ESTIMAREA NEDEPLASATĂ, CU DISPERSIE MINIMĂ, GENERALIZATĂ

Într-o problemă de estimare a unui parametru necunoscut, θ, dispunem de cele N eșantioane de date {x[0], x[1], …, x[N-1]}. Fiecare dintre acestea poartă informație despre parametrul necunoscut. Ne punem întrebarea dacă nu putem găsi un singur număr,T,care depinde de date și poartă toată informația despre θ

$$T\left(x[0],x[1],...,x[N-1]\right)=T\left(\mathbf{x}\right)$$

Ca și datele, T(**x**) este o variabilă aleatoare, numită "statistică". Vom considera modelul de semnal

$$x[n]=A+w[n]\ ;\quad n=0,1,...,N-1;\quad \mathbf{w}\sim\mathcal{N}\left(0,\quad \sigma^2\mathbf{I}_u\right)$$

Am arătat că media eșantion este un estimator MVU eficient Pentru componenta continuă, A

$$\hat{\mathrm{A}}=\bar{x}=\frac{1}{N}\sum_{n=0}^{N-1}x[n]$$

Pentru a estima A nu trebuie să cunoaștem toate valorile x[n]. Ne putem mărgini la cunoașterea unei singure valori, statistica T(**x**)

$$T\left(\mathbf{x}\right)=\sum_{n=0}^{N-1}x[n]$$

Ea se numește și "statistică suficientă"

1

Dacă pentru două seturi diferite de date

$$\mathbf{x}_1\neq\mathbf{x}_2$$

dar pentru care valorile statisticii T**(x)** sunt identice, adică

$$T\left(\mathbf{x}_1\right)=T\left(\mathbf{x}_2\right)$$

atunci valorile estimatului pentru **θ** determinate din cele două seturi de date sunt identice

Ne putem pune întrebarea firească "câte statistici suficiente există ?"

Dacă ne referim tot la exemplul în care se estimează componenta continuă, A, cele N date măsurate sunt suficiente pentru măsurare (de altele nici nu dispunem). Prin urmare, mulțimea

$$S_1=\left\{x[0],x[1],...,x[N-1]\right\}$$

constituie o statistică suficientă. În oricare situație, datele măsurate formează o statistică suficientă

Dar, în mod evident, o statistică suficientă o constituie și mulțimea

$$S_2=\left\{x[0]+x[1],x[2]+x[3],....\right\}$$

La început am văzut că mulțimea cu un singur element

$$S_3=\left\{\sum_{n=0}^{N-1}x[n]\right\}$$

este și ea o statistică suficientă dar pe care o vom numi și "minimală" deoarece are numărul minim de elemente între statisticile suficiente posibile

2

Pentru exemplul luat în considerare, avem

$$p(\mathbf{x};A)=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^{N}}\exp\left\{-\frac{1}{2\sigma^{2}}\sum_{n=0}^{N-1}\left(x[n]-A\right)^{2}\right\}$$

Dacă datele **x** măsurate determină o valoare fixă pentru statistica suficientă

$$T(\mathbf{x})=\sum_{n=0}^{N-1}x[n]=T_0$$

atunci densitatea de repartiție condiționată de aceasta nu mai poate fi funcție de parametrul necunoscut, A.

$$p\left(\mathbf{x}\middle|\sum_{n=0}^{N-1}x[n]=T_0;A\right)$$

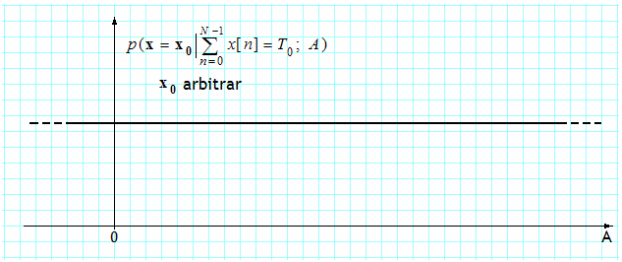
Dacă ar exista dependența de A, din date **x** diferite, dar care ar da aceeași valoare a statisticii suficiente, am mai putea obține informații privind parametrul necunoscut, A. Dar atunci T(**x**) nu ar fi statistică suficientă!

Pentru

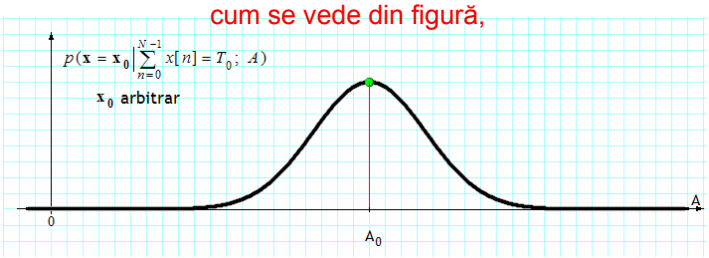
$$p\left(\mathbf{x}=\mathbf{x}_0\middle|\sum_{n=0}^{N-1}x[n]=T_0;A\right);\quad \mathbf{x}_0 \text{ fixat, dar arbitrar}$$

3

se arată în figură independența de valoarea parametrului necunoscut, A, dacă T(**x**) este o statistică suficientă



Dacă există variație funcție de valoarea parametrului necunoscut, A, deși pentru diverse seturi de date se obține aceeași valoare a statisticii T**(x)**, așa cum se vede din figură,



atunci nu avem de-a face cu o statistică suficientă!

