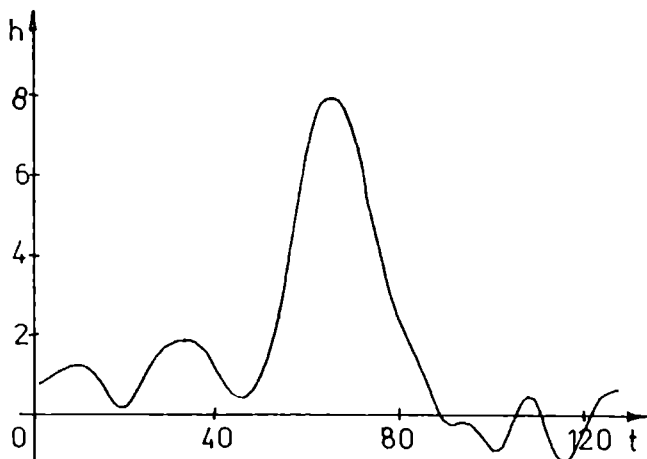


Fig. 5.18. Pic gaussian contaminat cu zgomot netezit utilizând filtrarea cu ajutorul transformatei Fourier.

Se observă că, deși semnalul filtrat este incomparabil mai ascuțit decât în zona maximului, cu un pic gaussian decât oricare din picurile simulate în anexa 5.3, el prezintă o cantitate mare de energie distribuită în lobii acestuia, fapt datorat trunchierii semnalului inițial, operație echivalentă cu mulțirea semnalului cu o fereastră de date dreptunghiulară. Prin utilizarea unei ferestre de date mai bune, acest inconvenient poate fi înlăturat satisfăcător.

Netezirea semnalelor reprezintă un ansamblu de tehnici matematice efectuate în domeniul de achiziție al semnalelor în scopul îndepărtării zgomotului. Ele prezintă o importanță mai redusă în comparație cu tehnicile de filtrare. Dintre acestea, de cea mai largă audiență [1, 5, 7] se bucură metoda aproximării semnalului prin polinomul ajustat prin metoda celor mai mici pătrate [22].

## V.5. Concluzii

Prelucrarea semnalelor analitice pentru valorificarea lor optimă constituie un domeniu vast și important al chemometriei, motiv pentru care s-a insistat mai mult asupra principiilor teoretice ale domeniului și mai puțin asupra gamei foarte largi de aplicații existente în literatură.

Prelucrarea statistică a semnalelor achiziționate numeric la un număr mare de momente de timp este de neconceput în afara utilizării calculatoarelor electronice cu putere mare de calcul și dotate cu software corespunzător [23].

## Bibliografie

1. R. Shiavi, **Introduction to Applied Statistical Analysis**, Aksen Publishers, Boston, 1991.
2. K. S. Booksh și B. R. Kowalski, *Anal. Chem.*, **66**,782A, 1994.
3. A. Aoki; T. Matsue și I. Uchida, *Anal. Chem.*, **64**, 44, 1992.
4. D. L. Andrews (Editor), **Perspectives in Modern Chemical Spectroscopy**, Springer-Verlag, Berlin, 1990.
5. C. Liteanu și I. Râcă, **Optimizarea proceselor analitice**, Editura Academiei Române, București, 1985.

6. I. Gh. Şabac; P. Cocârlan; O. Stănăşilă și A. Topală, **Matematici Speciale**, vol. II, E. D. P., Bucureşti, 1983.
7. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte și L. Kaufman, **Chemometrics: A Textbook**, Elsevier, Amsterdam, 1988.
8. P. R. Griffiths (Editor), **Transform Techniques in Chemistry**, Plenum Press, New York, 1978.
9. F. Harris, *Proc. IEEE*, **66**, 51, 1978.
10. L. Koopmans, **The Spectral Analysis of Time Series**, Academic Press, New York, 1974.
11. M. Priestley, **Spectral Analysis and Time Series**, vol 1, Academic Press, New York, 1981.
12. J. Ruzicka și E. H. Hansen, **Flow Injection Analysis**, Wiley, New York, 2nd Edition, 1988.
13. IUPAC, *Inf. Bull.*, **69**, 1978.
14. H. H. Rosenbrock, *Computer J.*, **3**, 174, 1960.
15. R. E. Kalman, *J. Basic Eng.*, **82**, 35, 1960.
16. S. D. Brown, *Anal. Chim. Acta*, **181**, 1, 1986.
17. R. L. Tranter, *Anal. Proc.*, **27**, 134, 1990.
18. V. David și M. Surmeian, Date experimentale în curs de publicare.
19. R. E. McKean și D. J. Curran, *Talanta*, **39**, 319, 1992.
20. J. Decker și M. Harwit, **Proc. Applications of Walsh Functions**, 1973. (AD-763-000).
21. A. Economu și P. R. Fielden, *Anal. Chim. Acta*, **305**, 165, 1995.
22. A. Savitzki și M. J. E. Golay, *Anal. Chem.*, **36**, 1627, 1964.
23. S. Wolfram, **Mathematica™. A System for Doing Mathematics by Computer**, Addison-Wesley, New York, 1988.

## Analiza spectrală a semnalelor determinate în domeniu de timp continuu

### A5.1.1. Funcții ortogonale

Seturile de funcții ortogonale,  $\{\Phi_m(t)\}$ , pot fi folosite pentru descrierea semnalelor continue într-un domeniu de timp finit cu durată  $P$ . Aceste semnale pot fi scrise, cu ajutorul unor astfel de funcții, ca serii de forma:

$$f(t) = A_0\Phi_0(t) + A_1\Phi_1(t) + \dots + A_m\Phi_m(t) + \dots = \sum_{m=0}^{\infty} A_m\Phi_m(t) \quad (\text{A1.1})$$

Setul de funcții este ortogonal dacă:

$$\int_0^P \Phi_m(t)\Phi_n(t) dt = \begin{cases} \lambda_n & \text{pentru } m = n \\ 0 & \text{pentru } m \neq n \end{cases} \quad (\text{A1.2})$$

iar domeniile uzuale pentru  $m$  sunt  $0 < m < \infty$  (ca și mai sus) și, respectiv,  $-\infty < m < \infty$ .

Coefficienții seriei (A1.1) se determină utilizând principiul celor mai mici pătrate al erorilor. Definind o aproximație a funcției  $f(t)$  folosind un număr finit de termeni,  $M+1$ , prin:

$$\hat{f}(t) = A_0\Phi_0(t) + A_1\Phi_1(t) + \dots + A_M\Phi_M(t) \quad (\text{A1.3})$$

eroarea pătratică a aproximării este:

$$E_M = \int_0^P (f(t) - \hat{f}(t))^2 dt = \int_0^P \left( f(t) - \sum_{m=0}^M A_m\Phi_m(t) \right)^2 dt = \int_0^P \left[ f(t)^2 - 2f(t) \sum_{m=0}^M A_m\Phi_m(t) + \left( \sum_{m=0}^M A_m\Phi_m(t) \right)^2 \right] dt \quad (\text{A1.4})$$

Derivata parțială în raport cu un anumit coeficient,  $A_m$ , trebuie să fie zero:

$$\frac{\partial E_M}{\partial A_m} = -2 \int_0^P f(t)\Phi_m(t) dt + 2 \int_0^P \left( \sum_{n=0}^M A_n\Phi_n(t) \right) \Phi_m(t) dt = 0 \quad (\text{A1.5})$$

Rearanjând ecuația precedentă și utilizând condițiile de ortogonalitate, se ajunge la:

$$\int_0^P f(t) \Phi_m(t) dt = \int_0^P \sum_{n=0}^M \Lambda_n \Phi_n(t) \Phi_m(t) dt = A_m \lambda_m \quad (A1.6)$$

Astfel se obține un mod simplu de calcul al coeficienților seriei:

$$\Lambda_m = \frac{1}{\lambda_m} \int_0^P f(t) \Phi_m(t) dt, \text{ pentru } 0 \leq m \leq \infty \quad (A1.7)$$

Conform teoremei lui Parseval, între energia pe o perioadă a semnalului  $f(t)$  și coeficienții dezvoltării în serie există relația:

$$E_{\text{tot}} = \sum_{m=0}^{\infty} A_m^2 \lambda_m \quad (A1.8)$$

În relațiile de mai sus, domeniul de variație al lui  $m$  este  $-\infty < m < \infty$  în cazul în care se utilizează un set extins de funcții ortogonale.

Există multe seturi de funcții ortogonale. Seturile continue includ funcțiile trigonometrice sinusoidale precum și funcțiile lui Hermite, ale lui Legendre și ale lui Laguerre iar cele discontinue - funcțiile de formă dreptunghiulară ale lui Walsh și Hadamard.

### A5.1.2. Serii Fourier

Pentru semnalele periodice cu perioada  $P$ ,  $f(t)$ , care satisfac condițiile lui Dirichlet, respectiv:

- a) au un număr finit de discontinuități într-o perioadă;
- b) au un număr finit de maxime și minime într-o perioadă;
- și c) sunt absolut integrabile, adică:

$$\int_0^P |f(t)| dt < \infty \quad (A1.9)$$

, funcțiile trigonometrice sinusoidale formează un set puternic de funcții ortogonale iar seriile formate cu ajutorul acestora se numesc serii Fourier.

Condițiile de existență ale seriilor Fourier sunt foarte generale iar semnalele analitice, dacă sunt limitate la un domeniu de timp finit, îndeplinesc aceste condiții. În mod tacit, vom considera că sunt îndeplinite și pentru funcțiile discutate în continuare.

Orice funcție pară,  $g(t)$ , admite o dezvoltare în serie Fourier de forma:

$$g(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos(n\omega_0 t) \quad (A1.10)$$

unde  $\omega_0 = 2\pi/P$  este frecvența fundamentală dată de relația

$$\omega_0 = \frac{2\pi}{P} = 2\pi f_0 \quad (\text{A1.11})$$

iar coeficienții  $A_n$ , pentru  $0 < n < \infty$ , se calculează cu o relație similară cu (A1.7) în care se verifică ușor că  $\lambda_m = P$ . Termenul  $A_0$ , corespunzând frecvenței zero, este egal cu media semnalului pe o perioadă și, prin analogie cu diferența dintre o tensiune alternativă și una continuă din electronică, se numește "termen DC".

Seriile Fourier ale funcțiilor impare,  $h(t)$ , au forma:

$$h(t) = \sum_{n=1}^{\infty} B_n \sin(n\omega_0 t) \quad (\text{A1.12})$$

iar cele ale unor funcții oarecare,  $f(t)$ :

$$f(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos(n\omega_0 t) + \sum_{n=1}^{\infty} B_n \sin(n\omega_0 t) \quad (\text{A1.13})$$

deoarece orice semnal poate fi reprezentat ca suma a două funcții, dintre care una pară și alta impară, alegând:

$$g(t) = \frac{f(t) + f(-t)}{2} \quad (\text{A1.14})$$

$$\text{și } h(t) = \frac{f(t) - f(-t)}{2} \quad (\text{A1.15})$$

Notând  $C_0 = A_0$  și, pentru fiecare  $n$  ( $1 < n < \infty$ ),  $C_n = (A_n^2 + B_n^2)^{1/2}$ , în domeniul  $[0, 2\pi)$  există un singur unghi,  $\theta_n$ , astfel încât  $\cos \theta_n = A_n / C_n$  și  $\sin \theta_n = -B_n / C_n$ .

Cu aceste notații, seria Fourier (A1.13) devine:

$$f(t) = C_0 + \sum_{n=1}^{\infty} C_n [\cos(n\omega_0 t) \cos \theta_n - \sin(n\omega_0 t) \sin \theta_n] \quad (\text{A1.16})$$

$$\text{sau } f(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n) \quad (\text{A1.17})$$

Comparând relațiile (A1.10) și (A1.12) cu (A1.17), se observă că toate au aceeași formă, cu deosebirea că, pentru funcții oarecare, armonicile din seria Fourier nu mai sunt toate în fază.

De obicei, în aplicațiile pe calculator disponibile comercial, se utilizează un alt set de funcții ortogonale pentru reprezentarea seriilor Fourier, și anume exponențialele complexe de forma  $\Phi_n = \exp(jn\omega_0 t)$  ( $-\infty < n < \infty$ ), care sunt legate de funcțiile trigonometrice prin formula lui Euler:

$$\exp(jn\omega_0 t) = \cos(n\omega_0 t) + j \sin(n\omega_0 t) \quad (\text{A1.18})$$

Pentru funcții complexe, condiția de ortogonalitate devine:

$$\int_0^P \Phi_m(t) \Phi_n^*(t) dt = \begin{cases} \lambda_n & \text{daca } m = n \\ 0 & \text{daca } m \neq n \end{cases} \quad (\text{A1.19})$$

unde  $\Phi_n^*(t)$  reprezintă conjugata funcției  $\Phi_n(t)$ .

Cu ajutorul exponențialelor complexe, seriile Fourier se scriu:

$$f(t) = \sum_{n=-\infty}^{\infty} Z_n \exp(jn\omega_0 t) \quad (\text{A1.20})$$

iar coeficienții  $Z_n$  reprezintă numere complexe și se obțin cu relația:

$$Z_n = \frac{1}{\lambda_n} \int_0^P f(t) \Phi_n^*(t) dt \quad (\text{A1.21})$$

Se verifică ușor că dacă  $f(t)$  reprezintă o funcție reală, atunci  $Z_n = Z_{-n}^*$ . Ei pot fi scriși și sub forma  $Z_n = Z_n \exp(j\theta_n)$  iar pentru funcții reale se verifică ușor că  $Z_{-n} = Z_n$  și  $\theta_{-n} = -\theta_n$ .

Funcția reală  $f(t)$  mai poate fi scrisă sub formă de serie Fourier astfel:

$$\begin{aligned} f(t) &= \sum_{n=-\infty}^{-1} Z_n \exp(j\theta_n) \exp(jn\omega_0 t) + z_0 + \sum_{n=1}^{\infty} Z_n \exp(j\theta_n) \exp(jn\omega_0 t) = \\ &= z_0 + \sum_{n=1}^{\infty} Z_n (\exp(-j\theta_n) \exp(-jn\omega_0 t) + \exp(j\theta_n) \exp(jn\omega_0 t)) = \\ &= z_0 + \sum_{n=1}^{\infty} 2Z_n \cos(n\omega_0 t + \theta_n) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n) \end{aligned} \quad (\text{A1.22})$$

Ultima relație exprimă corespondența dintre cele două moduri de reprezentare a seriilor Fourier. Termenii medii au aceeași mărime în timp ce magnitudinile coeficienților complecși reprezintă jumătate din cele ale coeficienților din reprezentarea trigonometrică. De asemenea, unghiul de fază al coeficienților complecși pentru  $n$  pozitivi sunt egali cu diferențele de fază ale funcțiilor cosinus.

Informația conținută în magnitudinile coeficienților și a unghiurilor de fază este foarte importantă în analiza de frecvență. Reprezentările acestora funcție de frecvență se numesc spectre de magnitudine și, respectiv, de fază. În cazul funcțiilor periodice, aceste spectre sunt spectre de linii deoarece valorile există numai la frecvențele armonice.



Seriile Fourier converg spre o funcție periodică cu perioada  $P$ . Ele există atâta timp cât integrarea se face pe o perioadă întreagă, fără ca timpul să aibă importanță. În general:

$$z_n = \frac{1}{P} \int_{t_1}^{t_1+P} f(t) \exp(-jnw_0t) dt, \text{ pentru } -\infty < t_1 < \infty \quad (\text{A1.23})$$

În orice punct de discontinuitate al funcției  $f(t)$ , seriile Fourier converg spre media sumei valorilor funcției din dreapta și din stânga aceluși punct. Deci, dacă funcția  $f(t)$  este discontinuă la timpul  $t_1$ , valoarea seriei la acest moment de timp este:

$$f(t_1) = \frac{1}{2} (f(t_1^-) + f(t_1^+)) \quad (\text{A1.24})$$

Caracteristicile energetice ale semnalului  $f(t)$  pot fi calculate și în funcție de parametrii reprezentării frecvențiale a acestuia. Astfel, energia totală pe o perioadă este dată de relația:

$$E_{\text{tot}} = \sum_{n=-\infty}^{\infty} P Z_n^2 \quad (\text{A1.25})$$

iar puterea medie de:

$$P_{\text{med}} = \frac{1}{P} \sum_{n=-\infty}^{\infty} Z_n^2 \quad (\text{A1.26})$$

### A5.1.3. Transformata Fourier continuă (CFT)

CFT se aplică la găsirea conținutului de frecvență al semnalelor aperiodice, semnale care nu pot fi reprezentate prin serii Fourier. Într-adevăr, astfel de semnale pot fi considerate ca având perioada infinită. Considerând diferența de frecvență dintre două armonice consecutive egală cu frecvența fundamentală, se obține:

$$\Delta f = f_0 = \lim_{P \rightarrow \infty} \frac{1}{P} = 0 \quad (\text{A1.27})$$

Pentru funcția de timp,  $x(t)$ , rezultă următoarele relații:

$$x(t) = \int_{-\infty}^{\infty} X(f) \exp(j2\pi ft) df = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(w) \exp(jwt) dw \quad (\text{A1.28})$$

De asemenea:

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft) dt \quad (\text{A1.29})$$

$$\text{sau } X(w) = \int_{-\infty}^{\infty} x(t) \exp(-jwt) dt \quad (\text{A1.30})$$

Ecuatiile (A1.29) și (A1.30) reprezintă expresiile CFT directe iar cu (A1.28) se calculează CFT inversă.  $x(t)$  și  $X(f)$  sau  $x(t)$  și  $X(\omega)$  reprezintă perechi de transformate Fourier, fapt care, de obicei, se notează pe scurt astfel:  $x(t) \Leftrightarrow X(f)$  sau  $x(t) \Leftrightarrow X(\omega)$ .

Proprietățile transformatei Fourier sunt similare cu cele ale coeficienților seriilor Fourier. Se verifică ușor că părțile reală și imaginara ale transformatei Fourier sunt funcții pare și, respectiv, impare, adică.  $\text{Re}[X(f)] = \text{Re}[X(-f)]$  și  $\text{Im}[X(f)] = -\text{Im}[X(-f)]$ , ceea ce se mai poate exprima prin relația:

$$X(f) = X^*(-f) \quad (\text{A1.31})$$

În forma polară, se poate scrie:

$$X(f) = |X(f)| \exp [j\theta(f)] \quad (\text{A1.32})$$

unde:  $|X(f)| = \sqrt{(\text{Re}^2[X(f)] + \text{Im}^2[X(f)])}$  (A1.33)

și  $\theta(f) = -\arctan \frac{\text{Im}[X(f)]}{\text{Re}[X(f)]}$  (A1.34)

$|X(f)|$  este o funcție pară iar  $\theta(f)$  - una impară.

Energia semnalelor aperiodice se definește prin:

$$E_{\text{tot}} = \int_{-\infty}^{\infty} x^2(t) dt = \int_{-\infty}^{\infty} x(t) \left( \int_{-\infty}^{\infty} X(f) \exp(j2\pi ft) df \right) dt \quad (\text{A1.35})$$

Scambînd ordinea de integrare, rezultă:

$$\begin{aligned} E_{\text{tot}} &= \int_{-\infty}^{\infty} X(f) \left( \int_{-\infty}^{\infty} x(t) \exp(j2\pi ft) dt \right) df = \int_{-\infty}^{\infty} X(f) X(-f) df = \\ &= \int_{-\infty}^{\infty} X(f) X^*(f) df = \int_{-\infty}^{\infty} |X(f)|^2 df \end{aligned} \quad (\text{A1.36})$$

Cu alte cuvinte, energia reprezintă aria de sub pătratul spectrului magnitudinilor. Funcția  $|X(f)|^2$  este energia pe unitatea de frecvență și se mai numește **spectrul densității de energie**.

## Noțiuni de simulare a proceselor aleatoare

Procesele aleatoare în domeniu de timp discret pot fi simulate utilizând algoritmi de generare a numerelor aleatoare presupunând că valorile generate succesiv reprezintă valori ale semnalului stochastic la timpi de prelevare succesivi. În acest scop, cel mai frecvent implementat pe calculatoarele personale este algoritmul congruențial liniar a cărui relație de recurență este:

$$I(n+1) = \{a I(n) + c\} [\text{modulo } m] \quad (\text{A2.1})$$

în care  $I(n)$  reprezintă numerele aleatoare generate și cuprinse între 0 și  $m - 1$ ,  $a$  și  $c$  - numere întregi iar  $m = 2^L$  ( $L$  - lungimea cuvintelor calculatorului). Secvența se inițiază cu un număr oarecare,  $I(0)$ , iar numerele următoare,  $I(n)$ , se calculează cu relația (A2.1). În final, se generează semnalul stochastic  $y(n)$ ,  $0 \leq n \leq N$ , în care fiecare  $y(n) = I(n)/m$  și  $0 \leq y(n) < 1$ . Valorile semnalului astfel generat sunt uniform distribuite în intervalul  $[0, 1)$  astfel încât  $m_y = 0,5$  și  $\sigma_y^2 = 0,0833$ . O porțiune dintr-un astfel de semnal este prezentată în figura A2.1. Serii foarte bune de numere aleatoare se pot obține alegând  $I(0)$  număr impar,  $c = 0$  și  $a = 8 \text{ INT} \pm 3$ , unde INT este un întreg. În astfel de cazuri, secvența periodică generată cuprinde  $2^{L-2}$  numere aleatoare independente într-o perioadă.

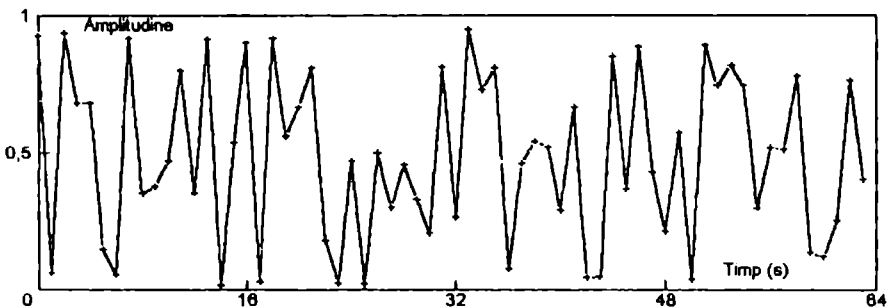


Fig. A2.1. Secvență de numere aleatoare distribuite uniform în intervalul  $[0, 1)$ .

Pentru obținerea de procese aleatoare cu medie și varianță dorite, se poate efectua o transformare liniară de forma:

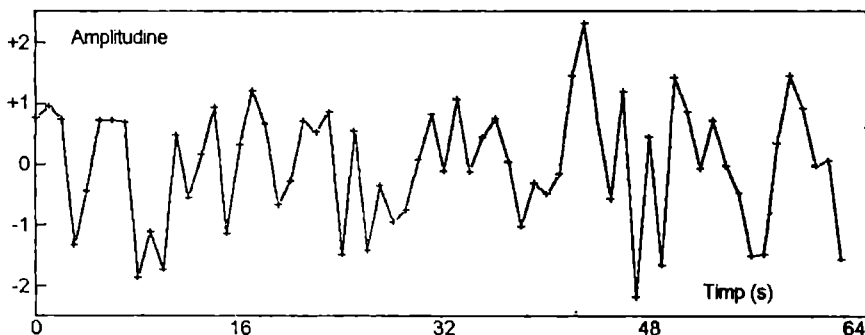
$$y(n) = a + b x(n) \quad (\text{A2.2})$$

care nu modifică forma funcției de probabilitate. Între mediile și varianțele proceselor aleatoare  $y(n)$  și  $x(n)$  există relațiile:

$$m_y = a + b m_x \quad \text{și} \quad \sigma_y^2 = b^2 \sigma_x^2 \quad (\text{A2.3})$$

Simularea unui proces aleator cu valori distribuite normal poate fi realizată în mai multe moduri. Unul dintre acestea se bazează pe teorema limită centrală care prevede că suma unui număr mare de variabile aleatoare independente, indiferent de distribuția acestora, reprezintă o variabilă aleatoare distribuită normal cu media și varianța egale cu suma mediilor și, respectiv, a varianțelor variabilelor aleatoare considerate. Simulările pe calculator [1] au arătat că însumarea a doar 12 variabile aleatoare distribuite uniform cu aceleași medii și varianțe produce deja o variabilă aleatoare distribuită normal. Mai mult chiar, variabilele însumate au fost extrase dintr-o singură secvență de numere aleatoare distribuite uniform,  $x(n)$ ,  $n = 1, 12 N$ , procesul aleator cu distribuție normală,  $y(n)$ ,  $n = 1, N$ , fiind generat în acord cu relațiile:

$$y(n) = \sum_{i=1}^{12} x_{12(n-1)+i}, \quad n = 1, N \quad (\text{A2.4})$$



**Fig. A2.2.** Secvență de numere aleatoare distribuite normal cu media zero.

Utilizând două secvențe de numere aleatoare independente, distribuite uniform în intervalul  $[0, 1)$  și notate  $x(n, 1)$  și  $x(n, 2)$ , o altă modalitate de generare a două procese aleatoare cu funcții de probabilitate normale,  $y(n, 1)$  și  $y(n, 2)$ , se bazează pe relațiile:

$$y(n, 1) = R(n) \cos(\text{ANG}(n)) \quad (\text{A2.5})$$

$$\text{și} \quad y(n, 2) = R(n) \sin(\text{ANG}(n)) \quad (\text{A2.6})$$

unde  $\text{ANG}(n)$  și  $R(n)$  sunt definite prin următoarele transformări unghiulare și, respectiv, de amplitudine:

$$\text{ANG}(n) = 2\pi x(n, 1) \quad \text{și} \quad R(n) = \sqrt{-2 \ln(x(n, 2))} \quad (\text{A2.7})$$

Cele două procese aleatoare astfel generate sunt independente între ele și fiecare dintre acestea are media  $m = 0$  și varianța  $\sigma^2 = 1$ . O secvență de numere aleatoare astfel simulate este prezentată în figura A2.2.

