1. **Definiti si caracterizati principalele concepte utilizate în analiza datelor (populatie, esantion, observatie, variabile etc.)**

Activitatea de analiză a datelor poate fi definită ca reprezentând o succesiune de operaţii de prelucrare şi interpretare, operaţii efectuate asupra unor informaţii primare referitoare la fenomene şi procese din realitatea economico-socială şi bazate pe o mare varietate de metode şi tehnici specifice, în scopul adâncirii cunoaşterii comportamentului

Informaţiile cantitative şi calitative referitoare la fenomenele şi procesele studiate exprimă o mulţime de stări şi evoluţii concrete din realitatea investigată şi sunt rezultatul unui laborios proces de observare, măsurare şi evaluare, proces în care intervin o serie de norme, principii, metodologii şi instrumente specifice procesului de măsurare. Informaţiile obţinute din realitatea investigată, în urma unor procese de observare şi de măsurare, sunt cunoscute sub numele de date. Datele reprezintă materialul brut, empiric, care stă la baza tuturor deciziilor din orice domeniu de activitate, iar de calitatea acestora depinde, în mod direct, calitatea respectivelor decizii

Fenomenul stohastic este acel fenomen observabil, ale cărui manifestări particulare sunt incerte, dar care evidenţiază o anumită regularitate a formelor de manifestare, o anumită legătură între aceste forme de manifestare. Informaţiile referitoare la un anumit fenomen supus studiului, informaţii necesare analizei comportamentului respectivului fenomen, reprezintă rezultatul unui proces de măsurare.Acest proces reprezintă, de fapt, o acţiune de atribuire de valori numerice pentru caracteristicile respectivului fenomen.

Prin proces de măsurare se înţelege totalitatea activităţilor de atribuire a unor valori numerice pentru caracteristicile fenomenului analizat. Alegerea unităţilor ce vor fi efectiv supuse procesului de măsurare şi înregistrare se face după criterii şi reguli foarte precise, fundamentate riguros din punct de vedere statistico-matematic. Valorile pe care le iau caracteristicile unităţilor studiate prin intermediul acestui procedeu alcătuiesc aşa-numitul eşantion de observaţii.

Datele reprezintă expresii cantitative şi calitative ale unor fenomene şi procese din realitatea înconjurătoare. Unul din conceptele fundamentale ale analizei datelor, de care este legată definirea multora dintre conceptele uzuale ale acestei discipline este cel de populaţie statistică.

Populaţia sau colectivitatea generală este reprezentată de mulţimea tuturor măsurătorilor efective sau conceptuale care prezintă interes pentru cercetător sau experimentator.

Eşantionul reprezintă o submulţime de măsurători selectate dintr-o populaţie, o submulţime a populaţiei statistice supusă investigaţiei ştiinţifice.

Variabila reprezintă o abstractizare a mulţimii de valori posibile pe care le poate înregistra o caracteristică a unui anumit fenomen.

**2. Ce sunt variabilele si cum se clasifica acestea**

Variabila reprezintă o abstractizare a mulţimii de valori posibile pe care le poate înregistra o caracteristică a unui anumit fenomen. Varietatea fenomenelor economico-sociale şi modalităţile diferite de exprimare a caracteristicilor acestora fac ca variabilele prin intermediul cărora sunt descrise aceste caracteristici să aibă natură diferită. Ca şi caracteristicile populaţiilor, după natura pe care o au, variabilele pot fi de două tipuri: variabile calitative şi variabile cantitative.

În analiza datelor apare necesitatea tratării diferenţiate a datelor de tip calitativ şi cantitativ deoarece există diferenţe substanţiale între aceste tipuri de date atât din punct de vedere al modalităţilor de abordare şi interpretare, cât şi din punct de vedere al metodelor şi tehnicilor utilizate în analiză. Din aceste motive, se face o deosebire netă între variabilele de tip calitativ şi variabilele de tip cantitativ.

Variabilele calitative sunt variabile ce diferă prin tip, se referă la proprietăţi nenumerice ale unităţilor elementare aparţinând unei populaţii şi nu pot fi exprimate numeric.

Variabilele cantitative sunt variabile care diferă prin mărime, se referă la proprietăţi numerice ale unităţilor elementare dintr-o populaţie şi sunt exprimate în unităţi numerice: de lungime, de greutate, valorice etc. În funcţie de natura valorilor pe care le iau, variabilele se împart în două categorii: variabile de tip discret şi variabile de tip continuu.

Variabilele de tip discret sunt variabile care pot lua o mulţime limitată, finită de valori şi care se mai numesc şi variabile categoriale. Valorile luate de variabilele discrete se numesc alternative, categorii, variante sau modalităţi. De regulă, variabilele calitative sunt variabile de tip discret. Variabile de tip discret pot fi însă şi unele variabile cantitative.

Variabilele de tip continuu sunt variabile care pot lua valori aparţinând unui interval continuu. Practic, mulţimea valorilor posibile ale variabilelor de tip continuu este o mulţime infinită. De regulă, variabilele calitative nu sunt variabile de tip continuu.

**3. Ce este scala de masurare si care sunt principalele tipuri de scale de masurare utilizate în analiza datelor**

Măsurarea reprezintă un proces prin intermediul căruia se asociază numere sau simboluri unor caracteristici sau proprietăţi ale unor obiecte sau ale unor subiecţi, care constituie obiectul studiului.

Atribuirea de numere sau simboluri pentru caracteristicile sau proprietăţile unor obiecte se face pe baza respectării unor reguli prestabilite şi prin utilizarea unor proceduri specifice. De exemplu, dacă obiectul studiului este reprezentat de indivizi care sunt potenţiali cumpărători ai unui anumit produs, atunci caracteristicile cărora este necesar a li se atribui numere sau simboluri pot fi: vârsta, venitul, sexul, profesia etc.

Măsurarea caracteristicilor sau proprietăţilor unor obiecte sau subiecţi este totdeauna caracterizată de o anumită specificitate, determinată de natura caracteristicii măsurate, şi presupune, cu necesitate, existenţa unor repere, a unor sisteme de referinţă, cunoscute sub numele de scală. Ca element fundamental al procesului de măsurare a caracteristicilor fenomenelor şi proceselor economice, scala poate fi definită sub forma următoare.

O scală reprezintă un etalon corespunzător, care stabileşte modul după care sunt atribuite valori variabilelor; a defini o scală de măsurare este echivalent cu:

* a stabili o mulţime de valori posibile ale variabilei, mulţime numită şi spaţiu de selecţie;
* a preciza regulile după care sunt atribuite simboluri pentru elementele unei realităţi date, adică a defini o struc­tură a spaţiului de selecţie.

În funcţie de natura variabilelor exprimate cu ajutorul lor, există patru tipuri de scale, pe care le von defini în cele ce urmează. Ca şi procesul de măsurare ca atare, scala sau sistemul de referinţă este, de asemenea, specifică naturii pe care o are caracteristica supusă procesului de măsurare. Din acest punct de vedere, există mai multe tipuri de scale de măsurare: scala nominală, scala ordinală, scala interval şi scala raport. Primele două tipuri de scale sunt scale de tip non-metric, iar ultimele două sunt scale de tip metric.

**4. Definiti si caracterizati scala nominala si scala ordinala. Evidentiati operatiile posibile pe aceste tipuri de scale**

**Scala nominală** este o scală non-metrică, pe baza căreia valorile variabilelor sunt definite prin intermediul simbolurilor nenumerice. Măsurarea variabilelor pe scala nominală este echivalentă cu procesul de codificare a variabilelor. Chiar în cazul în care pentru codificare sunt folosite numere, aceste numere sunt, totuşi, pur convenţionale.

Scala nominală este o scală non-metrică, prin intermediul căreia valorilor posibile ale caracteristicilor măsurate li se atribuie simboluri fără relevanţă numerică, în funcţie de natura acestor valori.

Scala nominală este utilizată pentru a măsura caracteristici ale căror valori sunt de natură calitativă, necuantificabilă. Valorile pe care pot să le ia caracteristicile de acest tip sunt cunoscute sub numele de categorii sau alternative. Variabilele măsurate pe scala nominală se numesc variabile nominale şi sunt variabile a căror formă de exprimare este de tip atributiv şi care pot fi folosite numai pentru stabilirea apartenenţei la o anumită clasă a entităţii descrise prin intermediul variabilei. O clasă specială a variabilelor de tip nominal o reprezintă variabilele binare, care sunt variabile ce pot să ia doar două valori de tip nenumeric.

Variabilele de tip nominal sunt variabile discrete şi pot fi utilizate numai în scopuri de clasificare de tip calitativ, natura nenumerică a acestor variabile făcând imposibilă utilizarea lor pentru comparaţii, ierarhizări sau ordonări. În cazul măsurării pe scala nominală, valorilor pe care pot să le ia caracteristicile supuse măsurării, respectiv categoriilor sau alternativelor, li se atribuie simboluri, care sunt de natură nenumerică.

Pe scala nominală, două valori diferite ale caracteristicii măsurate sunt evidenţiate prin intermediul a două simboluri diferite. Elementele scalei nominale, "diviziunile" acesteia, sunt reprezentate de simbolurile atribuite valorilor caracteristicii studiate, sau, mai exact, de categoriile respectivei caracteristici. Scala nominală este reprezentată chiar de mulţimea acestor simboluri. De exemplu, mulţimile: {"masculin","feminin"}, {"industrie","agricultură","construcţii",...}, {"muncitor","ţăran","intelectual"}, reprezintă scale de tip nominal utilizate pentru a măsura caracteristici cum ar fi sexul, domeniul de activitate, categoria socială, profesia.

Ceea ce este caracteristic scalei nominale este faptul că subiecţii studiaţi nu pot fi comparaţi din punct de vedere al valorii pe care o înregistrează la caracteristica măsurată pe această scală. Pe baza valorilor înregistrate pe scara nominală nu se poate afirma care subiect este "mai bine situat" din punct de vedere al caracteristicii studiate sau, cu atât mai puţin, "în ce măsură" un subiect este situat mai bine decât altul.

Tot pe această scală, caracteristicilor li se pot atribui şi numere, numai că aceste numere nu au sensul propriu-zis de număr, având practic aceeaşi semnificaţie ca şi simbolurile. Atât simbolurile propriu-zise, cât şi numerele cu rol de simbol, atribuite caracteristicilor pe această scală de măsurare, au numai rol de clasificare în anumite grupe a subiecţilor sau de contorizare a numărului de subiecţi din fiecare categorie, neputând fi folosite în nici un tip de calcul numeric. Prin intermediul valorilor măsurate pe scala nominală subiecţii se diferenţiază între ei doar din punct de vedere al apartenenţei la o anumită clasă sau al apartenenţei la o anumită categorie. Aceasta înseamnă că utilizarea scalei nominale pentru măsurarea caracteristicilor măsurabile pe această scală generează clase sau categorii de subiecţi.

Pentru caracteristicile măsurate pe scala nominală, poate fi calculat un număr limitat de indicatori statistici, care reprezintă, de fapt, contorizări ale simbolurilor apărute pe scala nominală. Aceşti indicatori sunt modulul şi frecvenţa. În cazul caracteristicilor măsurate pe scala nominală poate fi evidenţiată şi distribuţia de frecvenţă.

Într-o analiză de date, variabilele nominale pot fi reprezentate de o serie de variabile cum ar fi: sexul, categoria socială, tipul familiei, profesia, marca unui produs etc. Unica transformare de tip invarianta a scalei nominale este reprezentată de operaţia de recodificare, această operaţie neafectând apartenenţa la o anumită clasă a valorilor măsurate pe acestă scală.

**5. Definiti si caracterizati scala ordinala si scala raport. Evidentiati operatiile posibile pe aceste tipuri de scale**

**Scala ordinală** este o scală non-metrică, similară scalei nominale, adică o scală de codificare cu deosebirea că pe această scală este posibilă ordonarea valorilor variabilelor. Această scală este folosită cu precădere pentru măsurarea preferinţelor consumatorilor. Scala ordinală permite clasificarea valorilor unei variabile în funcţie de rangul acestora, însă diferenţele între ranguri nu sunt relevante şi nu au sens. Acest tip de scală nu dă posibilitatea stabilirii gradului în care caracteristicile a două entităţi  
distincte diferă între ele (mai mult, mai puţin).

Scala ordinală este o scală non-metrică, prin intermediul căreia valorilor posibile ale caracteristicilor li se atribuie numere de ordine sau ranguri, în funcţie de poziţia acestor valori într-o ierarhie.

Variabilele măsurate pe această scală se numesc variabile ordinale, sunt variabile calitative de tip discret şi nu pot fi exprimate sub o formă numerică reală. Ca exemple de variabile ordinale putem menţiona: categoria de venit (mic, mediu, mare), nivelul studiilor (elementare, medii, superioare), preferinţa consumatorilor pentru un anumit produs (foarte mare, mare, mică, foarte mică, deloc), nivelul calitativ al unui produs sau serviciu (inferior, mediu, superior), starea economică (recesiune, stagnare, expansiune) etc.

Scala ordinală este utilizată în cazul în care caracteristica subiecţilor supuşi analizei determină o diferenţiere a subiecţilor din punct de vedere al poziţiei pe care fiecare dintre aceştia o ocupă într-o ierarhie, într-o ordonare, adică în cazul în care caracteristica ia valori de tip ordinal. Valorile pe care pot să le ia caracteristicile măsurate pe scala ordinală sunt valori ordinale sau note, cunoscute şi sub numele de ranguri. Acestor valori li se atribuie fie numere de ordine, fie simboluri care evidenţiază o anumită ordine a valorilor caracteristicii.

Pe scala ordinală, două valori diferite ale unei caracteristici sunt evidenţiate prin intermediul a două ranguri diferite, adică prin intermediul a două poziţii diferite în cadrul ierarhiei. Elementele scalei ordinale, "diviziunile" acesteia, sunt reprezentate de numerele sau de simbolurile folosite pentru reprezentarea rangurilor, respectiv de poziţiile posibile în respectiva ordonare. Scala nominală este reprezentată chiar de mulţimea acestor numere sau simboluri.

Cu toate că valorile caracteristicilor de tip ordinal nu sunt numere propriu-zise, ele diferenţiază, totuşi, poziţia unui subiect în raport cu un alt subiect, "spun ceva" despre această poziţie. Valorile unei caracteristici măsurate pe scala ordinală permit doar ordonarea subiecţilor din punct de vedere al acestei caracteristici, determinând o ierarhizare a subiecţilor sau obiectelor.

Prin intermediul valorilor pe care le pot lua caracteristicile măsurate pe scala ordinală, indivizii se diferenţiază între ei doar din punct de vedere al rangului, al locului pe care îl ocupă în ierarhia generată de scala ordinală. Aceasta înseamnă că utilizarea scalei ordinale pentru măsurarea caracteristicilor măsurabile pe această scală generează ierarhii, ordonări ale subiecţilor.

Măsurarea pe scala ordinală permite comparaţii între subiecţi din punct de vedere al caracteristicii măsurate, dar aceste comparaţii se referă numai la modul în care un subiect "este situat" în raport cu altul, fără a se putea spune şi "în ce măsură" subiecţii diferă între ei după caracteristica respectivă. Diferenţele dintre două valori succesive de pe scala ordinală nu pot fi considerate ca fiind egale, ele nedeterminând o distanţare egală între indivizi, astfel încât să se poată afirma, de exemplu, că subiectul situat pe primul loc este "de trei ori mai bun" decât subiectul situat pe locul al treilea.

Pentru caracteristicile măsurate pe scala ordinală, pot fi calculaţi o serie de indicatori statistici cum ar fi: modulul, mediana, coeficientul de corelaţie a rangurilor, frecvenţa. De asemenea, pentru caracteristicile de tip ordinal se poate evidenţia şi distribuţia de frecvenţă. Este important să se facă, în acest context, precizarea că media şi diferenţele valorilor variabilelor ordinale sunt nerelevante, nu au sens informaţional şi nici sens logic.

Singura transformare invariantă a scalei ordinale este translaţia, adică transformarea care păstrează ordinea valorilor unei variabile. Analitic, acest tip de transformare invariantă a scalei ordinale poate fi definit astfel: y=a+x unde a este o constantă, pozitivă sau negativă, care dă sensul şi mărimea translaţiei valorilor scalei ordinale, valori reprezentate de x.

**Scala de tip raport** este scala care are toate proprietăţile scalei de tip interval, însă, în plus faţă de aceasta, are o origine naturală, neconvenţională, care nu poate fi schimbată. Este o scală metrică, pe care valorile sunt exprimate sub formă numerică, dar, spre deosebire de variabilele de tip interval, aceste valori sunt definite în raport cu o anumită origine. Originea scalei indică absenţa proprietăţii, caracteristicii. În plus faţă de scalele precedente, pe această scală este definit şiraportul valorilor, adică se poate compara de câte ori o valoare este mai mare decât alta. Scala raport este o scală metrică, prin intermediul căreia valorilor posibile pe care le pot lua caracteristicile măsurate li se atribuie numere definite in raport cu o origine prestabilită. Scala raport este invariantă până la o transformare proporţională pozitivă, adică până la transformarea: y = ax *.* Variabilele măsurate pe scala raport se numesc variabile tip raport şi sunt variabile cantitative. Cu aceste variabile sunt permise toate operaţiile definite pentru variabilele numerice. Ca exemple de variabile tip raport putem menţiona: preţul, venitul, vârsta, salariul,profitul, volumul vânzărilor, numărul cumpărătorilor etc.

**6. Care sunt principalele moduri de reprezentare (matriciala) a informatiilor în analiza datelor. Definiti si exemplificati fiecare dintre aceste moduri**

Pentru a se asigura o manipulare mai convenabilă şi mai eficientă, datele utilizate în analiza datelor sunt reprezentate sub o formă specifică, numită forma matricială. Această formă de reprezentare a datelor oferă atât avantajul unei structurări simple şi clare a datelor, cât şi avantajul de a oferi posibilitatea generalizării conceptului de mulţime de date.