4

Se pune întrebarea “cum poate fi determinată statistica suficientă (eventual minimală)?” Răspunsul este dat de teorema de factorizare Neyman-Fisher (Noiman-Fișer), al cărui enunț îl dăm fără demonstrație

Teorema de factorizare Neyman-Fisher

Dacă densitatea de probabilitate a datelor, **x**, p(**x**; θ), dependentă de parametrul necunoscut θ, poate fi factorizată sub forma

$$p(\mathbf{x};\theta) = g\left(T(\mathbf{x}),\theta\right)h(\mathbf{x})$$

în care g(.) este o funcție care depinde de datele **x** numai prin intermediul statisticii T(**x**) iar h(.) este o funcție numai de datele **x,** nu și de statistica T(**x**) sau de parametrul necunoscut θ, atunci T(**x**) este o statistică suficientă pentru θ

Reciproc, dacă T(**x** este o statistică suficientă pentru θ, atunci se poate obține factorizarea de mai sus

***

5

Statistica suficientă pentru estimarea unui nivel continuu, în zgomot alb, gaussian

Dacă în expresia densității de repartiție corespunzătoare se dezvoltă pătratul de la exponent, putem obține forma cerută de teorema de factorizare

$$p(\mathbf{x};A) = \underbrace{\frac{1}{\left(\sqrt{2\pi}\sigma\right)^N}\exp\left\{-\frac{1}{2\sigma^2}\left[NA^2 - 2A\sum_{n=0}^{N-1}x[n]\right]\right\}}_{g\left(T(\mathbf{x}),A\right)}\underbrace{\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right\}}_{h(\mathbf{x})}$$

și deducem că statistica suficientă pentru estimarea componentei continue este

$$T(\mathbf{x}) = \sum_{n=0}^{N-1}x[n]$$

Statistica suficientă pentru estimarea puterii zgomotului alb, gaussian

Pentru zgomotul alb gaussian, cu puterea (dispersia) necunoscută

$$x[n] = w[n];\ \ n = 0,1,...,N-1,\ \ w[n] \sim \mathcal{N}\left(0,\ \ \sigma^2\right)$$

factorizarea este evidentă dacă scriem densitatea de repartiție sub forma

$$p\left(\mathbf{x};\sigma^2\right) = \underbrace{\frac{1}{\left(2\pi\sigma^2\right)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right\}}_{g\left(T(\mathbf{x});\sigma^2\right)}\cdot\underbrace{1}_{h(\mathbf{x})}$$

6

și rezultă că statistica suficientă pentru estimarea dispersiei zgomotului este

$$T(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n]$$

Problema estimării fazei unei sinusoide

Pentru modelul de semnal sinusoidal cu faza inițială, Φ, necunoscută, afectată de un zgomot alb, gaussian datele x[n] au forma

$$x[n] = A\cos(2\pi f_0 n + \Phi) + w[n]; \quad n = 0,1,...,N-1; \quad \mathbf{w} \sim \mathcal{N}(0, \quad \sigma^2 \mathbf{I}_u)$$

iar densitatea de probabilitate a datelor **x** are forma

$$p(\mathbf{x};\Phi) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^N} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left[x[n] - A\cos(2\pi f_0 n + \Phi)\right]^2\right\}$$

Dezvoltăm pătratul de la exponent și obținem

$$\sum_{n=0}^{N-1}\left[x[n] - A\cos(2\pi f_0 n + \Phi)\right]^2$$

$$= \sum_{n=0}^{N-1} x^2[n] - 2A\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n + \Phi) + A^2\sum_{n=0}^{N-1}\cos^2(2\pi f_0 n + \Phi)$$

$$= \sum_{n=0}^{N-1} x^2[n] - 2A\cos\Phi\left[\sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n\right]$$

$$+ 2A\sin\Phi\left[\sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n\right] + A^2\sum_{n=0}^{N-1}\cos^2(2\pi f_0 n + \Phi)$$

7

Cu notațiile

$$T_1(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n$$

$$T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\sin 2\pi f_0 n$$

expresia anterioară devine

$$\sum_{n=0}^{N-1}\left[x[n] - A\cos(2\pi f_0 n + \Phi)\right]^2$$

$$= \sum_{n=0}^{N-1} x^2[n] - 2AT_1(\mathbf{x})\cos\Phi + 2AT_2(\mathbf{x})\sin\Phi$$

$$+ A^2\sum_{n=0}^{N-1}\cos^2(2\pi f_0 n + \Phi)$$

8

Densitatea de probabilitate a datelor **x** se poate factoriza acum

$$p(\mathbf{x};\Phi)=\frac{1}{\left(\sqrt{2\pi}\sigma\right)^{N}}$$

$$\underbrace{\cdot\exp\left\{-\frac{1}{2\sigma^2}\left[A^2\sum_{n=0}^{N-1}\cos^2\left(2\pi f_0 t+\Phi\right)-2A\cos\Phi\cdot T_1(\mathbf{x})+2A\sin\Phi\cdot T_2(\mathbf{x})\right]\right\}}_{g\left(T_1(\mathbf{x}),T_2(\mathbf{x}),\Phi\right)}$$

$$\underbrace{\cdot\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right\}}_{h(\mathbf{x})}$$

Apar două statistici în loc de una. Teorema de factorizare Neyman-Fisher se poate extinde, după cum urmează

9

Extinderea teoremei de factorizare Neyman- Fisher pentru un grup de r statistici suficiente

Dacă densitatea de probabilitate a datelor, **x**, dependentă de parametrul necunoscut θ, p(**x**; θ) poate fi factorizată sub forma

$$p(\mathbf{x};\theta)=g\left(T_1(\mathbf{x}),T_2(\mathbf{x}),...,T_r(\mathbf{x});\theta\right)h(\mathbf{x})$$

atunci

$$\left\{T_1(\mathbf{x}),T_2(\mathbf{x}),...,T_r(\mathbf{x})\right\}$$

formează un grup de r statistici mutual suficiente pentru estimarea parametrului necunoscut, θ. Reciproca teoremei este adevărată.