Simularea unui proces aleator cu o funcție de probabilitate dorită,  $f_y(\beta)$ , indiferent care este forma acesteia, se poate realiza prin aplicarea unei transformări care să modifice funcția de probabilitate cunoscută,  $f_x(\alpha)$ , a unui alt proces aleator generat în prealabil. Expresia acestei transformări este:

$$\int_{-\infty}^x f_x(\alpha) d\alpha = \int_{-\infty}^y f_y(\beta) d\beta \quad (\text{A2.8})$$

în care  $x$  semnifică, rând pe rând, toate valorile procesului aleator de pornire iar  $y$  - ale celui simulat.

Dacă se pornește de la un proces aleator având valorile distribuite uniform în intervalul  $[0, 1]$  și, deci, cu  $f_x(\alpha) = 1/\Delta x$  ( $\Delta x$  - lățimea intervalului), se obțin relațiile:

$$x = F_y(y) \quad \text{și} \quad y = F_x^{-1}(x) \quad (\text{A2.9})$$

Aplicarea lor pentru generarea unei variabile aleatoare cu o distribuție Rayleigh, pentru care:

$$f_y(\beta) = \frac{\beta}{a^2} \exp\left(-\frac{\beta^2}{2a^2}\right), \quad 0 \leq \beta < \infty \quad (\text{A2.10})$$

conduce la:

$$x = 1 - \exp\left(-\frac{y^2}{2a^2}\right) \quad \text{și} \quad y = \left(2a^2 \ln\left(\frac{1}{1-x}\right)\right)^{0,5} \quad (\text{A2.11})$$

Metodele de simulare a semnalelor aleatoare prezentate până în prezent conduc la procese pentru care  $\gamma_{x(\tau)} = 0$  pentru toți  $\tau \neq 0$ , aceste procese fiind denumite zgomot alb.

Zgomotul alb poate servi ca punct de pornire la simularea unor procese aleatoare,  $y(n)$ , care prezintă și corelație. Astfel, utilizând procesul aleator  $x(n)$ ,  $n = 0, N$  generat printr-una din metodele prezentate anterior, procese aleatoare prezentând corelație pot fi generate [1] utilizând fie modele ale mediei alunecătoare (notate MA(p)), fie modele autoregresive (MA(q)), fie modele mixte, având atât componente MA, cât și AR (ARMA(p, q)). Aceste modele sunt definite, în ordine, prin următoarele ecuații:

$$y(n) = \sum_{l=0}^p b(l)x(n-l) \quad (\text{A2.12})$$

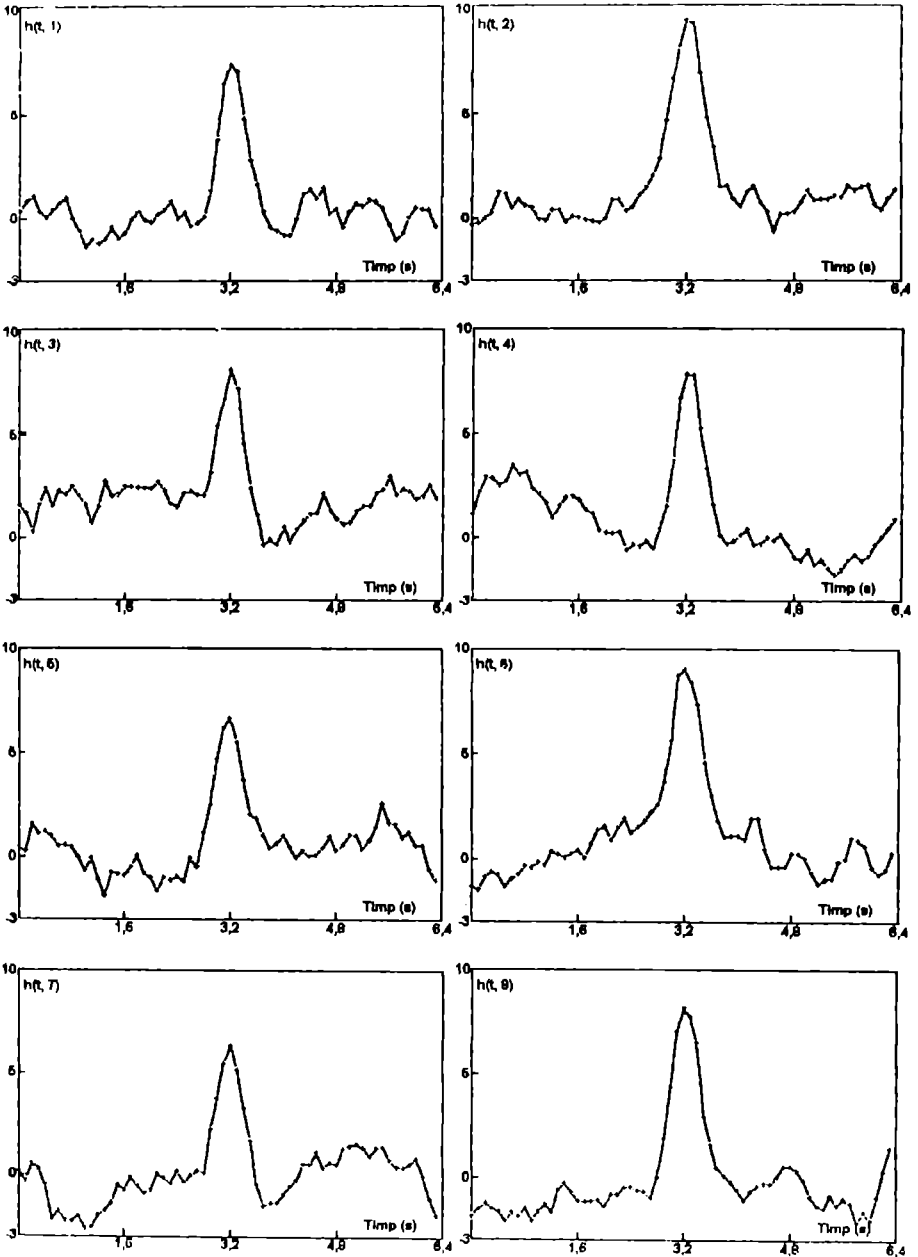
$$y(n) = -\sum_{l=1}^q a(l)x(n-l) \quad (\text{A2.13})$$

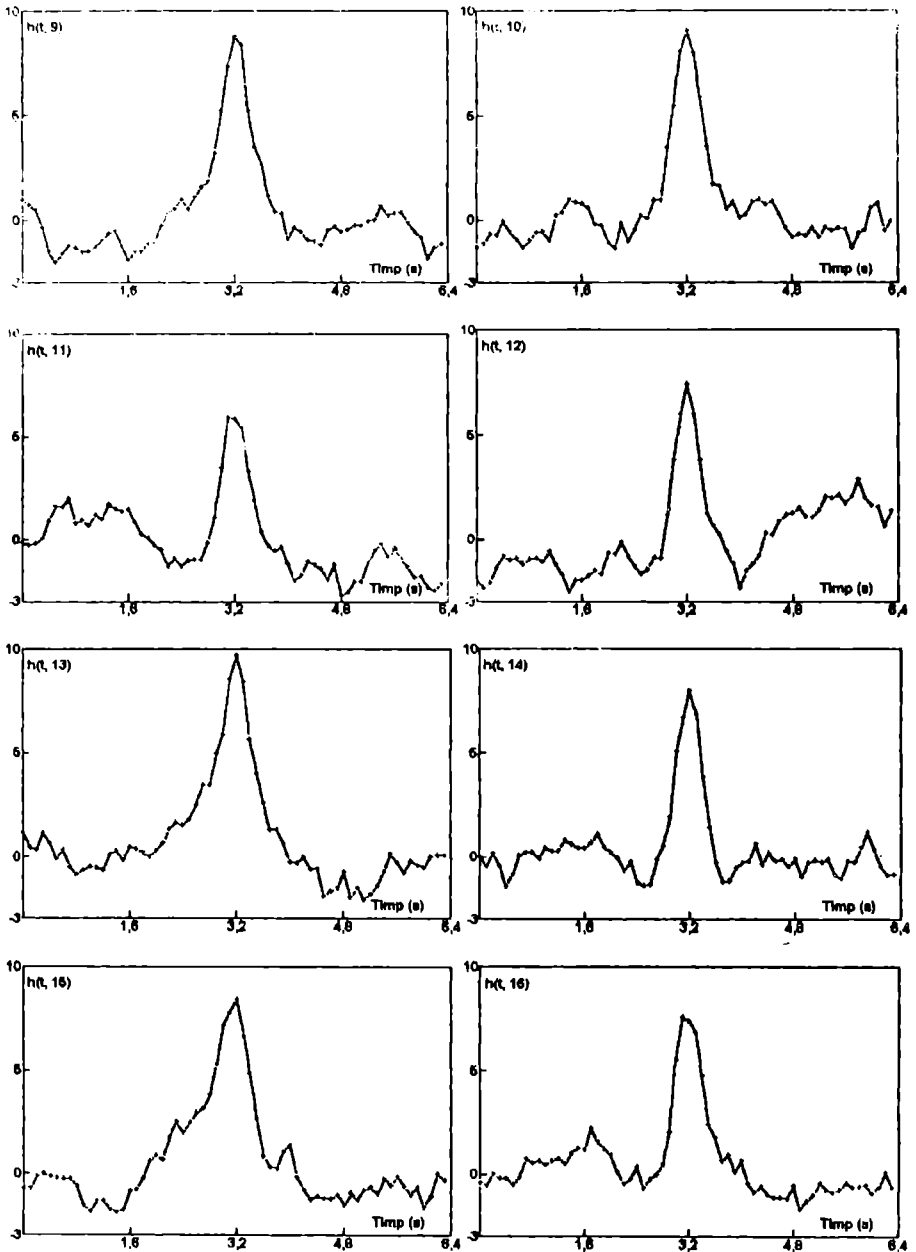
$$y(n) = -\sum_{l=1}^q a(l)x(n-l) + \sum_{l=0}^p b(l)x(n-l) \quad (\text{A2.14})$$

în care  $y(n)$  reprezintă valorile procesului aleator prezentând corelație;  $x(n)$  - valorile zgomotului alb considerat; iar  $a(l)$ , ( $l=1, q$ ) și  $b(l)$ , ( $l=0, p$ ) sunt coeficienți cu valori reale.

Se poate dovedi ușor că pentru astfel de procese aleatoare momentele de ordinul unu și doi se calculează cu relații simple în care intervin momentele de ordinul unu și doi ale zgomotului alb de pornire,  $x(n)$ ,  $n = 0, N$ , și valorile coeficienților  $b(l)$  și/sau  $a(l)$ .

## Ansamblu simulat reprezentat prin 16 funcții de probă





Funcțiile de probă simulate reprezintă un pic gaussian cu următoarele caracteristici: poziția maximumului picului  $m = 3,3$  s;  $h_{\max} = 8$  și  $\sigma = 0,2$  s, peste care s-a suprapus zgomot cu structură autoregresivă generat cu relația:  $y(n) = 0,9 y(n-1) + 0,5 x(n)$ . Raportul S/N = 6.



## Planuri experimentale de cântărire

Yates (1935) a arătat că prin cântărirea mai multor obiecte în grup este posibil să se determine mai precis masa obiectelor individuale. Pentru a demonstra acest lucru, presupunem că avem de cântărit patru obiecte cu masele necunoscute  $x_1, x_2, x_3, x_4$ , și că utilizăm în acest scop o balanță analitică cu două talere exactă a cărei precizie, cuantificată prin varianța erorilor experimentale,  $\sigma^2$ , nu depinde de masa care se află pe talere.

Considerăm trei planuri experimentale diferite de cântărire:

**A5.4.1. Cazul clasic (al cântăririlor secvențiale).** Pe unul din talerele balanței se pun pe rând cele patru obiecte. Balanța se echilibrează punând greutatea pe celălalt taler. Fie aceste greutăți  $y_1, y_2, y_3$  și  $y_4$ . Datorită erorilor experimentale care afectează cântările, între masele reale,  $x_i$ , și cele determinate experimental,  $y_i$ , există relațiile:

$$y_i = x_i + e_i, \quad i = 1, 2, 3 \text{ și } 4 \quad (\text{A4.1})$$

Pe baza cântăririlor efectuate, valorile estimate,  $\hat{x}_i$ , pentru masele obiectelor individuale sunt:

$$\hat{x}_i = y_i, \quad i = 1, 2, 3, \text{ și } 4 \quad (\text{A4.2})$$

Valorile de așteptare ale estimațiilor sunt egale cu masa exactă a obiectelor cântărite:

$$E[\hat{x}_i] = x_i, \quad i = 1, 2, 3 \text{ și } 4 \quad (\text{A4.3})$$

ceea ce este în acord cu faptul că balanța este exactă. Așadar, deși cântările individuale pot fi afectate de erori întâmplătoare, media unui mare număr de cântări repetate ale unui obiect nu diferă semnificativ de valoarea reală a masei obiectului respectiv.

Conform ipotezelor referitoare la precizia balanței utilizate, valorile de așteptare ale varianțelor erorilor sunt aceleași,  $\sigma^2$ :

$$E[(\hat{x}_i - x_i)^2] = E[e_i^2] = \sigma^2, \quad i = 1, 2, 3 \text{ și } 4 \quad (\text{A4.4})$$

Aceasta semnifică faptul că, deși nu se poate afirma cu certitudine care este mărimea erorii care afectează valoarea masei unui obiect determinată printr-un singur experiment de cântărire,  $e_i$ , cel mai probabil aceasta este apropiată, în valoare absolută, de deviația standard,  $\sigma$ .

O cale de micșorare a erorii care afectează determinarea masei unui obiect individual o constituie cântărirea repetată a obiectului și

medierea rezultatelor obținute. Eficiența acestei metode este însă limitată de faptul că sunt necesare un număr foarte mare de determinări individuale pentru o îmbunătățire semnificativă a preciziei. Astfel, din modul de calcul al varianței unei sume de variabile aleatoare, rezultă că, dacă erorile care afectează cântările individuale sunt independente între ele, pentru a micșora de  $n$  ori deviația standard a rezultatului, cântărirea trebuie repetată de  $n^2$  ori. De aceea, s-au căutat și alte metode mai eficiente de îmbunătățire a preciziei cântărilor. Ele se aplică însă doar atunci când trebuie determinate masele mai multor obiecte, fiind cu atât mai eficiente cu cât numărul obiectelor de cântărit este mai mare.

**A5.4.2. Cazul cântărilor în grup punând obiectele de măsurat pe ambele talere.** În prima cântărire, toate obiectele se așează pe un singur taler iar în următoarele trei - se combină pe rând obiectele între ele pentru a forma două grupuri de câte două obiecte. Considerând obiectele numerotate, un mod de grupare posibil este: întâi 1 cu 3 și 2 cu 4, apoi 1 cu 2 și 3 cu 4 și, în final, 1 cu 4 și 2 cu 3. Simultan, grupurile de obiecte se așează pe talerele balanței (câte un grup pe fiecare taler) iar diferența de masă se compensează adăugând greutatea pe talerul cu obiectele mai ușoare. În acest caz, între masele greutăților utilizate la echilibrarea balanței,  $y_i$ , și cele ale obiectelor individuale,  $x_i$ , există relațiile următoare:

$$\begin{aligned} y_1 &= x_1 + x_2 + x_3 + x_4 + e_1 \\ y_2 &= x_1 - x_2 + x_3 - x_4 + e_2 \\ y_3 &= x_1 + x_2 - x_3 - x_4 + e_3 \\ y_4 &= x_1 - x_2 - x_3 + x_4 + e_4 \end{aligned} \tag{A4.5}$$

Estimațiile maselor obiectelor, obținute prin rezolvarea sistemului de ecuații (A4.5), sunt următoarele:

$$\begin{aligned} \hat{x}_1 &= \frac{1}{4} (y_1 + y_2 + y_3 + y_4) = x_1 + \frac{1}{4} (e_1 + e_2 + e_3 + e_4) \\ \hat{x}_2 &= \frac{1}{4} (y_1 - y_2 + y_3 - y_4) = x_2 + \frac{1}{4} (e_1 - e_2 + e_3 - e_4) \\ \hat{x}_3 &= \frac{1}{4} (y_1 + y_2 - y_3 - y_4) = x_3 + \frac{1}{4} (e_1 + e_2 - e_3 - e_4) \\ \hat{x}_4 &= \frac{1}{4} (y_1 - y_2 - y_3 + y_4) = x_4 + \frac{1}{4} (e_1 - e_2 - e_3 + e_4) \end{aligned} \tag{A4.6}$$

iar ale varianțelor erorilor care afectează aceste rezultate:

$$1 \left[ (\hat{x}_i - x_i)^2 \right] = \frac{4 \sigma^2}{16} = \frac{\sigma^2}{4}, \quad i = 1, 2, 3 \text{ și } 4 \tag{A4.7}$$

deoarece varianța sumei sau diferenței unor variabile aleatoare independente este egală cu suma varianțelor variabilelor aleatoare individuale. Exemplificând deducerea relației (A4.7) pentru  $i = 1$ , rezultă într-adevăr că:

$$E\left[(\hat{x}_1 - x_1)^2\right] = E\left[\left\{\frac{1}{4} (e_1 + e_2 + e_3 + e_4)\right\}^2\right] = \quad (A4.8)$$

$$= \frac{1}{16} E\left[(e_1 + e_2 + e_3 + e_4)^2\right] = \frac{4 \sigma^2}{16} = \frac{\sigma^2}{4}$$

Se observă că, utilizând acest plan de cântărire, se reduce de 4 ori varianța erorilor de măsură iar raportul semnal/zgomot crește de 2 ori față de cazul în care obiectele sunt cântărite individual.

**A5.4.3. Cazul cântăririlor în grup punând obiectele pe un singur taler.** În acest caz vom considera că obiectele se grupează astfel: mai întâi obiectele 2, 3 și 4 și, după aceea, pe rând, obiectul 1 cu fiecare din celelalte 3, și că, la fel ca și în cazul A5.4.1, grupurile de obiecte se așează pe un singur taler, greutatea lor fiind compensată cu greutatea  $y_i$ , ( $i=1, 4$ ). Se poate acum scrie:

$$\begin{aligned} y_1 &= x_2 + x_3 + x_4 + e_1 \\ y_2 &= x_1 + x_2 + e_2 \\ y_3 &= x_1 + x_3 + e_3 \\ y_4 &= x_1 + x_4 + e_4 \end{aligned} \quad (A4.9)$$

de unde se obțin următoarele valori de estimare ale obiectelor de cântărit:

$$\begin{aligned} \hat{x}_1 &= \frac{1}{3} (y_1 + y_2 + y_3 + y_4) = x_1 + \frac{1}{3} (e_1 + e_2 + e_3 + e_4) \\ \hat{x}_2 &= \frac{1}{3} (2 y_2 + y_1 - y_3 - y_4) = x_2 + \frac{1}{3} (2 e_2 + e_1 - e_3 - e_4) \\ \hat{x}_3 &= \frac{1}{3} (2 y_3 + y_1 - y_2 - y_4) = x_2 + \frac{1}{3} (2 e_3 + e_1 - e_2 - e_4) \\ \hat{x}_4 &= \frac{1}{3} (2 y_4 + y_1 - y_2 - y_3) = x_2 + \frac{1}{3} (2 e_4 + e_1 - e_2 - e_3) \end{aligned} \quad (A4.10)$$

și următoarele valori de așteptare pentru varianțele acestora:

$$E\left[(\hat{x}_1 - x_1)^2\right] = \frac{4 \sigma^2}{9} \quad (A4.11)$$

$$E\left[(\hat{x}_i - x_i)^2\right] = \frac{7 \sigma^2}{9}, \quad i = 2 - 4$$

În concluzie, și în acest caz se îmbunătățește raportul semnal/zgomot față de cazul A.5.4.1., dar într-o măsură mai mică decât în cazul A.5.4.2.

**A.5.4.4. Generalizare** În general, dacă avem  $p$  necunoscute,  $(x_1, x_2, \dots, x_p)$ , și  $N$  ( $N > p$ ) măsurători,  $(y_1, y_2, \dots, y_N)$ , afectate de erorile  $(e_1, e_2, \dots, e_N)$ , putem scrie setul de ecuații:

$$y_i = w_{i1} x_1 + w_{i2} x_2 + \dots + w_{iN} x_N, \quad i = 1, N \quad (\text{A4.12})$$

care, în notații matriceale (vezi Capitolul IV) poate fi scris astfel:

$$\mathbf{y} = \mathbf{W} \mathbf{x} + \mathbf{e} \quad (\text{A4.13})$$

unde  $\mathbf{W}$  reprezintă o matrice cu  $N$  rânduri și  $p$  coloane. În cazul experimentelor de cântărire, coeficienții  $w_{ij}$  ( $i = 1, N$  și  $j = 1, p$ ) pot lua doar valorile 0, 1 și -1 iar specificarea lor completă se numește "plan experimental de cântărire".

## VI. METODOLOGIA SUPRAFETEI RĂSPUNSULUI

### VI.1. Suprafața de răspuns

Suprafața de răspuns este dată de reprezentarea grafică a răspunsului sistemului în funcție de unul sau mai mulți factori sau variabile independente. Suprafața de răspuns oferă chemometrului posibilitatea de a vizualiza modul în care este afectată măsurarea sistemului studiat de către diferiți factori. Studiul metodologiei suprafeței răspunsului a fost descrisă și dezvoltată pentru prima dată de către Box și Wilson[1,2].

Metodologia suprafeței răspunsului este o strategie experimental-statistică aplicată studiului răspunsului sistemelor. În general această metodologie se aplică în studiul proceselor industriale evident pentru îmbunătățirea performanțelor acestora dar sunt și posibilități de utilizare în chimia analitică privind răspunsul proceselor chimice.

Considerăm funcția de răspuns necunoscută în forma generală:

$$\hat{y} = f(x_1, x_2, \dots, x_n) \quad (6.1)$$

Dacă se dezvoltă funcția de răspuns în serie Taylor, în jurul unui punct, rezultă că orice funcție se poate aproxima printr-un polinom de ordinul I:

$$y = b_0 + \sum_{i=1}^n b_i x_i \quad (6.2)$$

de ordinul al II-lea:

$$y = b_0 + \sum_{i=1}^n b_i x_i + \sum_{\substack{i,j=1 \\ i \neq j}}^n b_{ij} x_i x_j + \sum_{i=1}^n b_{ii} x_i^2 \quad (6.3)$$

sau printr-un polinom de ordin superior. În concluzie, oricât de complicată ar fi funcția de răspuns, aceasta poate fi aproximată printr-un model liniar în raport cu parametrii pe care îi conține.

În acest caz, estimarea coeficienților prin metoda celor mai mici pătrate se face din condiția de minim pentru suma pătratelor deviațiilor (SPD):

$$SPD = \sum_{i=1}^n (y_i - b_0 + b_1 x_{1i} + \dots + b_{11} x_{1i}^2 + \dots)^2 = \text{minim} \quad (6.4)$$

adică, din sistemul de ecuații normale:

$$\begin{aligned} \frac{\partial (\text{SPD})}{\partial b_0} &= 0 \\ \frac{\partial (\text{SPD})}{\partial b_1} &= 0 \\ &\dots\dots\dots \\ \frac{\partial (\text{SPD})}{\partial b_{11}} &= 0 \end{aligned} \quad (6.5)$$

În cazul polinomului de ordinul întâi având următoarea expresie:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (6.6)$$

sistemul de ecuații normale obținut cu ajutorul metodei celor mai mici pătrate în care coeficienții modelului  $\beta_0, \beta_1, \dots, \beta_k$  au fost înlocuiți cu parametrii de estimare  $b_0, b_1, \dots, b_k$  este:

$$\begin{aligned} b_0 \sum_{i=1}^n x_{0i}^2 + b_1 \sum_{i=1}^n x_{0i}^2 x_{1i} + \dots + b_k \sum_{i=1}^n x_{0i} x_{ki} &= \sum_{i=1}^n x_{0i} y_i \\ b_0 \sum_{i=1}^n x_{1i} x_{0i} + b_1 \sum_{i=1}^n x_{1i}^2 + \dots + b_k \sum_{i=1}^n x_{1i} x_{ki} &= \sum_{i=1}^n x_{1i} y_i \\ &\dots\dots\dots \\ b_0 \sum_{i=1}^n x_{ki} x_{0i} + b_1 \sum_{i=1}^n x_{ki} x_{1i} + \dots + b_k \sum_{i=1}^n x_{ki}^2 &= \sum_{i=1}^n x_{ki} y_i \end{aligned} \quad (6.7)$$

Acest sistem de ecuații poate fi reprezentat compact sub formă matricială:

$$X^T X \beta = X^T y \quad (6.8)$$

unde am introdus matricea  $X$  și transpusa sa  $X^T$  pentru  $n$  experimente și  $k$  variabile independente:

$$X = \begin{pmatrix} x_{01} & x_{11} & \dots & x_{k1} \\ x_{02} & x_{12} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ x_{0n} & x_{1n} & \dots & x_{kn} \end{pmatrix}, \quad (6.9)$$

matricea valorilor răspunsurilor corespunzătoare fiecărui experiment reprezentat printr-o linie a matricii  $X$ :

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad (6.10)$$

precum și matricea coeficienților polinomului de aproximare:

$$\beta = \begin{pmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_k \end{pmatrix} \quad (6.11)$$

Atunci, în condițiile în care matricea  $X^T X$  este inversabilă (determinantul său asociat de ordinul  $k+1$  nenul), rezolvând ecuația (6.8) în raport cu matricea  $\beta$  se obțin valorile de estimare:

$$\beta = (X^T X)^{-1} X^T y \quad (6.12)$$

## 6.2. Experimente factoriale

Dacă în cazul unui experiment activ, matricea  $X$  de programare a experimentului se poate defini astfel încât matricea  $X^T X$  să devină o matrice diagonală de forma :

$$X^T X = \begin{vmatrix} \sum_{j=1}^n x_{0j}^2 & 0 & \dots & 0 \\ 0 & \sum_{j=1}^n x_{1j}^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sum_{j=1}^n x_{kj}^2 \end{vmatrix}$$

(6.13)

Un experiment cu o matrice  $X^T X$  de tip diagonală se numește experiment ortogonal [3,4]. În acest caz, coeficienții polinomului (6.2) se calculează astfel:

$$b_i = \left( \sum_{j=1}^n x_{ij}^2 \right)^{-1} \sum_{j=1}^n x_{ij} y_i \quad (6.14)$$

unde  $i=0,1,2, \dots, k$  și  $j=1,2, \dots, n$ .