În cele mai multe ipostaze din analiza datelor, matricea este entitatea care defineşte şi, în acelaşi timp, conţine totalitatea informaţiilor, totalitatea datelor, supuse procesului de analiză. În principiu, datele primare sunt reprezentate în analiza de date sub trei forme matriciale principale: matrici de observaţii, matrici sau tabele de contingenţă şi matrici sau tabele de proximitate.

2.3.1 Matrici de observaţii

O matrice de observaţii este un tablou rectangular în care liniile reprezintă obiectele supuse măsurătorilor, iar coloanele reprezintă caracteristicile obiectelor. Elementele tabloului reprezintă valori înregistrate în procesul de măsurare pentru caracteristicile obiectelor supuse măsurătorilor. Aceste valori mai poartă şi numele generic de scoruri. Matricile de observaţii se mai numesc şi matrici de tip "obiecte\*caracteristici".

Pentru o analiză de date în care numărul obiectelor supuse analizei este T, iar numărul de caracteristici ale obiectelor este n, matricea de observaţii are forma următoare:

Xll X12 — Xln

unde un element xij reprezintă valoarea înregistrată pentru cea de-a j-a caracteristică a obiectului i. O linie i a matricii de observaţii X defineşte un obiect Oi şi reprezintă valorile înregistrate de acest obiect la cele n caracteristici pe care le posedă. O coloană j a matricii de observaţii X reprezintă valorile înregistrate de caracteristica j pe mulţimea tuturor celor T obiecte supuse analizei. De regulă, în analiza de date, fiecare linie a matricii de observaţii X este numită observaţie şi fiecare coloană a acestei matrici este numită variabilă.

În multe situaţii, nu pot fi obţinute informaţii despre toate caracteristicile tuturor obiectelor supuse analizei. În cazul în care datele ce definesc obiectele nu sunt complete, matricea de observaţii definită mai sus poartă numele de matrice de observaţii cu valori omise.

2.3.2 Matrici de contingenţă

Sunt tablouri rectangulare de dimensiune mxn, utilizate pentru reprezentarea datelor referitoare la frecvenţele relative sau absolute înregistrate pe o mulţime de obiecte de valorile a două variabile de tip discret, prima variabilă, notată cu u, având m valori posibile, iar cea de-a doua variabilă, notată cu v, având n valori posibile. Liniile unei matrici de contingenţă reprezintă valorile posibile ale primei variabile discrete, iar coloanele acestei matrici reprezintă valorile posibile ale celei de-a doua variabile discrete. În analiza datelor, matricile de contingenţă se mai numesc şi matrici de tip "modalităţixmodalităţi".

Un element reprezintă frecvenţa, absolută sau relativă, a obiectelor pentru care prima variabilă ia valoarea U; şi cea de-a doua variabilă ia valoarea Vj. Acest element arată la câte obiecte cele două variabile analizate au simultan valorile Uj şi Vj.

2.3.3 Matrici de proximitate

Sunt matrici pătratice de dimensiune nxn, utilizate pentru reprezentarea datelor cu privire la similaritatea sau nesimilaritatea unor obiecte. Ordinul matricilor de proximitate este determinat de numărul obiectelor supuse studiului.Elementele unei matrici de proximitate reprezintă coeficienţi de similaritate, coeficienţi de nesimilaritate sau distanţe. Un element X- din această matrice

măsoară gradul de proximitate dintre obiectul i şi obiectul j.

Matricile de proximitate se mai numesc şi matrici de tip "obiecte\*obiecte" şi sunt utilizate în problemele de clasificare cu ajutorul tehnicilor de tip cluster şi în problemele de scalare multidimensională.

**7. Definiti principalii indicatori (unidimensionali) cu ajutorul carora este sintetizata tendinta centrala sau locatia sau pozitia (inclusiv relatii de calcul si proprietati). Aratati ca media este o sinteza optimala pentru o multime de observatii**

Una dintre masurile cele mai importante si mai relevante pentru descrierea valorilor unei caracteristici este cea reprezentata de *tendinta centrala*. Masurarea tendintei centrale are ca scop principal determinarea unei marimi care sa sintetizeze, sa rezume, multitudinea de valori reprezentate de observatiile efectuate asupra unor variabile, din punct de vedere al magnitudinii acestora. Este evident ca, pentru a fi relevanta, marimea utilizata pentru masurarea tendintei centrale trebuie sa fie un fel de “*centru* *de greutate*” al observatiilor disponibile, valorile observatiilor fiind repartizate în jurul acestei marimi.

Din punct de vedere geometric, determinarea unei masuri pentru exprimarea tendintei centrale este echivalenta cu a gasi un vector care sa aiba *acelasi sens* si *aceeasi directie* cu vectorul ale carui componente sunt egale cu unitatea si care sa fie *cât* *mai apropiat* de vectorul observatiilor. În acest sens, se poate spune ca, în cazul metricii euclidiene, marimea care exprima în mod *optimal* tendinta centrala este *media aritmetica*.

Tendinta centrala poate fi evidentiata prin intermediul unor indicatori statistici, între care cei mai importanti sunt: *media, mediana* si *modulul*. Fiecare dintre acesti indicatori exprima, într-un fel sau altul, mai mult sau mai putin sugestiv, nivelulcaracteristicii analizate de-a lungul obiectelor.

*Media* - se obţine împărţind suma valorilor individuale la efectivul populaţiei sau al eşantionului

*Mediana* - este valoarea care, în cadrul seriei statistice, separă efectivul populaţiei în două părţi egale. Nu are o formulă atît de simplă cum este cea a mediei; mai mult, o valoare mediană propriu-zisă nu există decît dacă numărul n este fără soţ, cînd există, de fapt, un individ mijlociu (al [n+1]/2 lea) a cărui valoare este mediana. Dacă n este par, se iau indivizii de rang n/2 şi n/2 + 1, cu valori, să zicem, Xi şi xi+1, iar mediana poate fi orice valoare din intervalul (Xi,Xi+1); de regulă se ia media aritmetică a celor două valori.

*Modulul* - se utilizează numai cînd se lucrează cu frecvenţe, fiind valoarea luată cu cea mai mare frecvenţă. Se poate vorbi şi de valori modale relative atunci cînd frecvenţele mai multor clase, neînvecinate, le întrec pe cele din imediata lor vecinătate; avem de a face atunci cu serii bimodale (cu două moduri) sau plurimodale.

**8. Definiti principalii indicatori (unidimensionali) cu ajutorul carora este sintetizata variabilitatea (inclusiv relatii de calcul si proprietati)**

O alta masura importanta pentru sintetizarea valorilor unei caracteristici este aceea a *variabilitatii* ce caracterizeaza observatiile variabilei, a *împrastierii*, a *dispersiei* acestor valori. Un indicator sintetic, utilizat pentru masurarea si exprimarea variabilitatii valorilor unei caracteristici, este ***varianta***.

Variabilitatea care caracterizeaza multimea observatiilor efectuate asupra unei anumite caracteristici este evidentiata prin diferentele care exista între valorile pe care le înregistreaza caracteristica pe multimea subiectilor, prin marimea variatiilor valorilor caracteristicii de la un subiect la altul. Variabilitatea este importanta atât din punct de vedere informational, cât si ca marime în contextul careia poate fi judecata relevanta mediei. Cu cât variabilitatea unei multimi de observatii este mai mica, cu atât media constituie o sintetizare, o rezumare mai potrivita si mai relevanta pentru multimea de observatii. Pe de alta parte, cu cât variabilitatea este mai mare, cu atât mai putin media poate fi considerata o expresie sintetica relevanta a valorilor observate. Prin urmare, se poate spune ca încrederea mai mare sau mai mica pe care o putem acorda mediei ca marime ce sintetizeaza valorile observate depinde de marimea variabilitatii acestor valori. Aceasta înseamna ca pentru a avea o masura a relevantei mediei este necesar sa se stabileasca o masura a variabilitatii.

Varianta este *direct proportionala* cu *marimea variatiei* valorilor caracteristicii masurate sau cu *marimea informatiei* care este continuta de observatiile disponibile pentru analiza de date. În conditiile notatiilor anterioare, *varianta* variabilei **Xi**, notata cu **Si2**, se determina cu ajutorul formulei urmatoare:



În mod concret, *varianta* reprezinta suma patratelor abaterilor valorilor individuale în raport cu media ce revine, *în medie*, pe fiecare valoare individuala, adica pe fiecare observatie efectuata asupra variabilei. Ca rezultat al faptului ca variabilitatea poate exista sau nu poate exista, varianta, ca masura a acestei variabilitati, este totdeauna o marime *nenegativa*.

Pornind de la modul în care varianta masoara variabilitatea si de la importanta pe care o are aceasta variabilitate în analiza datelor, se poate face afirmatia ca, într-un anumit sens, *varianta* reprezinta *o masura a informatiei* continute în datele analizate.

O deficienta majora a variantei, ca indicator de masurare a variabilitatii, a cantitatii de informatie continuta în datele primare, este legata de faptul ca variantele a doua caracteristici sau a doua variabile exprimate în unitati de masura diferite *nu* *pot fi comparate*. Comparatia variantelor este, totusi, posibila numai în cazul în care masuratorile caracteristicilor sunt exprimate în aceleasi unitati de masura. Tot în acest sens, exista si o alta deficienta importanta a variantei: aceea ca ea este o *marime nescalata*. Cu toate ca marimea variantei este limitata inferior, ea având o *margine inferioara* reprezentata de valoarea zero si evidentiind lipsa variabilitatii sau constanta, ea nu este limitata superior, *nu are o margine superioara*

O alta problema dificila, care apare în legatura cu varianta, este aceea ca unitatile de masura în care aceasta este exprimata sunt diferite de unitatile de masura ale caracteristicii a carei variabilitate o masoara. De fapt, varianta este masurata în unitati de masura care reprezinta *patrate* ale unitatilor de masura ale observatiilor efectuate asupra caracteristicii considerate. Aceasta trasatura a variantei creaza o serie de dificultati legate de interpretarea concreta a marimii acestui indicator al variatiei. Datorita lipsei de semnificatie a unitatilor de masura ale variantei, pentru masurarea variatiei se utilizeaza si un alt indicator, derivat din varianya si reprezentat de *radacina patrata* a variantei. Acest indicator este cunoscut sub numele de *abatere* *standard*



**9 Definiti varianta simpla, varianta totala si varianta generalizata. Deduceti si interpretati varianta generalizata. Aratati ca varianta generalizata este egala cu determinatul matricii de covarianta**

Varianta simpla este o masura pt devierea de la medie, deviatie standard fata de medie. **Vs=Σi=1n  (xi- xmediu)2** . Varianta totala masoara variabilitateace caracterizeaza observatiile unei multimi de variabile si se defineste ca suma a variantelor individuale ale variabilelor: **Vt=ΣSi2**

Varianta totala ofera o imagine cuprinzatoare asupra variabilitatii globale ce carac observatiile var analizate, ca masoara aceasta variabilitate doar in sens individual, neluand in considerare variabilitatea comuna, simultana a observatiilor, adica variabilitatea generalizata.

Varianta generalizata masoara variabilitatea ce carac observatiile multimii de variabile, atat din pct de vedere individual, cat si din pct de vedere al simultanietatii, al interaxtivitatii informationale variabilele. Varianta generalizata corespunzatoare spatiului observatiilor a 2 variabile considerate este data de rel: **Vg=(|x1|\*|x2| sin Φ)**

Se poate arata ca variatia generalizata este repre de determinantul matricii de covariant ace corespunde variabilelor supuse studiului, resprectiv: **Vg=|S|.**

Varianta generalizata este o masura extrem de importanta a variabilitatii totale, formata atat ca urmare a variabilitatii individuale ce caracterizeaza variabilele, cat si ca urmare a variabilitatii commune ce cracaterizeaza interactiunea variabilelor.

**10. Definiti principalii indicatori (unidimensionali) cu ajutorul carora sunt sintetizate legaturile (inclusiv relatii de calcul si proprietati)**

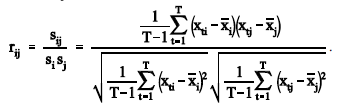
Intensitatea si sensul legaturii sau asocierii de tip liniar dintre doua caracteristici ale unor obiecte sau indivizi reprezinta o alta masura importanta utilizabila in sintetizarea numerica a datelor.

Masura asocierii de tip liniar poate fi exprimata prin intermediul corelarii variatiilor simultane sau covariatiilor a doua caracteristici pe o multime de obiecte sau indivizi. Marimea de baza utilizata pentru exprimarea variatiilor simultane e reprezentat de covarianta:



Covarianta este o masura a variatiei simultane a doua variabile, ea fiind, in valoare absoluta, cu atat mai mare cu cat valorile absolute ale variatiilor celor doua variabile in jurul mediei sunt mai apropiate ca magnitudine, evidentiind o anumita proportionalitate pe multimea subiectilor studiati. Covarianta este considerata a fi o expresie numerica a gradului de asociere a doua caracteristici ca urmare a faptului ca, in toate cazurile in care doua variabile sunt semnificativ legate intre ele, o variatie intr-un sens a uneia dintre ele va determina o variatie proportionala de acelasi sens (in cazul legaturii directe) sau de sens contrar (in cazul legaturii inverse) a celeilalte variabile.

O masura scalata a gradului de asociere liniara intre doua variabile, care elimina unele deficiente ale covariantei ca indicator de masurare a asocierii de tip liniar, o reprezinta coeficientul de corelatie Pearson:



Coeficientul de corelatie Pearson e o marime scalata in intervalul

O valoare nula a coeficientului de coralatie evidentiaza absenta legaturii de tip liniar intre cele doua variabile, dupa cum o valoare absoluta egala cu unitatea evidentiaza o legatura liniara perfecta, legatura care este directa daca valoarea este egala cu 1 si inversa daca valoarea este egala cu -1.

**11. Definiţi si interpretati corelatia si coeficientul de corelatie**

Corelatia este o metoda statistica utilizata pentru a determina relatiile dintre doua sau mai multe variabile. Exista mai multe tipuri de corelatii atât parametrice cât si neparametrice.

Coeficientul de corelatie este o valoare cantitativa ce descrie relatia dintre doua sau mai multe variabile. El variaza între [-1 si +1], unde valorile extreme presupun o relatie perfecta între variabile în timp ce 0 înseamna o lipsa totala de relatie liniara. O interpretare mai adecvata a valorilor obtinute se face prin compararea rezultatului obtinut cu anumite valori prestabilite în tabele de corelatii în functie de numarul de subiecti, tipul de legatura si pragul de semnificatie dorit.

[Image]

**12. Definiti datele de tip profil, de tip cronologic si de tip panel. Exemplificati fiecare dintre cele trei tipuri**

Datele de tip profil reprezinta informatii obtinute prin masuratori de natura static, efecuate asupra caracteristicilor unor unitati ale unei populatii, la acelasi moment de tip. O observatie in contextual datelor de tip profil este reprezentata de valoarea sau de valorile unei singure entitati, ale unei singure unitati din populatie. Nr de observatii coincide in acest caz cu nr de unitati observate si inregistrate. Datele de acest tip nu incorporeaza in semnificatia pe care acestea o poarta, influenta timpului asupra formarii caracteristicilor la niv pop si sensul scurgerii timpului, nici in mod explicit sau implicit. Datele de tip profil se refera la stare ape care o au la un anumit moment indivizi unei gospodarii firme etc. Ex date referitoare la salariul individual dintr-o luna al lucratorilor unei firme, datele referitoare la populatia medie a statelor lumii intr-un an etc.

Datele de tip serii de timp reprezinta informatia obtinuta prin masuratori de natura dinamica effectuate asupra carac unei unitati a unei populatii la moment sau interval successive de timp. Intervalele de timp pt care se fac masuratorile pot fi repre de : ore, luni ani etc. Acest tip de date se refera la evolutia in timp a starii unui individ, gospodarii etc. Datele de acest tip pot fi date de tip interval sau moment. Datele de tip interval sunt datele care se refera la carac care sunt marimi de tip stoc iar cele de tip moment sunt datele care se refera la carac care sunt marimi de tip flux

Datele de tip panel repre inf obt prin masuratori mixte, de natura static si de natura dinamica, effectuate asupra carac acelorasi unitati ale unei pop la moment sa in interval successive de timp.

Aceste date pot fi imaginate ca reprezentand “taieturi informationale mixte”, transversal si langitudinale, in raport cu axa timpului. Observarea acestor date se face intr-o nota de simultaneitate.

Ex bugetele familiilor: inregistrare ch si a vn a mai multor fam dintr-un esantion

**13. Definiti datele de tip observational si de tip experimental. Exemplificati fiecare categorie**

Datele de tip experimental rrepre inf obt prin org unor experimente controlate, in care influentele fact asupra efectului sunt controlate in mod direct, prin fixarea unor combinatii precise de influente. Sunt caracterizate doar unor domenii de cercetare, acelor domenii in care pot fi org experimente specific, necesare obt acestor date. Aceste domenii sunt: fizica chimie bilogie etc.

Aceste date sunt date de laborator, laborator insemnand o serie de cond special, restrictii si instrumente specific de masurare. In domeniul economic-social experimentarea este fie total imposibila, fie posibila dar numai foarte rar si in cond f restrictive si costisitoare.

Datele de tip observabil/non-experimentale repre inf obt prin observarea libera a miscarii fenomenelor si proceselor studiate, fara interventia directa a investigatorului asupra cond in care se desfasoara aceasta miscare. Obtinerea acestor date reprezinta rezultatul observatii pasive , constatarii. Interventia a celui ce face masuratorile este de tip ex post, are loc dupa ce desfasurarea fenomenelor si proceselor reale a avut loc. Datele observabile sunt specifice domeniului economic-ssocial.

