

Academia de Studii Economice

Facultatea de Cibernetica, Statistica si Informatica Economica

Prof. univ. dr. Gheorghe RUXANDA

DATA MINING

**Bucuresti
2013**

Tema 1. Probleme ale măsurării și cuantificării economico-sociale

1.1 Necessitatea și rolul analizei datelor

Cunoașterea științifică din orice domeniu de activitate umană presupune, indiferent de natura și specificul obiectivelor concrete urmărite, o complexă și riguroasă analiză cantitativă a fenomenelor și proceselor care fac obiectul cercetării. Este vizibil pentru oricine, și din ce în ce mai mult, că în epoca modernă aproape orice individ angrenat într-o activitate umană se ocupă, într-un fel sau altul, în mod direct sau indirect, cu date și informații, cu colectarea, prelucrarea și interpretarea acestora.

Desfășurarea oricărei activități umane implică o producție continuă de date sau informații, care se acumulează în timp și care pot fi folosite pentru cunoașterea structurală și evolutivă a fenomenelor la care se referă aceste informații, în scopul fundamentării corecte și eficiente a deciziilor care trebuie luate. Mai mult decât atât, desfășurarea activităților umane nici măcar nu poate fi concepută în zilele noastre fără un consum continuu, din ce în ce mai mare, de informație. Din acest punct de vedere, se poate spune că informația a devenit unul dintre factorii de producție importanți și activi, un factor de progres și civilizație.

Totdeauna, mulțimile de date conțin, într-un mod amalgamat și invizibil, atât aspecte semnificative, cât și aspecte nesemnificative, ale manifestării fenomenelor. Deoarece cunoașterea științifică vizează în mod exclusiv aspectele informaționale semnificative, apare necesitatea utilizării unor metode și tehnici specifice analizei datelor, cu ajutorul căror informația semnificativă să poate fi detectată, separată de informația nesemnificativă și exprimată sub o formă clară și interpretabilă.

Metodele și tehniciile de analiză a datelor sunt cele mai adecvate instrumente utilizabile pentru *identificarea unor structuri cauzale*, pentru *decelarea unor tendințe și configurații specifice* pe mulțimea datelor analizate și obținerea unor *reprezentări simplificate ale informațiilor de mare complexitate*.

Utilitatea și eficiența utilizării metodelor și tehniciilor de analiză a datelor sunt maxime în situațiile în care informațiile supuse studiului sunt în cantități foarte mari. Din acest punct de vedere, domeniul economic poate fi considerat ca fiind un domeniu privilegiat. În cea mai mare parte a lor, metodele și tehniciile de analiză a datelor au natură multidimensională, astfel încât, comparativ cu metodele și tehniciile de analiză statistică descriptivă, ele permit și investigarea *legăturilor și interdependențelor* evidențiate la nivelul mulțimilor de date.

Materia primă utilizată în orice activitate de analiză a datelor este reprezentată de o *colecție* sau *mulțime* de date sau informații cantitative, referitoare la stările sau evoluțiile unei mulțimi de fenomene. Aceste date pot fi obținute fie pe cale *observațională*, fie pe cale *experimentală*.

Din punct de vedere al analizei datelor, orice mulțime de informații supusă studiului este privită ca fiind o *reprezentare codificată*, într-o formă mai mult sau mai puțin *implicită*, a unor aspecte informaționale referitoare la *niveluri și variații* ale unor fenomene, *evoluții și tendințe* relevante, *legături și influențe* semnificative, *ierarhii și configurații structurale* specifice.

Datele supuse unui proces de analiză nu evidențiază, în mod direct și explicit, prin ele însese, informația utilă și semnificativă. De regulă, datele conțin informația utilă și semnificativă sub o formă mascată, ascunsă, amestecată într-un mod nediferențiat și fără o logică aparentă, cu informația nesemnificativă, rezultată din influențe accidentale și marginale. În acest sens, se poate spune că la nivelul datelor primare supuse analizei, informația semnificativă se găsește sub o formă *diluată și disipată* într-o mulțime informațională complexă, neordonată și nestructurată după vreun criteriu logic existent aprioric.

Rolul analizei datelor este acela de a prelucra și filtra informațiile conținute în datele supuse studiului, cu scopul de a *capta* sau de a *extrage esența informațională* conținută în aceste date și de a *evidenția această esență informațională într-o formă de reprezentare inteligeabilă, sugestivă, simplificată și sintetizatoare*. Atingerea acestui scop presupune realizarea unei succesiuni de transformări efectuate asupra datelor primare și implică utilizarea unor metode și tehnici specifice. Aceste transformări au scopul de a *maximiza relevanța și interpretabilitatea datelor* și presupun, printre altele, eliminarea informațiilor redundante sau lipsite de semnificație și generalitate, care au natură accidentală sau marginală. Din acest punct de vedere, procesul de analiză a datelor apare ca fiind un proces specific de *transformare informațională*, proces care are ca intrări datele primare, iar ca ieșiri informații sintetizatoare.

Avgând în vedere modul în care se efectuează, precum și natura instrumentelor pe care le folosește, analiza datelor este, prin excelență, o analiză de tip *multidimensional*, reprezentând, în comparație cu analiza simplă, unidimensională, o schimbare de natură calitativă. Analiza statistică descriptivă permite reprezentarea unor colecții foarte mari de date într-o manieră sugestivă și asimilabilă, simplificată și schematică. Spre deosebire de aceasta, analiza multidimensională constituie o generalizare naturală a logicii și informațiilor referitoare la mai multe variabile sau dimensiuni.

Activitățile de manipulare a datelor și informațiilor, de prelucrare și interpretare corectă și eficientă a acestora, presupun existența unui cadru conceptual adecvat și utilizarea unor metode și instrumente specifice. Atât cadrul conceptual necesar, cât și metodele și tehniciile utilizabile în prelucrarea, analiza și interpretarea datelor și informațiilor, sunt subsumate de obiectul unei discipline științifice numită *analiza datelor*.