***

În orice problemă de estimare, deoarece densitatea de repartiție a datelor se poate scrie sub forma

$$p(\mathbf{x};\theta)=\underbrace{p\left(x[0],x[1],...,x[N-1];\theta\right)}_{g\left(x[0],x[1],...,x[N-1];\theta\right)}\cdot\underbrace{1}_{h(\mathbf{x})}$$

rezultă că, la limită, datelepot fi asimilate cu un grup de N statistici mutual suficiente pentru estimarea parametrului necunoscut

$$\left\{x[0],x[1],...,x[N-1]\right\}$$

10

Determinarea estimatorilor MVU plecând de la o statistică suficientă

Dacă am determinat o statistică suficientă, T(x), pentru un parametru necunoscut, θ, se poate găsi un estimator MVU în două feluri, dintre care vom prezenta doar unul:

se caută o funcție g(.), atfel încât

$$\hat{\theta} = g\left(T(\mathbf{x})\right)$$

să fie un estimator nedeplasat pentru θ, adică

$$E\left\{\hat{\theta}\right\} = E\left\{g\left(T(\mathbf{x})\right)\right\} = \theta$$

Pentru exemplificare reluăm problema estimării componentei continue pentru care statistica suficientă este

$$T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]$$

Trebuie găsită funcția g(x) pentru care să avem

$$E\left\{g\left(\sum_{n=0}^{N-1} x[n]\right)\right\} = A$$

11

Funcția g(x) are forma evidentă

$$g(x) = \frac{x}{N}$$

astfel că estimatorul componentei continue A este, în mod evident

$$\hat{\mathrm{A}} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

Dacă funcția g(.) este unică, statistica suficientă T**(x)** se spune că este "completă"

Familia de repartiții exponențial-scalare, ce are forma

$$p(x;\theta) = \exp\left\{A(\theta)B(x) + C(x) + D(\theta)\right\}$$

are proprietatea de a genera statistici suficiente complete pentru parametrul necunoscut, θ. Repartiția gaussiană cu media μ necunoscută, repartiția Rayleigh cu dispersia necunoscută și repartiția exponențială cu parametrul λ necunoscut, fac toate parte din familia exponențial-scalară

12

Repartiția exponențial scalară poate fi factorizată conform teoremei Neyman-Fisher sub forma

$$p(x;\theta)=\underbrace{\exp\{A(\theta)B(x)+D(\theta)\}}_{g(T(x),\theta)}\underbrace{\exp\{C(x)\}}_{h(x)}$$

din care rezultă că statistica suficientă pentru parametrul necunoscut θ este

$$T(x)=B(x)$$

Considerăm că datele

$$\mathbf{x}=\begin{bmatrix}x[0] & x[1] & \cdots & x[N-1]\end{bmatrix}^T$$

sunt de tip IID, motiv pentru care densitatea de repartiție a vectorului **x** este

$$\begin{aligned}p(\mathbf{x};\theta)&=\prod_{n=0}^{N-1}p(x[n];\theta)\\&=\prod_{n=0}^{N-1}\left[\exp\{A(\theta)B(x[n])+D(\theta)\}\exp\{C(x[n])\}\right]\\&=\exp\left\{A(\theta)\sum_{n=0}^{N-1}B(x[n])+ND(\theta)\right\}\exp\left\{\sum_{n=0}^{N-1}C(x[n])\right\}\end{aligned}$$

13

Statistica suficientă și completă pentru parametrul necunoscut este, conform teoremei Neyman-Fisher

$$T(\mathbf{x})=\sum_{n=0}^{N-1}B(x[n])$$

Prezentăm câteva exemple

1) Pentru o repartiție gaussiană de medie μ necunoscută, densitatea de repartiție este

$$\begin{aligned}p(x;\mu)&=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\\&=\exp\left\{\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}-\ln\sqrt{2\pi}-\ln\sigma\right\}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\end{aligned}$$

Se observă că

$$B(x)=x$$

și deci statistica suficientă și completă, pentru cazul unui vector de date **x** este, pentru medie

$$T_1(\mathbf{x})=\sum_{n=0}^{N-1}x[n]$$

14

2) Pentru o repartiție Rayleigh de dispersie necunoscută, densitatea de repartiție este

$$p\left(x;\sigma^2\right)=\frac{x}{\sigma^2}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}u\left(x\right)$$

$$=\exp\left\{-\frac{1}{2\sigma^2}x^2-\ln\sigma^2\right\}\exp\left\{\ln\left[xu\left(x\right)\right]\right\}$$

în care u(x) este treapta unitară

$$u(x)=\begin{cases}1, & x>0\\ 0, & x<0\end{cases}$$

Prin identificare rezultă că

$$B\left(x\right)=x^2$$

și deci statistica suficientă și completă, pentru cazul unui vector de date **x** este, pentru dispersie

$$T_2\left(\mathbf{x}\right)=\sum_{n=0}^{N-1}x^2\left[n\right]$$

15

3) Pentru o repartiție exponențială parametru λ necunoscut, densitatea de repartiție este

$$p\left(x;\lambda\right)=\lambda\exp\left\{-\lambda x\right\}u\left(x\right)$$

$$=\exp\left\{-\lambda x+\ln\lambda\right\}\exp\left\{\ln u\left(x\right)\right\}$$

Prin identificare rezultă că

$$B\left(x\right)=x$$

și deci statistica suficientă și completă, pentru cazul unui vector de date **x** este, pentru parametrul λ

$$T_3\left(\mathbf{x}\right)=\sum_{n=0}^{N-1}x\left[n\right]$$

16

Pentru exemplul 1), am stabilit deja că funcția g(x) este

$$g(x)=\frac{x}{N}$$

și deci estimatorul mediei devine

$$\hat{\mu}=\frac{1}{N}\sum_{n=0}^{N-1}x[n]$$

Pentru exemplul 2), vom căuta forma funcției g(x). Media de ordinul doi a variabilei aleatoare Rayleigh este

$$E\{x^2\}=\int_0^\infty x^2\frac{x}{\sigma^2}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}dx=2\sigma^2\int_0^\infty ve^{-v}dv$$

$$=2\sigma^2 v\left(-e^{-v}\right)\Big|_0^\infty+2\sigma^2\int_0^\infty e^{-v}dv=2\sigma^2$$