**14. Care sunt principalele tipuri de transformari preliminare ale datelor. Interpretati marimile rezultate în urma acestor transformari si mentionati proprietatile acestora**

Inainte de utilizarea lor in analiza datelor, datele originale sunt supuse procesului de prelucrare preliminara, ce poate fi realizata prin operatia de centrare sau prin operatia de standardizare a datelor.

Operatia de **centrare** a datelor consta in substituirea valorii fiecarei observatii cu o noua valoare, reprezantata de abaterea valorii originale fata de media datelor initiale. Datorita faptului ca suma abaterilor valorilor originale este intotdeauna nula, operatia de centrare a datelor, va face ca variabilele centrate sa aiba media nula. In acest caz, variatia unei variabile este proportionala cupatratul lungimii vectorului compus din observatiile respectivei variabile, iar abaterea standard este proportionala cu lungimea aceluiasi vector.

In cazul variabilelor centrate, covarianta este proportionala cu produsul scalar al vectorilor ce reprezinta observatiile celor doua variabile, iar coeficientul de corelatie dintre doua astfel de variabile este raportul scalar dintre produsul scalar al vectorilor ce reprezinta observatiile asupra variabilelor si produsul lungimilor acestor vectori.

Operatia de **standardizare** a valorilor unei variabile consta in substituirea valorilor fiecarei observatii cu o noua valoare reprezentand raportul dintre valarea centrata a respectivei operatii si abaterea standard a respectivei variabile. De asemenea,variabilele standardizate au media aritmetica nula, in plus, varianta lor este egala cu unitatea, iar covarianta este scalata in intervalul [-1,1] (pt. cov=1 perfecta asociere liniara directa intre cele doua variabile, cov=0 nu exista asociere intre cele doua variabile, cov=-1 (perfecta asociere liniara indirecta). O consecinta a acestei proprietati o reprezinta faptul ca in cazul acestor variabile, covariantele sunt chiar coeficientii de corelatie Pearson.

Ca si in cazul variabilelor centrate, covarianta este proportionala cu produsul scalar al vectorilor ce reprezinta observatiile celor doua variabile, iar coeficientul de corelatie dintre doua astfel de variabile este identic cu covarianta si este proportional cu produsul scalar ce reprezinta observatiile asupra variabilelor.

**15. Definti principalele tipuri de matrici utilizate în analiza datelor (produse-încrucisate, covarianta, corelatie). Evidentiati relatiile de legatura dintre aceste tipuri de matrici**

Cele mai importante utilizate in analiza datelor sunt:

Matricea observatiilor centrate- poate fi obt ca diferenta intre mtricea e obs si matricea Xmediu ale carei coloane sunt mediile celor n variabile **Xc= X - Xmediu**

Matricea observatiilor standardizate – poate fi obt ca produs intre matricea var centrate si inversa matricii diagonal V, ale care elem sunt abaterile standard ale celor n variabile **Z = Xc\*V-1 = ()**

Matricea produselor incrucisate – poate fi determinate atat pt var originale, cat si pentru var centrate si standardizate. Pt cazul variabilelor originale, matricea prod incrucisate se obt ca produs intre mat transpusa X si matricea X. **C = Xt \* X**

In cazul in care variabilele sunt centrate, mat prod incrucisate poate fi det astfel: **Cc = . ….**

Matricea de covarianta – este una din cele mai frecvent uti matrici in analiza datelor, majoritatea tehnicilor de AD pp calculul acestei matrici. Pt situatia in care nr de var analizate este egal cu n , cov dintre orice 2 var pot fi aranjate sub forma unei matrici patrate si simetrice, de dim n\*n, numita matrice de cov :

**S= (S12 S12 .. S1n)** , unde Si= 1/(T-1) \* |xi – xmediui|2

**(S21 S22 … S2n)**si Sij=1/(T-1)\*(xi – xmediui)t \*(xj - xmediuj)

In cond notate anterior , matricea de covarianta pt var originale poate fi scrisa cu ajut mat prod incrucisate pt cazul var centrate **S= Cc**

Matricea de corelatie - aceasta este o mat imp in AD pentru faptul ca scrie o serie de metode si tehnici ale AD isi bazeaza procedurile pe analiza spectral a acestei matrici.Aceasta mat are urm forma:

**R= (1 r12 .. r1n)** , unde rij=(xi-xmediui)t \* (xj-xmediuj)/| xi-xmediui | \* | xj-xmediuj |

**(r21 1 .. r2n)**

Matricea de corelatie a var originale poate fi scrisa si cu ajutorul matt prod incrucisate pt cazul val standardizate **R= Zt \* Z**

**16. Ce este analiza componentelor principale. Evidentiati cinci categorii de probleme care pot fi solutionate cu ajutorul tehnicilor de analiza a componentelor principale**

ACP este o met de analiza multidim care are ca scop det unor noi var , numite componente principale si exprimate sub forma combinatiilor liniare de var originale, ai aceste var noi sa fie carac de o variabilitate maxima. Tehnica pe care se bazeaza ACP consta in calculul proiectiilor fiecarui pct din spatiul initial, det de var originale supuse analizei, pe axele unui nou spatiu , a carui dim este semnif mai redusa.

Din pct de vedre al principiilor pe care se bazeaza act de AD, prezinta interes numai acele combinatii liniare semnif dpdv informational. Cum numarul de combinatii liniare care pot fi formate cu var originare este necesara o triere a acestora, aceasta triere se realizeaza prin def unui criteriu care trebuie sa stea la baza deciziei de retinere sau de elim a unui anumite comb liniare. In cadru ACP acest criteriu este bazat pe magnitudinea var fiecareai combinatii liniare si poate fi formulat astfel : se elim combinatiile liniare cu var mica, nesem si se retin pt studio acele comb liniare care au o var maxima.

Principalele probleme care pot fi solutionate cu ajutorul tehnicilor de ACP sunt impartite in 2 mari categorii:simplificarea structurii dependentei cauzale si reducerea dimensionalitatii spatiului causal.

Prima categorie de prob se intelege simplificarea structurii dependentei cauzle, ai sa se obt un spatiu causal de dim mai mica si care sa permita o repre mai simpla si mai sugestiva a obiectelor. In urma acestei simplificari, este mult mai usor sa se evidentieze morfologia leg cauzale si sa se expr o struct adecvata a obiectelor. Un caz aparte de problem rezolvate prin ACP este si eliminarea redundantelor inf . In cazul in care o investigatie stiintifica are ca obiect direct spatial initial , este foarte dificil sa se deduca si sa se exprime o dependent structural care sa evidentieze cu claritate contrbuitile nete ale varanalizate la formarea variabilitatii intregului spatiu cauzal. Corelarea var cauzale det o struct complicate a dependentei , o struct redundant care include anumite suprapuneri inf ale influentelor var cauzale. Deoarece struct cauzala initiala este complicate si include numeroase suprapuneri inf genereaza niste dificultati in privinta intelegerii clare a raporturilor de cauzalitate si formularea unor conc pertinenete asupra datelor. In acest caz se poate renunta la includerea in analiza a inf coresp, cea care nu este semnificativa in toate directiile.

A 2-a categ de prob solutionata prin ACP este reducerea dimensionalitatii. La baza ACP sta idea ca repre unitatilor in sis initial de coordinate nu este intotdeauna cea mai potrivita, considerandu-se ca poate exista o alta modalitate de repre mai relevanta. Aceasta noua modalitate de repre poate fi obt considerand un nou spatiu de repre, care defineste noi carac ale obiectelor. Noile carac sunt numite comp principale iar val inreg se numesc scoruri. Problema repre intr-un spatiu maai redus este cunoscuta sub numele de prob a red dimensionalitatii. Din acest motiv, ACP este cunoscuta si ca tehnica de red a dimensionalitatii. De ex red dimensionalitati poate fi explicate si astfel : se considera un grafic bidimensional in care sunt illustrate niste valori, red dimensionalitatii in acest caz inseamna trecerea de la sis bidimensional la sis unidimesional, adica valorile sa fie repre printr-o dreapta. In urma acestei operatii se obt o noua enntitate inf, care poate fi interpretata ca repre o noua carac a datelor. Inf continuta in aceasta noua entitate este mai relevant decat cea continuta de observatiile facute anterioare.

Problem concrete rezolvate prin reducerea dimesionalitatii sunt:

* 1) selectarea var de influenta – deoarece nu toate var de influenta au aceiasi importanta in formarea caracteristicilor, este necesar ca var sa fie supuse unui process de filtrare, process prin care unele var sa fie elim iar altele pastrate in fct de semnificatia acestora. In afara procesului de filtrare a var indep in fct de imp acestora in AD apare frecvent necesitatea de a grupa var indep in fct de nat influentei
* 2) simplificarea modelelor matematice – exista multe motive care fac dificila si incomoda retinerea in cadrul unui model matem de analiza /predictive a unui nr f mare de var si care conduc la necesitatea unei anumite simpplificari a modelului din acest pct de vedere. In primul rand semnificatia fiecarei var intr-un model care include un nr prea mare de var este f mult diminuata. In al2-lea rand obt unor inf necesare estimarii unui model care contine un nr mare de var ar amplifica un effort si un cost de prohibitive.In al3lea rand este posibil ca in cazul unui nr mare de var sa avem o prob ridicata de existent a unor var f puternic intercorelate.in al 4-lea rand un nr f mare de var retinute intr-un model ar ridica serioase prob de complexitate a calculului
* 3) elim redundantelor informationale
* 4) vizualizarea relatiilor de cauzalitate complexe - prin intermediul ACP pot fi create conditiile necesare pt ca sa poate fi repre chiar si obiectele ce sunt carac de un nr f mare de var
* 5) compresia si restaurarea datelor in informatica - dupa elim redundantelor inf , noua repre a obiectelor asigurata prin aplicarea tehnicilor de ACP este insotita de o pierdere inf neglijabila, at in stocarea si transmiterea inf se pot uti noile carac ale obiectelor. Deoarece nr acestora este mult mai mic manipularea inf se poate face cu un consum mai mic de resurse.

**17. Interpretati logica analizei componentelor principale (inclusiv din punct de vedere geometric)**

Logica ACP se bazeaza pe idea fundamental ca se pot face anumite transformari asupra obs initiale, care sad et maximizarea var individuale pt anumite var individuale pt anumite var si minimizarea var pt alte variabile. Astfel se accentueaza semnif logico- informational a unor var si se diminueaza cea a altor var.Maximizarea variantei unor variabile, in detrimentula variantei celorlalte variabile, conduce la cresterea contributiei var a caror variante este maximizata la formarea variantei totale. Maximizarea variantei unor variabile si accentuarea semnificatiei inf a acestor var in raport cu celelalte este cu atat mai relevant cu cat intre var originale exista leg mai puternice.

Pentru a evidential modul in care pot fi deduse componentele principale, adica noile var care au propde a conserva variabilitatea ce carac spatial causal initial si care sunt necorelate, vom proceda la rotattii successive ale celor 2 axxe initiale, masurand variant ce carac cele 2 var pt fiecare pozitie modificata prin rotatie a sistemului de axe. Ca urmare a faptului ca sis de axe este rotit cu un anumit nr de grade, coordonatele celor 2 var se modif in mod cores psi astfel repre obs in noile coordonate este diferita.

Obs: - rotatia axelor initiale cu un anumit unghi in vedere maximizarii variantei de-a lungul unei axe nu modif pozitia sau configuratia pct repre de observatiile originale, ceea ce se schimba sunt doar coordonatele acestora in raport cu noile axe;

- noile axe rezultate in urma rotatiei care maximizaeaza variant dupa prima axa def 2 noi variabile, numite componenete principale si care au prop ca au media nula

- cele doua variabile sunt numite comp pprincipale repre comb liniare de var originale si sunt necorelate intre ele

- coordonatele noilor variabile sunt proiectiile pct repre de observatiile intiale si se num scoruri ale comp principale

- cele 2 comp principale conserva integral variant totala coresp variabilelor originale, adica suma variantelor celor 2 comp pricipale este egala cu suma variantelor celor 2 var originale

- prima componeneta principal are variant maximal, preluand maximum posibil din variant totala ce carac avr originale.

**18. Definiti componentele principale si mentionati proprietatile acestora**

Componentele principale sunt variabile vectoriale abstracte, definite sub forma unor comb liniare de var originale si care au urm 2 prop fundamentale:

- sunt necorelate 2 cate 2 si suma patratelor coef care def combinatia liniara ce coresp unei comp principale este egala cu unitatea

- prima comp principala este o combinative liniara normalizata a carei variant este maxima, cea de-a doua comp principal este o combinative liniara necorelata cu prima comp principal si care are o var cat mai mare posibila, insa mai mica decat cea a primei comp, etc

Dpdv geometric, var numite comp principale def un nou spatiu al obiectelor, in comtextul caruia sunt verif urm prop relevante pt def ACP:

- axele noului spatiu sunt ortogonale 2 cate 2 si def noile var numite comp principale

- coordonatele obiectelor in noul spatiu, adica proiectiile obiectelor pe axele acestuia, sunt evaluari ale obiectelor in raport cu noile variabile si se numesc scoruri ale comp principale sau scoruri principale

- dpdv teoretic, nr de comp principale este egal cu nr de var originale; nu toate comp principale au insa o semnificatie inf considerabila, astfel cele mai putin semnificative dpdv inf sunt eliminate

- comp principale sunt combinatii liniare de varianta maximal ale variabilelor originale

- comp principale sunt scalate in fct de magnitudinea variantei acestora, prima fiind componenta principal cu variant maxima, iar ultm fiind comp principal cu variant min

- comp principale sunt necorelate 2 cate 2

- suma variantelor comp principale coincide cu suma variantelor variabilelor originale , ai comp principale preiau in totalitate variabilitatea continuta in var originale.

Cu ajutorul comp principal se poate definii o struc a dependentei dintre variabile originale mai simpla si mai clara, deci mai usor de interpretat.

**19. Formulati modelul matematic al analizei componentelor principale, definiti si interpretati marimile definitorii ale acestuia**

In scopul formularii modelului mathematic care sta la baza ACP , vom considera ca spatiul cauzal initial supus investigarii este det de un nr de n variabile explicative notate x1…xn. . Aceste variabile simbolizeaza carac ale obiectelor supuse analizari, ceea ce inseamna ca fiecare obiect este pp a fi carac de n variabile.

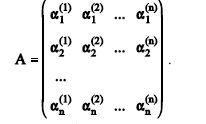
Activitatea de det a componenetelor principale poate fi descrisa prin intermediul unei transformari de tipul urm:  **Ψ: Rn →Rk** , unde Rn si Rk, sunt doua spatii vectoriale reale, iar dim celui de al 2-lea este mult mai mica decat dim primului.Problema de ACP din pct de vedere mathematic este echivalenta cu rezolvarea prob de extreme urm:

**OptA€Ma\*k Φ(x,w)**

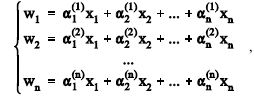
**SR: W=AT\*X**, unde opt poate fi de min sau de max.

O situatie specifica de solutionare a prob ACP este maximizarea variantei comp principale, ca masura a cantitatii de inf exprimata de fiecare dintre acestea.

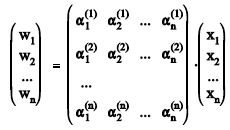
In scopul definirii modelului matematic al ACP, vom considera ca vectorii α(i) reprezinta coloanele unei matrici A de dimensiune n\*n de forma:



De asemenea , vom presupune ca x este vectorul ale carui coordinate sunt var originale x1, x2 …, si ca w este vectorul ale carui coordinate sunt comp principale w1, w2…. In aceste conditii combinatiile liniare care definesc comp principale pot fi scrise sub forma:



Sau sub forma matriceala :



Pe baza acestor notatii, modelul mathematic al ACP poate fi definit astfel:

mat4.JPG

Cele n coloane ale matricii A repre de fapt vectorii proprii normalizatti ai matricii de covarianta Σ, iar variant fiecarei component principale wi, care este o variant maximal in raport cu variantele comp prinncipale, este repre de val λi a aceleasi matrici de covarianta. Aceasta modalitate de determinare a elementelor matricii A este echivalenta cu calculul proiectiilor obiectelor de tip **x**€ **Rn** pe subspatiul liniar generat de vectorii coloanelor matricii A. Am vazut anterior ca cele n component principale ale spatiului causal det de variabilele originale x1, x2 … sunt definite de combinatiile liniare:

**Wi= α1(i)x1+…+ αn(i)xn, i=1..n**, ale caror ponderi **αj(i)** se det astfel incat sa maximizeze var comp principale Wi

**20. Ilustrati modul de deducere a componentelor principale**

Deducerea componentelor principale consta practic in a alege intre cele n valori proprii ale matricii de covarinta Σ pe cea mai mare si a determina componentele vectorului de ponderi α ce defineste respectiva componenta principala. Asdar, pentru fiecare valoare proprie din cele n valori proprii ale matricii de covarianta Σ, vom avea cate o solutie, adica cate un vector si deci cate o componenta principala .

**21. Definiti si justificati 3 dintre proprietatile componentelor principale.**

Componentele principale au o serie de proprietati extrem de interesante, care decurg chiar din modul lor de definire si care sunt importante pentru a intelege natura si continutul acestor constructii abstracte. Cunoasterea proprietatilor pe care le au componentele principale este deosibit de importanta in procesul de analiza a datelor, permitand stabilirea modificarilor induse asupra componentelor principale si asupra marimilor asociate acestora de catre transformarile aplicate asupra observatiilor variabilelor originale. Trei dintre acestea sunt:

1. **Distribuirea dupa legea normala**

In conditiile in care variabilele originale sunt repartizate normal, vectorul componentelor principale W este repartizat normal cu media **Atµ** si matricea de covarianta **Λ**, adica:

**w ~ N(Atµ, Λ)**, unde Λ este matricea diagonala ale carei elemente sunt valorile proprii **λ1, λ2,...., λn** ale matricii de covarianta Σ.