1.2 Specificitatea domeniului economico-social și modalități de abordare

Deși metodele și tehniciile de analiză a datelor sunt utilizate în majoritatea domeniilor activității umane, putem afirma că analiza datelor are cea mai largă utilizare în domeniul economico-social, iar eficiența utilizării ei în acest domeniu are o eficiență extrem de ridicată. Așa cum cum o să arătăm în continuare, activitatea de cunoaștere științifică din domeniul economico-social este caracterizată de anumite accente specifice, în comparație cu alte domenii ale cunoașterii umane. În virtutea acestor elemente de specificitate, domeniul economico-social pare a fi cel mai potrivit domeniu pentru utilizarea metodelor și tehniciilor de analiză

multidimensională a datelor.

Dintre toate elementele de specificitate a domeniului economico-social, două considerăm a fi mai importante și mai relevante din punct de vedere al utilizării analizei datelor: *complexitatea ridicată a fenomenelor economico-sociale și natura cantitativă* a acestor fenomene. Împreună cu multe alte caracteristici specifice, aceste două caracteristici fundamentale impun modalități specifice de abordare pentru cunoașterea științifică din domeniul economico-social.

Necesitatea de a sintetiza și de a simplifica în procesul de cunoaștere a realității, este impusă atât de faptul că datele utilizabile într-o analiză maschează, ascund, anumite aspecte, ci și de faptul că realitatea analizată este caracterizată de o complexitate foarte ridicată, care nu poate fi cuprinsă și înțeleasă numai pe baza intuiției.

În activitatea de analiză cantitativă, pe care se bazează în mod direct cunoașterea științifică din cele mai multe domenii de activitate, sunt implicate trei categorii esențiale de elemente: *teorii și principii teoretice generale și specifice domeniului investigat, informații cantitative și calitative* referitoare la fenomenele supuse studiului, *metode și tehnici* de cuantificare, evaluare, estimare și testare a mărimilor specifice și a relațiilor existente la nivelul realității investigate.

Teoriile și principiile teoretice care stau la baza oricărei analize cantitative sunt reprezentate de *mulțimea cunoștințelor științifice* acumulate în cursul timpului, de realizările obținute pe plan științific, atât la nivelul general al cunoașterii umane, cât și la nivelul cunoașterii în domeniul abordat.

Informațiile cantitative și calitative referitoare la fenomenele și procesele studiate exprimă o mulțime de stări și evoluții concrete din realitatea investigată și sunt rezultatul unui laborios proces de observare, măsurare și evaluare, proces în care intervin o serie de *norme, principii, metodologii și instrumente specifice procesului de măsurare*. Informațiile obținute din realitatea investigată, în urma unor procese de observare și 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.

Metodele și tehniciile reprezintă un set de *reguli, principii și proceduri de analiză, prelucrare și interpretare* a datelor. În analiza datelor, metodele și tehniciile se referă la cuantificare, evaluare, estimare și testare, și sunt reprezentate de o mulțime extinsă și variată de proceduri și instrumente statistică-matematice. Aceste proceduri sunt aplicate, sub o formă sau alta, informațiilor cantitative și calitative, datelor de intrare, în scopul deducerii anumitor rezultate și concluzii. De calitatea metodelor și tehniciilor utilizate într-o activitate de analiză și de eficiență utilizării lor depind, în mod direct și într-o proporție foarte mare, atât semnificația și validitatea concluziilor obținute, cât și calitatea rezultatelor obținute.

Combinarea, la nivelul analizei cantitative, a celor trei categorii de elemente menționate anterior, conduce la obținerea unor rezultate și concluzii, care contribuie la adâncirea procesului de cunoaștere a realității studiate și care se pot concretiza sub mai multe forme:

- obținerea de *informații relevante suplimentare* privind starea, evoluția și conexiunile componentelor realității investigate;
- relevarea unor noi principii și legități care guvernează mișcarea fenomenelor și proceselor din această realitate;
- formularea unor concluzii cu privire la existența unor legături și specificăți care caracterizează realitatea studiată;
- validarea unor ipoteze formulate cu privire la populațiile reprezentate de fenomenele și procesele studiate;
- identificarea unor tipologii și structuri specifice pe mulțimea de fenomene și procese analizate;
- estimarea unor efecte și influențe caracteristice interdependențelor dintre fenomene și procese;
- deducerea unor modele statistică-matematice, care să descrie comportamentul fenomenelor și proceselor;
- efectuarea de *predicții* cu privire la evoluția fenomenelor și proceselor;
- efectuarea de *simulații* privind evoluția fenomenelor și a interdependențelor manifestate între acestea.

Fenomenele economico-sociale și mișcarea acestora în timp și spațiu, adică procesele, au o caracteristică ce le face să se deosebească în mod esențial, fundamental, de fenomene și procese specifice altor domenii de activitate. Această caracteristică este dată de faptul că manifestarea la nivel observabil a acestor fenomene și a mișcării lor, este extrem de complexă și îmbracă o deosebit de mare varietate de forme, fiecare fenomen sau proces economico-social existând sub **formă multiplicată pe o scară foarte mare**. Proprietatea de multiplicitate la o scară foarte mare, pe care o au existența și manifestarea fenomenelor economice și sociale, este cunoscută în literatura domeniului sub numele de **caracter de masă al fenomenelor economico-sociale**.

