

# Metoda celor mai mici pătrate

## Aproximări în medie pătratică

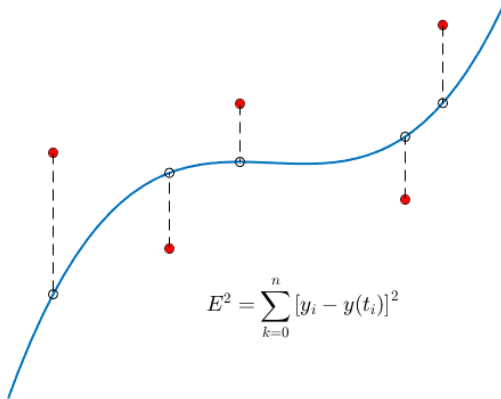
Radu T. Trîmbițaș

UBB

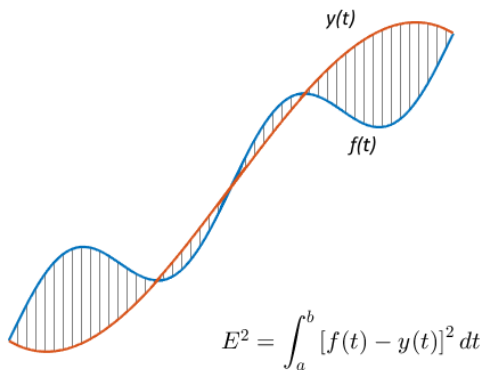
27 aprilie 2023

- Termenul *metoda celor mai mici pătrate* (MCMMP) (least squares method) sau *aproximare în medie pătratică* descrie o abordare utilizată frecvent la rezolvarea unor sisteme supradeterminate sau specificate inexact (în sens aproximativ). În loc să rezolvăm sistemul exact, vom încerca să minimizăm suma pătratelor reziduurilor.
- Interpretare statistică: dacă se fac ipoteze adecvate probabilistice asupra distribuției erorilor, MCMMP produce estimații de verosimilitate maximă ale parametrilor. Chiar dacă aceste ipoteze nu sunt satisfăcute, experiența a arătat că metoda produce rezultate utile.
- Algoritmii de rezolvare se bazează pe factorizări ortogonale ale matricelor.

# Aproximare MCMMP discretă



# Aproximare MCMMP continuă



# Modele și potrivirea curbelor I

- O sursă comună de probleme în sensul celor mai mici pătrate este *potrivirea curbelor* (*curve fitting*). Fie  $t$  variabila independentă și  $y(t)$  o funcție necunoscută de  $t$  pe care dorim să o aproximăm.
- Să presupunem că avem  $m$  observații ( $y_i$ ) măsurate în valorile specificate ( $t_i$ ):

$$y_i = y(t_i), \quad i = \overline{1, m}.$$

- *Modelul* nostru este o combinație de  $n$  funcții de bază ( $\pi_i$ ),  $m \gg n$

$$y(t) \approx c_1 \pi_1(t, \alpha) + \cdots + c_n \pi_n(t, \alpha).$$

- *Matricea de proiectare* (*design matrix*)  $X(\alpha)$  va fi matricea cu elementele

$$x_{i,j} = \pi_j(t_i, \alpha),$$

ale cărei elemente pot depinde de  $\alpha$ .

- În notație matricială, modelul se poate exprima ca:

$$y \approx X(\alpha)c.$$

- *Reziduurile* sunt diferențele dintre valorile observate și cele date de model

$$r_i = y_i - \sum_{j=1}^n c_j \pi_j(t_i, \alpha) \quad (1)$$

sau în notație matricială

$$r = y - X(\alpha)c. \quad (2)$$

# Modele și potrivirea curbelor III

- Ne propunem să minimizăm o anumită normă a reziduurilor. Cele mai frecvente alegeri sunt

$$\|r\|_2^2 = \sum_{i=1}^m r_i^2$$

sau

$$\|r\|_{2,w}^2 = \sum_{i=1}^m w_i r_i^2.$$

- O explicație intuitivă, fizică, a celei de-a doua alegeri ar fi aceea că anumite observații sunt mai importante decât altele și le vom asocia ponderi,  $w_i$ . De exemplu, dacă la observația  $i$  eroarea este aproximativ  $e_i$ , atunci putem alege  $w_i = 1/e_i$ . Deci, avem de a face cu o problemă discretă de aproximare în sensul celor mai mici pătrate. Problema este *liniară* dacă nu depinde de  $\alpha$  și *neliniară* în caz contrar.

# Modele și potrivirea curbelor IV

- Orice algoritm de rezolvare a unei probleme de aproximare în sensul celor mai mici pătrate fără ponderi poate fi utilizat la rezolvarea unei probleme cu ponderi prin scalarea observațiilor și a matricei de proiectare. În MATLAB aceasta se poate realiza prin  
 $X = \text{diag}(w) * X;$   
 $y = \text{diag}(w) * y$
- Dacă problema este liniară și avem mai multe observații decât funcții de bază, suntem conduși la rezolvarea sistemului supradeterminat

$$Xc \approx y,$$

pe care îl vom rezolva în sensul celor mai mici pătrate

$$c = X \backslash y.$$



- Dorim să rezolvăm

$$Xc \approx y \quad (3)$$

- Sistemul este supradeterminat (și în general incompatibil) — nu ne putem aștepta să îl rezolvăm exact. Îl vom rezolva în sensul celor mai mici pătrate:

$$\min_c \|y - Xc\|.$$

- Abordare teoretică: înmulțim ambii membri cu  $X^T$ . Aceasta reduce sistemul (3) la un sistem  $n \times n$ , pătratic, cunoscut sub numele de (sistem de) *ecuații normale*:

$$X^T Xc = X^T y. \quad (4)$$

## Ecuațiile normale II

- Matricea  $B = X^T X$  are elementele de forma

$$b_{ij} = (\pi_i(t), \pi_j(t)), \quad (5)$$

unde  $(\cdot, \cdot)$  este produsul scalar din  $\mathbb{R}^m$ .