Normalitatea celor n variabile reprezentand componentele principale rezulta din faptul ca acestea sunt combinatii liniare de cele n variabile originale, care, prin ipoteza, sunt variabile normale.

1. **Conservarea variantei generalizate**

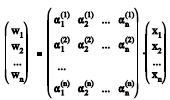
Componentele principale **w1 ,w2 ,......, wn**asigura conservarea integrala a variantei generalizate a variabilelor originale **x1,x2,....,xn**. Aceasta inseamna ca: **VG(x) = VG(w).** Aceasta proprietate evidentiaza calitatea informationala pe care o acu componentele principale de a reprezenta o reexprimare a variabilelor originale.

1. **Dependenta de unitatile de masura**

Componentele principale **w1,w2,....,wn** si variantele acestora depind de unitatile de masura in care sunt masurate variabilele originale **x1,x2,....,xn.** Aceasta inseamna ca, odata cu schimbarea unitatilor de masura ale variabilelor originale se schimba atat componentele principale, cat si variantele acestora.

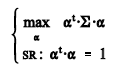
**22. Interpretati vectorii si valorile proprii ale matricii de covarianta**

Vom presupune ca x este vectorul ale carui coordonate sunt variabilele originale **x1,x2,....,xn** si ca w este vectorul ale carui coordonate sunt componentele principale **w1,w2,....,wn.** In aceste conditii, combinatiile liniare care definesc componentele principale pot fi scrise sub forma matriciala ca:

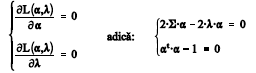


Cele n coloane alea matricii A reprezinta de fapt vectorii proprii normalizati ai matricii de covarianta Σ, iar varianta fiecarei componente principale **wi**, care este o varianta maximala in raport cu variantele componentelor principale anterioare, este reprezentata chiar de valoarea proprie **λi** a aceleeasi matrice de covarianta.

Problema determinarii componentelor vectorului α ce definesc combinatia liniara reprezentand componenta principala w se reduce la rezolvarea urmatoarea probleme de extrem cu legaturi:



Rezolvarea problemei se poate face cu ajutorul metodei multiplicatorilor lui Lagrange:

Rezulta ca solutia α a problemei de extrem este chiar unui din vectorii proprii ai matricii de covarianta Σ, anume cel asociat valorii proprii λ a aceleiasi matrici. Mai mult se observa ca valoarea maxima a formei patratice αt \* Σ \* α este, in punctul de extrem α, egala cu λ, respectiv:



Aceasta ultima relatie evedentiaza faptul ca varianta unei componente principala este egala cu o valoare proprie a matricii de covarianta.

**23. Ce sunt scorurile principale si cum se determina acestea. De ce este necesara determinarea scorurilor principale**

In analiza componentelor principale coordonatele obiectelor in spatiul redus se mai numesc si **scoruri principale ale obiectelor.** Daca vom presupune ca au fost retinute p componente si daca vom nota cu **Ū** matricea de dimensiune **n x p**, ale carei coloane sunt cei p vectori proprii care definesc cele p componente principale, atunci matricea scorurilor poate fi determinata astfel:

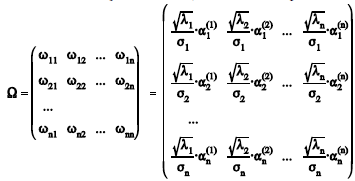


Liniile matricii W reprezinta scorurile corespunzatoare noilor variabile sau observatiile celor p componente principale. O data determinate, scorurile principale pot fi folosite in analiza ca substituit al observatiilor originale, simplificand, in acest fel, baza informationala initiala. In legatura cu aceasta problema, scorurile principale sunt mai potrivite pentru a fi folosite in analize deoarece sunt mai putin afectate de erori, in comparatie cu masuratorile originale. Faptul ca scorurile principale sunt mai robuste in raport cu perturbatiile introduse de erori, ca au o anumita invarianta in raport cu erorile, le face sa devina mai importante din punct de vedere informational decat observatiile originale.

**24. Ce este matricea factor (matricea de corelatie intre variabilele originale si componentele principale). Cum se calculeaza si cum se interpreteaza elementele sale**

O matrice importanta utilizata in contextul analizei componentelor principale, ale carei elemente ofera premize pentru interpretari interesante este matricea factor. Vom presupune ca cele n componente principale sunt reprezentate prin intermediul vectorului w, iar matricea de covarianta a componentelor principala este matricea diagonala Λ. De asemenea vom considera legatura dintre vectorul variabilelor originale si vectorul componentelor principale ca fiind data de relatia: **x = A \* w ,** unde A este matricea vectorilor proprii ai matricei de covarianta Σ.

Matricea Ω este o matrice foarte importanta pentru analiza componentelor principala si este cunoscuta sub numele de matrice factor. Aceasta matrice are forma:



un element generic ωij din matricea factor Ω fiind determinat de relatia:



Elementele matricii factor Ω se numesc intensitati ale factorilor si au o interpretare deosebit de interesanta din pdv al legaturii dintre variabilele originale **x1,x2,....,xn** si componentele principale **w1,w2,....,wn.** Astfel, elementul care se gaseste la intersectia liniei i cu coloana j in matricea factor Ω, adica elementul , reprezinta coeficientul de corelatie dintre cea de-a i-a variabila standardizata xi si cea de-a j-a componenta principala wj. Intensitatile factorilor sunt indicatori ai masurii in care variabilele originale participa la formarea componentelor principale sau, mai corect, ai masurii in care componentele principale sintetizeaza informatia continuta in variabilele originale. Cu cat este mai mare valoare coeficientului de corelatia dintre o variabila originala si o componenta principala, cu atat este mai adecvata si mai completa exprimarea informationala a variabilei originale prin intermediul componentei principale respective.

Matricea factor este foarte importanta deoarece, pe baza analizei valorilor elementelor ei, pot fi identificate o serie de partitii sau cluster-e pe multimea variabilelor, clustere care, asociate cu anumite componente principale, pot conduce la stabilirea unor semnificatii intuitive pentru acele componente. Aceasta inseamna ca analiza elementelor matricii factor Ω poate permite identificarea acelor variabile originale care sunt reprezentate prin intermediul unei anumite componente principale si, pe aceasta baza, crearea posibilitatii de atribuire a unei semnificatii concrete pentru fiecare componenta principala.

**25. Detaliati modul în care pot fi interpretate componentele principale în termeni cu semnificatie concreta. Exemplificati**

Interpretarea componentelor principale este facilitata de graficul “cercul corelatiilor” care este proiectia sferei unitate F, pe un plan delimitat de c1 si c2, doua componente principale in spatiul F, ale caror coordonate reprezinta coeficienti de corelatie ale variabilelor initiale cu componentele considerate doua cate doua.

*Criterii de alegere a numarului de componente principale*

Eficienta exprimarii variabilelor originale prin intermediul componentelor principale este strans legata de gradul de corelare a variabilelor originale si, mai ales, de felul in care sunt structurate aceste variabile din punct de vedere al corelarii. In legatura cu gradul de corelare a variabilelor originale, se poate face o observatie extrem de interesanta din punct de vedere teoretic si foarte utila din punct de vedere practic. Aceasta obseervatie se refera la faptul ca exista o puternica legatura intre gradul de corelare a variabilelor originale si numarul de componente principale cu ajutorul carora pot fi reexprimate in mod eficient variabilele originale.

Daca pe multimea variabilelor originale se evidentiaza cu claritate existenta unor submultimi formate din variabile care au proprietatea ca sunt foarte puternic corelate intre ele, pe de o parte, si foarte slab corelate cu variabile apartinand altor submultimi, pe de alta parte, atunci se poate face afirmatia ca variabilele originale pot fi reexprimate suficient de bine printr-un numar de componente principale egal cu numarul de submultimi de acest fel.

**27. Ce este analiza factoriala si ce tipuri de probleme pot fi rezolvate cu ajutorul acesteia**

Analiza factoriala este o analiza multivalenta care are ca scop explicarea corelatiilor manifestate intre o serie de variabile, numite indicatori sau teste, prin intermediul unui numar mai mic de facrori ordonati si necorelati, numiti factori comuni. Aceasta este folosita pentru rezolvarea problemelor legate de:

• studierea *nivelelor diferite de manifestare a interdependentelor* dintre variabilele explicative;

• *detectarea unei structuri simplificate si clare a relatiilor de interdependenta* existente între variabilele explicative;

• *obtinerea unei* "*cluster-izari*", *unei clasificari* a variabilelor explicative prin intermediul unor entitati numite *factori*;

• obtinerea unor informatii specifice, sub forma asa-numitilor *factori*, pe baza carora sa se poata face o interpretare sintetica a relatiilor de cauzalitate;

• verificarea unor ipoteze cu privire la existenta unei structuri factoriale particulare sau cu privire la existenta unui anumit numar de factori comuni;

• *sintetizarea potentialului cauzal comun* al mai multor variabile explicative sub forma unui numar cât mai redus de factori.

**28. Structura generala a modelului de analiza factoriala**

Procesul de analiza factoriala include urmatoarele etape esentiale:

• determinarea numarului *minimal* de factori comuni cu ajutorul caruia pot fi explicate în mod *optimal* corelatiile existente între variabilele indicator;

• efectuarea unor *rotatii* ale factorilor, în scopul determinarii *solutiei factor* sub cea mai simpla si mai clara forma;

• estimarea *intensitatilor factorilor*, *structurii legaturilor*, *comunalitatilor* si *variantelor factorilor unici*;

• deducerea unor *interpretari* adecvate pentru factorii comuni;

• estimarea *scorurilor factorilor*.

Dintre acestea, problema care ridica cele mai multe dificultati în executarea acestei analize este aceea de estimare a intensitatii factorilor comuni.

**29. Definiti si interpretati descompunerea variabilitatii în contextul analizei factoriale**

Analiza factoriala isi propune sa reexprime variabilitatea continuta in spatiul initial, intr-o maniera diferentiata, in functie de rolul pe care il au in formarea acesteia factorii comuni, pe de-o parte, si factorii unici, pe de alta parte.

Prin utilizarea tehnicilor de analiza multidimensionala care au ca scop reducerea dimensionalitatii, variabilitatea spatiului cauzal n-dimensional este conservata prin intermediul variabilitatii reduse, induse de factorii comuni. Acestia, impreuna cu factorul unic determina un spatiu numit spatiu test sau spatiu factor ale carui axe sunt ortogonale doua cate doua.

Variabilitatea ce caracterizeaza cele doua spatii este masurata prinintermediul ***variantei*** sau ***dispersiei***. În analiza datelor se considera ca o variabila este cu atât mai semnificativa cu cât variabilitatea sa este mai mare.

In analiza factoriala, variabilitatea spatiului cauzal initial este considerata a fi o compunere de variabilitati, formata sub influenta factorilor considerati. Din acest punct de vedere, varianta poate fi impartita in treicomponente importante: componenta comunalitate, componenta unicitate si componenta rezidualitate sau eroare. Astfel varianta variabilei indicator se scrie: Var()= Comunalitate+Unicitate+Rezidualitate. Acesta relatie defineste descompunerea variantei unei variabile indicator in functie de variantele celor trei categorii de factori care influenteaza variabila respectiva. Cu exceptia celei de-a treia componente a descompunerii, care este varianta factorului rezidual, primelor doua componente nu li se poate atribui calitatea de variante ale factorilor. Ele sunt determinate de coeficientii care pondereaza variantele factorilor, ceea ce inseaman ca ele reprezinta contributii ale varinatelor factorilor la formarea variantei variabilei indicator.

**30. Ce sunt scorurile factor, cum se calculeaza si cum se interpreteaza acestea**

Un scor factor reprezinta o forma de determinare a unei observatii corespunzatoare unui factor dat, acesta formandu-se pe baza contributiei variabilelor originale.

Scorurile pentru un anumit factor se exprima matriceal astfel: f=x. In mod practic, exprimarea sub forma scorurilor factor a celor T observatii efectuate asupra variabilei originale se bazeaza pe urmatoarea relatie: k=1...p, j=1...T

Unde: reprezinta scorurile factor, este elementul din linia k si coloana i a transpusei matricii factor, iar este cea de-a j-a observatie efectuata asupra celei de-a i-a variabile originale.

Considerand matricea de observatii X si matricea factor F, atuncimatricea Z definita Z=, se numeste matricea scorurilor factor, ce pot fi utilizate in analize ulterioare in locul variabilelor originale.

**31. Metode de estimare a modelului factorial**

Utilizarea analizei factoriale pentru rezolvarea unor probleme specific presupune si determinarea numarului de factori comuni ce vor fi retinuti in model. Desi decizia de a retine numarul de factori este subiectiva, exista o serie de criterii:

* **Criteriul procentului de acoperire**
  + - alegerea nr de factori care sa fie inclusi in modelul factorial depinde de proportia din variabilitatea comuna continuta in spatiul cauzal initial pe care utilizatorul doreste s-o exprime prin intermediul unei succesiuni de factori comuni
    - O estimatie aproximativa a acestei proportii, pentru cazul in care numarul de factori retinuti este egal cu k, poate fi obtinuta cu formula



k – nr de factori retinuti de modelul

n – nr de variabile originale

λi – valoarea proprie in raport cu care este definit factorul comun i

* + - Dezavantajul este dat de faptul ca marimea pk arata care este ponderea variantei primelor k component principale in varianta totala si nu ponderea variantei explicate de primii k factori in variant spatiului test, si reprezinta un inconvenient deoarece intre componentele principale si factorii comuni exista o deosebire de esenta
* **Criteriul lui Kaizer**
  + - Poate fi folosit atunci cand analiza factoriala este efectuata pe o matrice de corelatie, adica atunci cand se presupune ca variabilele originale sunt standardizate. Conform cu acesta, nr de factori necesari a fi inclusi intr-un model de analiza factoriala este egal cu nr de valori proprii mai mari sau egale cu 1
    - Justificarea acestuia ar fi ca pentru analiza prezinta importanta numai acei factori comuni a caror varianta este cel putin egala cu variant variabilelor originale, variabile care, fiind normalizate, au varianta unitara
    - Poate fi folosit doar in cazul in care se lucreaza cu variabile normalizate iar dezavantajul ar fi ca aplicarea acestuia conduce la retinerea in model a unui nr prea mare de factori
* **Criteriul “granulozitatii”**
  + - Dupa acest criteriu, numarul de factor ice vor fi retinuti in modelul de analiza factoriala se stabileste pe baza unei analize grafice a valorilor proprii. Graficul pe care se face analiza se construieste luand in abscisa nr d ordine al valorilor proprii iar in ordonata valorile acestor valori proprii. Astfel graficul va avea forma aprox a unei curbe de tipul exponentialei negative, deoarece valorile proprii sunt ordonate dupa magnitudinea lor descrescatoare
    - Nr de factori ce se vor retine in model este determinat de pct de pe graphic in dreapta caruia panta curbei devine neglijabila, nr de ordine al valorii proprii corespunzatoare acestui punct determinand nr de factori ce se vor retine.
    - Dezavantajul consta in faptul ca aplicarea sa conduce la retinerea unui nr prea mic de factori comuni dar in practica, construirea unui model cu unul sau doi factori comuni, are si avantajul ca faciliteaza reprezentarea grafica a marimilor analizei factoriale, utila in faza de interpretare

**32. Definiti recunoasterea formelor si exemplificati câteva dintre aplicatiile acesteia în domeniul economico-financiar.**

**Teoria recunoa**s**terii formelor** poate fi definita ca reprezentând totalitatea *normelor, principiilor*, *metodelor* si *instrumentelor* de analiza si decizie utilizate în scopul de a identifica *apartenenta* unor forme sau obiecte (unitati,fenomene, evenimente, actiuni, procese etc.) la anumite clase cu individualitate bine determinata.

Se poate spune ca recunoasterea formelor însumeaza toate încercarile de construire a acelor modele care simuleaza modul în care omul cuantifica, analizeaza, interpreteaza si anticipeaza comportamentul evolutiv al fenomenelor si proceselor. Din punct de vedere al teoriei sistemelor, recunoasterea formelor poate fi privita ca un sistem general în care *intrarile* reprezinta multimea caracteristicilor obiectelor ce urmeaza a fi clasificate, *iesirile* reprezinta multimea claselor posibile din care pot face parte obiectele analizate, iar *functia de transfer* exprima mecanismul decizional prin care un anumit obiect este identificat ca facând parte dintr-o anumita clasa.

În domeniul economico-social teoria recunoasterii formelor îsi gaseste o larga utilizare mai ales în procesul de *analiza a datelor* si în activitatea de *predictie.* Problema clasificarii unei multimi de obiecte este o problema standard, frecvent întâlnitaîn investigarea socio-economica, iar abordarea ei presupune utilizarea metodelor si tehnicilor specifice teoriei recunoasteriiformelor.

Numeroase probleme din domeniul analizei datelor, începând cu cele legate de *identificarea caracteristicilor definitorii* pentru cele mai diverse categorii de fenomene si terminând cu cele legate de *delimitarea functionala, ierarhizarea structurala* sau *sintetizarea informationala* a unor multimi de fenomene si procese economico-sociale.Metodele si tehnicile apartinând teoriei recunoasterii formelor sunt de neînlocuit în analizele care opereaza cu cantitati maride informatie, unde necesitatea de a *esentializa* si *sintetiza* interdependentele implica un proces continuu de clasificare sistructurare a informatiilor.

O utilizare si mai larga a teoriei recunoasterii formelor este întâlnita în domeniul predictiilor. Activitatea de realizare a predictiilor poate fi privita ca un proces ale carui caracteristici sunt foarte apropiate, mergând chiar pâna la identificare, de caracteristicile specifice ale unui proces de recunoastere a formelor. Evaluarea starilor pe care le poate avea în viitor un fenomen apartinând unei realitati date reprezinta, de fapt, un proces de recunoastere a acelor forme de evolutie a fenomenului care au cea mai mare probabilitate de realizare. Mai mult, atât în activitatea de predictie, cât si în procesul de clasificare sau de recunoastere a formelor, modalitatile de abordare au o natura preponderent probabilistica. Pe de alta parte, problema recunoasterii formelor este, ea însasi, o problema de predictie în care, pornind de la anumite caracteristici ale obiectelor analizate, obiecte numite si *forme*, se fac predictii cu privire la apartenenta acestor obiecte la anumite clase. De altfel, stabilirea apartenentei formelor la anumite clase reprezinta scopul principal al utilizarii tehnicilor de recunoastere a formelor.