Calculăm media statisticii suficiente pentru cazul 2). Avem

$$E\left\{\sum_{n=0}^{N-1}x^2[n]\right\}=\sum_{n=0}^{N-1}E\{x^2[n]\}=\sum_{n=0}^{N-1}\sigma^2[n]=2N\sigma^2\neq\sigma^2$$

Nu este greu de observat că

$$E\left\{\frac{1}{2N}\sum_{n=0}^{N-1}x^2[n]\right\}=\sigma^2$$

17

așa că funcția g(x), pentru cazul 2) este, evident

$$g(x)=x/(2N)$$

Am obținut, pentru acest caz estimatorul

$$\hat{\sigma^2}=\frac{1}{2N}\sum_{n=0}^{N-1}x^2[n]$$

Și pentru exemplul 3) trebuie căutată forma funcției g(x). Media variabilei aleatoare exponențiale este

$$E\{x\}=\int_0^\infty x\lambda\exp\{-\lambda x\}dx=\lambda x\left(-\frac{1}{\lambda}e^{-\lambda x}\right)\Big|_0^\infty+\lambda\int_0^\infty\frac{1}{\lambda}e^{-\lambda x}dx=\frac{1}{\lambda}$$

și deci

$$E\left\{\sum_{n=0}^{N-1}x[n]\right\}=\sum_{n=0}^{N-1}E\{x[n]\}=\frac{N}{\lambda}\neq\lambda$$

Această expresie a mediei ne sugerează să luăm ca parametru necunoscut

$$\theta=1/\lambda$$

Pentru noul parametru necunoscut, θ, avem densitatea de repartiție

$$p(x;\lambda)=p(x;\theta)=\frac{1}{\theta}\exp\left\{-\frac{x}{\theta}\right\}u(x)$$

$$=\exp\left\{-\frac{1}{\theta}\cdot x+\ln\frac{1}{\theta}\right\}\exp\{\ln u(x)\}$$

18

Rezultă că

$$B(x) = x$$

Dacă ținem seama că media datelor x[n] este, așa cum am stabilit mai înainte θ avem, pentru media statisticii suficiente din acest caz

$$E\{x[n]\} = \frac{1}{\lambda} = \theta$$

$$E\left\{\sum_{n=0}^{N-1} x[n]\right\} = \sum_{n=0}^{N-1} E\{x[n]\} = N\theta \neq \theta$$

Rezultă din

$$E\left\{\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right\} = \theta$$

estimatorul MVU pentru θ ca fiind media eșantion a datelor

$$\hat{\theta} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] = \frac{1}{\hat{\lambda}}$$

Ținînd seama de relația dintre θ și λ rezultă estimatorul căutat

$$\hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{\frac{1}{N}\sum_{n=0}^{N-1} x[n]}$$

19

Cazul existenței unui grup de statistici mutual suficiente

Reluăm problema estimării fazei inițiale Φ a unui semnal sinusoidal, afectat de un zgomot alb, gaussian

$$x[n] = A\cos(2\pi f_0 n + \Phi) + w[n];\quad n = 0,1,...,N-1;\quad w[n] \sim \mathcal{N}(0,\quad \sigma^2)$$

pentru care există două statistici mutual suficiente. Pentru a determina forma estimatorului MVU trebuie să găsim funcția g(.,.), astfel încât media ei statistică să fie chiar Φ

$$E\{g(T_1(\mathbf{x}), T_2(\mathbf{x}))\} = \Phi$$

Prima statistică suficientă poate fi aproximată cu relația

$$\begin{aligned} T_1(\mathbf{x}) &= \sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n \\ &= \sum_{n=0}^{N-1} \left[A\cos(2\pi f_0 n + \Phi) + w[n]\right]\cos 2\pi f_0 n \\ &= \sum_{n=0}^{N-1} \frac{A}{2}\left[\cos\Phi + \cos(4\pi f_0 n + \Phi)\right] + \sum_{n=0}^{N-1} w[n]\cos 2\pi f_0 n \\ &\cong \frac{NA}{2}\cos\Phi + \sum_{n=0}^{N-1} w[n]\cos 2\pi f_0 n \end{aligned}$$

20

care, pentru un raport semnal/zgomot (SNR) mare devine

$$T_1(\mathbf{x}) \cong \frac{NA}{2}\cos\Phi; \quad \text{SNR}= \frac{A^2}{2\sigma^2} \gg 1$$

O formă asemănătoare se poate stabili și pentru a doua statistică suficientă

$$T_2(\mathbf{x}) \cong -\frac{NA}{2}\sin\Phi; \quad \text{SNR}= \frac{A^2}{2\sigma^2} \gg 1$$

Raportul celor două statistici suficiente

$$\frac{T_1(\mathbf{x})}{T_2(\mathbf{x})} \cong -tg\Phi; \quad \text{SNR}= \frac{A^2}{2\sigma^2} \gg 1$$

ne permite să estimăm faza inițială a sinusoidei dar numai în cazul unui raport semnal/zgomot mare

$$\hat{\Phi} \cong -arctg\frac{T_2(\mathbf{x})}{T_1(\mathbf{x})}; \quad \text{SNR}= \frac{A^2}{2\sigma^2} \gg 1$$