- Sistemul (4) se scrie

$$\sum_{j=1}^n (\pi_i, \pi_j) c_j = (\pi_i, y), \quad i = 1, 2, \dots, n. \quad (6)$$

- Dacă există mii de observații și numai câțiva parametri, matricea de proiectare  $X$  este mare, dar  $X^T X$  este mică. **Am proiectat  $y$  pe subspațiul generat de coloanele lui  $X$ .** Dacă funcțiile de bază sunt linear independente, atunci  $X^T X$  este nesingulară și

$$c = (X^T X)^{-1} X^T y.$$

- Această formulă apare în majoritatea textelor de statistică și metode numerice.
- Caracteristici nedorite: ineficiența și mai ales proasta condiționare: ecuațiile normale sunt întotdeauna mai prost condiționate decât sistemul inițial

$$\text{cond}(X^T X) = \text{cond}(X)^2.$$

- în aritmetica cu precizie finită, ecuațiile normale pot deveni singulare și  $(X^T X)^{-1}$  ar putea să nu existe, chiar dacă coloanele lui  $X$  sunt liniar independente.
- Exemplu

$$X = \begin{bmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{bmatrix}$$

- Dacă  $\delta$  este mic, dar nenul, coloanele lui  $X$  sunt aproape paralele, dar liniar independente. Ecuațiile normale fac situația și mai proastă:

$$X^T X = \begin{bmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{bmatrix}^T \begin{bmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{bmatrix} = \begin{bmatrix} \delta^2 + 1 & 1 \\ 1 & \delta^2 + 1 \end{bmatrix}$$

- Dacă  $|\delta| < 10^{-8}$ , matricea  $X^T X$  calculată în dublă precizie este chiar singulară și inversa nu există.

# Factorizare QR I

- O metodă de a evita proasta condiționare este factorizarea QR:  
 $X = QR$ , unde  $Q$  este ortogonală, iar  $R$  este triunghiulară superior.

$$c = R^{-1}Q^T y.$$

- Abordarea numerică este să aplicăm ortogonalizarea atât matricei  $X$ , cât și membrului drept  $y$ . Vom obține astfel un sistem triunghiular superior, care se rezolvă prin substituție inversă.
- Operatorul `\` din MATLAB știe să rezolve sisteme în sensul celor mai mici pătrate prin factorizare QR. Factorizarea QR se poate obține prin funcția MATLAB `qr`.
- Există două versiuni de factorizare QR.
  - În versiunea completă,  $R$  are aceeași dimensiune ca  $X$  iar  $Q$  este pătratică cu același număr de linii ca  $X$ .
  - În versiunea redusă,  $Q$  are aceeași mărime ca  $X$ , iar  $R$  este pătratică cu același număr de coloane ca  $X$ .

- Procesul Gram-Schmidt din algebra liniară generează aceeași factorizare, dar este mai puțin stabil numeric.

# Factorizare QR III

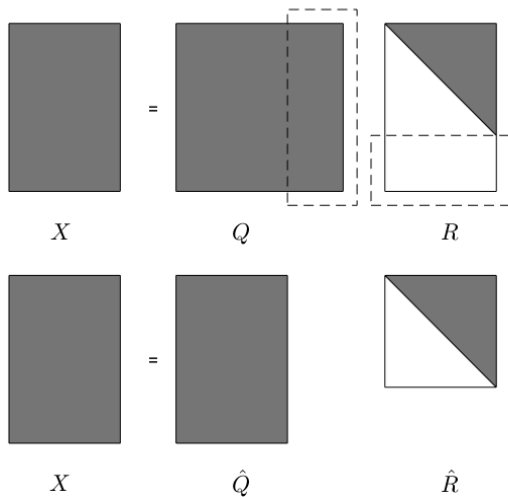


Figura: Factorizarea QR completă și redusă

- Scriem aproximanta sub forma

$$\varphi = \sum_{j=1}^n c_j \pi_j(t).$$

- Eroarea (reziduul) va fi  $y - \varphi$ , iar

$$E(\varphi)^2 = \|r\|^2 = \sum_{k=1}^m \left( y_k - \sum_{j=1}^n c_j \pi_j(t_k) \right)^2$$

- Pătratul erorii este o funcție quadratică de coeficienții  $c_1, \dots, c_n$  ai lui  $\varphi$ . Problema revine la a minimiza această funcție pătratică; ea se rezolvă anulând derivatele parțiale.



- Se obține

$$\frac{\partial}{\partial c_i} E(\varphi)^2 = 2 \sum_{k=1}^m \left( y_k - \sum_{j=1}^n c_j \pi_j(t_k) \right) \pi_i(t_k) = 0,$$

adică

$$\sum_{j=1}^n c_j \left( \sum_{k=1}^m \pi_i(t_k) \pi_j(t_k) \right) = \sum_{k=1}^m \pi_i(t_k) y_k.$$

- Cu ajutorul produsului scalar

$$\sum_{j=1}^n (\pi_i, \pi_j) c_j = (\pi_i, y), \quad i = 1, 2, \dots, n, \quad (7)$$

adică chiar ecuațiile normale (6).

# Ortogonalitate I

- Fie  $\hat{\varphi}_n$  aproximanta cu coeficienții  $(\hat{c}_k)$  soluții ale ecuațiilor normale. Observăm întâi că eroarea  $y - \hat{\varphi}_n$  este ortogonală pe  $\Phi_n = \langle \pi_1, \dots, \pi_n \rangle$ , adică

$$(y - \hat{\varphi}_n, \varphi) = 0, \quad \forall \varphi \in \Phi_n \quad (8)$$

- Deoarece  $\varphi$  este o combinație liniară de  $\pi_i$ , este suficient ca (8) să aibă loc pentru fiecare  $\varphi = \pi_i$ ,  $i = 1, 2, \dots, n$ .
- Se obține

$$(y - \hat{\varphi}_n, \pi_i) = \left( y - \sum_{j=1}^n \hat{c}_j \pi_j, \pi_i \right) = (y, \pi_i) - \sum_{j=1}^n \hat{c}_j (\pi_j, \pi_i) = 0,$$

sau

$$\sum_{j=1}^n (\pi_i, \pi_j) c_j = (\pi_i, y), \quad i = 1, 2, \dots, n, \quad (9)$$

adică chiar ecuațiile normale (6).

- Rezultatul din (8) are o interpretare geometrică simplă. Aproximanta în sensul celor mai mici pătrate  $\hat{\varphi}_n$  este proiecția ortogonală a lui  $y$  pe  $\Phi_n$ , vezi figura 2.