Ca un avantaj esențial al utilizării experimentului ortogonal la studiul suprafeței de răspuns o constituie faptul că asigură un volum de calcule mic pentru evaluarea coeficienților polinomului și analiza statistică a rezultatelor experimentale. Coeficienții obținuți cu ajutorul unui experiment factorial



ortogonal nu sunt interdependenți și deci în cazul corecturii date  
matematic, coeficienții deja calculați rămân valabili.

### **VI.2.1. Experiment factorial complet la două nivele**

#### **(Plan experimental $2^n$ )**

Experimentele factoriale sunt o clasă de experimente foarte economice care oferă o cantitate mare de informație dintr-un număr mic de experimente. Când numărul experimentelor efectuate sunt în număr mic, experimentul factorial oferă calea cea mai eficientă pentru a obține maximum de informație al experimentelor. Experimentele factoriale pot fi de două feluri:

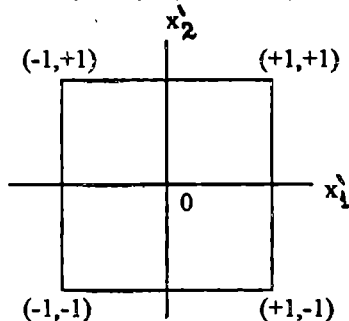
1) de tipul  $2^n$  unde 2 reprezintă numărul de nivele iar n, numărul factorilor;

2) de mai multe nivele, cu 2 sau mai mulți factori.

Cea mai cunoscută schemă de programare a experimentului factorial este aceea în care variabilele (factorii) iau numai două valori de o parte și de alta a unui nivel de bază. În chimia analitică factorii pot fi: pH, temperatură, timp, reactivi utilizați, aparat de măsură, mod de lucru, etc. Se poate programa experimental și la mai multe nivele dar în acest caz realizarea experimentului și prelucrarea datelor sunt mult mai grele. În cazul unui experiment factorial la două nivele, pentru toți factorii, numărul de combinații posibile este egal cu  $2^k$ . În acest caz avem un experiment factorial cu un plan experimental  $2^k$ . Se notează nivelul inferior și superior pentru fiecare factor în parte prin  $(x_i)_1$  și  $(x_i)_2$ , unde  $i = 1, 2, \dots, n$  și se introduc pentru fiecare variabilă în parte mărimile:

$$\begin{aligned}
 x_1^0 &= \frac{(x_1)_1 + (x_1)_2}{2}, & \text{nivel de baza} \\
 \Delta x_1 &= \frac{(x_1)_2 - (x_1)_1}{2}, & \text{unitatea intervalului} \\
 x_1' &= \frac{x_1 - x_1^0}{\Delta x_1}, & \text{valori codificate}
 \end{aligned}
 \tag{6.15}$$

Deci, în urma centrării și normării, pentru fiecare factor în parte, nivelul superior este codificat cu +1 iar nivelul inferior cu -1, coordonatele centrului fiind nule. În cazul particular a doi factori,  $k=2$  deci  $n=2^2=4$ ; toate combinațiile posibile între nivelele celor doi factori sunt date în figura 6.1.



*Fig. 6.1. Experiment factorial complet pentru doi factori la două nivele*

Matricea de programare a unui experiment complet pentru doi factori la două nivele fiecare este prezentată în tabelul 6.1.

*Tabel 6.1. Programul experimentului factorial  $2^2$*

Număr experiment	Factori			Răspuns
	$x_0'$	$x_1'$	$x_2'$	
1	+1	+1	+1	$y_1$
2	+1	-1	+1	$y_2$
3	+1	+1	-1	$y_3$
4	+1	-1	-1	$y_4$

În cazul a trei factori, numărul de combinații posibile între toți factorii pentru două nivele  $n = 2^3 = 8$  după cum se poate vedea în figura 6.2 iar matricea de programare a unui experiment complet la două nivele pentru trei factori este prezentată în tabelul 6.2.

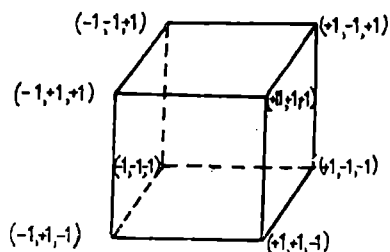


Fig. 6.2. Experiment factorial complet pentru 3 factori la două nivele.

Tabelul 6.2. Experiment factorial  $2^3$

Număr experiment	Factori				Răspuns
	$x_0$	$x_1$	$x_2$	$x_3$	
1	+1	+1	+1	+1	$y_1$
2	+1	-1	+1	+1	$y_2$
3	+1	+1	-1	+1	$y_3$
4	+1	-1	-1	+1	$y_4$
5	+1	+1	+1	-1	$y_5$
6	+1	-1	+1	-1	$y_6$
7	+1	+1	-1	-1	$y_7$
8	+1	-1	-1	-1	$y_8$

În acest caz, valorile de estimare  $b_0, b_1, \dots, b_n$ , pentru coeficienții polinomului liniar conform relației 6.14 se calculează astfel:

$$B = \begin{pmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{pmatrix} = (X^T X)^{-1} X^T Y \quad (6.16)$$

adică:

$$b_i = \frac{1}{n} \sum_{j=1}^n x'_{ij} y_j, \quad i = 0, 1, \dots, k; \quad j = 1, 2, \dots, n \quad (6.17)$$

În cazul unui experiment factorial la două nivele pentru trei factori, coeficientul  $b_1$  se calculează astfel:

$$b_1 = \begin{pmatrix} +1 \\ -1 \\ +1 \\ -1 \\ +1 \\ -1 \\ +1 \\ -1 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{pmatrix} = \frac{(y_1 + y_3 + y_5 + y_7) + (y_2 + y_4 + y_6 + y_8)}{8} \quad (6.18)$$

În mod similar se calculează și ceilalți coeficienți:  $b_0, b_2$  și  $b_3$ . Modelul experimentului factorial de tipul  $n=2^3$  care ia în considerare și interacțiunea factorilor are următoarea expresie:

$$y^n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 \quad (6.19)$$

pentru estimarea coeficienților  $\beta_{12}, \beta_{13}, \beta_{23}$  (efectul interacțiunii duble) și  $\beta_{123}$  (efectul interacțiunii triple) nu sunt necesare experimente suplimentare.

Valorile de estimare  $b_{12}, b_{23}, b_{13}, b_{123}$  pentru coeficienții care evaluează interacțiunile sunt determinate în același mod ca și coeficienții termenilor liniari. De exemplu, coeficientul  $b_{12}$  se calculează în felul următor:

$$b_{12} = \begin{array}{c} \begin{array}{|c|} \hline +1 \\ \hline \end{array} \begin{array}{|c|} \hline y_1 \\ \hline \end{array} \\ \begin{array}{|c|} \hline -1 \\ \hline \end{array} \begin{array}{|c|} \hline y_2 \\ \hline \end{array} \\ \begin{array}{|c|} \hline -1 \\ \hline \end{array} \begin{array}{|c|} \hline y_3 \\ \hline \end{array} \\ \begin{array}{|c|} \hline +1 \\ \hline \end{array} \begin{array}{|c|} \hline y_4 \\ \hline \end{array} \\ \begin{array}{|c|} \hline +1 \\ \hline \end{array} \begin{array}{|c|} \hline y_5 \\ \hline \end{array} \\ \begin{array}{|c|} \hline -1 \\ \hline \end{array} \begin{array}{|c|} \hline y_6 \\ \hline \end{array} \\ \begin{array}{|c|} \hline -1 \\ \hline \end{array} \begin{array}{|c|} \hline y_7 \\ \hline \end{array} \\ \begin{array}{|c|} \hline +1 \\ \hline \end{array} \begin{array}{|c|} \hline y_8 \\ \hline \end{array} \end{array} = \frac{(y_1 + y_4 + y_5 + y_8) - (y_2 + y_3 + y_6 + y_7)}{8}$$

(6.20)

Dispersiile pentru erorile de estimare ale coeficienților  $b_1, b_2, \dots, b_k$  care în cazul experimentului ortogonal se determină independent, sunt egale:  $\sigma_{b1}^2 = \sigma_{b2}^2 = \dots = \sigma_{bk}^2$ . Aceste dispersii au valoarea de estimație

$s_{b_i}^2 = \frac{s_e^2}{n}$ ,  $i = 1, 2, \dots, k$  unde  $s_e^2$  reprezintă valoarea de estimare pentru fluctuația răspunsului evaluată din măsurări repetate. Semnificația coeficienților  $b_1, b_2, \dots, b_k$ , adică semnificația efectului liniar al factorilor  $x_1, x_2, \dots, x_k$  asupra răspunsului se stabilește cu testul  $t$  (Student). Astfel se calculează pentru fiecare coeficient mărimea :

$$t_{\text{exp}} = \frac{s_i}{s_{b_i}} \quad (6.21)$$

Ipoteza semnificației coeficientului  $b_i$  se acceptă pentru un nivel de semnificație  $\alpha$ , dacă  $t_{\text{exp}} > t_{(1-\alpha/2; \nu)}$

Concordanța dintre polinomul evaluat și valoarea adevărată a funcției de răspuns necunoscută se verifică cu ajutorul testului  $F$ :

$$F_{\text{exp}} = \frac{s_i^2}{s_e^2} \quad (6.22)$$

Ipoteza că polinomul evaluat se deosebește semnificativ de funcția de răspuns adevărată necunoscută, adică cele două funcții nu concordă, se

acceptă dacă:  $F > F(\alpha, \nu_1, \nu_2)$ . Prin  $F(\alpha, \nu_1, \nu_2)$  s-a notat parametrul testului  $F$ , corespunzător unui nivel de semnificație  $\alpha$ , iar  $\nu_1$  și  $\nu_2$  grade de libertate.

### VI.2.2. Experimente factoriale fracționate

În cazul în care este suficientă aproximarea suprafeței de răspuns printr-un polinom liniar (hiperplan), nu sunt necesare toate măsurătorile unui experiment factorial  $n = 2^k$ . Atunci când interacțiunile lipsesc, se recurge la fracțiuni (jumătăți, sferturi) dintr-un experiment total, care prezintă însă proprietăți de ortogonalitate.

Experimentele factoriale fracționate nu sunt foarte eficiente deoarece interacțiunile dintre factori există și sunt în general numeroase. De asemenea, experimentele factoriale fracționate nu oferă date pentru calcularea dispersiei experimentelor. În consecință, experimentele factoriale fracționate sunt indicate atunci când dispersia experimentelor este cunoscută *a priori*.

### VI.2.3. Experiment factorial la trei nivele

Zona din jurul optimului funcției de răspuns, zonă care se cere evaluată mult mai corect, se aproximează de obicei cu un polinom de ordinul II de forma următoare:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \quad (6.23)$$

Sunt mai multe planuri speciale elaborate pentru evaluarea funcției de răspuns de ordinul II ca de exemplu : planul experimental compozițional [5] și planul experimental rotabil [6].

Planul experimental compozițional este alcătuit din :

- un experiment factorial complet la două nivele  $2^k$ , pentru  $k < 5$  și un experiment factorial fracționat  $2^{k-1}$  pentru  $n \geq 5$ ;
- $2k$  puncte experimentale:  $(\pm\alpha, 0, 0, \dots, 0)$ ,  $(0, \pm\alpha, 0, \dots, 0)$ , ... ,  $(0, 0, \dots, \pm\alpha)$  situate la distanța  $\alpha$  de centru.
- $k_0$  puncte experimentale în centrul planului.

În concluzie, pentru un experiment compozițional numărul total de măsurători este:

$$n = \begin{cases} 2^k + 2k + k_0, & k < 5 \\ 2^{k-1} + 2k + k_0, & k \geq 5 \end{cases} \quad (6.23)$$

În cazul experimentului compozițional, valorile parametrului  $\alpha$  se aleg astfel încât să se asigure ortogonalitatea rezultatelor experimentului.

Box și Hunter [7] elaborează un plan experimental compozițional rotabil pentru evaluarea suprafețelor de răspuns de ordinul II. Pentru ca experimentul compozițional să fie rotabil se alege parametrul  $\alpha$  astfel încât să satisfacă condiția de rotabilitate  $\alpha = 2^{r/4}$ , und  $r$  este exponentul tipului de experiment factorial la două nivele. De exemplu, pentru experimentul factorial fracționat de tip  $2^{k-1}$  parametrul  $\alpha$  ia valoarea:  $\alpha = 2^{(k-1)/4}$ . Matricea de programare a unui experiment rotabil nu este ortogonală și de aceea calcularea coeficienților termenilor pătratici ai polinomului de aproximare este mai complicată.

### ***Bibliografie***

1. G.E.P. Box and K.B. Wilson, *J. R. Statist. Soc.*, **B**, 13, 1, 1951.
2. G.E.P. Box and D.W. Behnken, *Technometrics*, **2**, 445, 1960.
3. N.R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley, New York, 1966.
4. W.G. Cochran and C.M. Cox, *Experimental Design*, John Wiley, New York, 1967.
5. G.E.P. Box and J.S. Hunter, *Ann. Math. Statist.*, **28**, 195, 1975.

# VII. OPTIMIZAREA IN CHIMIA ANALITICA

## VII. 1. Generalități

O preocupare frecventă a analiștilor este aceea de a adapta o metodă existentă unui anumit tip de probă sau unui anumit laborator. De exemplu trebuie ajustat volumul de reactiv, timpul de reacție sau temperatura în vederea atingerii performanțelor corespunzătoare problemei studiate deci trebuie optimizată metoda. Optimizarea nu presupune numai modificarea parametrilor pentru obținerea optimului metodei ci și modificarea însăși a sistemului. Problema alegerii optime a unei metode sau a unui grup de metode pe baza unor informații existente în literatură poate fi rezolvată cu ajutorul chemometriei. Deci ca o primă etapă în cadrul procesului de optimizare a unei metode de analiză o constituie alegerea caracteristicii sau caracteristicilor de performanță cu ajutorul cărora urmează ulterior să se definitiveze deciziile corespunzătoare.

În cadrul metodelor utilizate în analiza chimică se pot defini o serie de caracteristici după cum urmează: caracteristici tehnice (precizie, corectitudine, selectivitate, limită de detecție, cantitatea de informație, amplitudinea semnalului, raportul semnal/zgomot, eroarea totală de analiză, limita de determinare, rezoluția, etc.) și caracteristici economice (costul, timpul necesar efectuării unei determinări, etc.

## VII. 2. Reprezentarea matematică a sistemului de optimizat

Modelul matematic al unui proces este dat de relația dintre variabila de răspuns ( $y$ ) și variabilele independente ( $x_1, x_2, \dots, x_n$ ) model prezentat în următoarea formă :

$$y = f(x_1, x_2, \dots, x_n, b_1, b_2, \dots, b_n) \quad (7.1)$$

în care  $b_1, b_2, \dots, b_n$  reprezintă parametrii modelului.

În cazul în care procesul nu poate fi modelat matematic adecvat se folosesc metodele statistice ale experimentului activ, adică experimentul activ se desfășoară conform unui plan experimental elaborat în prealabil.

Reprezentarea matematică adecvată a sistemului de optimizat constituie obiectivul principal în rezolvarea optimizării privind dependența dintre criteriul de optimizare și variabilele de optimizat. Problema optimizării unui sistem, în general este abordată după descompunerea sistemului în subsistemele componente deoarece este mai ușor de rezolvat optimizarea pe subsisteme datorită numărului mai redus de variabile. În general, un sistem analitic poate fi descompus în următoarele subsisteme: eșantionarea materialului, tratamentul fizico-chimic al eșantionului, măsurarea semnalului analitic, prelucrarea semnalelor și calculul rezultatelor. Sistemul analitic cu funcție de detecție, identificare sau determinare este un lanț informațional în care la intrare se găsește materialul de analizat iar la



În cazul reprezentării matematice a unui sistem analizat în scopul optimizării acestuia se disting două cazuri [1] :

- sistemul de optimizat este riguros cunoscut deci poate fi reprezentat cantitativ în vederea evaluării funcției obiectiv (dependența dintre criteriul de optimizat și variabilele sistemului) cât și a restricțiilor pe care trebuie să le respecte funcția obiectiv și variabilele sistemului.

- sistemul de optimizat există în funcțiune iar reprezentarea sa nu este posibilă sau este greu de realizat. În acest caz, funcția obiectiv se evaluează pe cale experimentală în acord cu principiul black-box.

### VII.3. Funcția de răspuns și restricțiile

Funcția care exprimă dependența dintre caracteristica de calitate (criteriu de optimizare) și variabilele de optimizat se numește funcție de răspuns (funcție obiectiv sau funcție de optimizat).

În general, funcția de răspuns se notează prin expresia de mai jos :

$$y(X) = y(x_1, x_2, \dots, x_n) \quad (7.2)$$

În practică, variabilelor  $x_1, x_2, \dots, x_n$  le sunt impuse de regulă restricții, adică ele nu pot lua valori decât în anumite domenii. Se disting restricții de tip egalitate notate prin :

$$g_i = g(X) = 0, \quad i = 1, 2, \dots, m \quad (7.3)$$

și restricții de tip inegalitate prin :

$$h_j = h(X) > 0, \quad j = 1, 2, \dots, l. \quad (7.4)$$

În chimia analitică funcțiile obiectiv pot fi : eroarea de analiză, timpul afectat unei determinări, capacitatea de detecție, costul unei determinări, determinarea simultană a mai multor componenți dintr-o probă.

Funcția obiectiv în cazul optimizării capacității de detecție și determinare a unei metode este de fapt raportul semnal/zgomot. Deci trebuie alese acele condiții analitice cărora le corespunde un raport semnal/zgomot maxim. În general, problema optimizării răspunsului în chimia analitică se rezumă la găsirea acelor valori pentru condițiile analitice pentru care răspunsul capătă o valoare maximă.

Alegerea metodei de optimizare este determinată și de complexitatea matematică a funcției obiectiv și a restricțiilor. Chiar pentru sisteme relativ simple, funcția de răspuns (obiectiv) și restricțiile pot avea formulări matematice deosebit de complicate. Pentru rezolvarea problemelor de optimizare în acest caz sunt necesare metode numerice adecvate.

În cazul unei variabile de optimizat funcția obiectiv poate fi reprezentată grafic și examinată în vederea obținerii optimului. În acest caz se constată că optimul ales poate însemna o valoare bine determinată pentru variabila de optimizat sau un interval.

În cazul a două variabile  $x_1$  și  $x_2$ , funcția de răspuns este reprezentată în spațiul tridimensional de către o suprafață de forma următoare (figura 7.1) :

$$y = y(x_1, x_2) \quad (7.5)$$

sau printr-un sistem de curbe de nivel constant  $y = y(x_1, x_2) = \text{constant}$  (figura 7.2).

În cazul a  $n$  variabile independente de optimizat ( $n > 2$ ) funcția de răspuns este reprezentată de o hipersuprafață în spațiul  $n+1$  dimensiuni. Hipersuprafețele  $y = y(X) = \text{constant}$  sunt denumite hipersuprafețe de nivel constant al funcției de răspuns.

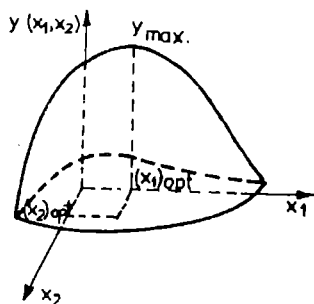


Fig. 7.1. Representarea grafică a funcției obiectiv de două variabile.

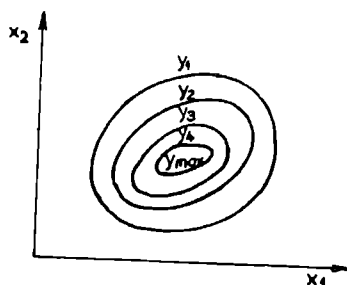


Fig. 7.2. Representarea grafică a funcției obiectiv de două variabile prin curbe de nivel constant.

Deci optimizarea se definește ca o cercetare a valorilor lui  $X_j$  care dau valoarea lui  $Y$  cea mai ridicată sau cea mai scăzută. Atât problema reprezentării matematice a sistemului de optimizat deci a stabilirii funcției obiectiv cât și problema găsirii maximumului sau minimumului funcției nu sunt probleme simple de abordat.

$\text{opt.}[y(X); g_i(X) = 0, i = 1, 2, \dots, m; h_j(X) > 0, j = 1, 2, \dots, l]$  (7.6)

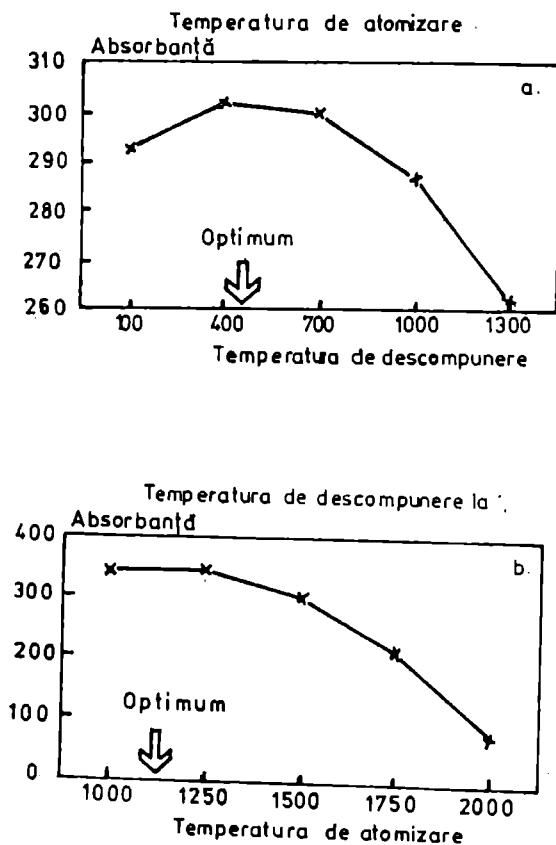
reprezintă maximumul sau minimumul funcției obiectiv  $y(X)$  și restricțiile corespunzătoare. Trecerea de la o problemă de maxim la una de minim se face conform relației :

$$\max. y(X) = \min. [-y(X)] \quad (7.7)$$

În concluzie problema optimizării este deplasată către aceea a explorării suprafeței de răspuns.

#### VII. 4. Strategia optimizării

În general punerea la punct a unei metode constă în modificarea unui singur factor, restul factorilor fiind menținuți constanți. În figura 7.3, a și b este prezentată grafic această strategie aplicată în cazul spectrometriei de absorbție atomică fără flacără în tub de grafit. Obiectivul acestui experiment [2] este de a găsi temperaturile de descompunere și respectiv de atomizare care dau o absorbantă maximă. Dacă se iau în considerare numai cei doi factori : temperatura de descompunere (TD) și temperatura de atomizare (TA) metodologia clasică constă de exemplu în aceea de a menține mai întâi factorul TA invariabil la o valoare convenabil aleasă și a modifica factorul TD (figura 7.3,a). Prin evaluarea răspunsurilor corespunzătoare se poate alege valoarea optimă pentru factorul TD , căreia îi corespunde o valoare maximă pentru răspuns . În continuare pentru a stabili valoarea optimă pentru factorul TA, se menține factorul TD invariabil la optimul găsit și se modifică factorul TA (figura 7.3,b).



**Fig. 7.3, a, b. Studiul izolat al factorilor**

Dacă se reprezintă cele două serii de măsurători simultan (figura 7.4), simulând absorbante teoretice, se poate vedea clar că optimul găsit aplicând metodologia variației unui singur factor, întâi TD apoi TA, nu corespunde optimului adevărat. Această eroare apare datorită faptului că metodologia clasică de a modifica un singur factor nu ține seama de interacțiunea care există între factorii TA și TD.

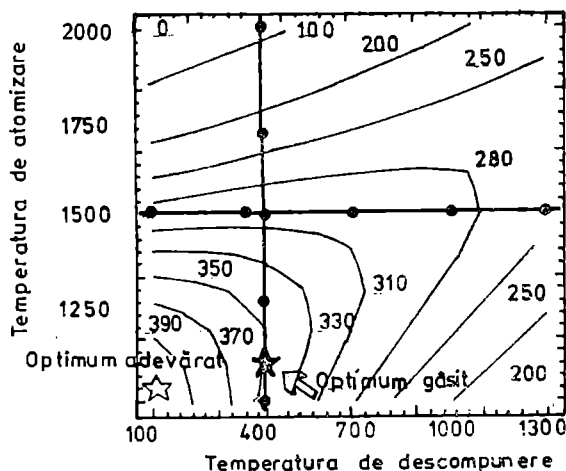
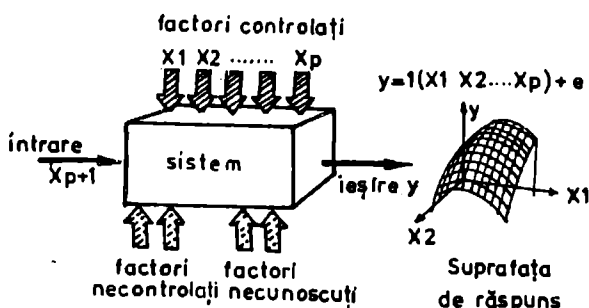


Fig. 7.4. Représentarea simultană a două serii de măsurători.