In prezent, cele mai moderne metode si tehnici din domeniul predictiei sunt cele bazate pe o noua clasa de modele, specifice conturarii unei noi modalitati de abordare în domeniul teoriei recunoasterii formelor, numite *retele neuronale*. Modalitatile de abordare stiintifica bazate pe retele neuronale sunt mult mai concordante cu pronuntata complexitate si imprevizibilitate ce caracterizeaza comportamentul fenomenelor si proceselor economico-sociale si ofera o serie de avantaje importante, în comparatie cu alte metode si tehnici utilizate în acelasi scop.

Tehnicile de recunoastere a formelor pot fi utilizate în domeniul economico-social pentru rezolvarea unor probleme cum ar fi: analiza datelor cu grad ridicat de eterogenitate, fundamentarea criteriilor de alegere a proiectelor de dezvoltare, clasificarea deciziilor în functie de impactul acestora asupra diverselor compartimente ale vietii economico-sociale, detectarea unor perioade cu caracter specific din evolutia unor sisteme economice, stabilirea politicilor de creditare în domeniul financiar-bancar, evaluarea eficientei activitatilor de promovare a unor produse, determinarea perioadelor cele mai potrivite pentru vânzarea anumitor sortimente de marfuri, identificarea celor mai profitabile domenii de afaceri, clasificarea si ierahizarea unor entitati economico-sociale etc.

**33. Definiti principalele concepte ale recunoasterii formelor**

Dintre numeroasele conceptele utilizate în teoria recunoasterii formelor, trei pot fi considerate ca fiind fundamentale si definitorii pentru esenta si scopurile teoriei recunoasterii formelor: **forma, clasa si clasificatorul**.**Forma**reprezinta expresia numerica a obiectului studiat în vederea clasificarii lui într-o anumita clasa si este rezultatul cuantificarii principalelor caracteristici posedate de obiectul respectiv. **Forma** sau **obiectul**este o entitate informationala individuala, caracterizata prin intermediul unui vector n-dimensional, ale carui componente definesc valorile caracteristicilor acesteia, si care face obiectul procesului de clasificare sau de predictie.

Una dintre ipotezele fundamentale pe care se bazeaza teoria recunoasterii formelor este aceea ca obiectele analizate sunt caracterizate de un anumit grad de *eterogenitate*. Aceasta înseamna ca se asuma, în mod implicit, existenta posibilitatii definirii unor clase distincte pe multimea obiectelor. Pe de alta parte, se mai presupune ca anumite obiecte apartinând multimii analizate au ceva comun, sunt caracterizate printr-un anumit grad de omogenitate. Cele doua cerinte impuse multimii obiectelor analizate sunt cunoscute sub numele de *similaritate* si *disimilaritate*.

**Clasa, grupa** sau **clusterul**reprezinta o submultime distincta de obiecte care verifica urmatoarele doua proprietati: obiectele care alcatuiesc o clasa sunt *omogene din punct de vedere al caracteristicilor* lor definitorii; doua obiecte între care exista *diferente semnificative* din punct de vedere al caracteristicilor definitorii fac parte din clase diferite. **Clasa, grupa** sau **clusterul**reprezinta o entitate informationala *distincta* si cu *semnificatie concreta,* formata din totalitatea obiectelor ale caror caracteristici sunt identice sau difera foarte putin si care sunt semnificativ diferite de caracteristicile obiectelor din alte clase sau grupe.

Numarul de clase care alcatuiesc multimea de iesire a unui sistem de recunoastere a formelor variaza în functie de specificul domeniului pentru care se foloseste acest sistem si de scopurile urmarite.

**Clasificatorul** este un model statistico-matematic care, pe baza informatiilor referitoare la caracteristicile unui anumit obiect, determina decizia de clasificare a obiectului într-o anumita clasa. Clasificatorul poate fi privit ca fiind setul de principii, reguli sau criterii, în functie de care obiectele analizate sunt atribuite unei clase sau alteia. **Clasificatorul** sau **criteriul de clasificare**reprezinta regula sau multimea de reguli pe baza carora obiectele care apartin multimii analizate sunt afectate sau atribuite unor clase sau grupe bine definite. În functie de natura regulilor utilizate în procesul de clasificare, exista mai multe categorii de clasificatori: *clasificatori ierarhici,* *clasificatori de cost minim, clasificatori de distanta minimala*, *clasificatori de tip Bayes-ian*, *clasificatori euristici* etc.

**34. Formulati problema generala a clasificarii**

Sub cea mai generala forma a sa, problema de clasificare poate fi formulata în termenii *teoriei deciziei*, iar metodele de clasificare pot fi definite sub forma unor *instrumente decizionale* specifice. Vom descrie în continuare modul în care problema de clasificare poate fi definita ca o problema decizionala. În acest scop, vom presupune existenta unei populatii de forme sau de obiecte, notata cu **Ω** si definita sub forma:



M reprezinta nr de unitati ale populatiei analizate

Fiecare obiect ce alcatuieste populatia **Ω** este definita prin intermediul unui numar N de caracteristici,  denumite variabile explicative. In acest fel populatia poate fi reprezentata caun vector N-dimensional:



Variabilele explicative, ce definesc caracteristicile obiectelor analizate, sunt marimile in functie de care se stabileste apartenenta unui obiect din populatie la una din clasele acestuia. Variabilele explicative pot fi de tip calitativ sau cantitativ. Ele pot fi masurate pe cele patru scale cunoscute, respectiv scala nominala, ordinala, interval sau raport.

Dintre elementele care reprezinta variabilele explicative unele pot sa aiba o *putere de discriminare mai redusa*, iar altele pot sa aiba o *putere de discriminare mai mare*. Din acest punct de vedere, în construirea algoritmilor de clasificare trebuie sa fie *selectate* acele variabile care au puterea de discriminare cea mai mare. Variabilele cu puterea de discriminare cea mai mare, definesc acele caracteristici ale obiectelor care permit o diferentiere mai puternica a claselor în care pot fi grupate respectivele obiecte si se numesc *variabile descriptor*. Pentru un anumit obiect, vectorul de valori ale variabilelor descriptor reprezinta chiar *forma* asociata respectivului obiect.

În raport cu o manifestare sau cu o actiune viitoare, elementele populatiei se pot gasi într-una din mai multe stari potentiale, numite *stari ale naturii*. Starile naturii reprezinta conjuncturi fizice, economice sau sociale, în raport cu care multimea de obiecte analizate se structureaza sub forma unor categorii bine individualizate. Starile naturii se caracterizeaza prin exhaustivitate si exclusivitate reciproca. Aceasta inseamna ca inafara acestor stari ale naturii, nu mai poate exista nicio alta stare posibila a naturii, respectiv ca doua stari diferite ale naturii nu se pot manifesta niciodata simultan.

Caracteristica principala a unei probleme de clasificare consta în faptul ca desi starile posibile ale naturii sunt cunoscute aprioric, ca numar, ca natura si ca plauzibilitate a manifestarii, iar fiecare element al populatiei se gaseste în mod sigur întruna, si numai într-una, din aceste stari, de obicei nu se cunoaste, cu precizie si în mod aprioric, în care dintre starile naturii se gaseste fiecare dintre unitatile populatiei.

***Problema generala******a clasific****a****rii****: Fiind data o multime de obiecte, se cere sa se determine criteriul sau regula care sa descrie apartenenta obiectelor la clasele sub forma carora se structureaza respectiva multime de obiecte.*

În functie de cunoasterea sau necunoasterea apriorica a apartenentei la clasele obiectelor care apartin esantionului extras din populatia , metodele de clasificare se împart în doua mari categorii: de *clasificare controlata* si de *clasificare* *necontrolata*. Odata ce criteriul de clasificare a fost stabilit, el poate fi folosit, în continuare, pentru efectuarea de *predictii* privind apartenenta la o anumita clasa a unor noi obiecte, din afara esantionului existent, obiecte a caror apartenenta nu este cunoscuta aprioric. Dupa ce criteriul de clasificare a fost identificat, si cu conditia ca apartenenta obiectelor apartinând esantionului disponibil sa fie cunoscuta, el poate fi utilizat si pentru verificarea corectitudinii cu care acesta poate face clasificarea, adica pentru *testarea* calitatii clasificatorului. Modul în care un clasificator asigura clasificarea obiectelor cu apartenenta cunoscuta poate fi descris prin intermediul unei matrici, numita *matricea corectitudinii clasificarii* sau, mai simplu, *matricea clasificarii*, care contine informatiile necesare pentru a aprecia corectitudinea clasificarii obiectelor.

**35. Definiti sistemele de recunoastere controlata si necontrolata**

Totalitatea activitatilor implicate într-un proces de recunoastere a formelor, ansamblul informatiilor manipulate în acest context si multimea procedurilor, algoritmilor, metodelor si tehnicilor utilizate în acest scop, sunt privite ca reprezentând un sistem, numit ***sistem de recunoastere a formelor***.

Exista doua tipuri fundamentale de sisteme de recunoastere a formelor: *sisteme de recunoastere necontrolata* si *sisteme de recunoastere controlata.* Aceste doua tipuri de sisteme de recunoastere a formelor sunt determinate de scopurile urmarite,de natura informatiilor pe care le prelucreaza, de specificitatea metodelor si intrumentelor utilizate, precum si de naturarezultatelor obtinute cu ajutorul acestora.

* **Sistemul de recunoastere necontrolata**

*Sistemele de recunoastere necontrolata a formelor* sunt sistemele în cadrul carora nu se dispune de informatii initiale referitoare la numarul de clase si la apartenenta formelor la anumite clase, construirea claselor facându-se progresiv, pe masura cresterii numarului de forme analizate, iar numarul de clase posibile fiind stabilit doar în faza finala a procesului de recunoastere.

Caracteristica principala a sistemelor de recunoastere necontrolata a formelor consta în faptul ca *nu se cunoaste apartenenta obiectelor analizate la o clasa sau alta*. Aceasta înseamna ca, în mod implicit, nu se cunoaste *cu precizie* nicinumarul de clase.

Principiile, procedurile, metodele si tehnicile apartinând sistemelor de recunoastere necontrolata a formelor sunt cunoscute sub denumirea generala de *tehnici de clasificare*, *clasificare nesupervizata* sau *analiza cluster*.

*Analiza cluster* este o tehnica de clasificare caracterizata prin faptul ca afectarea formelor sau obiectelor în clustere sau grupe se face progresiv si fara a cunoaste aprioric numarul de clase, în functie de verificarea a doua criterii fundamentale:

a. obiectele sau formele clasificate în fiecare clasa *sa fie cât mai similare* din punct de vedere al anumitor caracteristici;

b. obiectele clasificate într-o clasa *sa se diferentieze cât mai mult* de obiectele clasificate în oricare din celelalte clase.

Primul criteriu de afectare a formelor pe clase cere ca fiecare clasa sa fie cât mai omogena în raport cu caracteristicile luate în considerare pentru clasificarea obiectelor. Cel de-al doilea criteriu cere ca fiecare clasa sa difere cât mai mult din punct de vedere al caracteristicilor de clasificare. În functie de caracteristicile procedurilor pe care le utilizeaza, de ipotezele initiale pe care se bazeaza si de natura rezultatelor obtinute cu ajutorul lor, metodele de analiza cluster se împart în doua mari categorii: *metode de clusterizare ierarhica* si *metode de clasificare prin partitionare* sau *metode iterative*. Prima categorie include metodele de clusterizare prin *agregare* si metodele de clusterizare prin *divizare*. Pentru fiecare dintre cele doua tipuri de clusterizare exista mai multe proceduri specifice, între care mentionam: metoda *agregarii simple*, metoda *agregarii complete*, metoda *agregarii medii*, metoda lui *Ward* etc. Cea de-a doua categorie include o serie de algoritmi, între care mentionam: algoritmul celor *K-medii*, algoritmul celor *Kmedoizi*, algoritmul *CLARA*, algoritmul *fuzzy* etc.

În ceea ce priveste rezultatele furnizate de sistemele de recunoastere necontrolata a formelor, precizam ca iesirile acestor sisteme nu se reduc, de regula, la o unica si simpla configurare a obiectelor analizate pe clase, ci includ *mai multe variante* de configurare a obiectelor pe clase, variante continute într-o entitate informationala numita *structura cluster* sau *ierarhie cluster*. Ierahia cluster ofera posibilitatea cercetatorului de a alege o anumita configurare a obiectelor pe clase, ceea ce înseamna, implicit, si alegerea unui anumit numar de clase.

Sistemele de recunoastere necontrolata sunt utilizate mai mult pentru scopuri de sistematizare, grupare si sintetizare informationala, în situatiile în care sunt analizate cantitati foarte mari de date si aceste date se caracterizeaza printr-un grad ridicat de eterogenitate. În acest sens, tehnicile de recunoastere necontrolata a formelor sunt foarte utile si eficiente în activitatile de analiza preliminara a datelor. Utilizarea analizei cluster în aceasta faza a analizei datelor este importanta deoarece ea permite

organizarea mai eficienta a datelor eterogene. Regasirea informatiilor în cadrul masivelor de date structurate cu ajutorul tehnicilor de analiza cluster devine mult mai usoara, iar datele pot fi interpretate mult mai consistent.

* **Sistemul de recunoastere controlata**

*Sistemele de recunoastere controlata a formelor* sunt acele sisteme în cadrul carora se presupune existenta apriorica a unui *numar dat de clase* si a unui *set de forme*, numite *prototipuri* sau *referinte*, a caror apartenenta la aceste clase este cunoscuta. Acest set de forme este reprezentat de esantionul de obiecte extrase din populatia supusa studiului, esantion cunoscut si sub numele de *set de formare* sau *set de învatare*.

*Setul de formare* sau *setul de învatare* este un esantion de forme extrase din populatia studiata, forme a caror apartenenta la clasele populatiei este cunoscuta si pe baza carora sunt deduse criteriile formale de clasificare. În cadrul sistemelor de recunoastere controlata a formelor, datele reprezentate de setul de formare includ atât informatii referitoare la proprietatile esentiale ale obiectelor supuse analizei, cât si informatii referitoare la apartenenta acestor obiecte la clasele existente. Pe baza acestor informatii initiale, se deduc regulile si criteriile de decizie pentru partitionarea sub forma de *regiuni* sau *clase* a multimii de obiecte supusa studiului sau a spatiului în care iau valori caracteristicile obiectelor. De fapt, în cazul tehnicilor de acest fel informatiile continute în setul de formare sunt folosite pentru a face *inferente* cu privire la împartirea populatiei totale pe clase. Mai mult decât atât, din aplicarea tehnicilor de clasificare controlata rezulta si un set de reguli si criterii formale de clasificare, adica un *clasificator*. Aceste reguli si criterii sunt folosite, în continuare, pentru clasificarea unor noi forme neclasificate înca, forme a caror apartenenta este necunoscuta, adica pentru a face predictii cu privire la apartenenta noilor forme.

În mod uzual, setul initial de forme este împartit în doua subseturi folosite în scopuri diferite: primul subset este numit *set* *de formare* si contine acele forme utilizate pentru deducerea regulilor si criteriilor de clasificare, adica pentru construirea *clasificatorului* propriu-zis; al doilea subset este numit *set de predictie* si contine acele forme utilizate pentru testarea clasificatorului construit pe baza setului de formare.

*Sistemul de recunoastere controlata a formelor* reprezintatotalitatea activitatilor si procedurilor care au ca scop deducerea unor criterii de *partajare* a unei populatii de entitati informationale (obiecte sau variabile), sub forma unui numar *cunoscut* de clase, pe baza cunoasterii caracteristicilor si a apartenentei elementelor unui esantion provenit din respectiva populatie.

Spre deosebire de tehnicile de clasificare necontrolata, care se bazeaza, în principal, pe utilizarea conceptului de distanta, elementul fundamental al tehnicilor de clasificare controlata este un model formal, numit *clasificator*. În cazul analizei discriminante, clasificatorul este reprezentat de *functiile discriminat* sau de *functiile de clasificare*.

**36. Ce este analiza cluster, care sunt conceptele fundamentale ale acesteia si care sunt domeniile utilizarii ei**

Analiza cluster are ca scop cautarea si identificarea de *clase*, *grupe* sau clustere în cadrul unor multimi de obiecte sau forme, astfel încât elementele care apartin aceleiasi clase sa fie cât mai *asemanatoare*, iar elementele care apartin la clase diferite sa fie cât mai *deosebite* între ele. Altfel spus, analiza cluster este o modalitate de examinare a similaritatilor si disimilaritatilor dintre obiectele apartinând unei anumite multimi, în scopul gruparii acestor obiecte sub forma unor clase *distincte* între ele si *omogene* în interior.

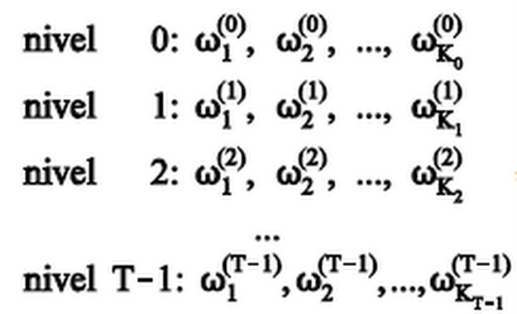
**Criteriu general de clasificare**: Clasificarea obiectelor în clase se face în asa fel încât sa se asigure o *variabilitate minima în interiorul claselor* si o *variabilitate maxima între clase*. Prin intermediul analizei cluster fiecare obiect din multimea analizata este atribuit *unei singure* clase, iar multimea claseloreste o multime *discreta* si *neordonabila*.