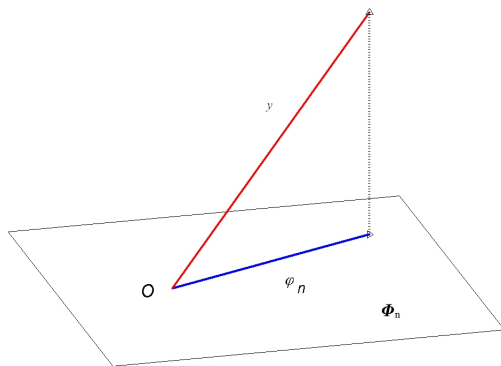


Figura: Interpretarea geometrică a aproximării prin MCMMP

# Exemplul 1 I

## Exemplu

*Dându-se punctele*

$$(0, -4), (1, 0), (2, 4), (3, -2),$$

*determinați polinomul de gradul 1 corespunzător acestor date prin metoda celor mai mici pătrate.*

**Soluție.** Aproximanta căutată are forma

$$\varphi(x) = c_0 + c_1x.$$

Sistemul de ecuații normale se determină din condițiile  $f - \varphi \perp 1$  și  $f - \varphi \perp x$ . Se obține

$$\begin{cases} c_0(1, 1) + c_1(x, 1) = (f, 1) \\ c_0(1, x) + c_1(x, x) = (f, x) \end{cases}$$

## Exemplul 1 II

Dar,  $(1, 1) = \sum_{i=1}^4 1 \cdot 1 = 4$ ,

$(1, x) = (x, 1) = \sum_{i=1}^4 1 \cdot x_i = 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 2 + 1 \cdot 3 = 6$ ,

$(x, x) = \sum_{i=1}^4 x_i^2 = 14$ . Pentru membrul drept avem  $(f, 1) = (y, 1) = -2$   
și  $(f, x) = (y, x) = 2$ . Am obținut sistemul

$$\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

cu soluția  $c_0 = -2$ ,  $c_1 = 1$ . Deci  $\varphi(x) = x - 2$ . ■

## Exemplul 2 I

### Exemplu

*Datele următoare dau populația SUA (în milioane) determinată la recensăminte de US Census, între anii 1900 și 2010. Dorim să modelăm populația și să o estimăm pentru anii 1975 și 2010.*

An	Populația	An	Populația
1900	75.995	1960	179.320
1910	91.972	1970	203.210
1920	105.710	1980	226.510
1930	123.200	1990	249.630
1940	131.670	2000	281.420
1950	150.700	2010	308.790

**Soluție.** Vom modela populația printr-un model polinomial de gradul 3

$$y = c_0 + c_1 t + c_2 t^2 + c_3 t^3,$$

## Exemplul 2 II

și printr-un model exponențial

$$y = Ke^{\lambda t}.$$

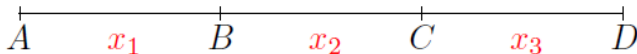




## Exemplul 3 I

### Exemplu (Măsurarea unui segment de drum - Stiefel)

*La măsurarea unui segment de drum, presupunem că am efectuat 5 măsurători*



$AD = 89m$ ,  $AC = 67m$ ,  $BD = 53m$ ,  $AB = 35m$  și  $CD = 20m$ ,  
și că dorim să determinăm lungimile segmentelor  $x_1 = AB$ ,  $x_2 = BC$  și  $x_3 = CD$ .

## Exemplul 3 II

**Soluție.** Conform observațiilor obținem un sistem cu mai multe ecuații decât necunoscute (sistem supradeterminat):

$$\begin{array}{rcl} x_1 + x_2 + x_3 & = & 89 \\ x_1 + x_2 & = & 67 \\ x_2 + x_3 & = & 53 \\ x_1 & = & 35 \\ x_3 & = & 20 \end{array} \Leftrightarrow Ax = b, \quad A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 89 \\ 67 \\ 53 \\ 35 \\ 20 \end{bmatrix}.$$

Din ultimele trei ecuații se obține soluția  $x_1 = 35$ ,  $x_2 = 33$  și  $x_3 = 20$ .  
Totuși, dacă înlocuim în primele două ecuații se obține

$$\begin{aligned} x_1 + x_2 + x_3 - 89 &= -1, \\ x_1 + x_2 - 67 &= 1. \end{aligned}$$

## Exemplul 3 III

Ecuatiile sunt contradictorii și se ajunge la un sistem incompatibil.  
Un remediu este să găsim o soluție aproximativă care satisface sistemul cât mai bine posibil. Considerăm vectorul reziduu

$$r = b - Ax.$$

Căutăm un vector  $x$  care să minimizeze într-un anumit sens vectorul reziduu.

$$x = \begin{bmatrix} 35.1250 & 32.5000 & 20.6250 \end{bmatrix}^T$$



# Aproximări continue I

- Dorim să aproximăm o funcție pe un interval  $[a, b]$ , mărginit sau nemărginit

$$f \approx \varphi = c_1\pi_1 + \cdots + c_n\pi_n$$

- Produsul scalar va fi

$$(u, v) = \int_a^b w(t)u(t)v(t)dt,$$

iar norma

$$\|u\| = \sqrt{\int_a^b w(t)u^2(t)dt}.$$

- Funcția  $w$  este o funcție nenegativă pe  $[a, b]$ , adică  $w(t) \geq 0$ ,  $\forall t \in [a, b]$  și neidentic nulă pe orice subinterval al lui  $[a, b]$ .

- Reziduul  $r = f - \varphi$  va avea norma minimă

$$\|r\|^2 = \int_a^b w(t) [f(t) - \varphi(t)]^2 dt \quad (10)$$

$$= \int_a^b w(t) \left[ f(t) - \sum_{j=0}^n c_j \pi_j(t) \right]^2 dt =: E^2[\varphi] \quad (11)$$

- Ecuațiile normale se pot obține minimizând pătratul normei cu ajutorul derivatelor parțiale, sau din condiția de ortogonalitate

$$\sum_{j=1}^n (\pi_i, \pi_j) c_j = (\pi_i, f), \quad i = 1, 2, \dots, n. \quad (12)$$

- Ele formează un sistem de forma

$$Ac = b \quad (13)$$

unde matricea  $A$  și vectorul  $b$  au elementele

$$A = [a_{ij}], \quad a_{ij} = (\pi_i, \pi_j), \quad b = [b_i], \quad b_i = (\pi_i, f). \quad (14)$$

# Existența și unicitatea I

- Datorită simetriei produsului scalar,  $A$  este o matrice simetrică. Mai mult,  $A$  este pozitiv definită, adică

$$x^T A x = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j > 0, \text{ dacă } x \neq [0, 0, \dots, 0]^T. \quad (15)$$