21

Mediile statistice ale celor două statistici suficiente se determină ușor

$$E\{T_1(\mathbf{x})\} = \frac{NA}{2}\cos\Phi + \sum_{n=0}^{N-1} \underbrace{E\{w[n]\}}_{=0}\cos 2\pi f_0 n$$

$$= \frac{NA}{2}\cos\Phi$$

$$E\{T_2(\mathbf{x})\} = -\frac{NA}{2}\sin\Phi$$

Media statistică a estimatorului fazei, chiar si pentru SNR mare, nu este Φ !

$$E\left\{-arctg\frac{T_2(\mathbf{x})}{T_1(\mathbf{x})}\right\} \neq -arctg\frac{E\{T_2(\mathbf{x})\}}{E\{T_1(\mathbf{x})\}} = -arctg\frac{-\frac{NA}{2}\sin\Phi}{\frac{NA}{2}\cos\Phi} = \Phi$$

Prin urmare

$$E\{\hat{\Phi}\} \neq \Phi$$

și deci estimatorul introdus pentru fază nu este un estimator MVU. Acest fapt se datorează neliniarității funcției arctg(x)

22

O extindere pentru cazul mai multor parametrii necunoscuți, organizați sub forma unui vector, $\boldsymbol{\theta}$ de dimensiune p.

Cele r statistici se organizează sub forma unui vector

$$\mathbf{T}_{r\times 1}(\mathbf{x}) = \begin{bmatrix} T_1(\mathbf{x}) & T_2(\mathbf{x}) & \cdots & T_r(\mathbf{x}) \end{bmatrix}^T$$

Teorema Neyman-Fisher se extinde și la cest caz. Astfel, dacă poate avea loc o factorizare de forma

$$p(\mathbf{x};\boldsymbol{\theta}) = g\left(\mathbf{T}(\mathbf{x});\boldsymbol{\theta}\right)h(\mathbf{x})$$

atunci vectorul **T(x)** este un vector statistică suficientă pentru parametrul vector $\boldsymbol{\theta}$

Reciproca teoremei este adevărată

23

Exemple de determinare a estimatorului MVU pentru cazul parametrului vector $\boldsymbol{\theta}$

Vom considera un exemplu de semnal sinusoidal, afectat de un zgomot alb, gaussian

$$x[n] = A\cos 2\pi f_0 n + w[n];\quad n = 0,1,\ldots,N-1;\quad w[n] \sim \mathcal{N}\left(0,\quad \sigma^2\right)$$

Parametrii necunoscuți sunt amplitudinea, frecvența (digitală) și puterea zgomotului. Cei trei parametri necunoscuți sunt componente ale vectorului

$$\boldsymbol{\theta} = \begin{bmatrix} A & f_0 & \sigma^2 \end{bmatrix}^T$$

Vectorul de date are componentele gaussiene, IID, cu densitatea de repartiție

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(x[n] - A\cos 2\pi f_0 n\right)^2\right\}$$

24

Paranteza de la exponent se dezvoltă sub forma

$$\sum_{n=0}^{N-1}\left(x[n]-A\cos 2\pi f_0 n\right)^2=\sum_{n=0}^{N-1}x^2[n]-2A\sum_{n=0}^{N-1}\cos 2\pi f_0 n+A^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n$$

Pentru determinarea estimatorului pentru vectorul $\boldsymbol{\theta}$ vom proceda la o abordare graduală

1) Dacă se cunoaște frecvența digitală, vectorul parametrilor necunoscuți va avea doar două compponente

$$\boldsymbol{\theta}_1=\begin{bmatrix}A & \sigma^2\end{bmatrix}^T$$

Vom proceda la factorizarea Neyman-Fisher pentru acest caz de complexitate mai redusă. Avem

$$p\left(\mathbf{x};\boldsymbol{\theta}_1\right)=\underbrace{\frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{n=0}^{N-1}x^2[n]-2A\sum_{n=0}^{N-1}x[n]\cos 2\pi f_0 n+A^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right]\right\}}_{g\left(\mathbf{T}(\mathbf{x}),\boldsymbol{\theta}\right)}\cdot\underbrace{1}_{h(\mathbf{x})}$$

25

Vectorul statisticilor suficiente are două componente (scalare)

$$\mathbf{T}(\mathbf{x})=\begin{bmatrix}\mathrm{T}_1(\mathbf{x})=\sum_{n=0}^{N-1}x[n]\cos 2\pi f_0 n\\ \mathrm{T}_2(\mathbf{x})=\sum_{n=0}^{N-1}x^2[n]\end{bmatrix}$$

1.1) Vom arăta că prima componentă scalară a vectorului **T**(**x**) corespunde unei statistici suficiente pentru amplitudinea A. Într-adevăr, dacă numai parametrul A ar fi necunoscut, factorizarea Neyman-Fisher ar lua forma

$$p\left(\mathbf{x};A\right)=\underbrace{\frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\left[A^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n-2A\sum_{n=0}^{N-1}x[n]\cos 2\pi f_0 n\right]\right\}}_{g\left(T'(\mathbf{x});A\right)}\cdot$$

$$\underbrace{\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right\}}_{h(\mathbf{x})}$$

Din care rezultă că statistica suficientă pentru estimarea amplitudinii A este

$$T'(\mathbf{x})=\sum_{n=0}^{N-1}x[n]\cos 2\pi f_0 n\equiv T_1(\mathbf{x})$$

statistică suficientă identică cu prima componentă a vectorului **T**(**x**)

26

Va trebui să găsim acea funție care face ca statistica suficientă pentru A să fie nedeplasată, adică să aibă media statistică egală cu A. Avem

$$E\{x[n]\} = A\cos 2\pi f_0 n$$

și apoi

$$E\{T'(\mathbf{x})\} = \sum_{n=0}^{N-1} E\{x[n]\}\cos 2\pi f_0 n$$

$$= A\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n$$

din care rezultă că

$$E\left\{\frac{T'(\mathbf{x})}{\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n}\right\} = A$$