Formele sub care se manifestă fenomenele economice și sociale apar, de regulă, ca o amalgamare de aspecte esențiale și neesențiale, legate direct sau indirect de conținutul fenomenului, de ceea ce este esențial și stabil în comportamentul acestuia, fiind caracterizate în timp, cel puțin la suprafață, de un grad relativ ridicat de instabilitate. Partea consistentă, semnificativă și stabilă a unui fenomen economico-social este, de obicei, ascunsă de această multitudine de manifestări, astfel încât cunoașterea acestora nu se poate face decât rareori prin observare directă, fiind necesar, de cele mai multe ori, un demers logic extrem de complex, în cadrul căruia se urmărește eliminarea a ceea ce este accidental, conjunctural, neesențial și nesemnificativ în manifestarea fenomenului, reținându-se ceea ce este trainic, cu caracter de regularitate, esențial și semnificativ.

O altă caracteristică a fenomenelor economico-sociale este cea legată de *multitudinea și eterogenitatea* acestor fenomene, de *dimensiunile* la care are loc desfășurarea acestora, de *numeroasele legături funcționale* existente între acestea și de *multiplele interdependențe și condiționări* dintre fenomenele economico-sociale și fenomenele specifice altor domenii. Gradul ridicat de interconectare a elementelor și multitudinea factorilor de influență, implică în manifestarea relațiilor de cauzalitate, reprezintă caracteristici definitorii pentru majoritatea fenomenelor și proceselor de natură socio-economică.

Amploarea complexității din domeniul economico-social rezultă și din faptul că *natura și intensitatea relațiilor de interdependență cunoște o mare variabilitate în timp și spațiu, că sensul relațiilor de cauzalitate este reversibil, chiar contradictoriu, și că, din cauza gradului ridicat de interconectare, manifestarea cauzalității are loc, în multe situații, sub forma unor "reacții în lanț"*. Specificitatea domeniului economico-social și caracteristicile care îl fac să se diferențieze fundamental de celelalte domenii de activitate, impun ca procesul de cunoaștere în acest domeniu, investigarea științifică a fenomenelor și proceselor de natură economico-socială să poarte o amprentă specifică, particulară.

Caracterul complex al manifestării fenomenelor economice și sociale face ca activitatea de cunoaștere în acest domeniu să fie puternic condiționată atât de **modalitatea concretă de abordare** a obiectului studiat, cât și de **natura și puterea metodelor, tehniciilor și instrumentelor** de investigare utilizate.

Faptul că fenomenele economice și sociale se caracterizează printr-un grad de complexitate foarte ridicat face ca activitatea

de investigare a comportamentului acestora să fie deosebit de dificilă și rezultatele acesteia să aibă un *grad ridicat de relativitate* din punct de vedere al semnificației și exactității. În aceste condiții, succesul investigațiilor socio-economice, valabilitatea și stabilitatea concluziilor rezultate din analizele efectuate, depind în mod direct de compatibilitatea dintre natura instrumentelor de investigare folosite și specificul fenomenelor investigate.

Alegerea modalităților de abordare și a celor mai potrivite metode și tehnici de analiză sau predicție reprezintă condiția de bază pentru obținerea unor rezultate satisfăcătoare în activitatea de cunoaștere a fenomenelor din domeniul economico-social. În ceea ce privește modalitățile de abordare, cele mai des utilizate în analiza datelor sunt: abordarea sistemică, abordarea statistică și abordarea bazată pe modelarea matematică și informatică.

1.2.1 Abordarea sistemică

O principală modalitate de abordare, esențială pentru orice proces de investigare științifică, impusă cu necesitate mai ales în cazurile în care domeniul vizat este caracterizat de un grad ridicat de complexitate, este cea cunoscută sub numele de **abordare sistemică**.

Atât în sens teoretic, cât și în sens practic, abordarea sistemică se detașează ca metodă generală, universală utilizată în investigarea științifică, indiferent de domeniul concret al investigării științifice. În cazul particular al analizei datelor, abordarea sistemică are o importanță specială, atât în faza de analiză propriu-zisă, cât și în faza de interpretare sa rezultatelor.

1.2.2 Abordarea statistică

Caracterul de masă pe care îl au fenomenele economice și sociale determină ca *manifestarea acestora, în timp și spațiu, să intre sub incidența unui principiu universal de regularitate și stabilitate*, specific fenomenelor cu frecvență foarte mare de existență și apariție, principiu sintetizat sub forma uneia din cele mai importante legi din domeniul cunoașterii umane: **legea numerelor mari**.

Regularitatea și stabilitatea comportamentului fenomenelor economice și sociale sunt asigurate, în virtutea acestui principiu, de **tendința de compensare a influențelor cu sensuri contrarii**, pozitive și negative, exercitate de factorii neesențiali, nesemnificativi și accidentalni implicați în determinarea relațiilor de cauzalitate dintre fenomenele socio-economice.

O consecință directă și imediată a acestui principiu constă în faptul că investigarea comportamentului fenomenelor economice și sociale și formularea de concluzii sintetizatoare cu privire la amplitudinea sau sensul evoluției acestor fenomene sunt cu mult mai ușor de realizat și caracterizate de un grad mult mai ridicat de semnificație și exactitate în cazul în care sunt supuse studiului colectivității mari de fenomene comparativ cu situația în care studiul vizează fenomene individuale, izolate.

Necesitatea de a fructifica în procesul cunoașterii fenomenelor și proceselor economice sau sociale facilitățile ce rezultă din faptul că acestea au caracter de masă, impune ca în întreaga activitate de analiză a datelor din domeniul economico-social să prevaleze **modalitatea de abordare statistică și folosirea metodelor și tehnicii oferite de teoria probabilităților și statistica matematică**.

Modalitatea de abordare statistică este impusă ca modalitate de investigare a fenomenelor de natură socio-economică și de faptul că manifestarea acestora este caracterizată de un **grad ridicat de incertitudine și imprevizibilitate**.

Elementele vizate în mod direct în cadrul investigațiilor științifice bazate pe utilizarea metodelor și tehnicii proprii analizei datelor, sunt reprezentate de fenomenele cu natură stohastică, adică de fenomenele al căror comportament are *caracter aleatoriu*, dar care manifestă o anumită *regularitate* mai mult sau mai puțin evidentă.