- Funcția (15) se numește **formă pătratică** (deoarece este omogenă de grad 2). Pozitiv definirea lui  $A$  ne spune că forma pătratică ai cărei coeficienți sunt elementele lui  $A$  este întotdeauna nenegativă și zero numai dacă variabilele  $x_i$  se anulează.
- Pentru a demonstra (15) să inserăm definiția lui  $a_{ij}$  și să utilizăm proprietățile produsului scalar

$$x^T A x = \sum_{i=1}^n \sum_{j=1}^n x_i x_j (\pi_i, \pi_j) = \sum_{i=1}^n \sum_{j=1}^n (x_i \pi_i, x_j \pi_j) = \left\| \sum_{i=1}^n x_i \pi_i \right\|^2.$$

- Aceasta este evident nenegativă. Ea este zero numai dacă  $\sum_{i=1}^n x_i \pi_i \equiv 0$  pe  $\text{supp} d\lambda$ , care pe baza liniar independenței lui  $\pi_i$  implică  $x_1 = x_2 = \dots = x_n = 0$ .
- Este un rezultat cunoscut din algebra liniară că o matrice  $A$  simetrică pozitiv definită este nesingulară. Într-adevăr, determinantul său, precum și minorii principali sunt strict pozitivi. Rezultă că sistemul de ecuații normale (7) are soluție unică.
- Corespunde această soluție minimului lui  $E^2[\varphi]$  în (10)? Matricea hessiană  $H = [\partial^2 E^2 / \partial c_i \partial c_j]$  trebuie să fie pozitiv definită. Dar  $H = 2A$ , deoarece  $E^2$  este o funcție quadratică. De aceea,  $H$ , ca și  $A$ , este într-adevăr pozitiv definită și soluția ecuațiilor normale ne dă minimul dorit.



- Problema de aproximare în sensul celor mai mici pătrate are o soluție unică, dată de

$$\hat{\varphi}(t) = \sum_{j=1}^n \hat{c}_j \pi_j(t) \quad (16)$$

unde  $\hat{c} = [\hat{c}_1, \hat{c}_2, \dots, \hat{c}_n]^T$  este vectorul soluție al ecuațiilor normale (7).

# Neajunsurile MCMMP I

- Ecuațiile normale rezolvă problema de aproximare în sensul celor mai mici pătrate complet în teorie. Dar în practică?
- Referitor la o mulțime generală de funcții de bază liniar independente, pot apărea următoarele dificultăți:
  - ① Sistemul de ecuații normale (7) poate fi **prost condiționat**. Un exemplu simplu este următorul:  $\text{supp } d\lambda = [0, 1]$ ,  $d\lambda(t) = dt$  pe  $[0, 1]$  și  $\pi_j(t) = t^{j-1}$ ,  $j = 1, 2, \dots, n$ . Atunci

$$(\pi_i, \pi_j) = \int_0^1 t^{i+j-2} dt = \frac{1}{i+j-1}, \quad i, j = 1, 2, \dots, n,$$

adică matricea  $A$  este matricea Hilbert. Proasta condiționare a ecuațiilor normale se datorează alegerii neinspirate a funcțiilor de bază. Acestea devin aproape liniar dependente când exponentul crește. O altă sursă de degradare provine din elementele membrului drept

$b_j = \int_0^1 \pi_j(t) f(t) dt$ . Când  $j$  este mare  $\pi_j(t) = t^{j-1}$  se comportă pe

$[0, 1]$  ca o funcție discontinuă. Un polinom care oscilează mai rapid pe  $[0, 1]$  ar fi de preferat, căci ar angaja mai viguros funcția  $f$ .

- ② Al doilea dezavantaj este faptul că toți coeficienții  $\hat{c}_j$  din (16) depind de  $n$ , adică  $\hat{c}_j = \hat{c}_j^{(n)}$ ,  $j = 1, 2, \dots, n$ . Mărirea lui  $n$  ne dă un nou sistem de ecuații mai mare și cu o soluție complet diferită. Acest fenomen se numește **nepermanența coeficienților**  $\hat{c}_j$ .

- Amândouă neajunsurile (1) și (2) pot fi eliminate (sau măcar atenuate) alegând ca funcții de bază un sistem ortogonal,

$$(\pi_i, \pi_j) = 0 \text{ dacă } i \neq j \quad (\pi_j, \pi_j) = \|\pi_j\|^2 > 0 \quad (17)$$

- Atunci sistemul de ecuații normale devine diagonal și poate fi rezolvat imediat cu formula

$$\hat{c}_j = \frac{(\pi_j, f)}{(\pi_j, \pi_j)}, \quad j = 1, 2, \dots, n. \quad (18)$$

Evident, acești coeficienți  $\hat{c}_j$  sunt independenți de  $n$  și odată calculați rămân la fel pentru orice  $n$  mai mare. Avem acum proprietatea de **permanență a coeficienților**. De asemenea nu trebuie să rezolvăm sistemul de ecuații normale, ci putem aplica direct (18).

- Orice sistem  $\{\hat{\pi}_j\}$  care este liniar independent pe  $\text{supp} d\lambda$  poate fi ortogonalizat (în raport cu măsura  $d\lambda$ ) prin **procedeul Gram-Schmidt**. Se ia

$$\pi_1 = \hat{\pi}_1$$

și apoi, pentru  $j = 2, 3, \dots$  se calculează recursiv

$$\pi_j = \hat{\pi}_j - \sum_{k=1}^{j-1} c_k \pi_k, \quad c_k = \frac{(\hat{\pi}_j, \pi_k)}{(\pi_k, \pi_k)}, \quad k = \overline{1, j-1}.$$

Atunci fiecare  $\pi_j$  astfel determinat este ortogonal pe toate funcțiile precedente.