Prin urmare, estimatorul MVU pentru amplitudinea A a sinusoidei este

$$\hat{A} = \frac{\sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n}{\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n}$$

27

1.2) Dacă nici A nici dispersia nu se cunosc, vom lua în considerare și a doua componentă a vectorului statistică suficientă, **T**(**x**). Media ei statististică este

$$E\{T_2(\mathbf{x})\} = E\left\{\sum_{n=0}^{N-1} x^2[n]\right\} = \sum_{n=0}^{N-1} E\{x^2[n]\}$$

$$= \sum_{n=0}^{N-1} E\left\{\left(A\cos 2\pi f_0 n + w[n]\right)^2\right\}$$

$$= \sum_{n=0}^{N-1} E\left\{A^2\cos^2 2\pi f_0 n + 2Aw[n]\cos 2\pi f_0 n + w^2[n]\right\}$$

$$= \sum_{n=0}^{N-1}\left[A^2\cos^2 2\pi f_0 n + 2AE\{w[n]\}\cos 2\pi f_0 n + E\{w^2[n]\}\right]$$

$$= A^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n + N\sigma^2 \neq \sigma^2$$

28

Avem deja estimatorul pentru amplitudinea A. Momentul de ordinul doi al acestuia se poate calcula aplicând definiția

$$E\left\{\hat{A}^2\right\}=\frac{1}{\left(\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right)^2}E\left\{\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}x[n]x[m]\cos 2\pi f_0 n\cos 2\pi f_0 m\right\}$$

$$=\frac{1}{\left(\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right)^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}E\left\{x[n]x[m]\right\}\cos 2\pi f_0 n\cos 2\pi f_0 m$$

În formulă intră corelația datelor, care se determină tot prin calcul direct, ținând seama de faptul că zgomotul w[n] este alb și are deci eșantioanele necorelate. Avem

$$E\left\{x[n]x[m]\right\}=E\left\{\left(A\cos 2\pi f_0 n+w[n]\right)\left(A\cos 2\pi f_0 m+w[m]\right)\right\}$$

$$=E\left\{A^2\cos 2\pi f_0 n\cos 2\pi f_0 m\right\}+E\left\{Aw[m]\cos 2\pi f_0 n\right\}$$

$$+E\left\{Aw[n]\cos 2\pi f_0 m\right\}+E\left\{w[n]w[m]\right\}$$

$$=A^2\cos 2\pi f_0 n\cos 2\pi f_0 m+\sigma^2\delta_{n,m}$$

29

care substituită în expresia momentului de ordinul doi al estimatorului amplitudinii conduce la relația

$$E\left\{\hat{A}^2\right\}=\frac{1}{\left(\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right)^2}$$

$$\cdot\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\left(A^2\cos 2\pi f_0 n\cos 2\pi f_0 m+\sigma^2\delta_{n,m}\right)\cos 2\pi f_0 n\cos 2\pi f_0 m$$

$$=\frac{1}{\left(\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right)^2}\left(A^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\sum_{m=0}^{N-1}\cos^2 2\pi f_0 m+\sigma^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right)$$

$$=\frac{1}{\left(\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right)^2}\left[A^2\left(\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right)^2+\sigma^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right]$$

$$=A^2+\frac{\sigma^2}{\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n}$$

30

Din ultima relație rezultă că

$$A^2 = E\left\{\hat{A}^2\right\} - \frac{\sigma^2}{\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n}$$

formă ce se substituie în E{T_2}. Obținem

$$E\left\{T_2(\mathbf{x})\right\} = \underbrace{\left(E\left\{\hat{A}^2\right\} - \frac{\sigma^2}{\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n}\right)}_{A^2}\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n + N\sigma^2$$

$$= E\left\{\hat{A}^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n + (N-1)\sigma^2\right\}$$

Comparând cei doi membri ai egalității de mai sus, rezultă că

$$T_2(\mathbf{x}) = \hat{A}^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n + (N-1)\sigma^2$$

31

Se explicitează, din această ultimă relație dispersia

$$T_2'(\mathbf{x}) = \frac{1}{N-1}\left\{T_2(\mathbf{x}) - \hat{A}^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right\} = \sigma^2$$

am definit astfel un estimator MVU pentru dispersie, deoarece

$$E\left\{T_2'(\mathbf{x})\right\} = \sigma^2$$

Grupăm cei doi estimatori MVU găsiți, pentru amplitudine și pentru dispersie, sub forma unui vector estimator cu două componente

$$\hat{\boldsymbol{\theta}}_1 = \begin{bmatrix}\hat{A}\\ \wedge\\ \sigma^2\end{bmatrix} = \begin{bmatrix}\dfrac{\sum_{n=0}^{N-1}x[n]\cos 2\pi f_0 n}{\left(\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right)^2}\\ \dfrac{1}{N-1}\left[\sum_{n=0}^{N-1}x^2[n] - \hat{A}^2\sum_{n=0}^{N-1}\cos^2 2\pi f_0 n\right]\end{bmatrix}$$

32

Rezultatul obținut se poate aplica imediat, dacă se cunoaște frecvența digitală a sinusoidei. Pentru frecvență nulă se ajunge la modelul de semnal componentă continuă, în care nu se cunosc amplitudinea acesteia și puterea zgomotului alb

$$x[n] = A + w[n];\quad n = 0,1,...,N-1;\quad w[n] \sim \mathcal{N}\left(0,\quad \sigma^2\right)$$