Evident au fost propuse și alte strategii care țin seama de interacțiile între factori. Una dintre ele ar fi cea deterministă în care valorile optime pentru variabila de optimizat se găsesc cu ajutorul metodelor de programare matematică. Aceste ecuații stabilite în baza legilor fizicii și chimiei sunt rezolvate algebric în așa fel încât sunt puse în evidență soluțiile cele mai favorabile. Altă strategie ar fi cea empirică și experimentală care constă în modelarea metodei studiate ca o cutie neagră (black box) și evaluarea comportamentului său cu ajutorul încercărilor planificate (figura 7.5). Este cazul metodelor planurilor experimentale [3] și al metodei simplex [4]. Pentru cea mai mare parte a chimiștilor soluția deterministă apare ca fiind cea mai comodă și ea este adesea utilizată în procesul de optimizare. Înșă, în cele mai multe cazuri ea este inaplicabilă deoarece nu întotdeauna fenomenul studiat poate fi evaluat teoretic riguros. De aceea este necesară de cele mai multe ori în special în cazul chimiei analitice o planificare a încercărilor pentru a evita erorile de interpretare.

Metoda optimizării este văzută de un experimentator ca un sistem în care se va selecționa o ieșire notată cu  $y$ , numită criteriu de optimizare.



**Fig. 7.5. Sistemul cutiei negre**

Se presupune că valorile  $y$ , variază datorită unui număr de factori controlați de experimentator care iau anumite valori fixe. Aceste valori fixe se numesc nivelele factorilor. Reprezentarea variațiilor lui  $y$  în funcție de nivelele factorilor formează suprafața de răspuns specifică sistemului studiat. Prin convenție se notează de obicei cu  $p$  factorii  $x_1 \dots x_p$  și nivelele lor pentru  $n$  încercări  $x_{11}, x_{21}, \dots, x_{n-1,p}, x_{np}$ . De exemplu, în cazul unei metode de dozare colorimetrică, ieșirea  $y$  va fi intensitatea reacției, măsura înălțimii picurilor, iar factorii controlați vor fi temperatura  $X_1$ , durata reacției  $X_2$  și pH-ul,  $X_3$ . De asemenea, întotdeauna trebuie avut în considerație faptul că există factori cunoscuți necontrolați ca de exemplu: fluctuațiile lămpii colorimetrului sau evaporarea soluției în cava de măsură precum și factori necunoscuți și necontrolați. Suma acestor factori necontrolați formează zgomotul de fond sau eroarea experimentală a metodei și se notează cu  $e$ .

Plecând de la aceste două concepte al factorilor controlați și al zgomotului de fond aleatoriu se poate scrie un model matematic astfel:

$$y = f(x_1, x_2, \dots, x_p) + e \quad (7.8)$$

### **Bibliografie**

1. C. Liteanu și I. Rică, **Optimizarea proceselor analitice**, Ed. Academiei Române 1985.
2. M. Feinberg, **L'actualite chimique**, 3, 201 (1991).
3. J. S. Hunter, **Experimental Designs: Fractionals**, in **Chemometrics** (Kowalski B., Ed.), Reidel Pub. Co., Dordrecht, 1984.
4. S. N. Deming, **Experimental Designs: Response Surfaces**, in **Chemometrics** (Kowalski B., Ed.), Reidel Pub. Co., Dordrecht, 1984.

## VIII. METODE DE EVALUARE A OPTIMULUI UNEI FUNCTII OBIECTIV

Există mai multe tipuri de metode care pot fi aplicate la stabilirea optimului unei funcții obiectiv :

**VIII.1. Metode analitice** sunt metodele care dau informații de ansamblu asupra funcției și ele presupun continuitatea funcției și a derivatelor ei. Aceste metode folosesc calculul diferențial pentru maximum sau minimum funcției. Totuși multe dintre sistemele de interes practic nu pot fi optimizate prin aceste metode.

**VIII.2. Metode numerice** sunt metodele care utilizează un algoritm pentru stabilirea optimului funcției obiectiv. În acest caz funcția obiectiv este tratată ca un sistem black-box [1]. În cazul aplicării metodelor numerice la optimizarea unui proces analitic, se disting următoarele două situații :

- funcția de optimizat este cunoscută în formă matematică exactă, caz în care pentru evaluarea optimului , sistemul se tratează ca un sistem black-box;
- funcția de optimizat nu este cunoscută , deci în acest caz se acționează asupra sistemului de optimizat aflat în funcțiune, acesta fiind tratat ca un sistem black-box.

**VIII.3. Metode mixte** sunt metodele în cazul cărora funcția obiectiv se tratează ca un sistem black-box. În baza unui plan experimental se face evaluarea numerică a funcției obiectiv pentru mai multe valori impuse variabilelor de optimizat. Cu ajutorul acestor date se poate rezolva analitic funcția obiectiv în scopul stabilirii optimului.

Metodele prezentate mai sus pot fi aplicate unui sistem riguros cunoscut. Dar cazurile în care se cunoaște sau se poate deduce teoretic funcția de răspuns cât și modelul matematic pentru aceasta sunt limitate. În majoritatea cazurilor în chimia analitică atât valoarea funcției de răspuns cât și modelul matematic pentru aceasta nu sunt cunoscute. În acest caz pentru rezolvarea problemelor de optimizare se aplică metodele experimental-statistice [2] conform cărora sistemul în funcțiune este studiat în acord cu principiul black-box.

Sistemul de optimizat există în funcțiune astfel încât se modifică controlat variabilele de optimizat evaluându-se răspunsul sistemului pentru fiecare stare nouă impusă variabilelor. Metodele de optimizare experimental-statistice sunt descrise în numeroase lucrări de specialitate care tratează probleme teoretice și practice în această direcție[3-6].

Există două posibilități de abordare a metodelor de optimizare experimental-statistice și anume ;

- una dintre posibilități consideră volumul de valori experimentale succesive (acestea cuprind valori ale variabilelor de optimizat cât și valori ale răspunsului sistemului, coresponzătoare) și dirijează cercetarea în direcția optimului;

- a doua posibilitate constă în desfășurarea optimizării după un plan experimental stabilit prealabil. Evident planul experimentului se poate modifica pe parcursul desfășurării experimentului.

Planul experimental este conceput astfel încât cu ajutorul mulțimii de date experimentale să se poată evalua răspunsul sistemului cât mai adecvat. Valorile optime pentru variabilele sistemului sunt prelucrate prin cercetare numerică sau analitică. Pentru optimizarea răspunsului în chimia analitică mai este utilizată și metoda simplex, metodă elaborată inițial de Spendley și colab.[7] și apoi modificată de Nedler și Mead [8].

Metoda simplex constă într-un algoritm iterativ de localizare a optimului funcției obiectiv pe suprafața de răspuns.

Cele două strategii de optimizare și anume cea a planului experimental și simplex au cunoscut multe aplicații în chimia analitică. Utilizarea unui plan experimental pentru explorarea unei suprafețe de răspuns și găsirea optimului se face în 6 etape a căror secvență trebuie respectată :

1. Alegerea unui criteriu de optimizare;
2. Selecția factorilor controlați și delimitarea exactă a domeniului experimental;
3. Alegerea unui model de preferință liniar;
4. Alegerea unui plan experimental;
5. Obținerea determinărilor, calculul coeficienților și localizarea optimului practic;
6. Efectuarea unei determinări complementare a optimului pentru validarea modelului.

### ***Bibliografie***

1. C. Liteanu și I. Rîcă, **Optimizarea proceselor analitice**, Ed.Academiei Române, 1985.
2. W.G. Cochran și G.M. Cox, **Experimental Design**, John Wiley, London, 1957.
3. J. Cea, **Optimization theory et algorythmes**, Dunod, Paris,1971.
4. E. Polak, **Computational Methods in Optimization**, Academic Press, London, 1971.
5. R.L. Fox, **Optimization Methods for Engineering Design**, Adison-Wesley Publishing Reading Mas.,1971.
6. W.J. Youden, **Statistical Methods for Chemists**, J.Wiley, N.York, 1962.
7. W. Spendley, G.R. Hext și F.R. Himsforth, **Technometrics**, **4**, 441,1962.
8. J.A. Nedler și R. Mead, **Comput.J.**, **7**, 308, 1965.

# IX. ELEMENTE DE TEORIA INFORMAȚIEI UTILIZATE ÎN CHIMIA ANALITICĂ

## IX.1. Generalități

Aplicarea teoriei informației în chimia analitică este datorată faptului că aceasta tratează metode de obținere a informației chimico-analitice prin intermediul unui experiment. Fresenius a dat următoarea definiție chimiei analitice: "Știința achiziționării și interpretării informațiilor despre sisteme de materiale referitoare la exploatarea și îmbunătățirea metodelor științelor naturale".

Ca în oricare alt domeniu al cunoașterii, și în chimia analitică este necesar să se măsoare și să se cunoască cantitatea de informație obținută în urma unui experiment analitic, lucru care se poate realiza prin introducerea conceptelor teoriei informației. Shannon [1] a pus bazele teoriei informației în lucrarea "A mathematical theory of communication" apărută în 1948. După 1960, centrul de greutate al lucrărilor de teoria informației s-a deplasat spre partea aplicativă.

Ca urmare a analogiilor existente între sistemele de transmisie a informației și sistemele analitice de măsură, noțiunile din teoria informației (codificare, perturbație, decodificare, entropie, cantitatea de informație, redundanță etc.) pot fi transpuse analizei chimice.

## IX.2. Procesele analitice ca sisteme de transmisie a informației

Un sistem de transmisie a informației se reprezintă ca o reuniune dintre sursa de informație, un canal ce include codificatorul și decodificatorul și receptorul. Prin analogie cu un sistem de transmisie a informației, un sistem analitic de măsură poate fi reprezentat ca un sistem informațional. Astfel, o probă de analizat trebuie privită ca un element al unei mulțimi bine definite, compoziția probei la un moment dat luând o anumită valoare dintr-o mulțime de valori posibile.

Procesul informativ se poate reprezenta ca un sistem cu următoarele elemente [2]:

INPUT → BLACK BOX → OUTPUT

Elementul INPUT (intrare), în chimia analitică este dat de proba analitică. Spre deosebire de alte probe, probei analitice i se atașează o problemă analitică. Problema analitică stabilește înaintea experimentului analitic, scopul urmărit prin efectuarea experimentului, adică natura și numărul componentelor care pot face parte din proba analitică [11].



BLACK BOX reprezintă experimentul utilizat (chimic, fizic, etc.) în efectuarea operației analitice. Tratamentelor fizico-chimic are drept scop realizarea condițiilor pentru realizarea unui semnal specific componentului de determinat, această operație analitică făcând posibilă detecția sau determinarea. Tratamentul fizico-chimic urmărește, în esență, același lucru ca și operația de codificare în cazul transmisiei informației, facilitarea transmisiei și înlăturarea efectului perturbațiilor. OUTPUT (ieșire) este reprezentat prin rezultatele finale.

Prin operația de calcul și prelucrare a rezultatelor se trece de la semnalul analitic la rezultatul de analiză, această operație fiind similară procesului de decodificare a semnalelor, întâlnit în cazul sistemelor de transmisie a informației. Atât transmisiile cât și măsurătorile analitice se efectuează în condițiile acțiunii perturbațiilor, atât în semnalele recepționate cât și rezultatele de analiză vor include întotdeauna o anumită nedeterminare.

Numai printr-o delimitare precisă, în urma unor cercetări analitice adecvate a mulțimii de materiale ce pot fi analizate în anumite condiții concrete, rezultatele de analiză nu vor îngloba erori cauzate de fenomene de interferență și interinfluență (efecte de matrice).

Pentru a se cunoaște valoare perturbațiilor din procesul de analiză trebuie să se cunoască probabilitățile condiționate ale rezultatelor de analiză față de conținuturile adevărate.

Notăm cu  $c \in C$  mulțimea probelor de material care pot fi analizate, cu  $c' \in C'$  mulțimea rezultatelor de analiză și cu  $p(c'/c)$  mulțimea probabilităților condiționate pentru fiecare element  $c' \in C'$  corespunzător fiecărui element  $c \in C$ . Rezultă că unei valori  $c$  la intrarea sistemului de măsură, ca urmare a acțiunii perturbațiilor, îi corespunde o mulțime de rezultate, marcată printr-o densitate de probabilitate.

Tripletul  $c \in C$ ,  $c' \in C'$  și  $p(c'/c)$  definește complet un sistem analitic și se poate nota:

$$\{c \in C, p(c'/c), c' \in C'\}$$

Această definiție probabilistică a unui sistem analitic de măsură corespunde definiției probabilistice a unui sistem de transmisie a informației (canal). Referindu-ne la un sistem de transmitere a informației,  $c \in C$  reprezintă mulțimea semnalelor de intrare,  $c' \in C'$  - mulțimea semnalelor recepționate iar  $p(c'/c)$  - probabilitățile condiționate care caracterizează perturbațiile pe canalul respectiv [3].

### IX.3. Entropie și informație

Pentru aplicarea conceptelor teoriei informației în chimia analitică este necesară înscrierea noțiunii de entropie considerată a fi o măsură a nedeterminării [3].

Dacă efectuăm un experiment, acesta ne va conduce la un rezultat alcătuit din mai multe rezultate (evenimente) posibile a căror repartiție discretă o notăm:

$$X = \begin{bmatrix} x_1, x_2, \dots, x_n \\ p_1, p_2, \dots, p_n \end{bmatrix} \quad (9.1)$$

Considerăm că sistemul de evenimente  $x_1, x_2, \dots, x_n$  este un sistem complet de evenimente, adică  $\sum_{i=1}^n p_i = 1$ , ceea ce înseamnă că prin efectuarea experimentului vom obține ca rezultat unul din evenimente. Deoarece ca rezultat al experimentului poate figura unul din evenimentele  $x_1, x_2, \dots, x_n$ , înseamnă că realizarea experimentului are drept scop înlăturarea unui anumit grad de nedeterminare.

Problema care se ridică este evident măsura gradului de nedeterminare (incertitudinea). Nedeterminarea depinde însă în mod esențial de probabilitățile evenimentelor.

Repartiția  $X = \begin{bmatrix} x_1, & x_2 \\ 0,5 & 0,5 \end{bmatrix}$  este formată din două evenimente  $x_1$  și  $x_2$  echiprobabile și conține o mai mare nedeterminare decât repartiția  $X = \begin{bmatrix} x_1, & x_2 \\ 0,95 & 0,05 \end{bmatrix}$  formată din aceleași evenimente, însă cu probabilități mult diferite.

Shannon [1] introduce noțiunea de entropie drept măsură a nedeterminării unui câmp de probabilitate discret:

$$H(p_1, p_2, \dots, p_n) = - \sum_{i=1}^n p_i \cdot \log p_i \quad (9.2)$$

În cazul în care se operează cu logaritmi în baza doi, nedeterminarea va fi măsurată în unități numite "bit".

Expresia entropiei în cazul unei repartiții continue de densitate  $p(x)$  este:

$$H[p(x)] = - \int_{-\infty}^{+\infty} p(x) \log p(x) dx \quad (9.3)$$

În cazul în care densitatea de probabilitate a unui eveniment este uniformă în intervalul  $(a, b)$  atunci expresia entropiei devine:

$$H[p(x)] = - \int_a^b \frac{1}{b-a} \log \frac{1}{b-a} dx = \log(b-a) \quad (9.4)$$

Funcția găsită a fi expresia entropiei trebuie să îndeplinească o serie de proprietăți, considerate a fi proprietățile entropiei [3]:

a) Entropia este o mărime nenegativă

$$H(p_1, p_2, \dots, p_n) \geq 0$$

(9.5)

2) Dacă este posibilă o singură stare (un singur eveniment), adică există o valoare  $p_i=1$ , atunci entropia este egală cu zero:

$$H(p_1, p_2, \dots, p_n) = 0 \quad (9.6)$$

3) Entropia  $H(p_1, p_2, \dots, p_n)$  are o valoare maximă pentru  $p_1=p_2=\dots=p_n$  adică atunci când evenimentele sunt echiprobabile:

$$H(p_1, p_2, \dots, p_n) \leq H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) \quad (9.7)$$

4) Fiind date două sisteme de evenimente oarecare  $X_1$  și  $X_2$ , entropia reuniunii celor două sisteme se scrie:

$$H(X_1, X_2) = H(X_1) + H(X_2/X_1) \quad (9.8)$$

unde prin  $H(X_2/X_1)$  s-a notat entropia sistemului  $X_2$  condiționată de sistemul  $X_1$ . Când evenimentele celor două sisteme sunt independente avem:

$$H(X_1, X_2) = H(X_1) + H(X_2) \quad (9.9)$$

Evident,  $H(X_2/X_1) \leq H(X_2)$ .

O altă noțiune, foarte generală, care trebuie introdusă este informația. Se poate spune însă că în urma unui experiment se obține o informație atunci când rezultatul înlătură o parte din nedeterminarea existentă la început. Cu cât nedeterminarea de la începutul experimentului este mai mare, cu atât mai mare va fi cantitatea de informație atunci când se obține rezultatul experimentului respectiv. Informația este egală cu nedeterminarea înlăturată și deci cantitatea de informație are aceeași măsură ca și gradul de nedeterminare înlăturat și evident că, drept urmare a acțiunii perturbațiilor, experimentul, în general, nu reușește să înlătore întreaga nedeterminare [2].

Notând, în general, prin  $H_0$  măsura nedeterminării apriori experimentului, iar prin  $H'$  măsură nedeterminării aposteriori experimentului, condiția obținerii unei informații în urma experimentului este:  $H' < H_0$ . Cu cât  $H' < H_0$ , cu atât cantitatea de informație furnizată de experiment este mai mare.

Cantitatea de informație furnizată de un experiment se exprimă prin:

$$I = H_0 - H' \quad (9.10)$$

Cantitatea de informație maximă,  $I_{\max}$ , care se poate obține în urma unui experiment este egală cu nedeterminarea apriori experimentului, deci cu entropia apriori experimentului:

$$I_{\max} = H_0 \quad (9.11)$$

Informații se pot obține în urma unor experimente de detecție, de identificare și de determinare [3].

Analiza de detecție reprezintă cercetarea analitică ce urmărește stabilirea prezenței sau absenței unuia sau mai multor componente într-un material de cercetat. În termeni analitici, aceasta reprezintă o analiză calitativă a materialului.

## IX.4. Aplicații ale teoriei informației în chimia analitică

### IX.4.1. Cuantificarea informației obținute în experimentele de identificare

Identificarea reprezintă cercetarea analitică în urma căreia se stabilește conformația structurală a moleculelor unui material de cercetat, adică se efectuează o analiză de structură a moleculei unei substanțe.

Considerăm cazul cel mai simplu al unui experiment de detecție analitică, adică stabilirea prezenței sau absenței unui component  $X_1$  într-un material de cercetat. Prin urmare materialul se poate găsi în două stări, cu și, respectiv, fără componentul  $X_1$ .

Notăm cu "0" starea în care materialul nu conține componentul de detectat și cu "1" starea cu care materialul conține componentul de detectat. În absența unei informații apriori experimentului, celor două stări le corespunde următoarea repartiție:

$$X_1 = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad (9.12)$$

Acestei informații discrete îi corespunde o nedeterminare  $H_0 = 1$  bit. Deci în urma unui experiment analitic de detecție a unui component într-un material de cercetat se poate obține o cantitate maximă de informație  $I_{\max} = 1$  bit.

Considerând un experiment de detecție prin care se urmărește stabilirea prezenței sau absenței a 3 componente  $X_1, X_2, X_3$  într-un material de cercetat se va nota cu "0" absența unuia din cei trei componente și prin "1" prezența sa. Stările posibile ale materialului de cercetat vor fi: 000, 100, 010, 001, 110, 101, 011, 111. În absența unei informații apriori experimentului de detecție, acestor stări echiprobabile le corespunde următoarea repartiție:

$$X_1, X_2, X_3 = \begin{bmatrix} 000 & 100 & 010 & 001 & 110 & 101 & 011 & 111 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{bmatrix} \quad (9.13)$$

În acest caz, nedeterminarea care urmează să fie înlăturată în urma detecției a 3 componente într-un material de cercetat este:

$$H_0 = \log_2 2^3 = 3 \text{ biți} = I_{\max} \quad (9.14)$$

În general, într-o cercetare analitică de detecție a "n" componente dintr-un material urmează să se înlătore o nedeterminare

$$H_0 = \log_2 2^n = n \text{ biți} = I_{\max} \quad (9.15)$$

În cazul unui experiment analitic de identificare, care urmărește stabilirea conformației structurale a moleculelor unui material, problema se pune în mod similar. În urma unui asemenea experiment, urmează să se

pune în mod similar. În urma unui asemenea experiment, urmează să se stabilească care din mai multe conformații structurale, ipotetic posibile, corespunde realității.

Notând conformațiile structurale posibile pentru moleculele materialului de cercetat prin  $s_i$ , experimentul urmează să stabilească care din ele ( $s_1, s_2, \dots, s_n$ ) corespunde realității, adică să înlăture nedeterminarea corespunzătoare repartiției:

$$S = \begin{bmatrix} s_1 & s_2 & \dots & s_n \\ p_1 & p_2 & \dots & p_n \end{bmatrix} \quad (9.16)$$

În absența unei informații apriori experimentului  $p_1=p_2=\dots=p_n=1/n$  și deci cantitatea maximă de informație care rezultă în urma cercetării va fi:  $I_{\max} = n$  biți.

Compoziția elementară a oricărui material cuprinde un număr de elemente chimice mai mic sau egal cu 104 (numărul de elemente chimice cunoscute). Deci, valoarea nedeterminării apriori oricărei cercetări analitice de detecție elementară trebuie să fie mai mică sau egală cu 104 biți ( $H_0 \leq 104$  biți) [3].

Cantitatea de informație care se obține în urma determinării conținutului unuia sau mai multor componente dintr-un material de cercetat se evaluează luând în considerare atât nedeterminarea apriori experimentului cât și nedeterminarea a posteriori experimentului.

#### IX.4.2. Cuantificarea informației obținute în experimentele de determinare

În absența oricărei informații prealabile experimentului, domeniul de incertitudine corespunzător concentrației procentuale a componentului de determinat este 0-100 %.

În condițiile unei informații prealabile domeniul de incertitudine real al concentrației procentuale a componentului de determinat este:

$$(c' \%, c'' \%) < (0, 100 \%)$$

Dacă se notează cu  $\Delta c$  lărgimea intervalului de concentrație în limitele căruia variația conținutului componentului de determinat nu este luată în considerare - pasul de cuantificare - intervalului de concentrație ( $c', c''$ ) îi corespund  $N=(c''-c')/\Delta c$  valori discrete de concentrație. În urma unui experiment de determinare, urmează să se stabilească care din aceste  $n$  valori de concentrație posibile corespund realității, adică se urmărește înlăturarea nedeterminării [4]:

$$H_0(c) = \log \frac{c''-c'}{\Delta c} \quad (9.17)$$

Datorită acțiunii perturbațiilor, sistemul nu înlătură întreaga nedeterminare. De aceea, pentru a evalua cantitatea de informație se ridică și problema evaluării nedeterminării aposteriori experimentului.

În condițiile absenței erorilor sistematice și pentru o distribuție normală a erorilor întâmplătoare, cu o abatere standard  $\sigma$ , nedeterminarea care rămâne după efectuarea unui experiment de determinare va fi:

$$H'(c) = \log \frac{\sigma \sqrt{2\pi e}}{\Delta c} \quad (9.18)$$

Cantitatea de informație care se obține în urma unui experiment analitic de determinare a unui component dintr-un material, atunci când erorile sistematice lipsesc iar cele întâmplătoare au o distribuție normală cu o abatere standard  $\sigma$ , este de forma:

$$I = H_0(c) - H'(c) = \log \frac{c'' - c'}{\sigma \sqrt{2\pi e}} \quad (9.19)$$

Prin luarea în considerație a unui interval de încredere pentru abaterea standard se poate atașa atât entropiei erorilor cât și cantității de informație un interval de încredere de forma:

$$\log \frac{c'' - c'}{\sqrt{2\pi e} \sqrt{\frac{s^2 v}{\chi^2(1-P)/2}}} < I < \log \frac{c'' - c'}{\sqrt{2\pi e} \sqrt{\frac{s^2 v}{\chi^2(1+P)/2}}} \quad (9.20)$$

Relația:

$$I = \log \frac{c'' - c'}{\sigma \sqrt{2\pi e}} \quad (9.21)$$

exprimă esența informațională a unei determinări analitice sau presupune îngustarea intervalului de incertitudine, numărătorul reprezentând mărimea intervalului de incertitudine apriori experimentului, iar numitorul - valoarea ponderală a intervalului de incertitudine aposteriori experimentului.

Atunci când se efectuează  $n$  determinări asupra materialului și rezultatul se exprimă ca medie a acestora, abaterea standard va fi  $\sigma/n$  și expresia pentru cantitatea de informație va deveni:

$$I = \log \frac{c'' - c'}{\sqrt{2\pi e} \cdot \frac{\sigma}{\sqrt{n}}} \quad (9.22)$$

Din ecuația de mai sus, rezultă că modul practic de a spori cantitatea de informație pe care o comportă rezultatul unei determinări analitice îl constituie micșorarea abaterii standard,  $\sigma$ , pentru rezultatele individuale atât prin reducerea perturbațiilor, cât și prin creșterea numărului de determinări care intră în alcătuirea rezultatului medie.

#### IX.4.3. Selectarea metodelor analitice pe baza informației furnizate

Cantitatea de informație poate fi utilizată drept criteriu pentru alegerea metodei de analiză. În acest scop se va fi:

metodă sau variantă a unei metode care ne furnizează cea mai bună variantă de informație referitor la evenimentul analitic de cercetat. Exemplul următor are drept scop să ilustreze modul de abordare a unei astfel de probleme [5].

În tabelul 9.1 sunt trecute valorile  $R_f$  pentru opt substanțe ipotetice (A, B, C, D, E, F, G, H) și trei eluenți (I, II, III). Acceptând că erorile de evaluare a valorilor  $R_f$  sunt neglijabile și folosind drept criteriu cantitatea de informație, să se stabilească care este cel mai eficient eluent, adică cea mai bună variantă pentru efectuarea identificării celor opt substanțe [6].