# Exemple de sisteme ortogonale

- 1 Sistemul trigonometric – cunoscut din analiza Fourier.
- 2 Polinoame ortogonale

# Sistemul trigonometric I

- *Sistemul trigonometric* este format din funcțiile:

$$1, \cos t, \cos 2t, \cos 3t, \dots, \sin t, \sin 2t, \sin 3t, \dots$$

- El este ortogonal pe  $[0, 2\pi]$  în raport ponderea  $w(t) = 1$ .

$$\int_0^{2\pi} \sin kt \sin \ell t dt = \begin{cases} 0, & \text{pentru } k \neq \ell \\ \pi, & \text{pentru } k = \ell \end{cases} \quad k, \ell = 1, 2, 3, \dots$$

$$\int_0^{2\pi} \cos kt \cos \ell t dt = \begin{cases} 0, & k \neq \ell \\ 2\pi, & k = \ell = 0 \\ \pi, & k = \ell > 0 \end{cases} \quad k, \ell = 0, 1, 2$$

$$\int_0^{2\pi} \sin kt \cos \ell t dt = 0, \quad k = 1, 2, 3, \dots, \quad \ell = 0, 1, 2, \dots$$

# Sistemul trigonometric II

- Aproximarea are forma

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kt + b_k \sin kt). \quad (19)$$

- Utilizând (18) obținem

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(t) \cos ktdt, \quad k = 1, 2, \dots \\ b_k &= \frac{1}{\pi} \int_0^{2\pi} f(t) \sin ktdt, \quad k = 1, 2, \dots \end{aligned} \quad (20)$$

numiți **coeficienți Fourier** ai lui  $f$ . Ei sunt coeficienții (18) pentru sistemul trigonometric.

- Prin extensie coeficienții (18) pentru orice sistem ortogonal  $(\pi_j)$  se vor numi **coeficienții Fourier** ai lui  $f$  relativ la acest sistem.



- În particular, recunoaștem în seria Fourier trunchiată pentru  $k = n$  aproximarea lui  $f$  în clasa polinoamelor trigonometrice de grad  $\leq n$  relativ la norma

$$\|u\|_2 = \left( \int_0^{2\pi} |u(t)|^2 dt \right)^{1/2}$$

# Polinoame ortogonale I

- Dându-se o măsură  $d\lambda$ , știm că orice număr finit de puteri  $1, t, t^2, \dots$  sunt liniar independente pe  $[a, b]$ , dacă  $\text{supp} d\lambda = [a, b]$ , iar  $1, t, \dots, t^{N-1}$  liniar independente pe  $\text{supp} d\lambda = \{t_1, t_2, \dots, t_N\}$ .
- Deoarece o mulțime de vectori liniar independenți a unui spațiu liniar poate fi ortogonalizată prin procedeul Gram-Schmidt, orice măsură  $d\lambda$  de tipul considerat generează o mulțime unică de polinoame ortogonale monice  $\pi_j(t, d\lambda)$ ,  $j = 0, 1, 2, \dots$  ce satisfac

$$\text{grad} \pi_j = j, \quad j = 0, 1, 2, \dots$$

$$\int_{\mathbb{R}} \pi_k(t) \pi_\ell(t) d\lambda(t) = 0, \text{ dacă } k \neq \ell \quad (21)$$

- Aceste polinoame se numesc **polinoame ortogonale** relativ la măsura  $d\lambda$ .

## Polinoame ortogonale II

- Vom permite indicilor să meargă de la 0. Mulțimea  $\pi_j$  este infinită dacă  $\text{supp}d\lambda = [a, b]$  și constă din exact  $N$  polinoame  $\pi_0, \pi_1, \dots, \pi_{N-1}$  dacă  $\text{supp}d\lambda = \{t_1, \dots, t_N\}$ . În ultimul caz polinoamele se numesc **polinoame ortogonale discrete**.
- Între trei polinoame ortogonale monice (un polinom se numește **monic** dacă coeficientul său dominant este 1) consecutive există o relație liniară. Mai exact, există constantele reale  $\alpha_k = \alpha_k(d\lambda)$  și  $\beta_k = \beta_k(d\lambda) > 0$  (depinzând de măsura  $d\lambda$ ) astfel încât

$$\pi_{k+1}(t) = (t - \alpha_k)\pi_k(t) - \beta_k\pi_{k-1}(t), \quad k = 0, 1, 2, \dots \quad (22)$$

$$\pi_{-1}(t) = 0, \quad \pi_0(t) = 1.$$

(Se subînțelege că (22) are loc pentru orice  $k \in \mathbb{N}$  dacă  $\text{supp}d\lambda = [a, b]$  și numai pentru  $k = \overline{0, N-2}$  dacă  $\text{supp}d\lambda = \{t_1, t_2, \dots, t_N\}$ ).

# Polinoame ortogonale III

- Pentru a demonstra (22) și a obține expresiile coeficienților să observăm că  $\pi_{k+1}(t) - t\pi_k(t)$  este un polinom de grad  $\leq k$ , și deci poate fi exprimat ca o combinație liniară a lui  $\pi_0, \pi_1, \dots, \pi_k$ . Scriem această combinație sub forma

$$\pi_{k+1} - t\pi_k(t) = -\alpha_k\pi_k(t) - \beta_k\pi_{k-1}(t) + \sum_{j=0}^{k-2} \gamma_{k,j}\pi_j(t) \quad (23)$$

(sumele vide se consideră nule).

- Înmulțim scalar ambii membri ai relației anterioare cu  $\pi_k$  și obținem

$$(-t\pi_k, \pi_k) = -\alpha_k(\pi_k, \pi_k)$$

adică

$$\alpha_k = \frac{(t\pi_k, \pi_k)}{(\pi_k, \pi_k)}, \quad k = 0, 1, 2, \dots \quad (24)$$

- La fel, înmulțind scalar cu  $\pi_{k-1}$  obținem

$$(-t\pi_k, \pi_{k-1}) = -\beta_k(\pi_{k-1}, \pi_{k-1}).$$

Deoarece  $(t\pi_k, \pi_{k-1}) = (\pi_k, t\pi_{k-1})$  și  $t\pi_{k-1}$  diferă de  $\pi_k$  printr-un polinom de grad  $< k$  se obține prin ortogonalitate

$(t\pi_k, \pi_{k-1}) = (\pi_k, \pi_k)$ , deci

$$\beta_k = \frac{(\pi_k, \pi_k)}{(\pi_{k-1}, \pi_{k-1})}, \quad k = 1, 2, \dots \quad (25)$$

- Înmulțind (23) cu  $\pi_\ell$ ,  $\ell < k-1$ , se obține

$$\gamma_{k,\ell} = 0, \quad \ell = 0, 1, \dots, k-1 \quad (26)$$

# Polinoame ortogonale V

- Formula de recurență (22) ne dă o modalitate practică de determinare a polinoamelor ortogonale. Deoarece  $\pi_0 = 1$ , putem calcula  $\alpha_0$  cu (24) pentru  $k = 0$  și apoi  $\pi_1$ , etc. Procedeu – numit **procedura lui Stieltjes** – este foarte potrivit pentru polinoame ortogonale discrete, căci în acest caz produsul scalar se exprimă prin sume finite.
- În cazul continuu, calculul produsului scalar necesită calcul de integrale, ceea ce complică lucrurile. Din fericire, pentru anumite măsuri speciale importante, coeficienții se cunosc explicit.
- Cazul special când măsura este simetrică (adică  $d\lambda(t) = w(t)$  cu  $w(-t) = w(t)$  și  $\text{supp} d\lambda$  simetrică față de origine) merită o atenție specială, deoarece în acest caz  $\alpha_k = 0$ ,  $\forall k \in \mathbb{N}$ , conform lui (19) căci

$$(t\pi_k, \pi_k) = \int_{\mathbb{R}} w(t)t\pi_k^2(t)dt = \int_a^b w(t)t\pi_k^2(t)dt = 0,$$

deoarece avem o integrală dintr-o funcție impară pe un domeniu simetric.