Vectorul parametrilor necunoscuți este

$$\boldsymbol{\theta}_1 = \begin{bmatrix} A & \sigma^2 \end{bmatrix}^T$$

Estimatorul vector MVU se obține din estimatorul stabilit anterior, punând valoarea zero pentru frecvența digitală. Obținem estimatorul vector

$$\hat{\boldsymbol{\theta}}_1 = \begin{bmatrix} \hat{A} \\ \wedge \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \dfrac{1}{N}\displaystyle\sum_{n=0}^{N-1} x[n] \\ \dfrac{1}{N-1}\left(\displaystyle\sum_{n=0}^{N-1} x^2[n] - N\hat{A}^2\right) \end{bmatrix}$$

33

în care

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] = \bar{x}$$

și

$$\overset{\wedge}{\sigma^2} = \frac{1}{N-1}\left(\sum_{n=0}^{N-1} x^2[n] - N\bar{x}^2\right)$$

Această ultimă relație se mai poate modifica

$$\frac{1}{N-1}\sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2 = \frac{1}{N-1}\left(\sum_{n=0}^{N-1} x^2[n] - 2\bar{x}\sum_{n=0}^{N-1} x[n] - \sum_{n=0}^{N-1}\bar{x}^2\right)$$

$$= \frac{1}{N-1}\left(\sum_{n=0}^{N-1} x^2[n] - 2\bar{x}N\bar{x} + N\bar{x}^2\right)$$

$$= \frac{1}{N-1}\left(\sum_{n=0}^{N-1} x^2[n] - N\bar{x}^2\right)$$

34

Estimatorul dispersiei poate fi utilizat și sub forma, in care se calculează puterea fluctuației în jurul mediei estimate, medierea fiind făcută prin împărțire cu N-1 și nu cu N

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2$$

O formă posibilă a vectorului estimator, larg utilizată în calculele statistice este

$$\begin{bmatrix} \hat{A} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \bar{x} \\ \frac{1}{N-1}\sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2 \end{bmatrix}$$

Semnal de tip componentă continuă necunoscută, afectată de un zgomot alb, gaussian cu puterea necunoscută. Reluare

$$x[n] = A + w[n];\quad n = 0,1,...,N-1;\quad w[n] \sim \mathcal{N}\left(0,\quad \sigma^2\right)$$

Am stabilit în modelul semnalului sinusoidal cu frecvența cunoscută că vectorul statistică suficientă este

$$\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \mathrm{T}_1(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\cos 2\pi f_0 n \\ \mathrm{T}_2(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n] \end{bmatrix}$$

35

Anulând frecvența digitală se obține vectorul statistică suficientă pentru cazul componentă continuă afectată de zgomot alb, gaussian

$$\mathbf{T}(\mathbf{x}) = \begin{bmatrix} \sum_{n=0}^{N-1} x[n] \\ \sum_{n=0}^{N-1} x^2[n] \end{bmatrix}$$

Mediind vectorul **T**(**x**) nu se obține vectorul **θ** adică componentele estimatorului nu sunt de tip MVU

$$E\left\{\mathbf{T}(\mathbf{x})\right\} = \begin{bmatrix} \sum_{n=0}^{N-1} E\left\{x[n]\right\} \\ \sum_{n=0}^{N-1} E\left\{x^2[n]\right\} \end{bmatrix} = \begin{bmatrix} \sum_{n=0}^{N-1} A \\ \sum_{n=0}^{N-1}\left(A^2+\sigma^2\right) \end{bmatrix} = \begin{bmatrix} NA \\ N\left(A^2+\sigma^2\right) \end{bmatrix} \neq \begin{bmatrix} A \\ \sigma^2 \end{bmatrix}$$

Prin mediere se stabilesc relațiile

$$E\left\{\frac{1}{N}\mathrm{T}_1(\mathbf{x})\right\} = E\left\{\bar{x}\right\} = A$$

$$\frac{1}{N}E\left\{\mathrm{T}_2(\mathbf{x})\right\} = A^2+\sigma^2$$

36

Din a doua relație rezultă

$$E\left\{\frac{1}{N}\mathrm{T}_2(\mathbf{x})-A^2\right\}=\sigma^2$$

relație ce sugerează că un estimator "bun" pentru dispersie ar putea fi expresia

$$\frac{1}{N}\mathrm{T}_2(\mathbf{x})-\bar{x}^2$$

Media ei se determină cu

$$E\left\{\frac{1}{N}\mathrm{T}_2(\mathbf{x})-\bar{x}^2\right\}=E\left\{\frac{1}{N}\mathrm{T}_2(\mathbf{x})\right\}-E\left\{\bar{x}^2\right\}$$

Media eșanion are repartiția normală

$$\bar{x}\sim N\left(A,\quad\frac{\sigma^2}{N}\right)$$

Drept urmare momentul de ordinul doi al mediei eșantion este

$$E\left\{\bar{x}^2\right\}=A^2+\frac{\sigma^2}{N}$$

37

Media estimatorului considerat a fi "bun" pentru dispersie devine deci

$$E\left\{\frac{1}{N}\mathbf{T}_2(\mathbf{x})-\bar{x}^2\right\}=\frac{1}{N}N\left(A^2+\sigma^2\right)-\left(A^2+\frac{\sigma^2}{N}\right)=\frac{N-1}{N}\sigma^2$$