Tabelul 9.1  
Valorile  $R_f$  pentru opt substanțe ipotetice și trei eluenți

Substanța	I	II	III
A( $x_1$ )	0,20	0,20	0,20
B( $x_2$ )	0,20	0,40	0,20
C( $x_3$ )	0,40	0,20	0,20
D( $x_4$ )	0,40	0,40	0,20
E( $x_5$ )	0,60	0,20	0,40
F( $x_6$ )	0,60	0,40	0,40
G( $x_7$ )	0,80	0,20	0,40
H( $x_8$ )	0,80	0,40	0,40

Tabelul 9.2  
Probabilități condiționate pentru trei eluenți și opt substanțe

X\Y	Y <sub>1</sub>			Y <sub>2</sub>			Y <sub>3</sub>			Y <sub>4</sub>		
	I	II	III	I	II	III	I	II	III	I	II	III
$x_1$	1	1	1	0	0	0	0	0	0	0	0	0
$x_2$	1	0	1	0	1	0	0	0	0	0	0	0
$x_3$	0	1	1	1	0	0	0	0	0	0	0	0
$x_4$	0	0	1	1	1	0	0	0	0	0	0	0
$x_5$	0	1	0	0	0	1	1	0	0	0	0	0
$x_6$	0	0	0	0	1	1	1	0	0	0	0	0
$x_7$	0	1	0	0	0	1	0	0	0	1	0	0
$x_8$	0	0	0	0	1	1	0	0	0	1	0	0

Pentru acest sistem complet de opt evenimente egal probabile, are loc repartiția:

$$P(X) = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{bmatrix} \quad (9.23)$$

Introducând notațiile:  $y_1=0,20$ ;  $y_2=0,40$ ;  $y_3=0,60$  și  $y_4=0,80$  rezultă că a posteriori experimentului are loc mulțimea de evenimente  $Y=(y_1, y_2, y_3, y_4)$ .

Cu ajutorul datelor din tabelul 1 se obțin, pentru cei trei eluenți, valorile probabilităților condiționate,  $p(y_i/x_j)$ , înscrise în tabelul 9.2. Calculând probabilitățile  $p(y_i)$  pentru fiecare eluent în parte, se obțin următoarele repartiții:

$$P(Y_I) = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \quad (9.24a)$$

$$P(Y_{II}) = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix} \quad (9.24b)$$

$$P(Y_{III}) = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix} \quad (9.24c)$$

Probabilitățile condiționate,  $p(x_i/y_j)$ , evaluate în acord cu relația lui Bayes [5]:

$$p(x_i / y_j) = \frac{p(x_i) \cdot p(y_j / x_i)}{\sum_{i=1}^n [p(x_i) \cdot p(y_j / x_i)]} \quad (9.25)$$

sunt înscrise în tabelul 9.3:

Tabelul 9.3  
Valorile probabilităților condiționate

Eluent	Y \ X	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
I	$y_1$	1/2	1/2	0	0	0	0	0	0
	$y_2$	0	0	1/2	1/2	0	0	0	0
	$y_3$	0	0	0	0	1/2	1/2	0	0
	$y_4$	0	0	0	0	0	0	1/2	1/2
II	$y_1$	1/4	0	1/4	0	1/4	0	1/4	0
	$y_2$	0	1/4	0	1/4	0	1/4	0	1/4
	$y_3$	0	0	0	0	0	0	0	0
	$y_4$	0	0	0	0	0	0	0	0
III	$y_1$	1/4	1/4	1/4	1/4	0	0	0	0
	$y_2$	0	0	0	0	1/4	1/4	1/4	1/4
	$y_3$	0	0	0	0	0	0	0	0
	$y_4$	0	0	0	0	0	0	0	0

Mediile cantităților de informație furnizate de fiecare eveniment  $y_j$  asupra evenimentelor mulțimii X, evaluate în acord cu relația:



$$I(X, y_i) = \sum_{i=1}^n [p(x_i / y_i) \cdot I(x_i, y_i)] = \sum_{i=1}^n \left[ p(x_i / y_i) \log \frac{p(x_i, y_i)}{p(x_i) \cdot p(y_i)} \right] \quad (9.26)$$

pentru fiecare eluent în parte sunt:

$$I(X, y_1)_I = I(X, y_2)_I = I(X, y_3)_I = I(X, y_4)_I = 2 \text{ biti}$$

$$I(X, y_1)_{II} = I(X, y_2)_{II} = I(X, y_3)_{II} = I(X, y_4)_{II} = 0 \quad (9.27)$$

$$I(X, y_1)_{III} = I(X, y_2)_{III} = I(X, y_3)_{III} = I(X, y_4)_{III} = 0$$

Mediile cantităților de informație furnizate de evenimentele mulțimii Y cu privire la evenimentele mulțimii X, evaluate pentru fiecare eluent în parte în acord cu relația:

$$I(X, Y) = - \sum_{j=1}^m p(x_i y_j) \sum_{i=1}^n \log p(x_i) - \sum_{i=1}^n p(x_i y_j) \sum_{j=1}^m \log p(y_j) +$$

$$+ \sum_{i=1}^n \sum_{j=1}^m [p(x_i y_j) \log p(x_i y_j)] \quad (9.28)$$

sunt:  $I(X, Y)_I = 2 \text{ biti}$ ;  $I(X, Y)_{II} = 1 \text{ bit}$  și  $I(X, Y)_{III} = 1 \text{ bit}$ .

Așadar, prin utilizarea oricăruia din cei trei eluenți, nedeterminarea apriori experimentului ( $H(X) = 3 \text{ biti}$ ) nu poate fi înlăturată. Cel mai eficient este eluentul I prin utilizarea căruia rezultatele de măsură furnizează o cantitate de informație egală cu 2 biți.

## IX.5. Alte mărimi informaționale utilizate în caracterizarea proceselor analitice

### IX.5.1. Redundanța materialului supus cercetării analitice

Redundanța materialului supus cercetării analitice [3] măsoară cât de mult se abate entropia (nedeterminarea materialului de cercetat) de la valoarea maximă posibilă:

$$R = H_{\max}(c) - H_0(c) \quad (9.29)$$

Redundanța relativă este:

$$\rho = 1 - \frac{H_0(c)}{H_{\max}(c)} \quad (9.30)$$

Redundanța este o măsură a gradului de informare prealabil experimentului analitic efectuat asupra materialului de cercetat.

În cazul unei cercetări analitice de detecție care urmărește stabilirea compoziției elementare a unui material, admitând că nu se cunoaște nimic despre material și că prezența oricărui element dintre cele 104 cunoscute este echiprobabilă, entropia maximă este  $H_{\max} = 104 \text{ biti}$ . Dacă însă informațiile referitoare la material existente apriori experimentului limitează incertitudinea asupra compoziției elementare a acestuia la  $n$  elemente posibile ( $n < 104$ ),

entropia reală a materialului de cercetat va fi  $H_0 = n \text{ biti} < 104 \text{ biti}$  iar redundanța acestuia este  $R = (104 - n) \text{ biți}$ .

### IX.5.2. Puterea informațională a unui procedeu analitic

Această mărime, referitoare la un procedeu analitic, a fost introdusă de către Kaiser [7]. La măsurarea unui semnal unidimensional (masă de precitat, intensitatea unei linii spectrale etc), puterea informațională se scrie:

$$P_{\text{inform}} = \log_2 S \quad (9.31)$$

unde  $S$  reprezintă numărul de valori care pot fi distinse pe scala instrumentului de măsură.

Pentru un procedeu analitic cu  $N$  canale de măsură, puterea informațională devine:

$$P_{\text{inform}} = \sum_{i=1}^N \log_2 S_i \quad (9.32)$$

În cazul în care fiecare canal în parte distinge același număr  $S$  de valori, relația precedentă se reduce la:

$$P_{\text{inform}} = N \log_2 S \quad (9.33)$$

Similar se poate deduce o relație de calcul a puterii informaționale și în cazul unui rezultat de măsură bidimensional (spectru, cromatogramă, polarogramă etc). Referitor la procedeele spectrale, s-a propus [7] următoarea formulă pentru puterea informațională:

$$P_{\text{inform}} = \int_{v_a}^{v_b} R(v) \log_2 S(v) \frac{dv}{v} \approx \bar{R} \log_2 \bar{S} \ln \frac{v_b}{v_a} \quad (9.34)$$

unde  $(v_a, v_b)$  reprezintă intervalul de frecvențe în limitele căruia se înregistrează spectrul,  $S(v)$  - numărul de valori distincte pentru amplitudinea semnalului iar  $R$  - rezoluția. Aproximarea puterii informaționale în relația anterioară s-a făcut introducând valorile medii pentru  $R(v)$  și  $S(v)$ .

În cazul semnalelor bidimensionale, o mărime analogă puterii informaționale este cantitatea de informație potențială [8, 9] a procedurii analitice. Considerând că un astfel de semnal este definit pe domeniul  $(x_{\min}, x_{\max})$  - corespunzător abscisei, și în domeniul de amplitudine  $(y_{\min}, y_{\max})$  - corespunzător ordonatei, cantitatea de informație potențială se definește prin mărimea:

$$I_p = \frac{x_{\min} - x_{\max}}{\Delta x} \log_2 \frac{(y_{\min} - y_{\max}) \sqrt{n}}{2 \bar{S} t(P, f)} \quad (9.35)$$

unde  $\Delta x$  este distanța minimă dintre două semnale care se pot distinge individual,  $2 \bar{S} t(P, f)$  - lățimea intervalului pentru fluctuațiile măsurătorilor individuale de amplitudine, iar  $n$  - numărul măsurătorilor care intră în calculul cantității de informație potențială.

Particularizată în cazul analizei spectrale [10], ultima expresie devine:

$$I_p = \frac{\lambda_{\min} - \lambda_{\max}}{\Delta\lambda} \log_2 \frac{\sqrt{n}}{2 S_f t(P, f)} \ln \frac{y_{\max}}{y_{\min}} \quad (9.36)$$

unde prin  $S_f$  s-a notat abaterea relativă standard. Raportul  $(\lambda_{\min} - \lambda_{\max}) / \Delta\lambda$  are semnificația numărului de valori distincte de pe abscisă iar

$$\log_2 \frac{\sqrt{n}}{2 S_f t(P, f)} \ln \frac{y_{\max}}{y_{\min}} - \text{de pe ordonată.}$$

### IX.5.3. Debitul și costul informației

Prin debit de informație [5] se înțelege cantitatea de informație obținută în unitatea de timp. Notând cu  $t$  timpul necesar efectuării unui proces analitic complet, în urma căruia se obține o cantitate de informație  $I$ , debitul sau fluxul de informație se definește prin:

$$J = \frac{I}{t} \quad (9.37)$$

Prin costul specific al unității de informație [5] se înțelege cheltuiala necesară producerii unei unități de informație, adică:

$$c = \frac{C}{I} \quad (9.38)$$

unde prin  $C$  s-au notat cheltuielile aferente producerii informației  $I$ .

### Bibliografie

1. C. E. Shannon, *Bell. Syst. Techn. J.*, **27**, 379; 623, 1948.
2. A. Spătaru, **Teoria transmisiunii informației**, Editura Tehnică, București, 1966.
3. C. Liteanu și I. Rică, **Teoria și metodologia statistică a analizei urmelor**, Editura Scrisul Românesc, Craiova, 1979.
4. L. Brillouin, **La science et la theorie de l'information**, Masson, Paris, 1959.
5. C. Liteanu și I. Rică, **Optimizarea proceselor analitice**, Editura Academiei Române, București, 1985.
6. D. L. Massart, *J. Chromatogr.*, **18**, 104, 1978.
7. H. Kaiser, *Pure Appl. Chem.*, **34**, 35, 1973.
8. K. Eckschlager, *Z. Chem.*, **16**, 111, 1976.
9. K. Danzer, *Z. Chem.*, **13**, 69, 1973.
10. K. Danzer, *Z. Chem.*, **15**, 158, 1975.
11. Victor David, *Rev. Chim. (Bucharest)*, **5**, 428, 1989.

## X. REȚELE NEURONALE

### X.1. Introducere

Importanța cercetărilor asupra sistemului nervos al diferitelor organisme aflate pe variate trepte de evoluție a materiei vii este atât teoretică, prin studii fundamentale asupra proceselor gândirii, cât și practică, prin transferul de tehnologie de la biosisteme spre tehnică și tehnologie [1]. Deși sistemul nervos central (SNC) procesează continuu informație relevantă situațiilor naturale mai rapid decât orice sistem cibernetic, experiențele de psihofiziologie umană arată că intervalul de timp scurs din momentul recepționării unui stimul și până în momentul declanșării răspunsului nu poate coborâ sub 100 ms, reacție care poate fi suplimentar întârziată de factori specifici ființelor vii. Sistemele vii sunt superioare sistemelor cibernetice, cu timpi specifici de răspuns sub 1 ms, în cel puțin următoarele privințe [2]:

- *capacitate de analiză și sinteză,*
- *flexibilitate,*
- *fiabilitate,*
- *redundanță,*
- *miniaturizare,*
- *grad înalt de conectivitate.*

Rezultă posibilitatea rafinării implementării proceselor de calcul prin revizuirea atât a arhitecturii sistemelor artificiale, cât și la nivel software.

Experiențe de neurofizică sugerează că elementele neuronele ale SNC se caracterizează printr-o viteză de calcul de ordinul milisecundelor, aceasta înseamnă un volum total de aproximativ 100 de calcule în intervalul de reacție. Cele mai recente și sofisticate programe de inteligență artificială (AI) cu destinații comparabile, nu sunt atât de generale și necesită milioane de calcule. Soluționarea rapidă a unor probleme computaționale complexe este fezabilă pe două căi. Prima constă în transmiterea cu o viteză mult mai mare a informației între elementele de procesare. Sistemele vii sunt însă extrem de lente, capacitatea de transmisie axonală fiind cuprinsă între 50 și 100 bit/s, iar sistemele tehnice, deși caracterizate de o capacitate de transmisie pe canal remarcabilă, au la dispoziție numai circuite integrate monolitice semiconductoare (IC) cu fan-out-ul (mai mic de 20) incomparabil cu cel al neuronilor din SNC (circa  $10^4$ ). A doua cale presupune ipoteza că structurile complexe de rezolvare a problemelor sunt codificate global, în configurația interconexiunilor unităților de procesare, astfel încât nu este necesară circulația unei cantități mari de informație. Acest mod de abordare a condus la conceptul de *conexionism* [3].

Calculatoarele secvențiale clasice (mașini *von Neumann*) nu sunt concepute pentru a opera asupra unui număr mare de variabile care se

influențează reciproc și, în plus, solicită date perfect definite. Rețelele neuronale formale, denumite aici alternativ modele cu *procesare distribuită paralel (modelele PDP)*, *modelele conexioniste* sau *modelele neuromorfice*, sunt concepute pentru rezolvarea unor asemenea probleme specifice *inteligenței artificiale*, dintre care cele mai semnificative sunt: *sistemele expert*, *recunoașterea de forme (configurații)* *procesarea imaginii*, *controlul automat al sistemelor*, etc.

Referitor la rețelele neuronale, se utilizează frecvent termenul de *paradigmă*. În acest context, paradigma cuprinde tipic descrierea formei și funcției unității elementare de procesare (celulă, nod, neuron,...), topologia rețelei, care specifică configurația interconexiunilor dintre unități, precum și o regulă de adaptare (învățare, instruire) a ponderilor interconexiunilor. Trei clase de sisteme pot constitui paradigme ale sistemelor inteligente: *sistemele continue*, *sistemele logice* și *sistemele simbolice*. Acestea le corespund în calculatoarele convenționale respectiv nivelele de *circuitistică*, de *logică* și de *programare*. Paradigmele rețelelor neuronale, deși distincte în detaliu, selectează o submulțime comună de atribute tipice rețelelor neuronale biologice [4]:

- *elemente simple de procesare*,
- *nivel ridicat de conectivitate*,
- *procesare distribuită paralel a datelor*,
- *funcție de transfer neliniară pentru neuroni*,
- *retroacțiune ("feed-back")*,
- *propagare neamortizată și cu viteză constantă energiei*,
- *procesare nonalgoritmă*,
- *adaptabilitate*,
- *autoorganizare*,
- *toleranță la defecte*,
- *răspunsuri relevante la date de intrare vagi ("fuzzy")*,
- *... putere de generalizare*.

Abilitatea procesării simbolice către care tind modelele performante de rețele neuronale actuale este o consecință tehnică nebanuită a calculatoarelor digitale, care nu au fost inițial concepute în acest scop. În mod similar poate fi privită facilitatea de a raționa logic și rezolva probleme matematice, care apare ca o "deficiență tehnică" inițială neașteptată a creierului uman, în mod indiscutabil inutilă supraviețuirii în zorii apariției speciei umane.

## **X.2. Elemente de teoria modelării**

În principiu, același sistem original poate fi conceptual echivalent cu modele matematice esențial diferite. Mai mult, există modele matematice mixte ale unor sisteme fizice care înglobează compuneri de funcții

caracteristice vagi și funcții de probabilitate discrete. Implementarea fizică a modelelor matematice oferă grade de libertate suplimentare în ceea ce privește suportul fizic de realizare a unui model matematic dat [5].

Rețelele neuronale formale pot fi împărțite în două clase pe baza măsurii în care utilizează modelul procesării distribuite [6]:

- *modele locale*, în care activitatea unei singure unități de procesare reprezintă gradul de participare a unei unități conceptuale cunoscute;

- *modele distribuite*, în care o configurație a activităților unui număr mare de unități determină același grad de participare.

Ca exemple, modelul activării interactive este un model PDP local, în timp ce rețeaua Hopfield este un model PDP distribuit.

Neuronul constituie elementul de bază morfologic și fiziologic al sistemului nervos, fiind o celulă excitabilă cu un înalt grad de specializare, care asigură recepția, codificarea, transmisia și prelucrarea informației conținută de stimuli. Conexiunile interneuronale se realizează prin contactele sinaptice, specificitatea rețelelor neuronale fiind determinate în esență, nu atât de numărul și natura neuronilor constituenți, cât de numărul extrem de mare al interconexiunilor și caracterul tridimensional al rețelei.

Au fost propuse numeroase modele pentru neuron și rețele neuronale, precum și tehnici de procesare paralelă a informației inspirate de mecanismele cerebrale. Modelarea și simularea rețelelor neuronale sunt importante atât pentru elaborarea de modele de procesare a informației în sistemul nervos animal ("*computational neuroscience*"), cât și pentru conceptele de origine neuronală operante în sistemele inteligente artificiale ("*neural computing*"). O teorie comună celor două discipline impune astfel dezvoltarea sistematică și în paralel de modele matematice și metode matematice de analiză, care să evidențieze posibilitățile și limitările intrinseci caracteristice mecanismelor neuronale.

Problema centrală în modelare este alegerea unui *model optimal*, care implică existența unor criterii de comparație. Aceste criterii sunt mai mult sau mai puțin semnificative în contextul unor aplicații particulare. Conceptul de *măsură a performanței* oferă o bază de comparație între diverse modele ale aceluiași sistem sau proces [12]. Alegerea unui model existent sau construcția unui model nou, depind de mai mulți factori cu ponderi diferite:

- disponibilitatea și tipul facilităților computaționale de prelucrare a datelor și simulare asistate de calculator;
- suport tehnic și financiar,
- experiența și nivelul științific al grupului de cercetare,
- antecedente istorice, etc.

Sursele de incertitudine (*nedeterminare*), care afectează modelarea, sunt clasificate în trei categorii distincte, având la origine următoarele [7]:

1. *măsurători imprecise sau nesigure* (cu o anumită lipsă de acuratețe);

2. *abateri aleatorii*;

3. *descrieri sau prezentări vagi (incerte)*.

Se pot considera că aceste trei clase calitativ diferite de la care poate proveni incertitudinea în procesul modelării, conduc la trei tipuri de modele distincte [1, 8]:

1. *modelul determinist*;

2. *modelul probabilist (stochastic)*;

3. *modelul vag ("fuzzy")*.

*Modelul matematic* al unui sistem original sau proces fizic este constituit din ansamblul supozițiilor și ecuațiilor matematice care descriu complet starea și evoluția subiectului modelării. *Modelul fizic* al unui sistem original sau proces fizic reprezintă realizarea materială concretă (tehnică) a modelului matematic ce constituie echivalentul conceptual al modelului fizic și pe care îl include intrinsec.

*Modelarea matematică* poate fi întreprinsă în cadrul unuia dintre cele trei moduri de tratare a incertitudinii prezentate mai sus.

*Modelul determinist* presupune că atât parametrii sistemului cât și variabilele de intrare sunt mărimi deterministe, așadar mărimi a caror evoluție în timp poate fi anticipată (determinată) dacă se cunoaște evoluția lor anterioară și condițiile inițiale. Acest model corespunde descrierii comportamentului rețelelor paucineuronale, cu un număr relative redus de neuroni.

*Modelul probabilist* presupune că atât parametrii sistemului cât și variabilele de intrare sunt mărimi stohastice. În general, acuratețea și utilitatea modelelor probabiliste sunt exclusiv dependente de valorile asociate acestor mărimi statistice și a funcțiilor de distribuție asociate a priori. Imprecizia provine astfel din incertitudinea intrinsecă a rezultatului final; odată efectuată însă experiența, acest rezultat este perfect cunoscut.

*Modelul vag* presupune vagi atât parametrii sistemului cât și variabilele de intrare ale sistemului original, pornind de la imprecizia datelor experimentale și încheiind cu imprecizia descriptivă a termenilor lingvistici. Variatele tipuri de nedeterminare pot fi specificate în cadrul teoriei mulțimilor vagi, astfel încât procesarea datelor se poate desfășura independent de natura investigațiilor. În comparație cu modelele statistice la care mulțimea rezultatelor posibile este calculată pornind de la o funcție de distribuție probabilistă acceptată iar incertitudinea asupra rezultatului dispare după desfășurarea experimentului, la modelele vagi, rezultatul însuși rămâne vag dar oferă maximum de informație în context. Conceptele teoriei mulțimilor vagi sunt transpuse în tehnici de raționament aproximativ

cu aplicații în cele mai diverse direcții ale AI. Raționamentele aproximative uni- și multidimensionale oferă posibilitatea de a deduce proprietăți relevante necunoscute din informația disponibilă a unei baze de date [8].

Modelele fizice (tehnice) neuronale se bazează pe fenomene electromecanice, termice, magnetice, electronice, etc., fiind cunoscute în general sub denumirea de neuristori. Așadar neuristorii sunt dispozitive active care simulează un set de caracteristici funcționale neuronale, care pot varia de la o aplicație la alta, păstrând însă un set comun de atribute specifice neuronului.

### **X.3. Modele de rețele neuronale**

#### **X.3.1 Rețele neuronale formale cu instruire supervizată (memorii asociative)**

Conceptul de *memorie asociativă* caracterizează abilitatea unui sistem dinamic de a parcurge o secvență de reprezentări interne sau de reconstituire a unei reprezentări complexe pornind de la o versiune distorsionată sau incompletă a sa. Memoria asociativă are la bază conceptul de reprezentare distribuită a informației prin stări colective ale rețelei neuronale. Dinamica interacțiilor poate fi făcută să corespundă structurii informaționale, fără un echivalent fizic direct în sistem, realizându-se prin efecte globale și reflectându-se în procesele de redare a informației memorate. Problemele fundamentale ridicate de memoria asociativă sunt: *diafonia* ("cross-talk"), *invarianța* și *inabilitatea* de a opera cu structuri.

Structurile interacțiilor sunt puse în corespondență cu structurile cunoașterii; chiar dacă acestea din urmă nu au un echivalent fizic direct în sistem, se pun în evidență prin efectele stărilor colective și se reflectă în procesele de *reproducere*, *redare* sau *reconstituire* ("recall") din memorie. Modul de reprezentare distribuită a informației prin stări colective ale neuronilor rețelei este numit *holografic* sau *hologic*, dar eleganța sa formală este estompată de implementarea pe calculatoare secvențiale convenționale. Modelele PDP concepute pentru explorarea posibilităților comportamentale ale structurilor modificabile, utilizează tipic reguli de corelație pre- și postsinaptice.

*Modelul Sutton-Barto* este reprezentativ în sensul influenței reciproce de condiționare dintre configurația așteptată și configurația stimulului, care este prea complexă pentru a fi exprimată printr-o regulă de corelație simplă de tip hebbian; în timp ce regula lui Hebb detectează corelații între semnalele de intrare și ieșire, algoritmul Sutton-Barto detectează corelații între "urmele" stimulilor de intrare și modificările produse la ieșire.



*Modelul heterostat* introdus de Klopfi și bazat pe analogia dintre sistemele sociale și rețelele neuronale, constituie un alt model clasic de memorie asociativă. Neuronii sunt presupuși înzestrați cu o caracteristică hedonistică, conform căreia activitatea neuronală se desfășoară în sensul obținerii excitației și evitării inhibiției.

*Modelul Hopfield-Tank*, bazat pe determinarea condițiilor de detonare a neuronilor, utilizează tehnici de relaxare pentru satisfacerea constrângerilor și *întărire simulată* ("*simulated annealing*") cu scopul de a evita fixarea în minime locale. În mod special se justifică aplicațiile rețelei Hopfield-Tank ca memorie asociativă și ca rezolvant în problemele de optimizare;

*Modelul Hoppenstead*, bazat esențialmente pe sincronizarea frecvențelor de detonare a neuronilor, poate fi analizat prin asocierea unei funcții de energie cu *calare pe fază* ("*phase-locking*") a unei rețele de oscilatori electrici și conduce la conceptul de *oscilator neuronal controlat de tensiune*. Ultimile două modele prezentate constituie exemple reprezentative de sisteme PDP clasice.

*Modelul Hamming* implementează calculul distanței Hamming, care măsoară diferența a doua configurații. În cazul procesării informației codificate binar, distanța Hamming este definită prin numărul de biți diferiți între cele două configurații comparate, având valoarea maximă egală cu numărul biților configurației de comparat (disjuncție) și valoarea minimă egală cu zero (identitate).