- Rădăcinile polinoamelor ortogonale sunt reale, distincte și situate în interiorul intervalului  $[a, b]$ .

Deoarece  $\int_{\mathbb{R}} \pi_n(t) d\lambda(t) = 0$ , trebuie să existe cel puțin un punct interior lui  $[a, b]$  pe care  $\pi_n(t)$  își schimbă semnul. Fie  $t_1, t_2, \dots, t_k, k \leq n$  toate aceste puncte. Dacă  $k < n$ , din ortogonalitate

$$\int_{\mathbb{R}} \pi_n(t) \prod_{j=1}^k (t - t_j) d\lambda(t) = 0,$$

imposibil, deoarece integrandul are semn constant. Deci,  $k = n$ .



Figura: Thomas Ioannes Stieltjes (1856-1894)



# Polinoamele lui Legendre I

- Se definesc prin așa-numita formulă a lui Rodrigues

$$\pi_k(t) = \frac{k!}{(2k)!} \frac{d^k}{dt^k} (t^2 - 1)^k. \quad (27)$$

- Exemple:

$$\pi_0(t) = 1,$$

$$\pi_1(t) = t,$$

$$\pi_2(t) = t^2 - \frac{1}{3},$$

$$\pi_3(t) = t^3 - \frac{3}{5}t.$$

- Verificăm întâi ortogonalitatea pe  $[-1, 1]$  în raport cu ponderea  $w(t) = 1$ .

- Pentru orice  $0 \leq \ell < k$ , prin integrare repetată prin părți se obține:

$$\begin{aligned} & \int_{-1}^1 \frac{d^k}{dt^k} (t^2 - 1)^k t^\ell dt \\ &= \sum_{m=0}^{\ell} \ell(\ell-1) \dots (\ell-m+1) t^{\ell-m} \frac{d^{k-m-1}}{dt^{k-m-1}} (t^2 - 1)^k \Big|_{-1}^1 = 0, \end{aligned}$$

ultima relație având loc deoarece  $0 \leq k-m-1 < k$ .

- Deci,

$$(\pi_k, p) = 0, \quad \forall p \in \mathbb{P}_{k-1},$$

demonstrându-se astfel ortogonalitatea.

- **Relația de recurență**
- Datorită simetriei, putem scrie

$$\pi_k(t) = t^k + \mu_k t^{k-2} + \dots, \quad k \geq 2$$

și observând (din nou datorită simetriei) că relația de recurență are forma

$$\pi_{k+1}(t) = t\pi_k(t) - \beta_k\pi_{k-1}(t),$$

obținem

$$\beta_k = \frac{t\pi_k(t) - \pi_{k+1}(t)}{\pi_{k-1}(t)},$$

care este valabilă pentru orice  $t$ .

# Polinoamele lui Legendre IV

- Făcând  $t \rightarrow \infty$ ,

$$\beta_k = \lim_{t \rightarrow \infty} \frac{t\pi_k(t) - \pi_{k+1}(t)}{\pi_{k-1}(t)} = \lim_{t \rightarrow \infty} \frac{(\mu_k - \mu_{k+1})t^{k-1} + \dots}{t^{k-1} + \dots} = \mu_k - \mu_{k+1}.$$

(Dacă  $k = 1$ , punem  $\mu_1 = 0$ .)

- Din formula lui Rodrigues rezultă

$$\begin{aligned}\pi_k(t) &= \frac{k!}{(2k)!} \frac{d^k}{dt^k} \left( t^{2k} - kt^{2k-2} + \dots \right) \\ &= \frac{k!}{(2k)!} (2k(2k-1) \dots (k+1)t^k - \\ &\quad k(2k-2)(2k-3) \dots (k-1)t^{k-1} + \dots) \\ &= t^k - \frac{k(k-1)}{2(2k-1)} t^{k-2} + \dots,\end{aligned}$$

aşa că

$$\mu_k = \frac{k(k-1)}{2(2k-1)}, \quad k \geq 2.$$

Deci,

$$\beta_k = \mu_k - \mu_{k+1} = \frac{k^2}{(2k-1)(2k+1)}$$

şi deoarece  $\mu_1 = 0$ ,

$$\beta_k = \frac{1}{4 - k^{-2}}, \quad k \geq 1. \quad (28)$$

- Ele se pot defini prin relația

$$T_n(x) = \cos(n \arccos x), \quad n \in \mathbb{N}. \quad (29)$$

- Din identitatea trigonometrică

$$\cos(k+1)\theta + \cos(k-1)\theta = 2 \cos \theta \cos k\theta$$

și din (29), punând  $\theta = \arccos x$  se obține

$$\begin{aligned} T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x) \quad k = 1, 2, 3, \dots \\ T_0(x) &= 1, \quad T_1(x) = x. \end{aligned} \quad (30)$$

# Polinoamele Cebîșev de speța I II

- De exemplu,

$$T_2(x) = 2x^2 - 1,$$

$$T_3(x) = 4x^3 - 3x,$$

$$T_4(x) = 8x^4 - 8x^2 + 1$$

ș.a.m.d.