Definiție: *Fenomenul stohastic* este acel fenomen *observabil*, ale cărui manifestări particulare sunt *incerte*, dar care evidențiază o anumită *regularitate* a formelor de manifestare, o anumită *legătură* între aceste forme de manifestare.

Necesitatea utilizării în procesul de investigare a fenomenelor economice și sociale a modalității de abordare statistică este determinată și de **caracterul de relativitate, parțialitate și inexactitate al informației** din acest domeniu.

Imposibilitatea de a obține informații cu caracter exhaustiv despre desfășurarea fenomenelor economice și sociale impune ca activitatea de cunoaștere a ansamblului acestei desfășurări să se bazeze pe informații cu caracter parțial, obținute pe bază de sondaj. Extinderea concluziilor rezultante din analiza informațiilor obținute prin sondaj la nivelul întregului ansamblu de fenomene investigate de sondaj se bazează pe metodele și tehniciile **inferenței statisticice**.

Una din cerințele de bază impuse de modalitatea de abordare statistică oricărui demers științific ce are ca obiect investigarea comportamentului unor fenomene și procese economice sau sociale poate fi exprimată prin *necesitatea de a studia variabilele ce descriu acest comportament, în interdependență și simultaneitatea lor*.

Analiza izolată a comportamentului fiecarei variabile este incompatibilă cu modalitatea de abordare sistemică și duce la o pierdere importantă de informație semnificativă în cazul în care variabilele sunt interdependente. Mai mult decât atât, concluziile parțiale desprinse în urma unei astfel de analize nu pot fi generalizate și nu au valabilitate la nivelul comportamentului "legat" al variabilelor, influența interacțiunii dintre acestea fiind, de cele mai multe ori, covârșitoare.

De regulă, analiza comportamentului individual al unui fenomen, în sensul său *unidimensional*, face obiectul *analizei statisticice descriptive*. Analiza statistică descriptivă permite reprezentarea unor colecții foarte mari de date într-o manieră sugestivă și asimilabilă, simplificată și schematizantă.

Considerând o mulțime de variabile aleatoare ce descriu un anumit fenomen economic sau social, vom obține rezultate total diferite dacă vom studia, de exemplu, repartitia fiecarei dintre ele *izolat, separat* de celelalte, comparativ cu situația în care în studiu ar fi vizată repartitia *comună, legată* a acestor variabile. Studiul comportamentului simultan, interdependent, al unei mulțimi de variabile presupune metode și tehnici de complexitate și rafinament mult mai ridicat decât studiul comportamentului izolat al acestora și face obiectul unei discipline aparte, și anume *analiza statistică multidimensională* sau *multivariată*.

1.2.3 Abordarea bazată pe modelarea matematică și informatică

Așa cum am arătat anterior, complexitatea ridicată ce caracterizează domeniul economico-social și celelalte elemente de specificitate ale fenomenelor aparținând acestuia, determină ca orice demers științific întreprins în scopul adâncirii cunoașterii

în acest domeniu să presupună, cu necesitate, utilizarea unor metode și tehnici de lucru corespunzătoare.

Din rândul instrumentelor de investigare științifică, a căror utilitate în contracararea dificultăților ridicate de gradul sporit de complexitate al domeniului studiat este de neînlocuit și a căror utilizare însوșește astăzi demersul științific din aproape oricare domeniu al cunoașterii umane, se detașează **modelarea matematică și tehnica de calcul**.

Ca unul dintre cele mai utile și eficiente instrumente de cunoaștere a realității, **modelarea matematică reprezintă procesul de descriere a comportamentului unor fenomene din natură și societate sub o formă sintetică, logică și formalizată matematic**. Descrierea sub o formă matematică a comportamentului fenomenelor și proceselor din cele mai diverse domenii de activitate a devenit astăzi dorință și, de cele mai multe ori, scopul oricărui cercetător, indiferent de domeniul în care acesta activează.

Surprinderea interdependentelor, a legităților și funcționalităților ce caracterizează fenomenele sau procesele din lumea ce ne înconjoară, sub forma sintetică a unui model matematic, este, probabil, cea mai relevantă măsură a nivelului la care a ajuns gradul de cunoaștere umană.

Din punct de vedere informațional, modelul matematic, în forma sa finală, poate fi privit ca un "**concentrat informațional**", având atât calitatea de acumulator al informației receptate din segmentul de realitate economică sau socială pe care îl descrie, cât și de generator de informație nouă, inexistentă în momentul construirii lui.

1.3 Măsurarea și cuantificarea în domeniul economico-social

Posibilitatea efectuării analizelor sau predicțiilor din domeniul economico-social este strict condiționată de existența unei baze informaționale care să cuprindă informațiile necesare cu privire la nivelul, structura sau evoluția în timp a fenomenelor și proceselor supuse investigării, sau cu privire la condiționările cauzale ale acestor fenomene și procese.

Obținerea informațiilor necesare pentru activitatea de analiză a datelor este, de regulă, rezultatul unor procese de observare și măsurare a fenomenelor și proceselor supuse studiului și presupune existența unor instrumente adecvate.

Informațiile referitoare la un anumit fenomen supus studiului, informații necesare analizei comportamentului respectivului fenomen, reprezintă rezultatul unui **proces de măsurare**. Acest proces reprezintă, de fapt, o acțiune de atribuire de valori numerice pentru caracteristicile respectivului fenomen.

Definiție: Prin **proces de măsurare** se înțelege totalitatea activităților de atribuire a unor valori numerice pentru caracteristicile fenomenului analizat.