Clasele sau grupele sub forma carora se structureaza multimile de obiecte se mai numesc si *clustere*. Un cluster este o submultime formata din obiecte similare, adica din obiecte care sunt suficient de asemanatoare între ele din punct de vedere al caracteristicilor care le definesc.

Din punct de vedere geometric, ca multimi de puncte dintr-un anumit spatiu, clusterele pot avea forme foarte diferite, mai mult sau mai putin regulate. Astfel, forma clusterelor poate sa fie de tip *convex* sau *concav*, de tip *compact* sau de tip *alungit* etc. De regula, analizele de tip cluster reprezinta proceduri de clasificare de tip *necontrolat*, în care nu este cunoscuta aprioric nici apartenenta anumitor obiecte la anumite clase, nici numarul de clase posibile. Numarul de clase sau clustere este variabil si este stabilit *concomitent* cu activitatea de clasificare propriu-zisa.

***Analiza cluster*** poate fi definita ca reprezentând o multime de principii, metode si algoritmi de clasificare, având ca scop organizarea datelor sub forma unor structuri informationale semnificative, relevante. Analiza cluster este o *analiza explorativa*, de tip multidimensional, care are ca scop gruparea unor entitati informationale, cu natura fizica sau abstracta, în clase sau clustere alcatuite din entitati informtionale cu grad ridicat de similaritate. Din punct de vedere concret, efectuarea unei clasificari cu ajutorul metodelor si tehnicilor de analiza cluster consta în obtinerea unor *solutii cluster* sau a unor *partitii*, reprezentate de o multime de clase sau clustere notate cu ώ1, ώ2, ..., ώk.IN cazul anumitor metode de clasificare, rezultatele clasificarii sunt reprezentate de *solutii cluster unice*, în timp ce în cazul altor metode de clasificare, cum ar fi metodele de clasificare ierarhica de tip aglomerativ, sunt reprezentate de *multimi de solutii cluster*, numite *ierarhii de solutii cluster* sau *ierarhii de partitii*.

În analiza cluster, *ierarhiile cluster* sunt formate dintr-un numar de T solutii cluster, fiecare solutie continând clustere din ce în ce mai mari, respectiv clustere cu *niveluri de agregare din ce în ce mai ridicate*. O ierarhie cluster are o structura de forma urmatoare:



unde **T** este numarul de obiecte, iar **Ki** este numarul de clustere din solutia cluster de la nivelul i.

În cazul metodelor ierarhice aglomerative, numarul de clustere din prima partitie este egal cu numarul de obiecte, adica K0=T. De asemenea, numarul de clustere dintr-o partitie de la un anumit nivel este mai mic cu 1 decât numarul de clustere din partitia de la nivelul inferior si mai mare cu 1 decât numarul de clustere din partiyia de la nivelul superior, respectiv:



Desi folosirea tehnicilor de analiza cluster nu este specifica doar pentru anumite domenii de activitate, totusi, utilizarea cea mai frecventa a acestora este întâlnita în domeniul marketingului, în investigatiile de natura psihosociala sau în evaluarile econo-micosociale la nivel teritorial. În domeniul marketingului, se detaseaza aplicatiile tehnicilor de analiza cluster în studierea comportamentului consumatorilor. Aceste aplicatii vizeaza evaluarea sanselor pe care poate sa le aiba lansarea unui produs nou, identificarea unor noi piete, modalitatile de segmentare a pietii sau identificarea pozitionarii pe piata a produselor diferitilor producatori. În cazul determinarii pozitionarii pe piata a diferitelor marci ale unui produs, analiza cluster este folosita pentru a clasifica marcile de fabricatie, în functie de similitudinea sau disimilitudinea perceptiilor pe care le manifesta consumatorii fata de aceste marci. Pe baza modului în care se clasifica marcile si a caracteristicilor consumatorilor care îsi manifesta preferintele, un producator poate identifica marcile concurente si trasaturile specifice ale categoriilor de consumatori care prefera produsul acestui producator.

**37. Definiti scopurile analizei cluster si descrieti tipul informatiilor utilizate în analiza cluster**

Analiza cluster se deosebeste în mod fundamental de procedurile de natura statistica, cum ar fi cele care au ca scop verificarea semnificatiei, prin faptul ca ea nu se bazeaza si nu presupune îndeplinirea apriorica a nici unei ipoteze specifice. În consecinta, prin esenta sa, analiza cluster constituie un important si eficient instrument de *analiza exploratorie*. Se poate spune ca scopul general al analizelor de tip cluster este acela de creare a asa-numitelor *taxonomii* sau *tipologii*.

Constructia tipologiilor este bazata pe analiza *asemanarilor* si *deosebirilor* existente între obiectele unei multimi date:

• *alegerea unui numar optimal de clustere*, în functie de natura problemei de clasificare si de scopurile care se urmaresc;

• *interpretarea semnificatiei* clusterelor;

Rezultatele unei analize cluster sunt reprezentate fie de *o singura solutie cluster*, fie de *ierarhii cluster*, care contin diferite modalitati de configurare a obiectelor pe clase, adica *mai multe solutii cluster*.

Analiza cluster poate fi privita, în general, ca un instrument care are ca scop *reducerea* unor multimi de obiecte, sau chiar de variabile, la un numar mai restrâns de entitati informationale, care sunt clasele sau clusterele. În sensul sau obisnuit, ca ansamblu de metode si tehnici de *clasificare a obiectelor*, analiza cluster este o analiza efectuata în *spatiul variabilelor*. Într-adevar, cele mai multe utilizari ale tehnicilor de analiza cluster sunt cele care au ca scop clasificarea obiectelor, si nu clasificarea variabilelor. Analiza cluster poate fi utilizata atât pentru *clasificarea obiectelor*, cât si pentru clasificarea *variabilelor* care definesc obiectele. Spre deosebire de utilizarea analizei cluster pentru clasificarea obiectelor, situatie în care specificitatea este reprezentata de faptul ca distantele sunt evaluate pentru *perechi de obiecte*, în cazul utilizarii analizei cluster pentru clasificarea variabilelor, evaluarea distantelor se face pentru *perechi de variabile*.

**38. Definiti analiza cluster si aratati cum se clasifica metodele de analiza cluster**

***Analiza cluster*** poate fi definita ca reprezentând o multime de principii, metode si algoritmi de clasificare, având ca scop organizarea datelor sub forma unor structuri informationale semnificative, relevante.

Problema cea mai importanta a oricarui tip de analiza cluster este aceea a modului în care poate fi masurata *proximitatea*, respectiv *gradul de apropiere* sau *gradul de departare*, dintre obiecte si dintre clustere. Orice proces de clasificare a obiectelor este definit în raport cu o anumita *masura* a gradului de apropiere sau de departare dintre obiectele analizate, indiferent de metoda sau algoritmul pe care se bazeaza acest proces. Aceasta masura poate fi reprezentata fie de un indicator de similaritate, fie de un indicator de disimilaritate. Fiecare dintre cele doua categorii de indicatori va fi definita si analizata în continuare.

În general, masurarea gradului de proximitate dintre obiecte se face cu ajutorul a doua grupe de indicatori, cunoscute sub numele de *indicatori de similaritate* si *indicatori de disimilaritate.* Indicatorii de similaritate si de disimilaritate pot fi utilizati ca baza informationala în orice proces de clasificare datorita

faptului ca ei pot induce o *relatie de ordine* pe multimea perechilor de obiecte sau de variabile si, în consecinta, pot contribui la clasificarea obiectelor sau variabilelor. Cu cât valoarea unui indicator de similaritate este mai mare, cu atât obiectele sau variabilele pentru care acest indicator se evalueaza pot fi considerate a fi mai asemanatoare, respectiv mai apropiate. De asemenea, o valoare foarte mica a

indicatorului de similaritate evidentiaza faptul ca cele doua obiecte sau cele doua variabile sunt mai departate între ele.

*Indicatorii de disimilaritate* sunt marimi numerice care exprima cât de *deosebite* sau cât de *departate* sunt doua obiecte sau doua variabile. Indicatorii de disimilaritate se mai numesc si indicatori sau coeficienti de *deosebire* sau de *distantare* a obiectelor sau variabilelor. Cu cât valoarea unui indicator de disimilaritate este mai mare, cu atât cele doua obiecte sau cele doua variabile pentru care se calculeaza sunt mai diferite, adica mai distantate între ele. Cea mai importanta si cea mai utilizata categorie de indicatori de disimilaritate este reprezentata de indicatorii de tip *distanta*. Spre deosebire de indicatorii de similaritate, care pot fi cel mai bine utilizati pentru exprimarea gradului de proximitate dintre obiectele cu caracteristici de tip calitativ, indicatorii de disimilaritate sunt marimi mai potrivite pentru masurarea proximitatii în cazul obiectelor cu caracteristici de tip *cantitativ*.

Informatiile utilizate, în ultima instanta, în analiza cluster sunt reprezentate sub forma unor matrici simetrice de tip *obiecte×obiecte*, numite, dupa caz, *matrici de proximitate*, *matrici de similaritate, matrici de asociere, matrici de incidenta*, *matrici de disimilaritate* sau *matrici de distante*. Atât liniile, cât si coloanele matricilor de acest fel se refera la obiectele analizate, astfel încât numarul lor este egal cu numarul de obiecte supuse analizei. Elementele acestor matrici sunt marimi numerice care exprima proximitatea dintre perechile de *obiecte* care eticheteaza rândurile si coloanele matricilor.

**39. Definiti conceptul de distanta si descrieti cateva modalitati de evaluare a distantelor dintre forme**

*Indicatorii de disimilaritate* sunt marimi numerice care exprima cât de *deosebite* sau cât de *departate* sunt doua obiecte sau doua variabile. Indicatorii de disimilaritate se mai numesc si indicatori sau coeficienti de *deosebire* sau de *distantare* a obiectelor sau variabilelor. Cu cât valoarea unui indicator de disimilaritate este mai mare, cu atât cele doua obiecte sau cele doua variabile pentru care se calculeaza sunt mai diferite, adica mai distantate între ele. Cea mai importanta si cea mai utilizata categorie de indicatori de disimilaritate este reprezentata de indicatorii de tip *distanta*.

Prin natura lor numerica, variabilele de tip cantitativ, adica variabilele masurate pe scalele de tip raport, interval si, eventual, ordinal, permit o definire mai naturala a conceptului de distanta. Pentru variabilele de tip nominal, inclusiv variabilele de tip binar, distantele se calculeaza într-un mod specific, compatibil cu natura acestor variabile. Pentru evaluarea disimilaritatilor dintre obiectele ale caror caracteristici sunt de tip cantitativ sau dintre variabile de tip cantitativ, pot fi folosite mai multe tipuri de distante, cum ar fi: distanta *Euclidiana* (simpla, ponderata sau patrata), distanta *Manhattan*, distanta *Cebîsev*, distanta *Minkovski*, distanta *Camberra*, distanta *Mahalanobis,* distanta *Pearson*, distanta *Jambu* etc.

***• Distan****t****a Euclidian****a*

Distanta *Euclidiana*, care mai este cunoscuta si sub numele de *norma de tip ,* este distanta cea mai frecvent utilizata în problemele de analiza cluster. Ea se calculeaza ca *radacina patrata a sumei patratelor diferentelor coordonatelor* celor doua obiecte sau variabile pentru care se evalueaza distansa.

Distanta Euclidiana exprima proximitatea dintre obiecte ca distanta între doua puncte din spatiul Euclidian, respectiv ca distanta masurata *în linie dreapta*.

***• Distanta Manhattan***

Distanta *Manhattan*, numita si distanta *rectangulara*, distanta “C*ity-Block*“ sau *norma de tip* , se calculeaza ca suma a valorilor absolute ale diferentelor coordonatelor celor doua obiecte sau celor doua variabile analizate Deoarece diferentele de coordonate utilizate în calculul sau nu sunt *amplificate* printr-o ridicare la o putere, distanta Manhattan este mai robusta în raport cu prezenta în date a valorilor aberante. Distanya Manhattan poate fi calculata si în varianta *ponderata*, calculul facându-se în mod similar cu cel al distantei Euclidiene ponderate. De asemenea, distanta Manhattan poate fi utilizata în cazul în care obiectele au caracteristici care sunt masurate pe scala de tip interval si pe scala de tip raport.

***• Distanta Cebî****s****ev***

Distanta *Cebîsev*, cunoscuta si sub numele de “*maxim al dimensiunilor*“ sau *norma de tip* , este o distanta de tip *valoare absoluta* si se determina ca fiind valoarea maxima a valorilor absolute ale diferentelor dintre coordonatele obiectelor sauvariabilelor.Distanta Cebîsev poate fi utilizata atunci când se doreste ca doua obiecte sau variabile sa apara ca fiind diferite, daca eledifera chiar si doar din punct de vedere al unei caracteristici, respectiv al unui obiect.

***• Distanta Mahalanobis***

Distanta *Mahalanobis* este una dintre cele mai cunoscute, mai importante si mai frecvent utilizate distante. Ea este o forma generalizata a conceptului de distanta. Distanta Mahalanobis reprezinta singurul tip de distanta care ia în considerare, într-o maniera completa, *gradul de* *dispersare* al multimii de obiecte sau al multimii de variabile analizate, precum si *gradul de corelare* al respectivelor entitati informationale. Utilizarea distantei Mahalanobis este recomandata, mai ales în situatiile în care variabilele care descriu obiectele sunt corelate între ele. Distanta Mahalanobis este utilizata si în cazul tehnicilor de clasificare controlata, pe baza acestei distante fiind dezvoltat chiar un criteriu operational de discriminare.

**40&41. Formulati criteriul general al clasificarii si aratati cum se evalueaza variabilitatea inter si intra cluster (cazul uni/multi-dimensional)**

**Criteriu general de clasificare**: Clasificarea obiectelor în clase se face în asa fel încât sa se asigure o *variabilitate minima în interiorul claselor(intra)* si o *variabilitate maxima între clase(inter)*

Distanta dintre doua clustere este, de fapt, o distanta dintre doua *multimi* de puncte, adica o distanta mai dificil de evaluat. Ca distanta între doua *multimi de puncte*, distanta dintre doua clustere poate fi masurata cu ajutorul uneia dintre mai multe metode posibile. Dintre metodele propuse pentru evaluarea distantelor dintre clustere mentionam: *metoda celor mai apropiati vecini*, *metoda* *celor mai departati vecini*, *metoda distantei medii între perechi*, *metoda centroidului* si *metoda lui Ward* etc.

*Metoda celor mai apropiati vecini* evalueaza distanta dintre doua clustere ca distanta între doua obiecte, unul din primul cluster, iar celalalt din cel de-al doilea cluster, care sunt *cele mai apropiate* între ele în sensul distantei utilizate.

*Metoda celor mai departati vecini* evalueaza distanta dintre doua clustere ca distanta între doua obiecte, unul din primul cluster, iar celalalt din cel de-al doilea cluster, care sunt *cel mai departate* între ele în sensul distantei utilizate.

*Metoda distantei medii dintre perechi* evalueaza distanta dintre doua clustere ca *medie* a distantelor dintre oricare doua obiecte care apartin celor doua clustere, unul primului cluster, iar celalalt din celui de-al doilea cluster.

*Metoda centroidului* evalueaza distanta dintre doua clustere ca *distanta între centroizii* celor doua clustere

***Metoda lui Ward*** este o metoda de evaluare a distantei dintre doua clustere, care se bazeaza pe *maximizarea gradului de omogenitate a* clusterelor sau, ceea ce este acelasi lucru, pe *minimizarea variabilitatii intracluster*. De regula, gradul de omogenitatea unui cluster se considera a fi *cu atât mai mare*, *cu cât suma totala a patratelor abaterilor intracluster este mai mica*. Înacest sens, se poate spune ca distanta Ward dintre doua clustere masoara *variabilitatea intracluster cumulata, pe care o induce comasarea celor doua* clustere la nivelul configuratiei cluster rezultate.

**42. Evaluarea distanţelor dintre clustere**

O problemă dificilă care apare în analiza cluster, este legată de necesitatea evaluării distanţelor dintre clase sau clustere. Dificultatea acestei probleme este dată de faptul că distanţele dintre clase sau clustere sunt, de fapt, distanţe între mulţimi de obiecte sau distanţe între mulţimi de variabile.

Problema evaluării distanţelor dintre clustere apare în special în cazul analizei cluster de tip ierarhic, în care construirea arborelui de clustere poate fi făcută pe baza comasării succesive sau divizării succesive a clusterelor. Comasarea clustere lor este numită amalgamare sau agregare, iar divizarea clusterelor este numită dezagregare.

Teoretic, procesul de agregare sau dezagregare succesivă a clusterelor se bazează pe definirea unei distanţe limită între clustere, distanţă numită şi prag de agregare, respectiv prag de dezagregare. În principiu, decizia de comasare a două clustere sau de divizare a unui cluster este luată numai dacă distanţa dintre aceste clustere este mai mică, respectiv mai mare decât distanţa limită fixată.

Dacă în cazul evaluării gradului de apropiere sau depărtare dintre două obiecte lucrurile sunt relativ simple, fiind suficient să se calculeze una din distanţele menţionate mai sus, în cazul în care este necesar a fi evaluat gradul de apropiere sau depărtare dintre două clustere lucrurile devin ceva mai complicate şi presupun existenţa unei metode specifice de evaluare. Distanţa dintre două clustere este, de fapt, o distanţă dintre două mulţimi de puncte, adică o distanţă mai dificil de evaluat.  
Ca distanţă între două mulţimi de puncte, distanţa dintre două clustere poate fi măsurată cu ajutorul uneia dintre mai multe metode posibile.