Rețeaua Hamming este superioară în raport cu rețeaua Hopfield în recunoașterea de configurații, în sensul că necesită mai puține conexiuni decât o rețea Hopfield similară. Modelele Hopfield și Hamming fac parte din clasa rețelelor neuronale cu instruire supervizată, adică învățarea se face într-o primă etapă în care sunt prezentate la intrare configurațiile de referință și are ca efect adaptarea corespunzătoare a ponderilor interconexiunilor. Ulterior, în etapa a doua, o configurație necunoscută este prezentată la intrare și este asociată cu cel mai "apropiat" exemplar din setul memorat, obținându-se fie exemplarul restaurat al configurației de intrare (rețeaua Hopfield), fie clasa corespunzătoare configurației de la intrare (rețeaua Hamming).

*Perceptronul* constituie o clasă reprezentativă de rețele neuronale introdusă de Rosenblat [11], care este capabilă de instruire supervizată, având astfel caracterul unei *memorii asociative*. Caracterul paralel al procesării informației este finalizat în modelele matematice ale *suprafetelor decizionale* (*modelul liniar* și *modelul descreșterii gradientului*), care susțin premiza fundamentală a lui Rosenblatt relativă la caracterul probabilistic al perceptronului. Dacă teorema de convergență a perceptronului oferă a priori succesul algoritmului recursiv de instruire, în cazul perceptronilor multistratificați pentru care nu există asemenea

teoreme de convergență, funcționează criteriile particulare de suficiență ("credit assignment problem").

Modelele matematice fundamentale ale rețelelor neuronale includ inevitabil, mai mult sau mai puțin voalat, modelul neuronului binar clasic, de inspirație biologică și introdus de McCulloch și Pitts [12]. Totuși, trebuie remarcat caracterul analogic (liniar) al comportamentului unui neuron biologic: funcționarea sa se bazează pe curenți ionici care variază continuu și mai puțin pe valori discrete cărora să li se poată asocia nivelele de 0 și 1. Digitizarea impune precizie la prelucrarea informației în sistemele biofizice esențialmente imprecise și totuși sistemele neuronale biologice sunt procesoare cu o excelentă eficiență. Pe de alta parte, comportamentul predominant determinist al rețelelor paucineuronale, pe măsura creșterii complexității rețelei, devine probabilistic în cazul rețelelor plurineuronale. O explicație calitativă posibilă o constituie activitatea rețelelor neuronale în consens cu legile fundamentale ale naturii (relații statistice) și mai puțin împotriva lor (relații deterministe). În acest sens este relevantă comparația energetică între cel mai perfecționat microprocesor actual, care consumă aproximativ  $10^{-9}$  J per operație, în timp ce neuronii biologici utilizează numai  $10^{-16}$  J pentru același proces. Mai mult, în sistemele digitale datele și procesele computaționale trebuiesc convertite în cod binar, proces care necesită aproximativ 10.000 modificari digitale de tensiune per operație. Sistemele analogice realizează aceeași operație într-un singur pas, micșorând astfel substanțial timpul dar mai ales consumul de energie cu un factor de 10.000. Se consideră că cea mai reprezentativă caracteristică a circuitelor neuronale analogice la funcționarea în medii cu parametri imprevizibili o constituie *abilitatea lor de a reprezenta informația în context*. Rețelele neuronale răspund la diferențele în amplitudine ale semnalelor și mai puțin la valoarea lor absolută, eliminând astfel necesitatea unei calibrări precise. Contextul poate fi constituit de o colecție complexă de configurații neuronale de referință acumulate în procesul de instruire. *Jocul dintre context și adaptabilitate este principiul fundamental al paradigmei rețelelor neuronale*, care impune anumite restricții specifice. Deoarece numai modificările și diferențele transmit informație, o permanentă variație a intensității stimulilor este o necesitate pentru sistemele neuronale și nu un impediment, ca în cazul sistemelor digitale. Această caracteristică definitorie a sistemelor vii, care impune variația stimulilor pentru o percepție optimă, plasează clar sistemele neuronale în interiorul lumii observabile, în contrast cu sistemele digitale (numerice) al căror design presupune implicit separarea dintre sistem și lumea exterioară [1].

Considerațiile de mai sus sunt cel puțin în parte susținute de realizarea relativ recentă pe suport semiconductor a reținei artificiale [13] sub forma unui circuit semiconductor monolitic VLSI în tehnologie

CMOS. În consecință, o pondere însemnată este acordată *modelelor statistice și modelelor spin glass*, dezvoltate prin asociație cu proprietățile statistice ale sistemelor magnetice, în care interacțiile dintre momentele magnetice se află "în conflict" reciproc datorită unei dezordini structurale impusă rețelei.

*Rețele Hopfield stohastice* cu straturi neuronale intermediare (ascunse), urmăresc posibilitatea realizării de memorii asociative capabile să reconstituie configurațiile de intrare prin utilizarea parțială a rețelei. Asemenea rețele constituie mașini de calcul bazate pe reguli computaționale statistice numite *mașini Boltzmann*, prin opoziție cu calculatoarele digitale secvențiale clasice numite *mașini van Neumann*.

### **X.3.2. Rețele neuronale formale cu instruire nesupervizată (clasificatori vectoriali)**

Rețelele neuronale cu instruire nesupervizată și autosupervizată sunt denumite generic *clasificatori vectoriali*.

*Rețeaua Carpenter-Grosberg* este reprezentativă prin comparație cu rețeaua Hamming; în sensul că asigură clasificarea exemplarelor pe măsura prezentării lor la intrare prin identificarea unor clase de configurații asemănătoare specificate prin câte un exemplar reprezentativ.

*Modelul instruirii competitive*, model de învățare nesupervizată corespunzător detectorului de regularitate, este în esență un algoritm de instruire statistică neasociativă și este analizat în legătură cu abilitatea rețelelor neuronale adaptive de a descoperi caracteristici relevante ale stimulilor de intrare. Cele mai reprezentative arhitecturi structurale și mecanisme de instruire competitivă sunt: *modelul Linsker și modelele Fukushima*.

*Modelul Linsker* este inspirat din proprietățile celulelor primelor etape de procesare a informației în sistemul vizual la mamifere care, dezvoltându-se la unele animale înainte de naștere, nu pot constitui rezultatul unei experiențe structurate. În acest context, sunt consacrate *metoda analizei componentei principale* pentru identificarea structurii de interes neanticipate în seturi mari de date (de exemplu *partiționarea*) și *principiul infomax* de preservare la ieșire a informației maxime (în sensul lui Shannon) relativ la activitatea de intrare supusă la constrângeri.

*Modelele Fukushima*: cognitronul și, ulterior, neocognitronul, care constituie variante perfecționate de instruire a rețelelor neuronale multistratificate, sunt dezvoltate în principal ca răspuns la observația că rețelele neuronale din centrul vizual ai creierului evoluează în direcția recunoașterii caracteristicilor configurațiilor prezentate la intrare.

*Modelul activării interactive* este calitativ un algoritm capabil să explice elementele fundamentale ale percepției vizuale și auditive.

*Teoria rezonanței adaptive* continuă seria modelelor de instruire competitivă, care încearcă să modeleze stabilitatea algoritmilor competitivi în cazul rețelelor plurineuronale.

Un grup important de modele caracterizate prin instruire autosupervizată cuprinde *algoritmii dezvoltați pe baza principiului darwinist de evoluție a speciilor prin selecție*, reflectat în rețelele neuronale prin competiția dintre neuroni în procesele de învățare. În acest sens și în perspectiva suportului biologic de care se bucură, următoarele modele pot fi considerate cele reprezentative [1]:

*Modelul Malsburg*, care modifică ușor principiul evoluționist al selecției printr-o combinație inteligentă de inhibiție și excitație locală, explicând astfel structura cortexului vizual la pisică;

*Modelul Kohonen*, care propune un mecanism de autoorganizare a rețelelor neuronale prin care numărul neuronilor alocați unei regiuni din spațiul stimulilor de intrare este proporțional cu nivelul ei de activitate;

*Modele "genetice" de instruire*, care se bazează pe un ansamblu de opțiuni inițiale (populația de cromozomi) în procesele iterative de instruire;

*Modelul Hecht-Nielsen*, care combină regula lui Hebb cu conceptele instruirii competitive, în scopul identificării aspectelor structurale caracteristice câmpului datelor de intrare.

Multitudinea modelelor teoretice cuprinse în literatura de specialitate și care reflectă, în esență, intuițiile autorilor asupra funcționării creierului, au fost selecționate pentru prezentare în bună măsură subiectiv, dar ordonate după două criterii considerate de noi fundamentale. Primul corespunde clasificării propuse de Rumelhart [14], având la bază caracteristicile perechii intrare-ieșire, criteriu conform căruia se deosebesc patru clase:

1. *autoasociatorul*,
2. *asociatorul de configurații*,
3. *clasificatorul de configurații*,
4. *detectorul de regularitate*.

Al doilea criteriu, corelat însă cu primul, se bazează pe natura asistării procesului de învățare a rețelei, conform căruia se disting următoarele modalități:

- *instruire supervizată*,
- *instruire nesupervizată*,
- *instruire autosupervizată*.

## X.4. Modele matematice, reguli de propagare și algoritmi de instruire a rețelelor neuronale formale

### X.4.1. Modele matematice generale

Spațiul de intrare  $X = \mathbf{R}^n$  și spațiul de ieșire  $Y = \mathbf{R}^m$  sunt considerate spații vectoriale finite, produsul cartezian  $X \times Y \ni \{x, y\}$  constituie spațiul stărilor sistemului;  $X = \{0, 1\}^n$  pentru rețele care operează cu funcții booleene și  $X = [0, 1]^n$  pentru funcții vagi ("fuzzy"). Setul configurațiilor de instruire se definește ca o submulțime  $K \subset X \times Y$ .

Fie  $N$  mulțimea neuronilor formali ai rețelei indexați cu  $j=1, \dots, n$ , atunci  $P(N) = 2^n$  constituie conjuncții sau coaliții neuronale; o conexiune neuronală leagă un neuron postsinaptic  $j$  de  $S \subset P(N)$  conjuncții de neuroni presinaptici. Fiecare conjuncție neuronală  $S$  preprocesează semnalele aferente  $x_i$  produse de neuronii presinaptici:

$$x = (x_i)_{i=1, \dots, n} \xrightarrow{\varphi_S} \varphi_S(x) \quad (10.1)$$

Dacă  $S = \{i\}$ , atunci rolul de control îl joacă matricea sinaptică:

$$W = (w_j^S)_{\substack{S \in 2^N \\ j=1, \dots, n}} \quad (10.2)$$

unde  $w_j^S$  reprezintă ponderile sinaptice de la conjuncția  $S$  la neuronul  $j$ . Se poate considera astfel că neuronul  $j$  este excitat de semnalul:

$$\sum_{S \in 2^N} w_j^S \varphi_S(x) \quad (10.3)$$

atunci pentru neuronii sincroni (sisteme dinamice discrete) ieșirea  $y_j$  a neuronului  $j$  este:

$$y_j = f_j \left( \left\{ (w_j^S)_S(x) \right\}_{S \in 2^N} \right) \quad (10.4)$$

iar pentru neuronii asincroni (sisteme dinamice continue) avem:

$$x'_j(t) = f_j \left( \left\{ (w_j^S)_S(t), j_S(x(t)) \right\}_{S \in 2^N} \right) \quad (10.5)$$

ultimile două expresii se numesc *reguli de propagare* și caracterizează *dinamica rețelei*, unde în general putem exprima mai simplu:

$$f_j \left( \left\{ \left( x_j^{\mathbf{S}} \right)_{\mathbf{S} \in 2^N} \right\} \cdot j_{\mathbf{S}}(x) \right) = g_j \left( \sum_{\mathbf{S} \in 2^N} w_j^{\mathbf{S}} j_{\mathbf{S}}(x) \right) \quad (10.6)$$

1. *Memorii asociative booleene* (absența preprocesării în conjuncțiile neuronale  $j_{\mathbf{S}}(x) = 0$ ,  $|\mathbf{S}| > 1$ ,  $j_{\{i\}}(x) = x_i$  și funcțiile  $g_j$  afine):

$$y_j = \sum_{i=1}^n w_j^i x_i + c_j \quad (10.7)$$

2. *Memorii asociative vagi* ( $X = \mathbf{R}^n$ ,  $Y = \mathbf{R}$ ,  $K = [0, 1]^n \times [0, 1]$ ):

$$y = \sum_{\mathbf{S} \in 2^N} w^{\mathbf{S}} \left( \prod_{i \in \mathbf{S}} x_i \right)^{\frac{1}{|\mathbf{S}|}} \quad (10.8)$$

3. *Neuronii McCulloch-Pitts*:

$$y_j = \begin{cases} 0 & \text{dacă } \sum_{\mathbf{S} \in 2^N} w_j^{\mathbf{S}} j_{\mathbf{S}}(x) < b_j \\ 1 & \text{dacă } \sum_{\mathbf{S} \in 2^N} w_j^{\mathbf{S}} j_{\mathbf{S}}(x) \geq b_j \end{cases} \quad (10.9)$$

4. *Perceptronul* ( $X = \mathbf{R}^n$ ,  $Y = \mathbf{R}$ ,  $\varphi_{\mathbf{S}}(x) = 0$  dacă  $|\mathbf{S}| > 1$ ):

$$y = \begin{cases} 0 & \text{dacă } \sum_{i=1}^n w^i x_i < \beta \\ 1 & \text{dacă } \sum_{i=1}^n w^i x_i \geq \beta \end{cases} \quad (10.10)$$

5. *Rețele neuronale adaptive*. Să considerăm funcția:

$$\Phi: X \times U \longrightarrow Y \quad (10.11)$$

unde  $X$  este spațiul de intrare,  $Y$  este spațiul de ieșire iar  $U$  este spațiul de control (spațiul parametrilor). O pereche intrare-ieșire  $(a^p, b^p)$  se numește *configurație*. O configurație  $(x, y) \in K$  este recunoscută (descoperită, generalizată) de sistemul adaptiv programat de un semnal de control  $u$ , dacă  $y = \Phi(x, u)$  este răspunsul sistemului la excitația  $x$ . Alegerea unui asemenea control se face prin învățarea unui număr dat de perechi intrare-ieșire  $(a^p, b^p)_{p \in \mathbf{P}} \subset K$ , unde  $\mathbf{P}$  este numit *setul (mulțimea) de instruire (învățare)*, determinând un control  $u^p$  care satisface:

$$\Phi(a^p, u^p) = b^p, (\forall) p \in \mathbf{P} \quad (10.12)$$

### 6. Rețele neuronale liniare. Matrici sinaptice optimale.

Fie spațiile vectoriale  $X$  și  $Y$  finit dimensionale, înzestrate cu produsele scalare  $l$  și  $m$ , iar spațiul operatorilor liniari  $L(X, Y)$  cu produsul tensorial  $l_* \otimes m$ . Fie  $\mathbf{P}$  o mulțime finită de perechi  $(a^p, b^p) \in X \times Y$  care definește un subspațiu  $L_p$  al operatorilor  $W \in L(X, Y)$  ce satisface constrângerile:

$$W a^p = b^p, (\forall) p \in \mathbf{P} \quad (10.13)$$

Să considerăm matricea cu intrările  $(l(a^p, a^q))_{p, q \in \mathbf{P}}$  care este inversabilă atunci când vectorii  $a^p$  sunt liniari independenți în  $X$ ; inversa sa o vom nota cu  $(g_{pq})_{p, q \in \mathbf{P}}$ , avem:

$$\sum_{q \in \mathbf{P}} g_{pq} l(a^q, a^r) = 0 \text{ dacă } p \neq r \text{ și } \sum_{q \in \mathbf{P}} g_{pq} l(a^q, a^p) = 1 \quad (10.14)$$

Fie  $U \in L(X, Y)$  dat și să presupunem că vectorii  $a^p$  sunt liniar independenți, atunci există o soluție unică  $\bar{W} \in L_{\mathbf{P}}$  care minimizează distanța  $(\lambda_* \otimes \mu)(W - U)$  relativ la  $U$  dată de expresia:

$$\bar{W} = U - \sum_{p, q \in \mathbf{P}} g_{pq} L a^q \otimes (U a^p - b^p) \quad (10.15)$$

Dacă  $a^p$  sunt 1-ortogonali, expresia de mai sus devine:

$$\bar{W} = U - \sum_{p \in \mathbf{P}} L a^p \otimes \frac{U a^p - b^p}{(\lambda(a^p))^2} \quad (10.16)$$

Mai mult, dacă  $\mathbf{P} = \mathbf{P}_0 \cup p_1$  și  $U a^p = b^p$ ,  $(\forall) p \in \mathbf{P}_0$ , atunci:

$$\bar{W} = U - q_1 \otimes (U a^{p_1} - b^{p_1}), \text{ unde } q_1 = L \left( \sum_{q \in \mathbf{P}} g_{qp_1} a^q \right) \quad (10.17)$$

## X.4.2. Algoritmi de instruire generalizați

### X.4.2.1. Algoritmul instruirii exhaustive ("Heavy Learning Algorithm")

Rețelele neuronale multistratificate pot fi caracterizate alternativ de:

$$U = L(Z_1, Z_2)$$

$$x \longrightarrow y = y_1 + y_2,$$

$$\text{unde } y_1 = c(x), \quad c: X \longrightarrow Y$$

$$y_2 = \Phi(x) W \Psi(x), \quad \Psi: X \longrightarrow Z_1, \quad W \in L(Z_1, Z_2), \quad \Phi: X \longrightarrow L(Z_2, Y)$$

Algoritmul trebuie să găsească o matrice sinaptică  $\mathbf{W}_P$  care să învețe setul de instruire în sensul că:

$$(\forall) p = 1, \dots, P, \quad c(a_p) + \Phi(a_p) \mathbf{W}_P \Psi(a_p) = b_p \quad (10.18)$$

Dacă sunt satisfăcute următoarele condiții:

$$c: X \longrightarrow Y \text{ și } \Psi: X \longrightarrow Z_1 \text{ continue}$$

$$\Phi: X \longrightarrow L(Z_2, Y) \text{ continuă și surjectivă}$$

$$\Psi(a_p) \text{ ortonormale între ele}$$

atunci :

$$\mathbf{W}_P = \mathbf{W}_{P-1} - \Psi(a_p) \otimes \Phi(a_p)^+ (c(a_p) + \Phi(a_p) \mathbf{W}_{P-1} \Psi(a_p) - b_p) \quad (10.19)$$



### X.4.2.2. Algoritmul gradientului

Fie  $f: X \longrightarrow \mathbf{R}$  o funcție pe care dorim să o minimizăm, atunci:

1. (Inițializare) Se alege  $x_n \in X$ ,  $n = 0$ ;

2. Dacă  $f'(x_n) = 0$ , procesul se încheie, altfel  $x_{n+1} = x_n - \delta_n \frac{f'(x_n)}{\|f'(x_n)\|}$ ;

3.  $n \longrightarrow n+1$  și se revine la pct. 2.

Se poate arăta [15] că dacă  $f: X \longrightarrow \mathbf{R}$  este o funcție convexă diferențiable și mărginită inferior, iar dacă considerăm o secvență de "pași"

$$\delta_n > 0, \quad \lim_{n \longrightarrow +\infty} \delta_n = 0, \quad \sum_{n=0}^{\infty} \delta_n = +\infty$$

atunci următoarea secvență descrescătoare de scalari este convergentă:

$$\theta_k = \min_{n=0, \dots, k} f(x_n) \xrightarrow{k \longrightarrow +\infty} v = \inf f(x). \quad (10.20)$$

#### **Rețele neuronale neliniare.**

Să considerăm în continuare următoarea rețea:

$X = \mathbf{R}^n$ ,  $Y = \mathbf{R}^m$ ,  $W \in L(X, Y)$ ,  $f(x, W) = g(Wx)$  unde  $g: Y \longrightarrow Y$ .

Se cere determinarea unei matrici sinaptice care să optimizeze rețeaua neuronală pentru un set de configurații dat:

$$\Phi(x, W) = y \quad (10.21)$$

unde am notat cu  $W$  matricea asociată operatorului  $W$ . Fiind date o funcție de evaluare  $E$  definită pe spațiul de ieșire  $Y$  și o configurație  $(a, b)$ , se caută o matrice sinaptică  $W$ , care să minimizeze funcția:

$$W \longrightarrow H(W) = E(g(Wa) - b) \quad (10.22)$$

Pentru cazul simplu al unei rețele neuronale cu un singur strat descrisă de funcția de propagare  $g: Y \longrightarrow Y$  și  $\Phi(a, W) = g(Wa) - b$ , și având în vedere modul de evaluare a expresiei gradientului unei funcții  $y \xrightarrow{\Psi} \Psi(y) = E(g(y))$  care conduce la  $g'(y)^* E'(g(y))$ , avem:

$$H'(\mathbf{W}) = a \otimes g'(\mathbf{W}a)^* E'(g(\mathbf{W}a) - b) \quad (10.22)$$

având intrările egale cu:

$$\frac{\partial H}{\partial w_{ij}} = a_{ij} \left( \sum_{k=1}^m \frac{\partial g_k}{\partial y_i}(\mathbf{W}a) \frac{\partial E}{\partial y_k}(g(\mathbf{W}a) - b) \right) \quad (10.23)$$

Rezultă că dacă matricea sinaptică  $\overline{\mathbf{W}}$  determină minimumul lui  $H$ , atunci ea este soluția sistemului de  $mn$  ecuații:

$$a_j \left( \frac{\partial g_i}{\partial y_i}(\overline{\mathbf{W}}a) \frac{\partial E}{\partial y_i}(g(\overline{\mathbf{W}}a) - b) \right) = 0, \quad (\forall)_{i,j} \quad (10.24)$$

și regula de instruire a rețelei va fi:

$$w_{ij}^{n+1} = w_{ij}^n - \varepsilon_n a_j \left( \frac{\partial g_i}{\partial y_i}(\mathbf{W}^n a) \frac{\partial E}{\partial y_i}(g(\mathbf{W}^n a) - b) \right) \quad (10.35)$$

#### X.4.2.3. Algoritmul lui Newton

Să considerăm rețele neuronale cu unul, mai multe, sau o infinitate de straturi (continuă):  $x \xrightarrow{\Phi} y = \Phi(x, \mathbf{W})$ , fie  $(a^j, b^j)_{1 \leq j \leq J}$  o familie de configurații; atunci să determinăm matrice sinaptice  $\overline{\mathbf{W}}$  care să învețe aceste configurații, deci:

$$\Phi(a^j, \overline{\mathbf{W}}) = b^j, \quad (\forall)_{j=1, \dots, J} \text{ sau mai compact } \Phi(\overline{\mathbf{W}}) = (g(\overline{\mathbf{W}}a^j) - b^j)_{1 \leq j \leq J}$$

Algoritmul lui Newton în acest context pornește de la:

$$\Phi(\mathbf{W}^n)(\mathbf{W}^{n+1} - \mathbf{W}^n) = -\delta \Phi(\mathbf{W}^n), \quad (\forall)_{n \geq 0}. \quad (10.36)$$

unde se presupune date matricea inițială arbitrară  $\mathbf{W}^0$  și  $\delta \geq 0.5$ .

Să considerăm o rețea neuronală neliniară care asociază spațiului de intrare  $X = \mathbf{R}^n$  spațiului de ieșire  $Z = \mathbf{R}^m$ , prin stratul intermediar  $Y$ . Semnalelor de intrare  $x \in X$ , ponderate prin intermediul matricii sinaptice  $\mathbf{W}$  asociat operatorului  $W \in L(X, Y)$ , le corespund la ieșire răspunsurile  $f(x, \mathbf{W}) = g(\mathbf{W}x)$ ,  $g: Y \rightarrow Z$ ; atunci dacă funcția de transfer  $g$  este

diferențiabilă, funcțiile  $g'(W a^j)$ ,  $(\forall)_{j=1, \dots, J}$ , sunt surjective  $(\forall) W \in L(X, Y)$ , iar intrările  $a^j$  sunt ortogonale pentru produsul scalar  $l$  definit pe  $X$ , avem:

$$W^{n+1} = W^n + \delta L a^j \otimes \sum_{j=1}^J (g'(W^n a^j))^+ \cdot \left( \frac{b^j - g(W^n a^j)}{l(a^j, a^j)} \right) \quad (10.37)$$

## X.5. Aplicații ale rețelelor neuronale

### X.5.1. Concepte fundamentale

Rețelele neuronale constituie, în esență, un mediu specific de procesare a informației în sisteme care operează cu date de intrare multivariate și răspunsuri multiple. *Datele multivariate* sunt asociate cu un obiect caracterizat de mai multe variabile; ca exemplu, analiza unor probe de apă potabilă pentru determinarea poluanților Pb, Cl<sub>2</sub> și CO constituie o problemă cu date de intrare multivariate. *Răspunsurile multiple* sunt asociate cu date multivariate studiate în funcție de mai mulți factori care pot influența răspunsul sistemului; în exemplul de mai sus factorii ar putea fi temperatura, presiunea, etc.

Relația generală dintre intrarea  $m$ -variata  $X_s = (x_{s1}, x_{s2}, \dots, x_{sm})$  și răspunsul  $n$ -variata  $Y_s = (y_{s1}, y_{s2}, \dots, y_{sn})$  al sistemului, unde  $s = 1, 2, \dots, r$  indică eșantioane (probe) diferite, poate fi explicitată compact sub forma:

$$Y_s = A X_s \quad (10.38)$$

unde  $A$  este în general un operator complex. În particular,  $A$  poate fi o matrice  $(m \times n)$  care transformă liniar vectorul  $X_s$  în vectorul  $Y_s$ , astfel încât ecuația (10.38) constituie *modelul matematic al problemei multivariate multirăspuns*.

În principiu, problema este următoarea: se dă un set de date  $m$ -variate  $\{X_s\}_{s=1, 2, \dots, r}$  care conduce la setul răspunsurilor  $n$ -variate  $\{Y_s\}_{s=1, 2, \dots, r}$ , în timp ce operatorul  $A$  este de cele mai multe ori necunoscut. Mai mult, în majoritatea aplicațiilor se solicită estimarea chiar a răspunsurilor posibile  $\{Y_s\}_{s=1, 2, \dots, r}$ . Sarcina fundamentală este de a trage concluzii relevante din probe necunoscute pe baza răspunsurilor obținute pornind de la un set de date multivariate selecționate,  $\{Y_s\}_{s=1, 2, \dots, p}$ , unde  $s \in \wp$ , card  $\wp = p < r$ ,  $\wp$  fiind denumit *setul de antrenament, învățare sau instruire*. Nu există un algoritm universal care să selecționeze din

mulțimea perechilor de vectori intrare-ieșire  $\{(X_s, Y_s)\}$ , numite pe scurt configurații, acele perechi care sunt cele mai relevante prin conținutul informațional în procesul de instruire a rețelei.