În cea mai mare parte a lor, informațiile necesare efectuării analizelor sau predicțiilor din domeniul economico-social trebuie să aibă natură cantitativă, să reprezinte **exprimarea sub formă numerică** a caracteristicilor specifice fenomenelor analizate. Această condiție presupune, în mod implicit, necesitatea existenței unor **unități de măsură**, prin intermediul cărora diferențele caracteristici ale fenomenelor economice sau sociale să poată fi exprimate, a unor **instrumente** adecvate pentru măsurarea caracteristicilor fenomenelor și a unor **modalități de exprimare** numerică.

Din nefericire însă, atât unitățile de măsură, cât și instrumentele utilizate pentru exprimarea sub o formă cantitativă a caracteristicilor fenomenelor economice și sociale, sunt caracterizate de inexactitate și instabilitate, iar utilizarea acestora este generatoarea unei multitudini de erori cu natură extrem de variată.

Exprimarea sub o formă numerică a caracteristicilor fenomenelor și proceselor economice sau sociale presupune un proces de **observare** a formelor individuale prin intermediul cărora se manifestă acestea și de **înregistrare** a valorilor pe care le iau aceste caracteristici în momentul observării.

Odată cu activitatea de înregistrare a mărimilor caracteristicilor fenomenelor și proceselor economico-sociale, adică a valorilor luate de variabilele supuse studiului, are loc și un proces implicit de măsurare, de cuantificare, proces ce presupune folosirea unor unități de măsură specifice și instrumente de lucru adecvate. De regulă, informația rezultată în urma acestui proces este o informație brută, primară, care este cunoscută sub numele de **dată**.

Cele mai dificile probleme cu care este confruntată investigarea fenomenelor economice sau sociale, de-a lungul întregului șir de procese de culegere, prelucrare și interpretare a informațiilor referitoare la aceste fenomene, apar tocmai în cadrul procesului de observare și înregistrare, de măsurare și cuantificare a caracteristicilor fenomenelor. În general, putem spune că măsurarea și cuantificarea din domeniul socio-economic au ca obiective principale:

- stabilirea nivelului sau volumului fenomenelor economice și sociale;
- evidențierea alcătuirii lor structurale;
- caracterizarea evoluției lor în timp sau spațiu;
- exprimarea legăturilor acestora cu alte fenomene sau procese.

Din punct de vedere conceptual, toate aspectele fenomenelor sau proceselor care fac obiectul măsurării și cuantificării sunt reunite sub termenul generic de **caracteristici** ale acestor fenomene sau procese.

În activitatea de măsurare și cuantificare din domeniul economico-social apar o serie de dificultăți, determinate de specificitatea acestui domeniu, cum ar fi: problema unităților de măsură, parțialitatea informațiilor, problema erorilor, problema imposibilității măsurării directe etc.

1.3.1 Unitățile de măsură

O problemă dificilă și complexă care apare în mod frecvent în procesul de măsurare din domeniul economico-social, este cea dată de faptul că unitățile de măsură utilizate pentru exprimarea sub formă numerică a caracteristicilor unui anumit fenomen nu sunt stabile. Cu o problemă de acest fel este confruntată, în special, activitatea de măsurare și cuantificare din domeniul economic.

Spre deosebire de cele mai multe domenii ale științei, în domeniul economic, unitățile de măsură cunosc modificări substanțiale în timp, modificări ce conferă rezultatelor măsurătorilor un puternic caracter de instabilitate. De cele mai multe ori, variația unităților de măsură este indusă chiar de modificări, de un anumit tip, ale fenomenelor și proceselor studiate, ale fenomenelor cu care acestea sunt, direct sau indirect, legate sau chiar a unor fenomene independente de fenomenele studiate.

1.3.2 Parțialitatea informațiilor

O altă problemă ce ridică o serie de dificultăți în măsurarea și cuantificarea din domeniul economico-social, este cea legată de *parțialitatea informațiilor* disponibile pentru activitățile de analiză a datelor.

Complexitatea fenomenelor economice și sociale, multitudinea a formelor de manifestare în timp și spațiu a acestora, fac ca obținerea de informații printr-o observare completă a desfășurării acestor fenomene să fie de cele mai multe ori imposibilă. În aceste condiții, modalitatea cea mai comodă, dar și cea mai eficientă, de obținere a informațiilor necesare efectuării analizelor sau realizării prediciilor este cea a *observării selective* a manifestării fenomenelor.

Această modalitate de investigare se bazează pe faptul că, în anumite condiții, cunoașterea caracteristicilor întregii populații poate fi obținută prin studierea directă numai a unui anumit număr de unități ce intră în compoziția acesteia, număr care este mult mai redus în comparație cu volumul total al populației. Alegerea unităților ce vor fi efectiv supuse procesului de măsurare și înregistrare se face după criterii și reguli foarte precise, fundamentate riguroz din punct de vedere statistică-matematică. Valorile pe care le iau caracteristicile unităților studiate prin intermediul acestui procedeu alcătuiesc așa-numitul *eșantion de observații*.

Investigarea fenomenelor și proceselor economice sau sociale pe baza informațiilor provenite din observări selective și concluziile rezultate în urma acestei investigări sunt puternic influențate de gradul în care informațiile ce alcătuiesc eșantionul sunt reprezentative la nivelul întregii populații de la care au fost obținute. De regulă, gradul de reprezentativitate al unui eșantion este invers proporțional cu gradul de heterogenitate a populației și direct proporțional cu numărul de unități incluse în eșantion.

În condițiile în care informațiile obținute de la unitățile selectate în eșantion nu produc într-o măsură acceptabilă comportamentul fenomenului la nivelul întregii populații, rezultatele analizelor efectuate vor fi incorecte, iar concluziile formulate pe baza acestora vor fi eronate. De aceea, în toate cazurile în care baza informațională utilizată în procesele de analiză a datelor este obținută pe calea observării selective, este necesară o laborioasă activitate de verificare a semnificației rezultatelor și de validare a concluziilor desprinse pe baza acestor rezultate.