Dintre metodele propuse pentru evaluarea distanţelor dintre clustere menţionăm: metoda celor mai apropiaţi vecini, metoda celor mai depărtaţi vecini, metoda distanţei medii între perechi, metoda centroidului şi metoda lui Ward etc.

**43.Analiza cluster de tip ierarhic**

Analiza cluster de tip ierarhic sau arboreseent este o metodă de clasificare bazată pe gruparea obiectelor pe bază de agregare succesivă în clase din ce în ce mai largi de obiecte sau de dezagregare succesivă în clase din ce în ce mai mici.

Analiza cluster presupune o organizare arborescenta pe grupe sau clase a structurilor de tip latent. Acesta organizare se face cu algoritmi de tip euristic sau algoritmi de tip model formal care generează structurile cluster pe baza maximizării verosimilităţii.

Rezultatul utilizării analizei cluster de tip ierarhic îl reprezintă o mulţime de structuri particulare de clustere, numită arbore al clasificării sau arbore ierarhic.

Structurile cluster de tip ierarhic sunt caracterizate prin nivele diferite de agregare, cuprinse între un nivel minim şi un nivel maxim. Pe masura ce nivelul ierarhic creste numarul de clustere scade

Structura cluster cu cel mai înalt nivel de agregare este formată dintr-un singur cluster, care include toate obiectele supuse clasificării. Structura cluster cu cel mai redus nivel de agregare este formată dintr-un număr de clustere egal cu numărul de obiecte analizare, fiecare cluster incluzând un singur obiect.

Cu cât nivelul de agregare al structurilor cluster este mai ridicat, cu atât asemanarile dintre obiectele unui cluster scad deaoare un cluster de la un nivel mai inalt contine mai multe obiecte decat un cluster de la un nivel inferior.

Algoritmii de clasificare de tip ierarhic pot fi împărţiţi în două mari categorii:

• algoritmi de clasificare prin agregare, amalgamare sau combinare;

• algoritmi de clasificare prin dezagregare sau divizare.

Algoritmii de dezagregare construiese clusterele într-o manieră descendentă, pornind cu toate obiectele într-un singur cluster şi continuând, prin divizarea succesivă a acestuia, până la obţinerea unor clustere care conţin câte un singur obiect.

Algoritmii de agregare sau de amalgamare construiese clustere într-o manieră aseendentă, pornind de la clustere care conţin câte un singur obiect şi continuând, prin comasare succesivă a clusterelor, până la obţinerea unui cluster care include toate obiectele.

În cazul procedurilor de clasificare prin agregare, în fiecare pas se comasează într-un singur cluster fie două obiecte, fie un obiect şi un cluster, fie două clustere diferite. În fiecare etapă a procedurilor divizative, un cluster este divizat fie sub forma a două clustere, fie sub  
forma unui cluster şi unui obiect, fie sub forma a două obiecte.

Numărul de paşi necesari pentru obţinerea unei soluţii cluster de tip ierarhic depinde de numărul de obiecte supuse clasificării şi este diferit pentru cele două categorii de metode de clasificare ierarhică.

Procesele de agregare şi de dezagregare a clusterelor, specifice celor două categorii de proceduri de clasificare ierarhică, presupun utilizarea unor metode specifice de evaluare a distanţelor dintre clustere.

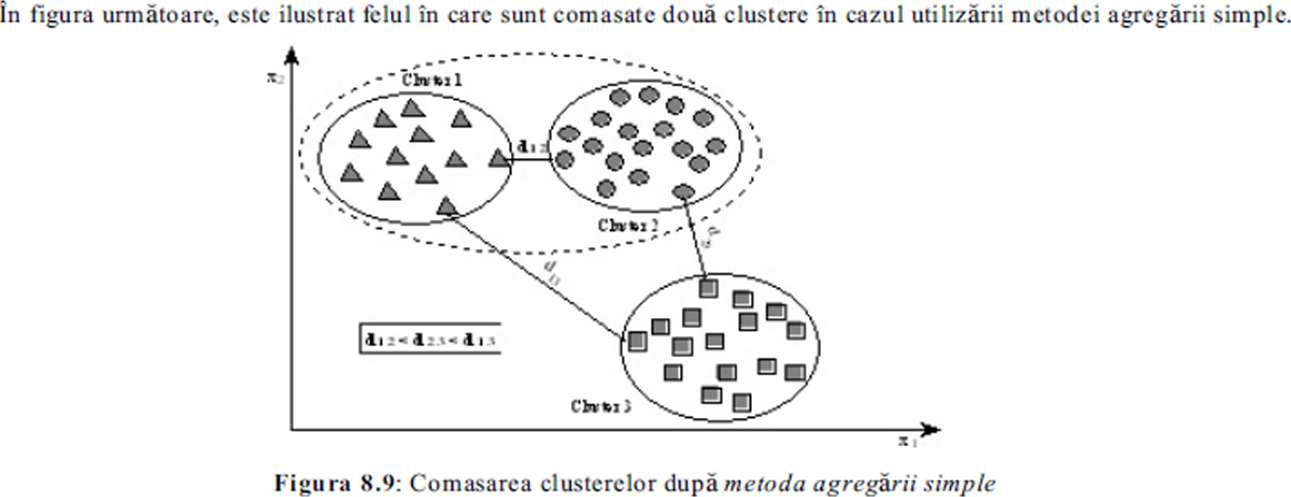
**44.Metoda agregării simple**

In analiza cluster bazată pe agregare simpla afectarea unui obiect la un cluster se face numai dacă acel obiect are un anumit grad de disimilaritate cu unul dintre obiectele care aparţin deja clusterului. Clusterizarea de acest tip se mai numeşte şi analiza cluster de distanţă minima sau analiza cluster de tip MIN.

Metoda agregării simple se bazează pe exprimarea proximităţii dintre două clustere prin intermediul distanţei dintre cele mai apropiate obiecte din cele două clustere. Evaluarea acestei distanţe se face cu ajutorul metodei celor mai apropiati vecini.

**Metoda agregării simple este o metodă de clasificare ierarhică de tip aseendent, care comasează în fiecare etapă a clasificării acele două clustere pentru care distanţa dintre cei mai apropiaţi vecini este cea mai mică, în comparaţie cu alte perechi de clustere**

.



Cea mai mică distanţă între cei mai apropiaţi vecini din cele trei perechi posibile de clustere este distanţa care corespunde perechii de clustere (clusterl,cluster2). Ca urmare, clusterul 1 va fi comasat cu clusterul2, rezultând un nou cluster, care va conţine obiectele din cele două clustere.

**45.Metoda agregării complete**

Această metodă de clusterizare este similară cu metoda agregării simple, cu deosebirea că agregarea a două clustere se face pe baza unei distanţe de agregare care este distanţa dintre cele mai depărtate obiecte din acele clustere. Clusterizarega de acest tip se mai numeşte şi analiza cluster de distanţa maxima sau analiza cluster de tip MAX.

În cazul metodei agregării complete evaluarea distanţelor dintre clustere se face cu ajutorul metodei celor mai depărtaţi vecini. Aceasta înseamnă că distanţa dintre două clustere este considerată a fi în acest caz distanţa cea mai mare dintre oricare două puncte aparţinând celor două clustere.

**Metoda agregării complete este o metodă de clasificare ierarhică de tip ascendent, care comasează în fiecare etapă a clasificării acele două clustere pentru care distanţa dintre cei mai departati vecini este cea mai mica, în comparaţie cu alte perechi de clustere.**

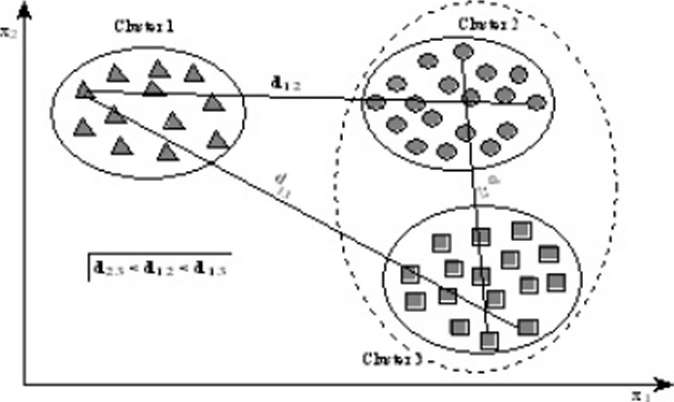
e

Figura 8.10: Comasarea clusterelor după metoda agregării complete

**46.Metoda agregării medii**

Metoda agregării medii este o metodă de clusterizare similară celor două metode menţionate anterior, cu deosebirea că evaluarea distanţei dintre două clustere este considerată a fi media distantelor care separă obiectele aparţinând celor două clustere.

Agregarea clusterelor cu ajutorul metodei agregării medii se face pc baza determinării unui grad de conectivitate medie dintre clustere, grad evaluat ca distanţa medie corespunzătoare unei perechi de obiecte, primul obiect aparţinând unui cluster, iar al doilea obiect aparţinând celuilalt cluster.

**Metoda agregării medii este o metodă de clasificare ierarhică de tip ascendent, care comasează în fiecare etapă a clasificării acele două clustere pentru care distanţa medie dintre toate perechile formate cu obiecte din cele doua clustere este cea mai mica, în comparaţie cu alte perechi de clustere.**

**47.Metoda centroidului**

Metoda centroidului este o metodă de clasificarc ierarhică ascendentă, în care distanţele dintrc clustere sunt evaluate cu ajutorul metodei centroidului. Ideea de bază a metodei centroidului este aceea de obţinere a unui nou cluster prin comasarea a două clustere existente, în funcţie de distanţa cea mai mică dintre centroizii clusterelor care sunt verificate în scopul comasării.

**Definiţie: Metoda centroidului este o metodă de clasificare ierarhică de tip ascendent, care comasează în fiecare etapă a clasificării acele două clustewre pentru care distanta dintre centroizii celor doua clustere este cea mai mica, în comparaţie cu alte perechi de clustere.**

Două clustere sunt comasate într-un nou cluster dacă şi numai dacă distanţa dintre centroizii lor este cea mai mică dintre toate distanţele dintre centroizii oricăror două clustere care aparţin configuraţiei cluster disponibile. în figura următoarc este

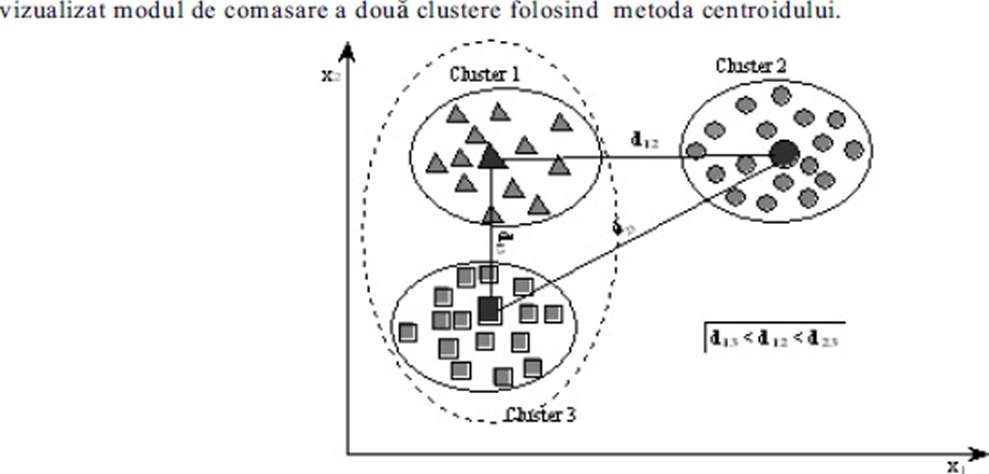


Figura 8.11: Ilustrarea metodei centroidului

**48.Metoda lui Ward**

Metoda lui Ward, cunoscuta şi sub numele de metoda minimei varianţe intraclustere, este una dintre cele mai cunoscute şi mai eficiente metode de clasificare ierarhica prin agregare.

Un obiect poate fi atribuit unui cluster doar daca se realieaza minimizarea sumelor elementelor de pe diagonala matricii comune de covarianţa a clusterelor.

Metoda lui Ward este o metoda de evaluare a distanţei dintre doua clustere care se bazeaza pe maximizarea gradului de omogenitate a clusterelor.

**Definiţie: Metoda lui Ward este o metoda de clasificare ierarhica de tip ascendent, care comaseaza în fiecare etapa a clasificarii acele doua clustere pentru care suma patratelor abaterilor la nivelul clusterului rezultat din comasare este cea mai mica, în comparaţie cu alte perechi de clustere.**

Metoda lui Ward nu este o metoda propriu-zisa de calcul a distanţelor dintre clustere, ci o metoda de formare a clusterelor pe baza maximizarii gradului de omogenitate a clusterelor.

Ca masura a gradului de omogenitate a clusterelor este utilizata suma patratelor abaterilor, numita suma patratelor abaterilor intraclustere. Gradul de omogenitatea unui cluster se considera a fi cu atât mai mare cu cât suma abaterilor intraclustere este mai mica.

Distanţa Ward se evalueaza pentru toate combinaţiile posibile de comasare într-un singur cluster a oricaror doua clustere din configuratia initiala

**49. Algoritmul k-means**

Clusteringul partitional este total diferit de cel ierarhic. Principala diferenta este ca se cunoaste de dinainte numarul de clustere si se incearca crearea unui algoritm din care sa reiasa numarul de clustere stabilit anterior si acestea sa fie cat mai diferite unele de altele.

Algoritmul k means presupune alegerea arbitrara a unui nr k intreg si apoi organizarea obiectelor in k clustere avand in vedere maximizarea variantei intercluster si minimizarea variantei intra cluster. Fiecare cluster are un centru(centroid) ce reprezinta media tuturor punctelor de date din cluster.

Pasii algoritmului k means

1. Selectează aleator k puncte ca şi centre de clustere.

2. Se determină distanţa dintre fiecare obiect şi centroizi.

3. Atribuie instanţele/obiectele acelor clustere care au cel mai apropiat centru faţă de

acestea, în raport cu o anumită măsură de similaritate.

4. Dupa ce toate obiectele au fost atribuite, recalculează poziţia fiecăruia din cei k - centroizi (media tuturor instanţelor din fiecare cluster).

5. Se repetă paşii 2 şi 3, până când un anumit criteriul de oprire este satisfăcut.

Criteriul de oprire (de convergenţă) poate fi unul din urmatoarele:

- nicio reatribuire (minimum de reatribuiri) ale punctelor de date pentru diferite clustere

- nicio schimbare (sau minimum de schimbari) de centroizi

Remarca 1: Cu toate că, în principiu, metoda k-means clustering produce exact k clustere care divizează mulţimea iniţială de obiecte cât mai distinct posibil, rămâne deschisă problema estimării numărului optim de clustere care să conducă la separarea cea mai bună a obiectelor.

Remarca 2: O problemă majoră în utilizarea metodei k-Means este ca datorita faptului ca trebuie sa calculam media nu o putem aplica doar pe date numerice (e.g. categoriale).

Avantaje si dezavantaje k-means

• (+) Mai puţin sensibil la valori extreme (outliers) şi la atribute irelevante;

• (+) Aplicabil la mulţimi mari şi foarte mari de date;

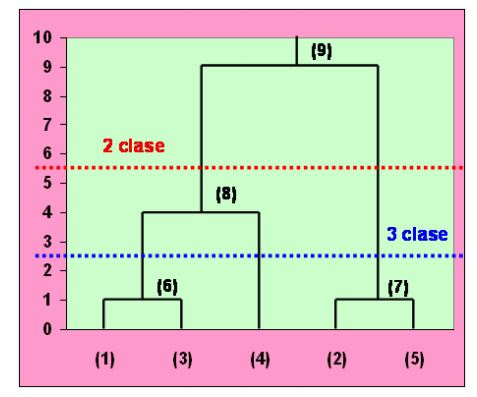
• (-) Are nevoie de alegerea prealabilă a numărului de clustere, ceea ce se dovedeşte a

fi uneori greu de stabilit.

**50. Dendrograma (arborele de clasificare ierarhica)**

Dendograma este un arobore ce este folosit la ilustrarea organizarilor ierarhice a clusterelor. Dendograma se poate realiza prin aplicarea algoritmilor ce organizeaza clusterele pe grupe, intre un nivel minim si un nivel maxim. Numarul de pasi necesari pentru o astfel de organizare depinde de numarul de clustere, pe masura ce avanasm in ierarhie, asemanarea dintre clustere scade si numarul de obiecte continute de acestea creste.

Prin secţionarea orizontală a dendrogramei se obţine o partiţie a mulţimii elementelor clasificate. Componentele partiţiei sunt clasele căutate.

În figura alăturată este prezentată o dendogramă. Pe axa orizontală sunt elementele iniţiale (ordinea este cea care permite desenarea arborelui). Pe axa verticală sunt distanţele dintre obiecte, de exemplu, între obiectele 4 şi 6 este o distanţă egală cu 4.

Deoarece într-o problemă de clusterizare nu se cunoaşte nimic apriori (numărul de clase în special), evaluarea calităţii partiţiei obţinute este o etapă foarte importantă. Evaluarea trebuie să ia în considerare atât faptul că, poate, mulţimea iniţială nu are o structură bine determinată de clase, cât şi faptul că diferite metode conduc la clase diferite.

Procedurile uzuale de evaluare:

• Vizualizarea partiţiei (dendrograme, profiluri, proiecţii).

• Indicatori de calitate

Coeficienţii de divizare (divisive coefficient - DC) şi de aglomerare  
(agglomerative coefficient - AC) care oferă indicatori (medii) globali.

o Indici de siluetă (Silhouette) care se pot defini atât global, cât şi local pentru fiecare cluster.

**51. Cum se alege numarul de clustere în cazul clasificarilor de tip ierarhic**

Analiza claselor ierarhice (*Hierarhical Cluster Analysis* – HCA) este o metodă de grupare „ierarhică” în care fiecare clasă este în totalitate conţinută în altă clasă. Nu este cerută nici o informaţie *a priori* despre numărul claselor, iar odată ce un individ a fost asociat unei clase, el va rămâne acolo. Nu este recomandată a fi utilizată pentru baze de date mari, cu mulţi indivizi.