Alegerea unei strategii de instruire *supervizată* sau *nesupervizată* a rețelei este o chestiune care depinde de specificul problemei și datele disponibile. Deoarece instruirea nesupervizată este mai flexibilă prin multitudinea răspunsurilor posibile, este recomandată ca etapă inițială de acumulare a informației. Ulterior, se trece la instruirea supervizată, proces în care datele multivariate cunoscute se împart în trei clase:

- *setul de instruire (învățare)*  $\mathcal{D}$ , care conține mulțimea configurațiilor de referință;
- *setul de control*, care determină încheierea procesului de instruire;
- *setul de testare*, care permite verificarea predictibilității rețelei.

Deși alegerea aleatorie a configurațiilor pentru cele trei clase considerate nu este cea mai indicată strategie, totuși este singura opțiune rezonabilă în absența unor date sau informații prealabile.

Importanța remarcabilă a selecției configurațiilor în setul de instruire a condus la apariția unui domeniu de specializare denumit *proiectare experimentală ("experimental design")* sau *proiectare factorială ("factorial design")*. Este foarte important pentru setul de instruire să acopere integral și cât mai reprezentativ spațiul variabilelor. Un mod consacrat de abordare a acestei probleme constă în selecția variabilelor relevante și divizarea domeniilor lor de valori în nivele fixe sau intervale limitate (de exemplu: scurte, medii și lungi). Apoi datele sunt selecționate astfel încât setul de instruire să conțină valori ale tuturor variabilelor reprezentând combinații ale tuturor nivelelor sau intervalelor. Reluând exemplul de mai sus, dacă 3 nivele (intervale) posibile de valori sunt asignate la 4 variabile, atunci datele de instruire trebuiesc selecționate astfel încât fiecare din cele  $3^4$  subspații să conțină cel puțin datele unui experiment. Când numărul de variabile este prea mare pentru o proiectare experimentală cu trei nivele, numărul datelor se reduce, dar se menține acoperirea integrală a spațiului cât mai reprezentativ posibil. În acest sens, se ține seama că variabilele strâns corelate conțin informație relativ similară și contribuie numai la creșterea redundanței, astfel încât un mod efectiv de a reduce datele de instruire este de a reduce numărul variabilelor la cele care conțin o cantitate maximă de informație nouă.

Limitările majore ale proiectării experimentale sunt două:

- reducerea numărului variabilelor este laborioasă și conduce inevitabil la pierderea de informație,
- definirea intervalelor fixe pentru domeniul valorilor variabilelor este în buna măsură subiectivă și arbitrară. Rețelele Kohonen oferă o alternativă optimă în cazul în care selecția setului de instruire se realizează

prin antrenamentul rețelei Kohonen utilizându-se toate datele de intrare, în timp ce la proiectarea experimentală numărul variabilelor este micșorat pentru a reduce numărul subspațiilor dependente exponențial de numărul variabilelor.

În cele mai multe aplicații, arhitectura care se dovedește optimă corespunde rețelelor multistratificate de tip multiperceptron cu un singur strat intermediar (ascuns), la care numărul unităților intermediare se stabilește experimental. O regulă empirică determinată pe mai multe clase de aplicații sugerează că numărul optim de unități intermediare este primul întreg care depășește logaritmul în baza doi din numărul unităților stratului de intrare al rețelei.

La alegerea strategiei adecvate de instruire, factorul imediat important îl constituie numărul elementelor (cardinalul) celor trei mulțimi considerate. Instruirea supervizată poate implica parcurgerea de câteva mii de ori a configurațiilor de instruire. Mai mult, pot fi necesare corecții a mii de coeficienți sinaptici pentru fiecare obiect de intrare. În consecință, eficiența procesului de instruire depinde de numărul configurațiilor cunoscute, de numărul variabilelor asociate fiecărei configurații care definește dimensiunea spațiului (stratului) de intrare, și de numărul straturilor intermediare (ascunse), care permit reprezentări interne ale informației crescând semnificativ puterea de separare a rețelei.

Doi factori de merit sunt considerați esențiali în caracterizarea performanțelor rețelelor neuronale:

- *reproducerea din memorie ("recall")*, care se exprimă prin procentajul configurațiilor recunoscute corect din *setul de instruire*, după ce procesul de învățare s-a încheiat,

- *predicția ("prediction")*, care se exprimă prin procentajul configurațiilor recunoscute corect din *setul de testare*, configurații neutilizate în procesul de instruire a rețelei.

### **X.5.2. Clase de aplicații specifice ale rețelelor neuronale**

**Clasificarea** datelor multivariate sortează configurațiile în clase simple, sau ierarhizează în interiorul claselor. **Ierarhizarea** implică o secvență a nivelelor decizionale care să clasifice adecvat o configurație într-un grup sau subgrup, categorie sau subcategorie, clasă sau subclasă.

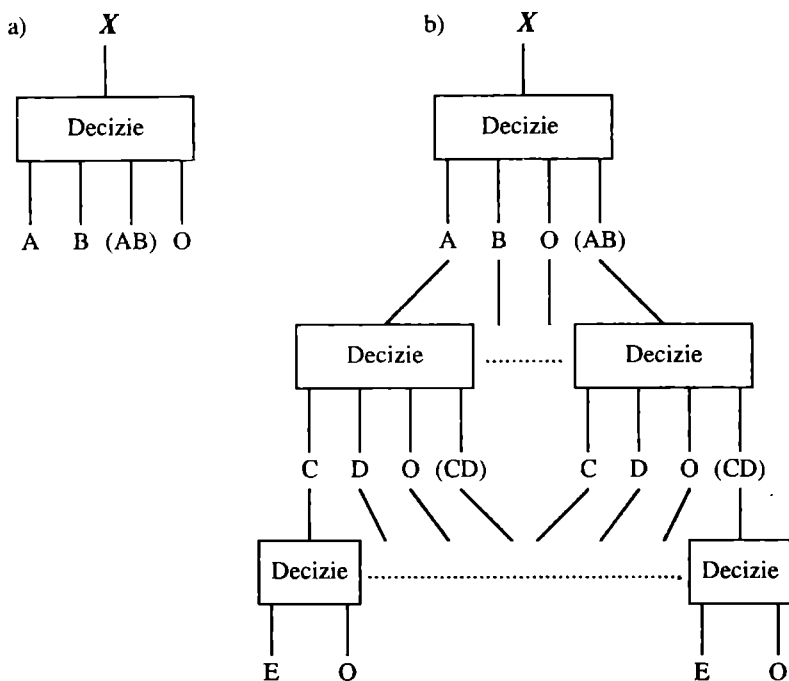
Specific aplicațiilor în chimie, un obiect poate aparține simultan mai multor clase distincte, astfel încât rezultatul este produsul logic al mai multor decizii (Fig. 10.1).

**Modelarea** urmărește elaborarea de sisteme formale capabile să genereze răspunsuri continue pentru datele de intrare, în sensul că la modificări mici la intrare corespund modificări mici ale răspunsului sistemului. Prin esența sa, modelarea necesită întotdeauna instruire

supervizată și, în consecință, implică algoritmi de propagare inversă sau contra-propagare a erorii ("back-propagation" și "counter-propagation").

În general, criteriul alegerii parametrilor de modelare este minimizarea diferenței dintre setul răspunsurilor determinate experimental,  $\{Y_s^{\text{experimental}}\}_{s \in \mathcal{S}}$ , și răspunsurile generate de model,  $\{Y_s^{\text{model}}\}_{s \in \mathcal{S}}$ , corespunzătoare aceluiași set de configurații de intrare  $\{X_s\}_{s \in \mathcal{S}}$ , adică:

$$\min (Y_s^{\text{experimental}} - Y_s^{\text{model}})^2 \quad (10.39)$$

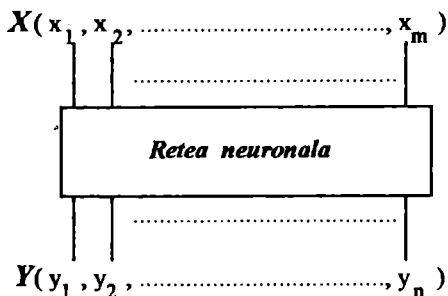


**Fig. 10.1.** Clasificare pe un nivel decizional (a) și clasificare ierarhică (b)

Avantajul major al rețelelor neuronale în problemele multirăspuns asupra modelării clasice cu funcții analitice constă în faptul că "țintele",  $\{Y_s^{\text{model}}\}_{s \in \mathcal{S}}$ , pot fi multidimensionale. Cu alte cuvinte, modelele clasice necesită predefinirea unei funcții pentru fiecare răspuns, în timp ce rețelele neuronale pot modela un vector multirăspuns fără informații apriorice.

Modelarea este eficientă pentru configurații cu un număr relativ redus de variabile de intrare (în general mai mic de 10), care conduc la arhitecturi de rețele relativ limitate și putere de calcul solicitată inferioară celor implicate de tehnicile clasice de clasificare (Fig. 10.2).

Alegerea algoritmului de instruire este consecința directă a distribuției obiectelor de intrare în spațiul variabilelor. Metoda propagării inverse a erorii ("*error back-propagation algorithm*") este indicată în cazul aplicațiilor cu un număr relativ redus de obiecte și repartizate relativ egal. Metoda contra-propagării erorii ("*error counter-propagation algorithm*") este adecvată în cazul unui număr mare de obiecte de intrare repartizate inegal în spațiul variabilelor, după ce au fost în prealabil reduse la o acoperire regulată a spațiului prin intermediul rețelelor Kohonen sau evaluări statistice ale intervalelor variabilelor.



**Fig. 10.2** Modelarea cu rețele neuronale

**Mapping-ul** este în esență o metodă de transformare a unui spațiu de dimensiune mare într-unul cu dimensiune relativ mai mică. În timp ce clasificarea sau determinarea de clusteri se concentrează asupra clasei sau clusterului la care aparține o anumită configurație, mapping-ul privește întreg spațiul și efectiv obține clusteri din datele de intrare. Specificitatea rețelelor neuronale constă în reprezentarea celui mai activ neuron prin coordonatele sale și nu prin valoarea ieșirii sale. Elementele esențiale în determinarea corespondenței se referă la topologia spațiului întreg, numărul regiunilor care pot fi distinse, forma lor și corelația dintre caracteristicile spațiului și cele ale configurațiilor de intrare care le-au generat.

Mapping-ul este în particular recomandat ca o metodă de preprocesare și selecție a celei mai adecvate submulțimi de configurații ale unei mulțimi mari, cum ar fi selecția spectrelor pentru corelații spectru-spectru, sau eliminarea combinațiilor redundante de parametri aparținând unei serii de spectre. Rețelele Kohonen rezolvă orice problemă posibilă de mapping.

**Auto- și hetero-asocierea**, deși apropiate ca tehnici de clasificare, pot fi considerate distinct ca metode specifice de identificare corectă a perechilor intrare-ieșire. Rețelele *auto-asociative* furnizează răspunsuri corecte la date de intrare incomplete, distorsionate sau zgomotoase,

asociind datelor de intrare  $\{X_s\}_{s=1,2,\dots,m}$  răspunsuri adecvate  $\{Y_s\}_{s=1,2,\dots,n}$ . Rețelele *hetero-asociative* realizează corespondența "unu-la unu" dintre mulțimea obiectelor de intrare  $\{X_s\}_{s=1,2,\dots,m}$  și răspunsurile corespunzătoare  $\{Y_s\} = \{X_s\}$ , unde  $s = 1, 2, \dots, m$ .

Memoriile asociative se instruesc prin exemple, astfel încât numai sistemele neuronale formale cu instruire supervizată sunt adecvate pentru implementarea funcțiilor asociative.

**Proгноza secvențelor temporale** modelează comportamentul sistemelor multirăspuns corespunzător unor serii de date înregistrate în timp. Variabilele de intrare și ieșire sunt esențial similare în modelarea temporală; deosebirea este că datele de intrare conțin valorile variabilelor de proces din trecut și prezent, în timp ce datele de ieșire prognozează valorile viitoare ale aceluiași variabile. Seturile consecutive de date privite ca vectori de intrare și seturile consecutive de date viitoare privite ca vector de ieșire sunt denumite *orizonturi trecute* și respectiv *orizont prezent* ale procesului de instruire [16]. Deși orizonturile pot principial avea orice lungime, totuși orizontul viitor acoperă numai un pas înainte în majoritatea aplicațiilor de control al proceselor.

Fie  $P$  un vector  $m$ -dimensional care conține informație asupra unui proces dat la un moment de timp  $t$ :

$$P_t = (x_{1t}, x_{2t}, \dots, x_{mt}) \quad (10.40)$$

Dacă se consideră trei pași consecutivi ai procesului, atunci vectorul de intrare este:

$$X = (x_{1t}, x_{2t}, \dots, x_{mt}, x_{1t+1}, x_{2t+1}, \dots, x_{mt+1}, x_{1t+2}, x_{2t+2}, \dots, x_{mt+2}) \quad (10.41)$$

care se mai poate scrie compact, în termeni vectoriali temporali:

$$X = (P_t, P_{t+1}, P_{t+2}) \quad (10.42)$$

Vectorul țintă corespunzător  $Y$  poate fi constituit de al patrulea vector în secvență, și anume:

$$Y = (x_{1t+3}, x_{2t+3}, \dots, x_{mt+3}) \quad (10.43)$$

sau vectorial:

$$Y = P_{t+3} \quad (10.44)$$

Deși modelul complet al unui proces necesită ca toate variabilele de proces să fie rulate, numai un număr limitat de variabile semnificative sunt utilizate la instruire pentru un proces dat. Datele "viitoare" solicitate la citirea instruirii sunt obținute fie din modele teoretice sau realmente din experiențe anterioare.



Investigarea proceselor temporale poate fi utilizată în diverse probleme care implică secvențe temporale, începând cu prognoza parametrilor de mediu în uzinele chimice și până la deducerea structurii secundare a proteinelor bazată pe secvența aminoacizilor.

### X.5.3. Abordarea modelării neuronale

Utilizarea rețelelor neuronale este deschisă unei largi varietăți de aplicații. În general, tipuri diferite de probleme necesită diferite arhitecturi neuronale și strategii de instruire adaptate [17].

Ca modalitate principală de abordare a unei probleme utilizând rețele neuronale, este în general indicat a se clasifica mai întâi problema de interes într-una din clasele de aplicații prezentate mai sus. În al doilea rând se alege și se adaptează topologia rețelei care corespunde aplicației identificate. În al treilea rând se optează pentru un algoritm de instruire adecvat capabil să rezolve rețeaua. Indicații generale privind clasificarea tipurilor de probleme care se pot aborda cu tehnici de inspirație neuronală sunt prezentate pe larg în literatura de specialitate [18-20] și sunt sumarizate în tabelul 10.1.

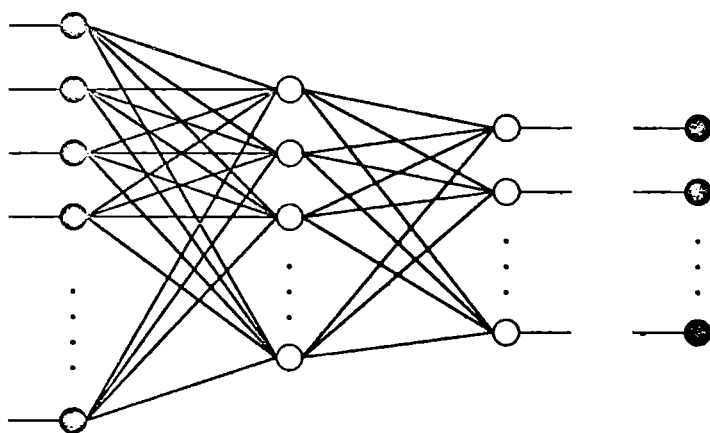
**Tabelul 10.1.** *Tipuri de probleme rezolubile cu rețele neuronale și algoritmi de instruire recomandați*

Tipuri de probleme	Algoritmi de instruire		
	Back-propagation	Counter-propagation	Kohonen map
Clasificare	Da	Da	Da
Modelare	Da	(Da)	
Mapping		(Da)	Da
Asociere	Da	Da	
Secvențe temporale	Da	Da	

Problemele de chimie conțin adesea elemente care aparțin la tipuri diferite de probleme, astfel încât rezolvarea globală se desfășoară utilizând mai multe tipuri de rețele simultan. Deoarece rețelele neuronale sunt metode generale de rezolvare, un accent deosebit trebuie pus pe reprezentarea informației chimice în context. Cu toate acestea, cele mai multe aplicații practice în chimie publicate în literatura de specialitate [21,22] dovedesc succesul rețelei directe, complet conectate, de tip

perceptron, cu un singur strat intermediar (Fig. 10.3) și care se instruieste utilizând algoritmul de propagare inversă a erorii.

Atunci când rețelele directe clasice sunt instruite pentru mai multe configurații cu reguli de propagare derivate din algoritmul de propagare inversă a erorii, pot apare interferențe puternice la stabilirea ponderilor sinaptice. Aceste interferențe au ca efect descreșterea convergenței procesului de instruire și scăderea puterii de generalizare a rețelei. O cale de a evita interferențele pentru anumite subclase de configurații constă în utilizarea simultană a mai multor *rețele locale diferite* și o *rețea-poartă* care să decidă rețeaua locală activă pentru fiecare subclasă în parte [23]. Ultimile realizări în această direcție se referă la elaborarea unor noi *funcții obiectiv* specializate pentru asociații adaptive de rețele neuronale locale. Mai clar, în locul unei combinații liniare a ieșirilor furnizate separat de rețelele locale, este sugerată abordarea unei rețele-poartă care decide statistic rețeaua locală responsabilă pentru clasificarea fiecărei dintre configurațiile de intrare [24].



Unitati de intrare    Unitati intermediare    Unitati de iesire    Unitati tinta

**Fig. 10.3.** Rețea neuronală de tip perceptron cu un singur strat intermediar

#### X.5.4. Aplicații ale rețelelor neuronale în chimie

##### *Reactivitatea legăturilor chimice*

Una din problemele fundamentale în chimie constă în anticiparea evoluției și a produșilor unei reacții chimice. Deoarece reacțiile chimice sunt inițiate prin desfacerea unor legături dintre molecule, cunoașterea

reactivității legăturilor chimice este esențială. Procesele polare sunt larg implicate în reacțiile organice, ruperea legăturilor generând particule încărcate (*heteroliza*).

Parametrii care influențează desfacerea legăturilor chimice sunt limitați în modelare la următorul set reprezentativ:

- $\Delta q_{tot}$  : diferența de sarcină totală,
- $\Delta q_{\pi}$  : diferența în sarcină  $\pi$ ,
- $\Delta \chi_{\sigma}$  : diferența în electronegativitate  $\sigma$ ,
- $Q_{\sigma}$  : măsura polarității legăturii,
- $\alpha_b$  : polarizabilitatea legăturii,
- $BDE$  : energia de disociere a legăturii,
- $R^{\pm}$  : măsura stabilității rezonanței disponibilă pentru particulele generate prin heteroliză.

Valorile acestor parametri sunt calculate și asignate diverselor legături utilizând un pachet de programe numit *PETRA* (*Parameter Estimation for the Treatment of Reactivity Applications*) [20]. Modelul trebuie să coreleze valorile parametrilor cu reactivitatea legăturilor de interes. Clasificarea reactivității legăturilor este o problemă multivariată din moment ce date particulare atestă insuficiența separării modului reactiv (+) de cel nereactiv (-) de desfacere a legăturilor. Corespunzător parametrilor relevanți de control al reactivității considerați mai sus, numărul intrărilor rețelei va 7, fiecare valoare a variabilelor de intrare fiind scalată între 0 și 1. Codificarea ieșirii poate fi 0 pentru modurile de desfacere nereactivă a legăturilor și 1 pentru cele reactive; în consecință este necesar un singur neuron (unitate) de ieșire. În cele mai multe aplicații, un singur strat intermediar se dovedește suficient, dar numărul neuronilor constituienți se stabilește experimental. Astfel, o rețea neuronală multistratificată de tip perceptron care să realizeze clasificarea modurile de desfacere a legăturilor chimice ca reactive și nereactive are arhitectura (7×3×1), care presupune deci 21 de ponderi sinaptice adaptabile. Regula de instruire recomandată este algoritmul de propagare inversă a erorii, regulă consacrată acestei topologii de rețea.

O abordare mai rafinată și eficientă constă în combinația unei rețele Kohonen pentru selectarea setului de instruire, cu o rețea multistratificată echipată cu instruire supervizată și antrenată pentru clasificare. Ansamblul celor doua rețele este capabil să prezică în moleculele alifatiche care dintre legăturile simple se vor desface preferențial polar și care nu, ba mai mult, indică și direcția în care se vor deplasa sarcinile după heteroliză. Deși rețeaua poate fi antrenată numai pe legături simple, ea dispune de caracteristica de generalizare, în sensul că asociază reactivitatea corectă inclusiv la unele legături duble (de exemplu  $C=O$ ). Răspunsurile rețelei

pot fi interpretate ca *probabilități* de desfacere a legăturilor, ca și o *măsură cantitativă* a reactivității.

### *Optimizarea proceselor HPLC*

Una din cele mai frecvente sarcini ale chimiei analitice este identificarea componentelor individuale ale unui amestec complex. Deoarece componentele care unnează să fie identificate și estimate cantitativ trebuie să fie inițial separate, *metodele cromatografice* joacă un rol important în *tehnicele de separare*. Cromatografia de lichide de înaltă performanță (HPLC) constituie o metodă răspândită de analiză în care un amestec este distribuit între o fază staționară, lichidă sau solidă, pe un suport solid, și o fază lichidă mobilă supusă la presiune ridicată. Diversele componente se distribuie diferit între cele două faze și astfel sunt separate. Principalul factor de merit îl constituie aici *factorul de performanță (FP)* al analizei, care include două aspecte: puterea de separare a coloanei și timpul necesar pentru desfășurarea integrală a procesului de separare. Evident, factorul de performanță este un compromis între separabilitate și timp.

Orice analiză de proces HPLC implică modelarea fazei mobile, în sensul specificării factorului de performanță ca funcție de parametrii săi. Apoi se alege cea mai potrivită fază mobilă pentru o analiză particulară.

În majoritatea aplicațiilor se consideră numai doi parametri de care depinde faza mobilă: *concentrația de etanol ( $C_e$ )* și *aciditatea fazei (pH)*. Orice modelare, clasică sau cu rețele neuronale, necesită date experimentale. În modelarea clasică, pentru calculul valorii factorului de performanță se utilizează în general o formă biliniară:

$$FP = ax^2 + by^2 + cxy + dx + ey = f \quad (10.45)$$

unde  $x$  semnifică valorile concentrației  $C_e$ , iar  $y$  valorile pH-ului. Parametrii  $a, b, c, d$  și  $f$  sunt obținuți prin orice tehnică de optimizare, cum ar fi *regresia liniară*, metoda *Simplex*, etc.

Aceleași date pot fi utilizate pentru proiectarea unei rețele neuronale formale cu un singur strat intermediar, echipată cu o regulă de instruire de tipul propagării inverse a erorii. Alegerea acestui algoritm de instruire este consecința arhitecturii multistratificate a rețelei, precum și consecința necesității instruirii supervizate reclamată de modelare. Corespunzător proprietăților avute în vedere la modelare, rețeaua trebuie să conțină doi neuroni de intrare, un neuron de ieșire și un număr de neuroni intermediari.

Experiențe de optimizare a proceselor HPLC în analiza vinului au relevat ca cele mai performante rețelele de tipul  $(2 \times 6 \times 1)$  conținând  $(2 \times 1) \times 6 + (6 + 1) \times 1 = 25$  coeficienți sinaptici. Rezultatele obținute s-au dovedit în bun acord cu cele furnizate prin metode clasice.

### *Relația structură-activitate*

Unul din domeniile cele mai promițătoare de aplicații ale tehnicilor de modelare îl constituie determinarea relațiilor cantitative dintre structura chimică a unui compus și activitatea sa fizică, chimică sau biologică. În mod deosebit, corelarea structurii chimice a medicamentelor cu efectele farmacologice este de mare interes în industria farmaceutică. Datorită complexității care poate fi atinsă în modelarea rețelelor neuronale, acestea joacă un rol din ce în ce mai important în studiul relației cantitative structură-activitate, problemă denumită în literatura de specialitate QSAR ("Quantitative Structure-Activity Relationships").

Un exemplu remarcabil în acest sens îl constituie lucrările lui Aoyama, Suzuki și Ichikawa [21], care urmăresc prognozarea dozei medicamentoase minime necesară pentru prelungirea cu 40% a vieții cobailor test, cărora li s-au inoculat celule cancerigene L-1210. Doza efectivă minimă depinde de concentrația  $C$  a substanței care produce efectul dorit și se exprimă în termeni logaritmici ( $\log(I/C)$ ), consecință a plăjei de mai multe ordine de mărime a concentrațiilor solicitate.

Setul de date cuprinde modificările caracteristicilor structurale ale carbochinonelor, o clasă de compuși care manifestă activitate anticancerigenă cu diverse intensități. Această activitate este în primul rând influențată de identitatea substituenților  $R^1$  și  $R^2$ . În analiza multiliniară standard, influența combinată a substituenților  $R^1$  și  $R^2$  este descrisă de următoarele variabile fizico-chimice:

- $MR_{1,2}$  : refractivitatea molară,
- $\pi_{1,2}$  : contribuția substituenților la hidrofobicitate,
- $F$  : suma constantelor substituenților la efectul de câmp,
- $R$  : suma constantelor substituenților la efectul de rezonanță; și două variabile locale responsabile de influența exclusivă a lui  $R^1$  :
  - $MR_1$  : refractivitatea molară a lui  $R^1$ ,
  - $\pi_1$  : contribuția lui  $R^1$  la hidrofobicitate.