Implicațiile negative ale caracterului de parțialitate a informațiilor din domeniul economico-social asupra valabilității concluziilor formulate în urma prelucrării acestor informații, pot fi mult diminuate, dacă organizarea cercetării selective este făcută înținând cont de o serie de principii de bază ale teoriei selecției, dacă în procesul de prelucrare a informațiilor sunt utilizate metodele și tehniciile statistică-matematice cele mai potrivite și dacă în extinderea rezultatelor nu sunt încălcate o serie de principii statistice elementare.

Respectarea cu strictețe a cerințelor de acest tip poate conduce la obținerea unor rezultate cu adevărat excepționale, chiar atunci când cantitatea de informație disponibilă este neglijabilă în raport cu dimensiunea și proporțiile fenomenului studiat.

1.3.3 Erorile

Impactul numeroaselor imperfecțiuni legate de măsurarea și cuantificarea din domeniul economico-social este concretizat, din punct de vedere al proceselor de analiză a datelor, în manifestarea unor distorsiuni și inexactități informaționale.

În mai toate cazurile, informațiile disponibile pentru efectuarea unor analize sau programe sunt afectate de erori. Faptul că informațiile disponibile sunt puternic afectate de erori, ale căror surse sunt numeroase, variate și dificil de controlat, impune utilizarea unor modalități specifice de abordare și a unor instrumente de lucru corespunzătoare.

Erorile ce însoțesc observațiile rezultate din măsurarea și cuantificarea fenomenelor și proceselor economice sau sociale pot fi grupate în două mari categorii: *erori cu caracter sistematic* și *erori cu caracter accidental, aleator*.

Erorile cu caracter sistematic

Sunt erorile care afectează în mod substanțial, semnificativ, natura și semnificația informațiilor și care pot determina denaturarea drastică a rezultatelor obținute și formularea unor concluzii fundamental eronate.

Sursele erorilor cu caracter sistematic sunt numeroase și extrem de variate: *gradul redus de reprezentativitate a informațiilor obținute pe bază de eșantionare, imperfecțiunile metodologice specifice calculului unor indicatori economici, raportările incorecte determinante de interesele de natură fiscală ale agenților economici, falsificările informaționale la nivel guvernamental etc.*

Definiție: *Eroarea cu caracter sistematic* poate fi definită ca reprezentând diferența dintre valoarea adevărată a unei mărimi la nivel de populație și valoarea obținută pentru această mărime în urma măsurării tuturor unităților elementare ale populației.

Problema cea mai dificilă legată de impactul negativ al erorilor cu caracter sistematic este aceea că, spre deosebire de erorile cu caracter întâmplător, în cazul acestor erori nu se manifestă tendința de compensare, astfel încât distorsionarea informației determinată de aceste erori are loc numai într-un sens.

În cazul erorilor de natură sistematică, influența negativă a acestora nu mai poate fi estompată de caracterul de masă al fenomenelor economico-sociale, care în cazul altor tipuri de erori poate conduce la compensarea efectelor erorilor.

Erorile cu caracter accidental

Sunt erorile a căror influență asupra semnificației informațiilor este neglijabilă; distorsiunile provocate de acest tip de erori la nivelul datelor ce descriu comportamentul fenomenelor economice sau sociale sunt, de regulă, de sens opus, astfel încât în numeroase situații are loc o compensare a influențelor acestora.

Definiție: *Eroarea cu caracter accidental, aleator* poate fi definită ca reprezentând diferența dintre valoarea unei mărimi calculate pe baza unităților aparținând unui eșantion și valoarea aceleiași mărimi calculată pe baza tuturor unităților populației.

Faptul că erorile de acest tip sunt distribuite după legea de probabilitate normală oferă posibilitatea abordării statisticice a influențelor cu caracter distorsionant pe care le pot avea acestea asupra datelor.

Pe de altă parte, în virtutea legii numerelor mari, are loc un efect de compensare a erorilor de tip accidental, impactul negativ al acestora asupra calității informațiilor utilizate în analiză devenind neglijabil.

În principal, sursele erorilor cu caracter accidental, întâmplător, ţin de o serie de factori cum ar fi: *imperfecțiunile instrumentelor de măsurare, deficiențe în înregistrarea valorilor caracteristicilor urmărite, efectuarea inexacă a unor calcule* etc.

O însemnatate cu adevărat excepțională pentru procesele de măsurare și cuantificare în care intervin erori cu caracter întâmplător, o are faptul că erorile de acest tip sunt distribuite după *legea normală*, de medie nulă.

Această proprietate este deosebit de utilă atât pentru măsurarea impactului pe care erorile îl au asupra informațiilor obținute în urma proceselor de măsurare, de cuantificare și de analiză a datelor, cât și pentru dezvoltarea unor proceduri specifice care au ca scop minimizarea influenței erorilor asupra rezultatelor obținute în analiza datelor.

Formarea valorii unei măsurători individuale, sub influența erorilor sistematice și accidentale, poate fi descrisă cu ajutorul relației următoare:

$$\text{Valoare măsurată} = \text{Valoare adevărată} + \text{Eroare sistematică} + \text{Eroare accidentală}$$

Efectuarea analizelor de date în condițiile existenței și manifestării erorilor de tip accidental, determină necesitatea utilizării unor metode și tehnici adecvate, de natură statistică-matematică.

1.3.4 Măsurarea indirectă

În numeroase situații din investigarea științifică, este imposibilă măsurarea directă a unor fenomene sau procese specifice domeniului economico-social. Situațiile de acest fel sunt întâlnite în cazul existenței factorilor latenți sau ascunși, care au caracter neobservabil și care, în mod implicit, nu pot fi supuși unui proces direct de măsurare.