Metodele ierarhice de formare a claselor se caracterizează prin faptul că numărul de clase nu este cunoscut dinainte, ci este determinat pe parcurs, prin algoritmul de clasificare. Se disting două categorii de algoritmi ierarhici de clasificare, şi anume algoritmii ascendenţi (sau de agregare) şi algoritmii descendenţi.

În cele ce urmează vom prezenta paşii principali pentru un algoritm de agregare. Presupunem că avem *n* indivizi pe care dorim să-i clasificăm.

**Pasul 1.** Se consideră *n*0 = *n*, adică partiţia cea mai fină, formată iniţial din clase cu un singur individ fiecare. În această mulţime de indivizi/clase se selecţionează doi, cei mai apropiaţi după indicele de proximitate folosit. Aceştia vor forma primul grup.

**Pasul 2**. Se calculează o nouă matrice de proximitate ce conţine *n*0 - 1 linii, corespunzătoare celor *n*0 – 2 obiecte/clase încă negrupate şi primului grup creat. Pe baza acestei noi matrice se identifică alte două obiecte, cele mai apropiate între ele, şi cu acestea se va forma un nou grup. Iterativ, aceste obiecte pot fi fie doi indivizi, fie un individ şi un grup deja constituit, fie două grupuri deja constituie.

Descreştem *n*0 (= *n*0 – 1) şi repetăm pasul 2 până când toţi indivizii au fost grupaţi.

**52. Formulati problema generala a recunoasterii supervizate a formelor si mentionati cateva domenii de utilizare**

Tehnicile de invatare supervizata au ca scop construirea unui model al datelor initiale in care o parte dintre variabile sunt explicative (variabile predictor) si una sau mai multe variabile sunt variabile raspuns. Printre tehnicile supervizate sunt: analiza canonica, analiza discriminarii, regresia liniara multipla, regresia logística.

Pentru a evalua legătura (liniară) între două variabile cantitative, putem alege să calculăm coeficientul de corelaţie Pearson şi să interpretăm valoarea obţinută. Dar dacă am dorim să evaluăm legătura liniară între două seturi de variabile (cantitative) o posibilitate este cea a evaluării corelaţiei între două combinaţii liniare, care reprezintă optimal cele două seturi de variabile.

De obicei, **analiza canonică** este folosită în următorul context: asupra unor indivizi ai populaţiei s-au făcut atât măsurători obiective, cât şi aprecieri subiective (exprimate însă cantitativ, sub forma unor note). De exemplu, indivizii ar putea fi un ansamblu de firme, variabilele obiective ar putea fi indicatorii financiar-contabili, iar variabilele subiective ar putea fi nota acordată (de către un panel de specialişti) politicii de promovare a produselor, preferinţa acţionarilor pentru active etc.

Metodele de **analiză a discriminării** se aplică unei populaţii de indivizi caracterizaţi prin variabile continue sau categoriale care sunt *a priori*  impartiti in grupuri. în analiza discriminării populaţia de indivizi care au fost cercetaţi este împărţită în grupuri şi că dispunem de datele observate pentru aceşti indivizi. (În unele situaţii grupurile apar în mod natural, în altele ele sunt rezultatul unei analize anterioare.)

**53. Definiti scopurile recunoasterii supervizate a formelor si descrieti tipul informatiilor utilizate în recunoasterea supervizata**

Analiza discriminantă are două scopuri bine precizate, şi anume:

– **Un scop decizional**, destul de frecvent, ce are în vedere construirea unei reguli de afectare a indivizilor la un grup, regulă ce poate fi aplicată şi în viitor. Această regulă se construieşte în funcţie de ansamblul de variabile predictor observate asupra indivizilor. O regulă bună de afectare este aceea care va conduce în viitor la erori de clasare a observaţiilor viitoare cât mai mici posibile.

– **Un scop explicativ,** prin care se urmăreşte să se descopere variabilele cele mai pertinente în descrierea diferenţelor dintre grupurile formate *a priori.*

Tipurile de informatii utilizate: OBIECTIVE si SUBIECTIVE.

Asupra unor indivizi ai populaţiei s-au făcut atât măsurători obiective, cât şi aprecieri subiective (exprimate însă cantitativ, sub forma unor note). Aşadar, primul set de variabile este format din cele „obiective”, fie acestea . Datele obţinute de la cei *n* indivizi vor forma matricea *X* (de dimensiuni ). Al doilea set de variabile este format din cele „subiective”, fie acestea , iar datele obţinute de la cei *n* indivizi vor forma matricea *Y* (de dimensiuni *pxxx*,...,,21*pn*×*qyyy*,...,,21*qn*×).

**54. Ce sunt clasificatorii de tip liniar. Descrieti logica discriminarii liniare si spatiul discriminat**

*Clasificatorul* sau *criteriul de clasificare* reprezinta regula sau multimea de reguli pe baza carora obiectele care apar in multimii analizate sunt afectate sau atribuite unor clase sau grupe bine definite.

Prima modalitate de abordare a problemelor de clasificare cu ajutorul tehnicilor de analiza discriminanta dateaza din anul 1933 si a fost propusa de Fisher. Ulterior abordarile de acest tip s-au dezvoltat în mod constant, iar aplicaTiile bazate pe analiza discriminanta s-au extins la din ce în ce mai multe domenii de activitate si s-au diversificat din ce în ce mai mult.

Cele mai multe si cele mai utile aplicatii ale analizei discriminant bazata pe criteriul lui Fisher sunt întâlnite în domeniul financiar-bancar, domeniu în care tehnicile de tip se numesc *tehnici de credit-scoring; ele* constituie cele mai importante instrumente pentru fundamentarea deciziilor privind acordarea de credite.

Metoda de analiza discriminanta propusa de Fisher este o metoda parametrica, caracterizata prin simplitate si robustete, si care ofera posibilitati de interpretare foarte utile pentru analiza. Simplitatea acestei metode decurge din faptul ca utilizarea sa nu necesita decât evaluarea unor estimatii pentru parametrii populatiei si claselor acesteia, parametri reprezentati de medii,variante sau covariante. aceasta reprezinta un avantaj foarte important al analizei discriminante de tip fisher, în comparatie,de exemplu, cu tehnicile de analiza discriminanta bazate pe criteriul bayesian, tehnici a caror utilizare presupune cunoasterea probabilitatilor apriorice.

**55. Definiti functiile discriminant liniare, variabilele discriminant si scorurile discriminant**

Fundamentul teoretic al analizei discriminante de tip Fisher este reprezentat de analiza variantei. Criteriul lui Fisher defineste o modalitate de deducere a functiilor discriminant pe baza analizei comparative dintre *variabilitatea intragrupala* si *variabilitatea intergrupala*, la nivelul claselor sau grupelor populatiei analizate. Functiile discriminant deduse pe baza criteriului lui Fisher se mai numesc si *functii scor* si sunt functii *liniare*.

Criteriul fundamental care sta la baza împartirii multimii de obiecte în submultimi, este un criteriu mixt, care urmareste *minimizarea variabilitatii intragrupale* si *maximizarea variabilitatii* *intergrupale*. Utilizarea acestui criteriu combinat asigura cea mai buna diferentiere a claselor sau grupelor populatiei.

O functie discriminat de tip Fisher se determina ca o combinatie liniara de variabilele discriminant, combinatie ai carei coeficienti sunt componente ale unui vector propriu al matricii . Din aceasta modalitate de definire rezulta, în mod implicit, ca pot fi identificate mai multe functii discriminant.

Numarul maxim posibil de functii discriminant care pot fi identificate pe baza criteriului lui Fisher este egal cu numarul de valori proprii *distincte* si *strict pozitive* ale matricii.

**Determinarea functiilor discriminant** este echivalenta cu gasirea unor directii, sau vectori,în raport cu care variabilitatea intragrupala sa fie minima, iar variabilitatea intergrupala sa fie maxima. aceste directii vor define axele spatiului discriminat si pot fi identificate sub forma unor combinatii liniare de variabilele descriptor selectate în analiza.

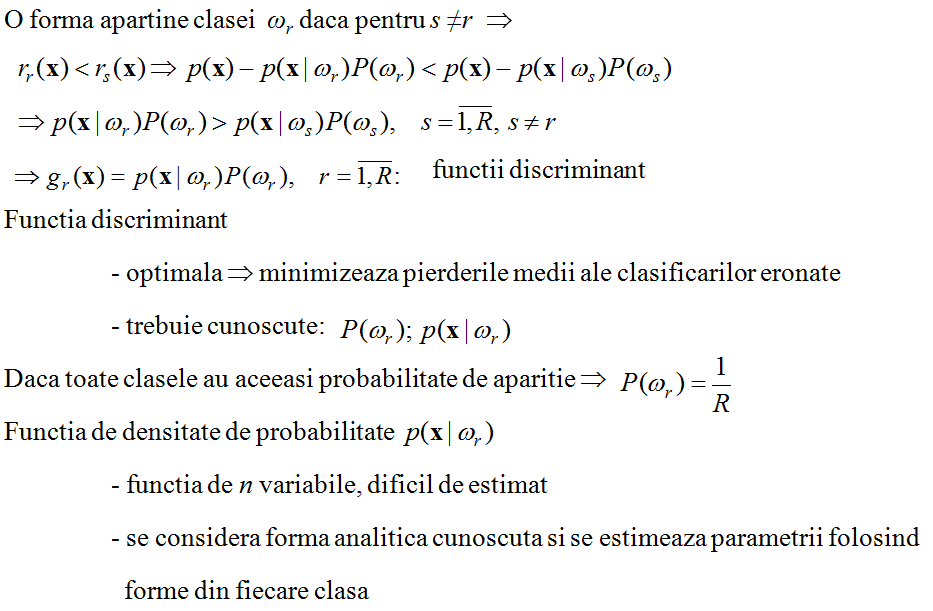
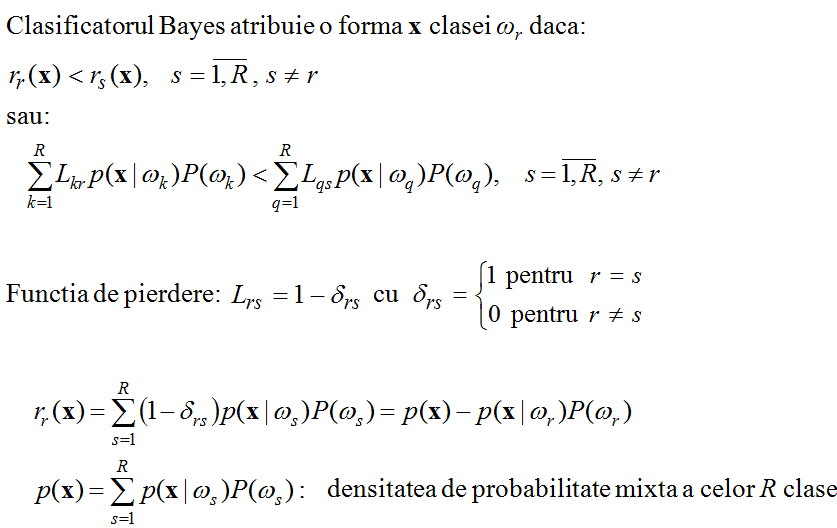
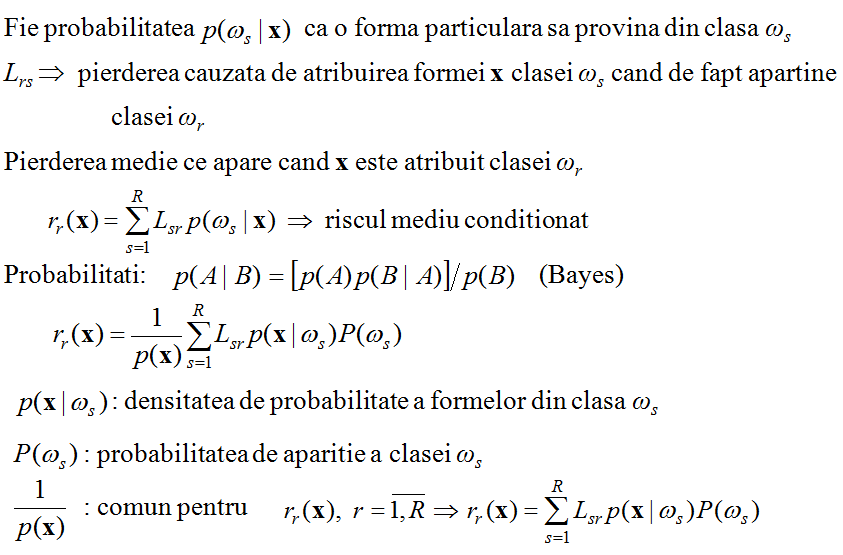
Ideea care sta la baza criteriului lui Fisher este aceea a determinarii unor *direc*t*ii* sau *axe*, astfel încât, de-a lungul acestora, clasele multimii sa se *diferen*t*ieze cât mai mult între ele* si, în acelasi timp, fiecare clasa sa aiba un *grad de omogenitate cât* *mai mare*. Cu alte cuvinte, criteriul lui Fisher are ca scop determinarea unor *direc*t*ii* de-a lungul carora variabilitatea intergrupala sa fie cât mai mare, iar variabilitatea intragrupala sa fie cât mai mica.

Proiectiile obiectelor pe axele definite de aceste directii reprezinta sunt noi coordonate ale obiectelor Si se numesc ***scoruri discriminant*.**

În analiza discriminanta noile directii care trebuie identificate nu trebuie sa fie în mod obligatoriu ortogonale, spre deosebire de analiza componentelor principale în care direcTiile de variabilitate maxima trebuie sa verifice proprietatea de ortogonalitate.

**Variabilele discriminant**sunt combinatii liniare de variabilele descriptor, de forma: **d=β0+ βt \*x**

**56&57 Descrieti clasificatorul Bayesian si aratati cum poate fi utilizat acesta in predictia apartenentei formelor. Descrieti forma clasificatorului Bayesian in cazul normalitatii si homoscedasticitatii claselor**

****

**58. Descrieti clasificatorul liniar Fisher si aratati cum poate fi utilizat acesta in predictia apartenentei formelor**

Clasificatorii au natura predictiva, avand scopul de a stabili apartenenta unor forme cu apartenenta necunoscuta la o anumita clasa.

Baza informational necesara pentru construirea clasificatorilor este reprezentata de o matrice de observatii extinsa de dimensiune T+1 linii si n+1 coloane, unde primele n coloane sunt reprez de caracteristicile propriu-zise ale formelor, iar ultima coloana este asociata cu o variabila suplimentara, care este explicata prin sisteme de functii fi(x1,….xi), numite clasificatori.

Ideea ce sta la baza clasificatorului Fisher este aceea a determinarii unei axe/directii de-a lungul careia clasele multimii Ω sa se diferentieze cat mai mult intre ele si, in acelasi timp, fiecare clasa sa aiba un grad cat mai mare de omogenitate in interiorul ei. Cu alte cuvinte, scopul clasificatorului Fisher e sa gaseasca o directie, de-a lungul careia variabilitatea intraclasa este cat mai mare, iar variabilitatea interclasa cat mai mica.

La final, functia discriminant o sa fie un produs de variabile initiale si vectori proprii ai produsului dintre inversa matricei de covarianta intraclasa si matricea de covarianta interclasa.

**59. Descrieti clasificatorul Mahalanobis si aratati cum poate fi utilizat acesta in predictia apartenentei formelor**

Clasificatorii au natura predictiva, avand scopul de a stabili apartenenta unor forme cu apartenenta necunoscuta la o anumita clasa.

Baza informational necesara pentru construirea clasificatorilor este reprezentata de o matrice de observatii extinsa de dimensiune T+1 linii si n+1 coloane, unde primele n coloane sunt reprez de caracteristicile propriu-zise ale formelor, iar ultima coloana este asociata cu o variabila suplimentara, care este explicata prin sisteme de functii fi(x1,….xi), numite clasificatori.

**Primii doi pasi in cadrul clasificatorului Mahalanobis, sunt urmatorii:**

1. Se estimeaza vectorii mediilor claselor xmed1,….,xmedk
2. Se estimeaza matricea de covarianta Σ1,…,Σk.

Apoi se evalueaza distanta Mahalanobis de la o forma la centroidul fiecarei clase. **D(x,xmedk)=( x-xmedk)T Σk-1( x-xmedk)**

Forma se considera ca apartine clasei unde se obtine cea mai mica distanta Mahalanobis.

**60. Descrieti modul de stabilire a abilitatii predictive a unui clasificator si matricea corectitudinii clasificarii**

Se consideră că un clasificator determinat asigură o capacitate mare de clasificare dacă exista un compromis între viteza de recunoaştere şi abilitatea predictivă a acestuia.

Prin viteza de recunoastere se intelege acea fractiune din formele din setul de formare clasificate corect de clasificatorul determinat.

Abilitatea predictiva reprezinta numarul de obiecte din setul de predictie, a caror apartenenta la clase este presupusa necunoscuta, sunt clasificate corect de catre clasificator.

Fie functia de eroare c:XxXxY->[0;∞] astfel incat c(x,y,y)=0, {x,y,f(x)} apartin XxXxY, iar x este o forma din X, y este clasa reala din care face parte si f(x) clasa in care a fost incadrata de clasificator.

In general, functia de eroare are forma D(x,y,f(x))=0.5 |f(x)-y|

Daca forma x este corect clasificata, atunci D(x,y,f(x))=0, altfel, este egala cu 1. Astfel obtinem matricea corectitudinii clasificarii. Daca formam o matrice cu elementele x pe linie si elementele f(x) pe coloane, atunci numarul de elemente ale matricei care sunt egali cu 0, dau abilitatea predictiva a unui clasificator.