- Din relația (30) se obține pentru coeficientul dominant al lui  $T_n$  valoarea  $2^{n-1}$  (dacă  $n \geq 1$ ), deci polinomul Cebîșev de speța I monic este

$$\overset{\circ}{T}_n(x) = \frac{1}{2^{n-1}} T_n(x), \quad n \geq 0, \quad \overset{\circ}{T}_0 = T_0. \quad (31)$$

- Din (29) se pot obține rădăcinile lui  $T_n$

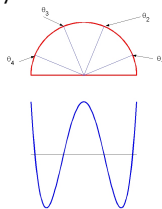
$$x_k^{(n)} = \cos \theta_k^{(n)}, \quad \theta_k^{(n)} = \frac{2k-1}{2n} \pi, \quad k = \overline{1, n}. \quad (32)$$

# Polinoamele Cebîșev de speța I III

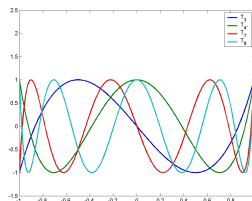
- Ele sunt proiecțiile pe axa reală ale punctelor de pe cercul unitate de argument  $\theta_k^{(n)}$ .
- Pe intervalul  $[-1, 1]$   $T_n$  oscilează de la  $+1$  la  $-1$ , atingând aceste valori extreme în punctele

$$y_k^{(n)} = \cos \eta_k^{(n)}, \quad \eta_k^{(n)} = \frac{k\pi}{n}, \quad k = \overline{0, n}.$$

$T_4$  și rădăcinile sale



$T_3, T_4, T_7, T_8$  pe  $[-1, 1]$





- Polinoamele Cebîșev de speța I sunt ortogonale pe  $[-1, 1]$  în raport cu ponderea

$$w(x) = \frac{1}{\sqrt{1-x^2}}.$$

- Se verifică ușor din (29) că

$$\begin{aligned} \int_{-1}^1 T_k(x) T_\ell(x) \frac{dx}{\sqrt{1-x^2}} &= \int_0^\pi T_k(\cos \theta) T_\ell(\cos \theta) d\theta \\ &= \int_0^\pi \cos k\theta \cos \ell\theta d\theta = \begin{cases} 0 & \text{dacă } k \neq \ell \\ \pi & \text{dacă } k = \ell = 0 \\ \pi/2 & \text{dacă } k = \ell \neq 0 \end{cases} \end{aligned} \quad (33)$$

# Polinoamele Cebîșev de speța I V

- Dezvoltarea în serie Fourier de polinoame Cebîșev este dată de

$$f(x) = \sum_{j=0}^{\infty} c_j T_j(x) = \frac{1}{2} c_0 + \sum_{j=1}^{\infty} c_j T_j(x), \quad (34)$$

unde

$$c_j = \frac{2}{\pi} \int_{-1}^1 f(x) T_j(x) \frac{dx}{\sqrt{1-x^2}}, \quad j \in \mathbb{N}.$$

- Păstrând în (34) numai termenii de grad cel mult  $n$  se obține o aproximare polinomială utilă de grad  $n$

$$\tau_n(x) = \sum_{j=0}^n c_j T_j(x), \quad (35)$$

având eroarea

$$f(x) - \tau_n(x) = \sum_{j=n+1}^{\infty} c_j T_j(x) \approx c_{n+1} T_{n+1}(x). \quad (36)$$

- Aproximanta din (35) este cu atât mai bună cu cât coeficienții din extremitatea dreaptă tind mai repede către zero. Eroarea (36) oscilează în esență între  $+c_{n+1}$  și  $-c_{n+1}$  și este deci de mărime „uniformă”. Acest lucru contrastează puternic cu dezvoltarea Taylor în jurul lui  $x = 0$ , unde polinomul de grad  $n$  are eroarea proporțională cu  $x^{n+1}$  pe  $[-1, 1]$ .

## Teoremă

Pentru orice polinom monic  $\overset{\circ}{p}_n$  de grad  $n$  are loc

$$\max_{-1 \leq x \leq 1} \left| \overset{\circ}{p}_n(x) \right| \geq \max_{-1 \leq x \leq 1} \left| \overset{\circ}{T}_n(x) \right| = \frac{1}{2^{n-1}}, \quad n \geq 1, \quad (37)$$

unde  $\overset{\circ}{T}_n(x)$  este dat de (31).

**Demonstrație.** Se face prin reducere la absurd. Presupunem că

$$\max_{-1 \leq x \leq 1} \left| \overset{\circ}{p}_n(x) \right| < \frac{1}{2^{n-1}}. \quad (38)$$

## Minimalitatea normei II

Atunci polinomul  $d_n(x) = \overset{\circ}{T}_n(x) - \overset{\circ}{p}_n(x)$  (de grad  $\leq n-1$ ) satisface

$$d_n\left(y_0^{(n)}\right) > 0, \quad d_n\left(y_1^{(n)}\right) < 0, \quad d_n\left(y_2^{(n)}\right) > 0, \dots, (-1)^n d_n\left(y_n^{(n)}\right) > 0. \quad (39)$$

Deoarece  $d_n$  are  $n$  schimbări de semn, el este identic nul; aceasta contrazice (39) și astfel (38) nu poate fi adevărată. ■

Rezultatul (37) se poate interpreta în modul următor: cea mai bună aproximare uniformă din  $\mathbb{P}_{n-1}$  pe  $[-1, 1]$  a lui  $f(x) = x^n$  este dată de  $x^n - \overset{\circ}{T}_n(x)$ , adică, de agregarea termenilor până la gradul  $n-1$  din  $\overset{\circ}{T}_n$  luați cu semnul minus. Din teoria aproximațiilor uniforme se știe că cea mai bună aproximare polinomială uniformă este unică. Deci, egalitatea în (37) poate avea loc numai dacă  $\overset{\circ}{p}_n(x) = \overset{\circ}{T}_n(x)$ .

# Polinoamele Cebîșev de speța a II-a

- Se definesc prin

$$Q_n(t) = \frac{\sin[(n+1) \arccos t]}{\sqrt{1-t^2}}, \quad t \in [-1, 1]$$

- Ele sunt ortogonale pe  $[-1, 1]$  în raport cu măsura  $d\lambda(t) = w(t)dt$ ,  $w(t) = \sqrt{1-t^2}$ .
- Relația de recurență este

$$Q_{n+1}(t) = 2tQ_n(t) - Q_{n-1}(t), \quad Q_0(t) = 1, \quad Q_1(t) = 2t.$$



Figura: Pafnuti Lvovici Cebîșev (1821-1894)

# Polinoamele lui Laguerre I

- Sunt ortogonale pe  $[0, \infty)$  în raport cu ponderea  $w(t) = t^\alpha e^{-t}$ .
- Se definesc prin

$$\ell_n^\alpha(t) = \frac{e^t t^{-\alpha}}{n!} \frac{d^n}{dt^n} (t^{n+\alpha} e^{-t}) \text{ pentru } \alpha > -1$$