Factorii latenți sunt expresia manifestării unor fenomene de maximă generalitate, cu semnificație și consistență foarte importante, caracterizate printr-o mare stabilitate a manifestărilor lor. Ei reprezintă *agregări* ale manifestărilor unei multitudini complexe de fenomene cu natură particulară. Datorită semnificației și importanței prin care factorii latenți se detașează de fenomenele care reprezintă forme particulare de manifestare ale acestora, în numeroase investigații științifice scopurile urmărite sunt direct și strâns legate într-o măsură mult mai mare de existența factorilor latenți, în comparație cu manifestările particulare ale acestora. Deși au caracter neobservabil, factorii latenți se manifestă totdeauna în mod indirect, prin intermediul unor forme particulare de manifestare, forme care pot avea caracter observabil și care pot fi supuse unui proces de măsurare.

În analiza datelor, se consideră că formele particulare prin care se manifestă factorii latenți reprezintă *indicatori* ai factorilor latenți. În acest fel, informațiile disponibile cu privire la acești indicatori pot fi folosite pentru a deduce informații cantitative și calitative despre factorii latenți.

Problema măsurării indirecte, prin intermediul unor indicatori specifici, a factorilor cu natură latentă, apare în mod frecvent în cazul investigării unor caracteristici psihosociale, în cercetările legate de cuantificarea nivelului de dezvoltare economică, nivelului de dezvoltare socială sau nivelului progresului tehnic, în construirea unor indicatori agregări etc. În aceste cazuri, se apelează la proceduri speciale de *estimare*, la metode și tehnici de măsurare *indirectă* sau la *metode de analiză multidimensională*, cum ar fi: *analiza componentelor principale*, *analiza factorială* sau *analiza corespondențelor*.

1.4. Definirea analizei datelor

Analiza datelor are ca obiectiv fundamental *extragerea informației relevante, semnificative*, care este conținută în date, în informația primară. Această informație este utilizată, în continuare, pentru rezolvarea unor probleme specifice ale analizei datelor: testare, estimare, interpretare, predicții etc.

Conceptul de analiză a datelor este un concept extrem de cuprinsător și dificil de definit din punct de vedere al sferei de cuprindere, iar metodele și tehniciile utilizate în analiza datelor variază de la cele mai simple modalități de prezentare grafică a datelor și calcul al unor indicatori statistici specifici, până la cele mai sofisticate metode de analiză multidimensională.

Definiție: *Activitatea de analiză a datelor* poate fi definită ca reprezentând o succesiune de operații de prelucrare și interpretare, operații efectuate asupra unor informații primare referitoare la fenomene și procese din realitatea economico-socială și bazate pe o mare varietate de metode și tehnici specifice, în scopul adâncirii cunoașterii comportamentului acestor fenomene și al formulării unor concluzii cu privire la specificitatea manifestării lor.

În figura următoare, este evidențiată o reprezentare simplificată a procesului general de analiză cantitativă, a modului în care intervin cele trei elemente menționate anterior și a principalelor faze ale acestui proces.

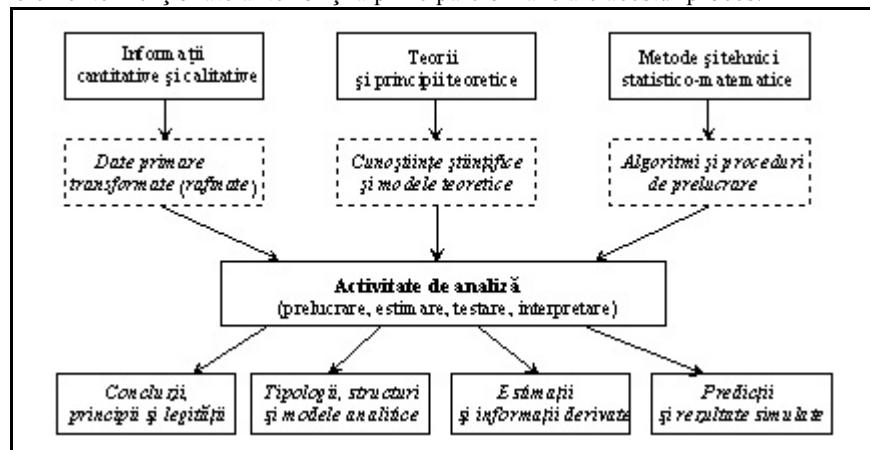


Figura 1.1: Reprezentarea simplificată a unui proces general de analiză cantitativă

Privit într-un mod foarte general, procesul de analiză a datelor poate fi prezentat ca o succesiune de operații sintetizate sub forma următoarelor activități:

- **formularea ipotezelor** cu privire la comportamentul fenomenului ce constituie obiectul studiului;
- **organizarea experimentelor** necesare măsurării caracteristicilor fenomenului studiat;
- **culegerea datelor** privind comportamentul fenomenului ;
- **analiza și interpretarea datelor** disponibile;
- **formularea concluziilor, efectuarea predicțiilor și luarea deciziilor.**

Scopul urmărit în cele mai multe probleme de analiză a datelor este legat de crearea condițiilor informaționale necesare pentru **efectuarea de predicții** cu privire la comportamentul fenomenelor investigate.

Deși între activitatea de analiză a datelor și activitatea de predicție există numeroase deosebiri, legate în primul rând de natura celor două activități, totuși, între cele două genuri de activități există o importanță suprapunere. Această suprapunere este determinată de *existența unor modalități comune de abordare, de utilizarea unor concepte teoretice identice și de folosirea aceleiași game de instrumente de lucru.*

Cu toate că obiectivele intermediare urmărite în activitatea de predicție sunt similare celor urmărite în analiza datelor, totuși, scopul final al oricărei activități de predicție este legat în mod direct de prefigurarea evoluției viitoare a fenomenelor și proceselor economice sau sociale, a modificărilor structurale ce pot fi înregistrate în viitor de aceste fenomene și procese.