Au fost selecționate 35 de carbochinone diferite pentru care, fie au fost măsurate valorile celor 6 variabile ( $MR_{1,2}$ ,  $\pi_{1,2}$ ,  $MR_1$ ,  $\pi_1$ ,  $F$  și  $R$ ), fie au fost preluate date din literatura de specialitate ( $\pi_1$ ,  $F$  și  $R$ ), apoi doza efectivă minimă a fost măsurată și reprezentată ca  $\log(I/C)$ . În consecință, a fost utilizată o rețea neuronală cu 6 unități de intrare corespunzătoare celor 6 variabile ( $MR_{1,2}$ ,  $\pi_{1,2}$ ,  $MR_1$ ,  $\pi_1$ ,  $F$  și  $R$ ), o unitate de ieșire și o unitate țintă, care specifică doza minimă efectivă  $\log(I/C)$ . În majoritatea aplicațiilor s-a dovedit suficient un singur strat intermediar al cărui număr de unități s-a stabilit experimental. În aceste condiții, a fost utilizată o variantă a regulii de instruire cu propagare inversă a erorii. Rezultatele

obținute cu rețele neuronale sunt superioare, deși nu în mod dramatic, celor furnizate de tehnica clasică MLRA ("*Multi-Linear Regression Analysis*").

### **Caracteristicile generale ale structurii secundare a proteinelor**

Determinarea structurii macromoleculelor biologice este o problemă fundamentală în biologia moleculară și constituie un prim pas în elucidarea proprietăților funcționale complexe care le caracterizează. Forma elementară a problemei privește prognoza structurii secundare a proteinelor, adică conformația spațială la nivel inferior (de exemplu  $\alpha$ -elicii,  $\beta$ -plane și spirale aleatorii) pe baza structurii lor primare [24]. În această problemă, legea care determină structura secundară a lanțului proteic în funcție de proprietățile colective ale aminoacizilor individuali este principala necunoscută. Modurile de abordare a problemei se încadrează principial în două categorii: *tehnicile clasice* și *metode specifice inteligenței artificiale*.

Metodele clasice bazate pe reguli *ad hoc* sau *euristice* sunt în general specifice pe domeniu de aplicabilitate și au în consecință un spectru îngust. Tehnicile clasice dezvoltate pentru studiul *proteinelor globulare* conduc la rezultate slabe dacă se aplică *proteinelor membranare*, și reciproc, metodele concepute pentru proteinele membranare prognozează imprecis eliciile cu hidrofobicitate scăzută în proteinele globulare. Mai mult, limitările sunt generate și de faptul că o creștere în precizia unei metode anumite îngustează gama structurilor secundare asupra cărora pot fi aplicate cu un nivel de încredere rezonabil.

Abordarea neuronală a modelării structurii secundare din perspectiva tehnicilor aparținând domeniului inteligenței artificiale este superioară deoarece rețelele neuronale sunt caracterizate de o procesare holistică a datelor de intrare și, mai mult, nu necesită reguli de instruire *ad hoc* sau informații extrem de *specifice*, ca o consecință directă a procesului de instruire prin exemple și a puterii de generalizare a rețelei.

Rețelele neuronale recomandate în acest gen de aplicații sunt de tip perceptron multistratificat, iar algoritmul de instruire de tipul cu propagare inversă a erorii. Configurația unei secvențe primare de reziduuri de lungime  $l$  este aplicată la intrarea perceptronului, care urmează să asocieze aminoacidului din mijlocul configurației ( $l$  este presupus impar) structura sa secundară. Dimensiunea  $l$  a ferestrei este astfel aleasă încât reziduurile configurațiilor de intrare să cuprindă cele mai multe din interacțiile care prezumtiv ar stabiliza structura. Experimental, s-a determinat că dimensiuni ale ferestrei mai mari de 5 reziduuri nu aduc în mod necesar o îmbunătățire sensibilă, dar s-a preferat în aplicații valoarea 7. Studiul statistic al proteinelor a relevat că lungimea unei configurații afectează moderat performanța metodei. A fost utilizat un cod expandat, în care grupuri de câte 20 neuroni au fost folosiți la codificarea fiecărui aminoacid din cei 20. Dintre aceștia, 19 aminoacizi sunt inactivi, iar poziția celui de-al

20-lea neuron activ în interiorul grupului identifică reziduurile considerate. Astfel se evită corelațiile nedorite între diverse reziduuri care sunt simultan prezente la intrarea perceptronului. În consecință numărul neuronilor de intrare este 7-1. Fiecare neuron de ieșire reprezintă una din posibilele structuri a fi prognozate și ieșirea rețelei este făcută să corespundă cu neuronul având activitate maximă. La instruire au fost necesare cel puțin 100 de cicluri complete de parcurgere a setului de configurații de referință [25].

Concluziile principale care se pot desprinde din modelările neuronale aplicate în prognoza caracteristicilor generale ale structurii secundare a proteinelor sunt următoarele. Mai întâi, rețelele neuronale testate pe clase *omogene* de proteine (setul de instruire și cel de testare sunt alese din aceeași clasă), fie globulare, fie membranare, manifestă o putere de generalizare superioară metodelor clasice de prognoză. Mai mult, rețelele de tip perceptron sunt capabile de a identifica și alte structuri secundare decât elicii transmembranare (de exemplu elicii extramembranare sau  $\beta$ -spirale). În al doilea rând, experiențele efectuate pe clase de proteine presupuse *eterogene* (setul de instruire și cel de testare sunt alese din clase diferite) au evidențiat abilitatea perceptronilor de a învăța și recunoaște configurații și proprietăți statistice cu o eficiență comparabilă cu cea obținută pe clase omogene. În al treilea rând, superioritatea sensibilă a performanței sistemelor neuronale de tip perceptron în raport cu metodele clasice aplicate proteinelor membranare, sugerează că metodele de abordare cu rețele neuronale sunt mai eficiente în identificarea regularităților statistice conținute în setul de instruire, precum și mai eficiente decât metodele statistice în extrapolarea datelor structurale de la proteinele globulare la proteinele membranare [21, 26].

### **Corelația spectru-structură**

Elucidarea structurii compușilor organici se bazează masiv pe metode spectroscopice, deși relațiile dintre structură și datele spectrale sunt în general prea complexe pentru a fi exprimate sub forma unor ecuații explicite. Se cunosc numeroase tehnici de abordare a interpretării automate a spectrelor infraroșii printre care: căutarea și fitarea cu spectrele aparținând unei baze de date, utilizarea tabelelor de corelații, diferite metode statistice, teoria mulțimilor, precum și diverse metode de recunoaștere de forme ("*pattern recognition*") și sisteme expert ("*expert systems*") [27]. Corelațiile precise dintre spectrele complete de înaltă rezoluție și structurile chimice asociate, precum și disponibilitatea unor largi baze de date computerizate, fac extrem de atractiv acest domeniu pentru aplicații de interes a rețelelor neuronale.

Sistemele neuronale formale constituie o clasă de metode elegante de clasificare a configurațiilor în care un vector de intrare, ale cărui valori reprezintă obiectul de intrare, este transformat într-un vector de ieșire, ale

căruia valori reprezintă clasificarea dorită a obiectului de intrare. Transformarea este realizată prin intermediul unei formule de propagare, care combină vectorul de intrare cu coeficienții interni ai rețelei. Obiectele de intrare, care constituie subiectul transformării, sunt spectrele infraroșii ale unor compuși diverși, în timp ce ieșirea furnizează serii de substructuri care sunt conținute (sau nu) de compusul reprezentat la intrare. Procesul include *interpretarea spectrală*, adică exprimarea proprietăților spectroscopice ca fragmente substructurale, și *generarea structurii*, adică reuniunea fragmentelor substructurale alcătuind toate structurile moleculare complete compatibile cu ele.

O tehnică convenabilă de a opera cu date complexe multivariate este *clasificarea*, care sortează obiecte în clase de echivalență sau ierarhizează subclase în clase. Uzual, stabilirea unei clasificări ierarhice în decizii complexe începe cu un studiu statistic prealabil al relațiilor dintre reprezentarea obiectelor și clasele lor de apartenență. Clasificările ierarhice implică mai multe nivele decizionale și, mai mult, sunt aplicatii în care obiectul de intrare este caracterizat imprecis sau vag ("*fuzzy*"). Modelele clasice necesită funcții analitice predefinite pentru fiecare răspuns, în timp ce rețelele neuronale prezintă avantajul că pot modela vectori multirăspuns fără cunoștințe apriorice și procesa atât funcții de apartenență boolene ("*crisp membership functions*"), cât și funcții de apartenență vagi ("*fuzzy membership functions*"). Analiza corelației spectru-structură constituie o problemă tipică de clasificare multidecizională: obiectul de intrare este configurația unui spectru, iar răspunsul este o listă de clase prezente (sau absente) în compusul care a produs spectrul de la intrare. Atât instruirea supervizată, cât și cea nesupervizată, pot fi utilizate în proiectarea rețelelor neuronale destinate clasificării; în primul caz se recomandă o arhitectură cu un singur strat intermediar instruită cu regula de propagare inversă a erorii, în cel de-al doilea caz se face invariabil apel la rețele de tip Kohonen.

Munk, Madison și Rob [28] sugerează diviziunea spectrului infraroșu de la 3960 la 400  $\text{cm}^{-1}$  în 640 intervale de 5,6  $\text{cm}^{-1}$  lățime. Intensitatea  $x_i$  a radiației transmise de intervalul  $i$  este scalată la:

$$x_i = 1.00 - (\%T)/100.0 \quad (10.46)$$

unde  $T$  este factorul de transmisie. Numărul mare de unități de intrare a fost redus experimental la 256, pentru a elimina răspunsurile neconcludente. O altă îmbunătățire a metodei constă în îngustarea lățimii intervalelor la joasă frecvență și lărgirea lor la înaltă frecvență, cu scopul de a le adapta la puterea de discriminare variabilă de la un capăt al altuia al spectrului. În consecință, dependența de frecvență a lungimii intervalului  $i$  este dată de următoarea expresie:

$$i = 1; (6.0\sqrt{v} - 120.0) \quad (10.47)$$



unde  $E(\cdot)$  este funcția parte întreagă, iar  $v$  este frecvența. Domeniul de frecvență asignat intrării  $i$  este apoi explorat și sunt înregistrate vârfurile în amplitudine. Dacă un maxim este prezent într-un interval, atunci amplitudinea sa este scalată între 0.000 și 1.000, în caz contrar valoarea intrării este setată la zero.

Experimental, compușii au fost descriși în termenii a 36 grupări funcționale reprezentative (alcooli primari, fenoli, amine, aminoacizi, esteri,...), care au fost alese din 140 grupări funcționale studiate anterior [27]. Fiecărei grupări  $i$  s-a repartizat o unitate de ieșire arbitrară, astfel încât răspunsul țintă devine un vector 36-variat cu componente binare în care valoarea 1 indică prezența grupării funcționale asociate iar 0 indică absența sa. Pe scurt, esența acestei probleme 256-variate cu 36-răspunsuri este de a stabili prezența sau absența a oricărei grupări funcționale din cele 36 considerate relevante pe baza unor spectre diferite alcătuite din 256 puncte. Majoritatea compușilor analizați contin mai multe grupări funcționale, astfel încât multe unități de ieșire sunt simultan active. Prin încercări s-a ajuns la arhitectura optimă a rețelei:  $(256 \times 34 \times 36)$ . Biblioteca de spectre a cuprins 2915 compuși, dintre care au fost aleși aleatoriu 2499 pentru setul de instruire și restul de 416 pentru setul de test. A fost respectată astfel o regulă empirică generală referitoare la instruirea rețelelor mari utilizând algoritmul cu propagarea inversă a erorii, și anume ca numărul datelor alese pentru instruire să depășească numărul ponderilor sinaptice modificabile ale rețelei. În cazul studiat, setul de instruire conține aproximativ 650.000 valori (2.499 spectre  $\times$  256 intensități), în timp ce numărul coeficienților modificabili este puțin sub 350.000 ( $256 \times 34 \times 36$ ).

Rețelele neuronale s-au dovedit, deși pe un număr restrâns de experimente publicate până în prezent, a fi extrem de promițătoare în competiția cu alte sisteme expert dedicate elucidării structurale a compușilor chimici prin prelucrare automată a datelor. Mai mult, rețelele neuronale pot fi combinate cu sisteme expert și implementa o combinație de metode care, adesea, se dovedește mai tare decât fiecare în parte. Domeniul de explorare a sistemului expert se limitează la compușii pentru care răspunsurile rețelei nu sunt concludente [29, 30].

### ***Sisteme automate de control al proceselor și detecție a defectelor***

Două clase de probleme caracterizează desfășurarea proceselor chimice. Prima clasă privește recunoșterea și detecția defectelor de o manieră calitativă, atunci când este necesar să se identifice numai disfuncționalitățile sistemului și/sau părțile sale defecte. A doua clasă constă în controlul efectiv al procesului însuși, din moment ce valorile variabilelor de proces trebuie monitorizate și menținute în limitele prescrise de tehnologic aplicând reacție negativă ("*feed-back techniques*").

Controlul fiabil al unui proces dat se realizează dacă sunt satisfăcute două cerințe fundamentale. Prima privește concepția unui model  $M$  al procesului în măsură să prezică parametrii critici ai procesului pe durata unei secvențe de câteva unități de timp înainte; a doua cerință impune inversabilitatea modelului  $M$ , adică determinarea modelului  $M^{-1}$ , care să permită ajustarea variabilelor de corecție pe baza prognozei datelor furnizate de modelul direct  $M$ , astfel încât sistemul (procesul) să revină la normal dacă ar fi aplicate aceste ajustări. Atât algoritmul cu propagare inversă, cât și cel cu contra-propagare a erorii se dovedesc comparabile ca rezultate. Totuși, algoritmul cu contra-propagare a erorii poate fi privit ca un procedeu de netezire neliniară ("*nonlinear smoothing*"), astfel încât solicită o cantitate de date de intrare mai mare pentru o acuratețe dată, dar timpul de instruire total nu crește rapid cu dimensiunea setului de intrare. Deoarece numărul ciclurilor complete de antrenament este similar cu cel solicitat de instruirea rețelelor Kohonen, convergența în rețelele cu contra-propagare se atinge cu ordine de mărime mai rapid decât în rețelele cu propagare inversă a erorii.

Implementarea sistemelor de control al proceselor și detecția defectelor se bazează masiv pe modelarea matematică a proceselor, fapt care atrage dificultăți suplimentare:

- lipsa unor modele sigure pentru un proces dat adecvat tuturor circumstanțelor care pot apare,

- relațiile neliniare complexe dintre variabilele de control și desfășurarea procesului face extrem de dificilă, dacă nu imposibilă dezvoltarea de modele analitice,

- datele de control reactiv zgomotoase, imprecise sau incomplete, afectează negativ eficiența și siguranța monitORIZĂRII datelor,

- lipsa unei relații generale de izomorfism între setul de simptome observabile și diagnoza corectă.

Mai multe metode de abordare au fost propuse în cadrul teoriilor de control automat al proceselor:

- analiză statistică,

- modele liniare care utilizează tehnici clasice,

- modele de recunoaștere de configurații ("*pattern recognition*"),

- sisteme expert,

- rețele neuronale echipate cu reguli de instruire cu propagare inversă a erorii.

Hoskins and Himmelblau [31] au tratat exhaustiv diversele aplicații ale rețelelor neuronale în ingineria chimică, cum ar fi detecția defectelor ("*fault detection*"), diagnoza ("*diagnosis*"), controlul proceselor ("*process control*"), proiectarea proceselor ("*process design*") și simularea proceselor ("*process simulation*"). Deși este oarecum dificil a separa net

aceste tipuri de aplicatii, totuși rețelele neuronale prezintă global o serie de avantaje semnificative asupra tehnicilor clasice:

-nu sunt necesare modele analitice deoarece rețelele neuronale învață din exemple;

-modelele neliniare sunt agreate datorită răspunsurilor inerent neliniare ale rețelelor neuronale;

-funcțiile de asociere și clasificare caracteristice rețelelor neuronale le recomandă pentru procesarea datelor vagi ("*fuzzy*"), în mod special a celor aflate în afara regiunilor de instruire,

-date de intrare de orice natură pot fi procesate simultan (continue, discrete și vagi),

-rețelele neuronale pot modela funcții inverse, o caracteristică esențială în generarea semnalelor de control reactiv.

Un proces chimic este monitorizat prin măsurarea continuă a unui număr de variabile,  $x_i$ , în diferite puncte ale procesului și supervizat de semnale de reacție care acționează asupra unor elemente de control. Sarcina *elementelor de control* este de a menține variabilele măsurate într-o stare dată (*activ/inactiv*), fie la o valoare fixă ("*set-point*"), fie în interiorul unui interval definit ("*recommended range*"). Ansamblul tuturor variabilelor  $x_i$  măsurate și al stărilor la un moment dat  $t$ , este denumit *vectorul de proces*  $P_t$ :

$$P_t = (x_{1t}, x_{2t}, \dots, x_{1t}, \dots, x_{mt}) \quad (10.48)$$

Vectorii de instruire,  $X$ , formează mulțimi de variabile măsurate care sunt cunoscute că asigură funcționarea normală a procesului, iar vectorii țintă,  $Y$ , specifică stările anormale la care poate conduce intrarea  $X$ . În general, cauza unei disfuncționalități este o combinație de stări anormale care adesea sunt *binare* sau *discrete*. Modul cel mai elegant de a trata *stările discrete multiple* ale unei variabile este de a transforma variabila într-o *reprezentare distribuită*, adică fiecare variabilă multistare este transformată în tot atâtea intrări binare câte stări are variabila.

Modul de variație a variabilei de proces  $x_i$  este prognozat dacă  $X$  conține cel puțin două valori consecutive în timp (de exemplu  $x_{it}$  și  $x_{it+1}$ ). Așadar *orizontul trecut* al vectorului de instruire  $X$  trebuie să aibă o lungime minimă de două evenimente în tehnica ferestrei ("*moving window*"), în timp ce următoarea valoare a variabilei,  $x_{it+2}$ , trebuie să fie conținută în *orizontul viitor*,  $Y$  (vectorul țintă). Dacă previziunile se fac pentru mai multe unități temporale înainte,  $Y$  poate avea o dimensiune mai mare de o unitate de timp. Se recomandă ca orizontul trecut să fie cel puțin de două ori mai mare decât cel viitor.

Toate aplicațiile legate de controlul automat al proceselor necesită instruire supervizată bazată pe regula propagării inverse a erorii.

Dimensiunea rețelei este în primul rând impusă de numărul variabilelor de intrare și ieșire. O simplificare poate fi introdusă, mai ales în sistemele de control cantitativ, în sensul că nu este necesar ca toți vectorii  $P_i$  din setul de instruire să conțină toate variabilele de proces, ci numai pe acelea prin care se realizează efectiv controlul procesului.

### N.5.5. Concluzii finale privind aplicațiile rețelelor neuronale în chimie

Aplicațiile rețelelor neuronale în științele naturale, și în mod special în chimie, oferă noi metode de cercetare care pot îmbunătăți sau releva noi perspective în modelarea, prelucrarea și interpretarea datelor experimentale. Creșterea continuă a publicațiilor anuale dedicate cercetărilor în domeniul rețelelor neuronale, ca și a diverselor aplicații raportate, dovedeste interesul crescând al comunității științifice pentru sistemele artificiale de inspirație neuronală.

Particularitățile care fac rețelele neuronale atractive sunt multiple și cu ponderi variabile în aplicații specifice. Se impune însă reprezentarea proprie a informației în interiorul rețelei multistratificate, fără echivalent conceptual direct, dar care crește considerabil eficiența rețelei ca rezolvant extrem de general al problemelor de clasificare și clustering. În aceste condiții, suntem forțați să reconsiderăm reprezentarea și interpretarea datelor experimentale pentru a extrage maximul de informație utilă.

Abilitatea de a procesa date indiferent de caracterul lor (discrete, continui sau vagi), fără specificarea unor corelații prealabile, recomandă rețelele neuronale în aplicații complexe, indiferent de natura datelor. Mai mult, acolo unde nu există modele analitice liniare, rețelele neuronale se dovedesc eficiente datorită caracterului neliniar al răspunsurilor furnizate.

Abordarea datelor de intrare la nivelul întregii rețele le recomandă în aplicații cu date multivariate corelate, adică exact unde modelarea clasică este mai greoaie și mai puțin transparentă.

Una din tendințele actuale de a rezolva problemele hipercomplexe este splitarea lor în subprobleme mai simple reunind în final soluțiile parțiale. Rețelele neuronale sunt capabile de a realiza clasificarea și asocierea pe un singur nivel și, mai mult, procesează seturi mari de date eficient prin clasificare ierarhizată, utilizând rețele-poartă.

Condiția primordială a succesului în aplicații cu sisteme artificiale neuronale rămâne de departe reprezentarea informației în context. Sensibilitatea rețelei la datele de intrare conținute în setul de instruire este deosebită și a condus pe mulți adepți ai tehnicilor de rezolvare a problemelor cu rețele neuronale la o "regulă de aur" formulată extrem de simplu și sugestiv: "*garbage in - garbage out*"!

## **Bibliografie**

1. R. Mutihac and L. Mutihac, Artificial Neural Networks Applied in Chemistry, Part I: Foundations of Neural Networks, *Roum. Chem. Quart. Rev.*, **4**, 1995 (in press).
2. P. Smolensky, Neural and Conceptual Interpretation of PDP Models in **Parallel Distributed Processing, Exploration in the Microstructure of Cognition** (D. Rumelhart and J McClelland, eds.), The MIT Press, 1986.
3. F. Rosenblatt, **Principles of Neurodynamics**, Spartan Books, 1962.
4. R. Mutihac and V. Gheorghe, *Rev. Roum. Phys.*, **37**, No. 5, 535, 1992.
5. R. Mutihac, *Romanian Reports in Physics*, **47**, No. 1, 76, 1995.
6. J.J. Hopfield, *Proc. Nat. Acad. Sci.*, USA, **79**, 2554, 1982.
7. J.C. Bezdek, **Pattern Recognition with Fuzzy Objective Function Algorithms**, Plenum Press, New-York, 1981.
8. R. Mutihac, *St. Cerc. Fiz.*, **43**, Nos. 9-10, 603, 1991.
9. R. Mutihac, *Proc. 13th World Congress on Computation and Applied Mathematics*, Dublin, Ireland, 1009, 1991.
10. R. Mutihac, *St. Cerc. Fiz.*, **44**, No. 5, 361, 1992.
11. M. Minsky and S. Papert, *Perceptron: An Introduction to Computation Geometry*", Cambridge MA, The MIT Press, 1969.
12. W.S. McCulloch and W. Pitts, *Bull. Math. Biophys.*, **5**, 115, 1943.
13. M.A. Mahowald and C. Mead, *Scientific American*, **5**, 76, 1991.
14. D. Rumelhart and J. McClelland, PDP Models and General Issues in Cognitive Science, Chapter 4, in **Parallel Distributed Processing, Exploration in the Microstructure of Cognition** (D. Rumelhart and J McClelland, eds.), The MIT Press, 1986.
15. J.-P. Aubin, *SIAM J. on Control and Optimization*, **28**, 749, 1990.
16. J. Zupan, **Algorithms for Chemists**, John Wiley, Chichester, UK, 1989.
17. R. Mutihac, **Modelarea și simularea rețelelor neuronale**, Teza de doctorat, Universitatea București, 1994.
18. S.M. Deming and S.L. Morgan, **Experimental Design: A Chemometric Approach**, Elsevier, Amsterdam, NL, 1987.

19. K. Varmuza, **Pattern Recognition in Chemistry**, Springer-Verlag, Berlin, Germany, 1980.
20. J. Zupan and J. Gasteiger, **Neural Networks for Chemists**, VCH, 1991.
21. R. Mutilhac and L. Mutilhac, Artificial Neural Networks Applied in Chemistry, Part II: Applications in Chemistry, *Roum. Chem. Quart. Rev.*, **2**, 1996 (in press).
22. D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, and L. Kaufmann, **Chemometrics: A Textbook**, Elsevier, Amsterdam, NL, 1988.
23. V. Kvasnička, *Neur. Net. World*, **2**, 171, 1993.
23. R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton, *Neur. Comp.*, **3**, 79, 1991.
24. W. Kabsch and C. Sander, **Dictionary of Protein Secondary Structure: Pattern of Hydrogen-Bonded and Geometrical Features**, *Biopolymers*, **22**, 2577, 1983.
25. M. Compiani, P. Fariselli, and R. Casadio, Neural Networks Extracting General Features of Protein Secondary Structures, in **Third Italian Workshop on Parallel Architectures and Neural Networks** (E.R. Caianiello, ed.), World Scientific Publishing Co., 1990.
26. J. Garnier, D.J. Osguthorpe, and B. Robson, *J. Mol. Biol.*, **120**, 97, 1978.
27. E.W. Robb and M.E. Munk, *Mikrochim Acta*, **I**, 131, 1990.
28. M.E. Munk, M.S. Madison, and E.W. Robb, Neural Networks Models for Infrared Spectrum Interpretation, *Mikrochim Acta II*, 505, 1991.
29. W.J. Melssen, J.R.M. Smits, G.H. Rolf, and G. Kateman, *Chemom. Intell. Lab. Syst.*, 1993, in press.
30. M. Novic and J. Zupan, *Vestn. Slov. Kem. Drus.*, **39**, 195, 1992.
31. J.C. Hoskins and D.M. Himmelblau, *Comput. Chem. Eng.*, **12**, 881, 1988.

VERIFICAT  
2007

VERIFICAT  
2017

---



---

Tiparul s-a efectuat sub c-da nr. 242/1996, la  
Tipografia Editurii Universității București

---



---



# DATA RESTITUIRII

16 FEB. 2007

~~\_\_\_\_\_~~

**ISBN 973 – 575 – 085 – 6**

**Lei 4900**