Din aceasta deducem că

$$E\left\{\frac{1}{N-1}\mathbf{T}_2(\mathbf{x})-\frac{N}{N-1}\bar{x}^2\right\}=\sigma^2$$

ceeace înseamnă că estimatorul cu adevărat bun pentru dispersie este

$$\hat{\sigma}^2=\frac{1}{N-1}\sum_{n=0}^{N-1}x^2[n]-\frac{N}{N-1}\bar{x}^2$$

$$=\frac{1}{N-1}\sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2$$

estimator găsit și mai înainte

38

Se specifică în literatura de specialitate că dacă datele x[n] au o repartiție normală

$$x[n] \sim N\left(\mu, \quad \sigma^2\right)$$

atunci variabila aleatore

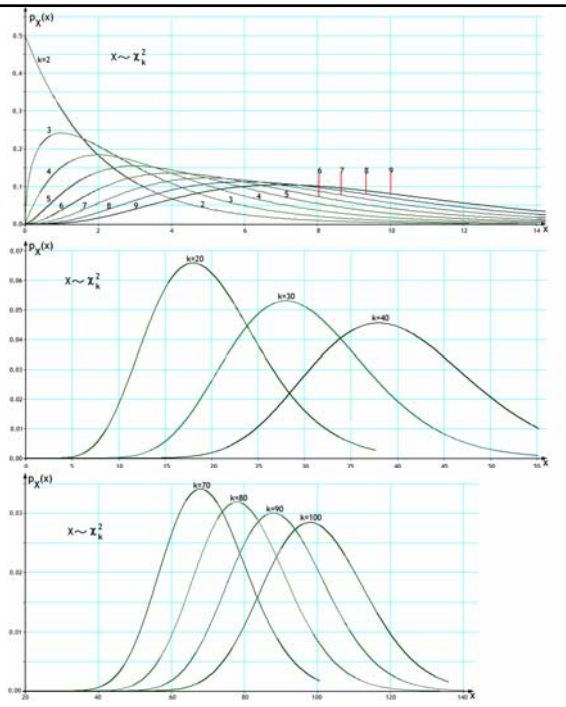
$$y = \sum_{n=0}^{N-1}\left(\frac{x[n]-\overline{x}}{\sigma}\right)^2 \sim \chi^2_{N-1}$$

are o repartiție "hi-pătrat" cu k=N-1 grade de libertate

***

În figura următoare se dau curbele densității de probabilitate pentru repartiții hi-pătrat cu k=2,3...9 grade de libertate și pentru repartiții hi-pătrat cu k=70, 80, 90 și 100 grade de libertate

39



40

Pe măsură ce crește numărul gradelor de libertate ale unei repartiții hi-pătrat, aceasta se apropie tot mai mult de o repartiție normală, așa cum se poate vedea și din figură

Media variabilei alatoare cu repartiție hi- pătrat, cu k grade de libertate este k iar dispersia ei este 2k

Din expresia estimatorului dispersiei

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2 = \frac{\sigma^2}{N-1}\sum_{n=0}^{N-1}\left(\frac{x[n]-\bar{x}}{\sigma}\right)^2$$

se poate deduce că

$$\frac{N-1}{\sigma^2}\hat{\sigma}^2 = \sum_{n=0}^{N-1}\left(\frac{x[n]-\bar{x}}{\sigma}\right)^2 \sim \chi^2_{N-1}$$

41

Se poate scrie, ținând seama că dispersia repartiției hi-pătrat este 2(N-1) că

$$Disp\left\{\frac{N-1}{\sigma^2}\hat{\sigma}^2\right\} = \frac{(N-1)^2}{\sigma^4}Disp\left\{\hat{\sigma}^2\right\} = \underbrace{2(N-1)}_{Disp\left\{\chi^2_{N-1}\right\}}$$

din care rezultă că dispersia estimatorului pentru puterea zgomotului alb gaussian este

$$Disp\left\{\hat{\sigma}^2\right\} = \frac{2\sigma^4}{N-1}$$

Eșantioanele de zgomot fiind necorelate elementele din matricea de covarianță a estimatorului sunt nule, cu excepția celor de pe diagonala principală, care sunt dispersiile celor doi estimatori, dispersii determinate de noi. Avem deci

$$\mathbf{C}_{\hat{\theta}} = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N-1} \end{bmatrix}$$

Anterior am stabilit că CRLB pentru puterea zgomotului alb este

$$CRLB_{\sigma^2} = \frac{2\sigma^4}{N}$$

42

Dispersia estimatorului pe care l-am stabilit este ușor mai mare decât CRLB. El este deci neeficient. Asimptotic însă poate fi considerat eficient

$$\frac{2\sigma^4}{N-1} > \frac{2\sigma^4}{N} = CRLB_{\sigma^2}$$

Trebuie să menționăm că se putea găsi un vector statistică suficientă și direct, pornind de la repartiția

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left(x[n]-A\right)^2\right\}$$

Suma de la exponent se dezvoltă după cum urmează

$$\begin{aligned}\sum_{n=0}^{N-1}\left(x[n]-A\right)^2 &= \sum_{n=0}^{N-1}\left[\left(x[n]-\bar{x}\right)+\left(\bar{x}-A\right)\right]^2 \\ &= \sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2 - 2\left(\bar{x}-A\right)\sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)+\sum_{n=0}^{N-1}\left(\bar{x}-A\right)^2 \\ &= \sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2 - 2\left(\bar{x}-A\right)\left(N\bar{x}-N\bar{x}\right)+\sum_{n=0}^{N-1}\left(\bar{x}-A\right)^2 \\ &= \sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2 + N\left(\bar{x}-A\right)^2\end{aligned}$$

43

Substituim această dezvoltare în expresia repartiției gaussiene și obținem factorizarea

$$p(\mathbf{x};\boldsymbol{\theta}) = \underbrace{\frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2 + N\left(\bar{x}-A\right)^2\right]\right\}}_{g\left(T'(\mathbf{x}),\boldsymbol{\theta}\right)} \cdot \underbrace{1}_{h(\mathbf{x})}$$

din care se deduc imediat cele două statistici suficiente, pentru media A și pentru dispersie. Vectorul statisticilor suficiente este

$$T'(\mathbf{x}) = \begin{bmatrix} \bar{x} \\ \sum_{n=0}^{N-1}\left(x[n]-\bar{x}\right)^2 \end{bmatrix}$$

44