- Relația de recurență pentru polinoamele monice este

$$\ell_{k+1}^\alpha(t) = (t - 2k - \alpha - 1)\ell_k^\alpha(t) - \beta_k \ell_{k-1}^\alpha(t),$$

unde

$$\beta_k = \begin{cases} \Gamma(1 + \alpha), & \text{pentru } k = 0; \\ k(k + \alpha), & \text{pentru } k > 0. \end{cases}$$



- Exemple pentru  $\alpha = 0$ :

$$\ell_0^{(0)}(t) = 1,$$

$$\ell_1^{(0)}(t) = t - 1,$$

$$\ell_2^{(0)}(t) = t^2 - 4t + 2,$$

$$\ell_3^{(0)}(t) = t^3 - 9t^2 + 18t - 6$$



Figura: Edmond Laguerre (1834-1886)

- Se definesc prin

$$H_n(t) = (-1)^n e^{t^2} \frac{d^n}{dt^n} (e^{-t^2}).$$

- Ele sunt ortogonale pe  $(-\infty, \infty)$  în raport cu ponderea  $w(t) = e^{-t^2}$  și verifică relația de recurență

$$H_{k+1}(t) = tH_k(t) - \beta_k H_{k-1}(t)$$

unde

$$\beta_k = \begin{cases} \sqrt{\pi}, & \text{pentru } k = 0; \\ \frac{k}{2}, & \text{pentru } k > 0. \end{cases}$$

- Exemple:

$$H_0(t) = 1,$$

$$H_1(t) = t,$$

$$H_2(t) = t^2 - \frac{1}{2},$$

$$H_3(t) = t^3 - \frac{3}{2}t.$$



Figura: Charles Hermite (1822-1901)

- Sunt ortogonale pe  $[-1, 1]$  în raport cu ponderea

$$w(t) = (1 - t)^\alpha (1 + t)^\beta.$$

- Formula lui Rodrigues:

$$P_n^{(\alpha, \beta)}(t) = \frac{(-1)^n}{2^n n!} (1 - t)^{-\alpha} (1 + t)^{-\beta} \cdot \frac{d^n}{dt^n} \left\{ (1 - t)^\alpha (1 + t)^\beta (1 - t^2)^n \right\}.$$

- Coeficienții din relația de recurență sunt

$$\alpha_k = \frac{\beta^2 - \alpha^2}{(2k + \alpha + \beta)(2k + \alpha + \beta + 2)}$$

și

$$\beta_0 = 2^{\alpha+\beta+1} \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+1)},$$

$$\beta_1 = \frac{4(1+\alpha)(1+\beta)}{(2+\alpha+\beta)^2(3+\alpha+\beta)},$$

$$\beta_k = \frac{4k(k+\alpha)(k+\alpha+\beta)(k+\beta)}{(2k+\alpha+\beta-1)(2k+\alpha+\beta)^2(2k+\alpha+\beta+1)}, \quad k > 1.$$

- Exemple pentru  $\alpha = 1/2$  și  $\beta = -1/2$

$$\pi_0^{(\alpha,\beta)}(t) = 1,$$

$$\pi_1^{(\alpha,\beta)}(t) = t,$$

$$\pi_2^{(\alpha,\beta)}(t) = t^2 + \frac{1}{2}t - \frac{1}{4},$$

$$\pi_3^{(\alpha,\beta)}(t) = t^3 + \frac{1}{2}t^2 - \frac{1}{2}t - \frac{1}{8}.$$





**Figura:** Carl Gustav Jacob Jacobi (1804-1851)

## Exemplu

Pentru funcția  $f(t) = \arccos t$ ,  $t \in [-1, 1]$ , obțineți aproximanta în sensul celor mai mici pătrate,  $\hat{\varphi} \in P_n$  a lui  $f$  relativ la funcția pondere  $w(t) = (1 - t^2)^{-\frac{1}{2}} = \frac{1}{\sqrt{1-t^2}}$  adică, găsiți soluția  $\varphi = \hat{\varphi}$  a problemei

$$\min \left\{ \int_{-1}^1 [f(t) - \varphi(t)]^2 \frac{dt}{\sqrt{1-t^2}} : \varphi \in P_n \right\}.$$

Exprimați  $\varphi$  cu ajutorul polinoamelor Cebîșev  $\pi_j(t) = T_j(t)$ .

**Soluție.**  $\hat{\varphi}(t) = \frac{c_0}{2} + c_1 T_1(x) + \dots + c_n T_n(x)$







$$\begin{aligned} c_k &= \frac{(f, T_k)}{(T_k, T_k)} = \frac{2}{\pi} (f, T_k) = \frac{2}{\pi} \int_{-1}^1 \frac{\arccos t}{\sqrt{1-t^2}} \cos(k \arccos t) dt \\ &= \frac{2}{\pi} \int_0^\pi u \cos ku du = \frac{2}{\pi} \left[ \frac{u \sin ku}{k} \Big|_0^\pi - \frac{1}{k} \int_0^\pi \sin kudu \right] \end{aligned}$$




$$= \frac{2}{\pi} \left[ \frac{1}{k} \frac{\cos ku}{k} \Big|_0^{\pi} \right] = -\frac{2}{\pi k^2} [(-1)^k - 1]$$

$k$  par  $c_k = 0$

$$k \text{ impar } c_k = -\frac{2}{\pi k^2}(-2) = \frac{4}{\pi k^2} \quad \blacksquare$$

# Bibliografie I

-  Å. Björk, *Numerical Methods for Least Squares Problem*, SIAM, Philadelphia, 1996.
-  E. Blum, *Numerical Computing: Theory and Practice*, Addison-Wesley, 1972.
-  P. G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris, Milan, Barcelone, Mexico, 1990.
-  Gheorghe Coman, *Analiză numerică*, Editura Libris, Cluj-Napoca, 1995.
-  W. Gander, M. Gander, F. Kwok, *Scientific Computing. An Introduction using Maple and MATLAB*, Springer, 2014
-  W. Gautschi, *Numerical Analysis. An Introduction*, Birkhäuser, Basel, 1997.

-  W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sidney, 1996, disponibilă prin [www, http://www.nr.com/](http://www.nr.com/).
-  D. D. Stancu, *Analiză numerică – Curs și culegere de probleme*, Lito UBB, Cluj-Napoca, 1977.
-  J. Stoer, R. Burlisch, *Introduction to Numerical Analysis*, 2nd ed., Springer Verlag, 1992.