Atingerea acestui scop este însă condiționată de desfășurarea unei laborioase activități de analiză a datelor cu privire la realitatea fenomenului care face obiectul predicției, activitate care, de regulă, precede procesul de predicție propriu-zis.

Conținutul activității de analiză a datelor din domeniul economico-social este strâns legat de natura claselor de probleme supuse rezolvării. În funcție de această natură, analiza datelor poate presupune o anumită succedare de activități specifice, utilizarea unei anumite game de metode și tehnici de lucru, urmărirea anumitor obiective.

În rândul categoriilor de probleme a căror rezolvare conduce la efectuarea unui proces de analiză a datelor menționăm:

- identificarea legităților ce guvernează mișcarea fenomenelor și proceselor economice sau sociale, a principalelor tendințe și regularități specifice evoluției acestora;
- identificarea principalilor factori sub a căror influență se formează comportamentul unor fenomene și procese;
- stabilirea sensului și intensității legăturilor cauzale manifestate între diferite fenomene și procese;
- determinarea gradului în care o mulțime de factori de influență contribuie la formarea unui anumit efect;
- verificarea unor ipoteze cu privire la existența unor legături de tip cauzal, la posibilitatea ca anumite caracteristici ale unor fenomene să înregistreze anumite valori specifice etc;
- ierarhizarea factorilor ce determină un anumit efect în funcție de importanța și semnificația influenței acestora;
- stabilirea modului în care comportamentul fenomenelor este afectat de anumite decizii sau măsuri de politică economică și socială;
- identificarea principalelor posibilități prin intermediu cărora comportamentul fenomenelor să poată fi influențat într-un sens dorit;
- determinarea sensurilor în care comportamentul unui anumit fenomen poate evoluă, a stărilor posibile în care acesta se poate afla în spațiu și timp sau efectuarea de predicții privind stările și evoluția acestui fenomen.

În funcție de specificul fiecăreia din aceste clase de probleme, activitatea de analiză a datelor se poate concretiza printr-un anumit gen de operații de modelare statistică-matematică și prelucrare informațională vizând:

- simplificarea și sintetizarea relațiilor de cauzalitate;
- măsurarea interdependențelor, cuantificarea influențelor și verificarea semnificației acestora;
- descrierea statistică-matematică a comportamentului fenomenelor;
- efectuarea de predicții;
- măsurarea gradului de omogenitate sau eterogenitate pentru anumite categorii de fenomene;
- clasificarea și ierarhizarea fenomenelor.

Fiecare din aceste tipuri de operații presupune utilizarea unor instrumente de lucru corespunzătoare, eficiente în raport cu specificul problemei considerate. Dintre principalele metode și tehnici proprii analizei multidimensionale a datelor și utilizate pentru rezolvarea problemelor enumerate anterior menționăm: *analiza componentelor principale, analiza factorială, analiza corespondențelor, tehniciile de scalare multidimensională, metodele și tehniciile de estimare, analiza corelațiilor canonice, analiza varianței, analiza regresiei liniare și neliniare, metodele și tehniciile de recunoaștere a formelor.*

1.5 Clasificarea metodelor de analiză a datelor

În general, în analiza de date, obiectul studiului este reprezentat de către o mulțime de date structurate sub formă unui anumit număr de variabile și a unui anumit număr de observații. Această structură informațională este reprezentată prin intermediu unei entități cunoscute sub numele de matrice de observații.

În funcție de natura variabilelor analizate și de informațiile existente aprioric cu privire la natura legăturilor cauză-efect în care sunt implicate variabilele, metodele de analiză a datelor pot fi grupate în două mari categorii: *metode de analiză a dependenței și metode de analiză a interdependenței.*

Metodele de analiză a dependenței sunt metode statistice utilizate în cazul în care variabilele considerate în cadrul unei analize pot fi grupate în două submulțimi și au ca scop *testarea prezenței sau absenței unor interdependențe* între cele două submulțimi de variabile.

În situația în care există informațiile necesare, astfel încât, variabilelor dintr-un set li se poate atribui semnificația de *variabile dependente*, iar variabilelor din celălalt set li se poate atribui semnificația de *variabile independente*, metodele de analiză a dependenței au ca scop de a determina dacă și cum variabilele independente influențează variabilele dependente, atât în manieră *individuală*, cât și în manieră *comună, simultană*.

De regulă, dându-se o mulțime de date reprezentate de observațiile efectuate asupra unor variabile, cu ajutorul metodelor statistice se poate verifica doar dacă există legături, interdependențe între variabile, putându-se, eventual, determina sensul

legăturilor (directe sau inverse) sau chiar măsura intensitatea acestor legături. Ceea ce nu se poate stabili cu ajutorul metodelor și tehnicilor statistice este natura, sensul relațiilor de cauzalitate, respectiv, care sunt variabilele dependente și care sunt variabilele independente ce determină evoluția, mișcarea celor dependente. Pentru a stabili cu exactitate relația *cauză-efect* este nevoie de informații suplimentare, apriorice, exogene în raport cu metodele și tehniciile statistice utilizate.

Există un anumit context al analizei datelor, în cadrul căruia delimitarea variabilelor dependente de variabilele independente este foarte clară, neechivocă. Acest context este întâlnit în cadrul *experimentelor controlate*, când, în urma stabilirii unor combinații de nivele ale variabilelor independente, sunt măsurate nivelele înregistrate pentru variabilele dependente, ca urmare a influenței exercitate de fiecare combinație de nivele ale variabilelor dependente.

În multe situații, analiza datelor este efectuată pe mulțimi de variabile pentru care nu există nici interesul și nici posibilitatea conceptuală de a separa din mulțimea variabilelor analizate o submulțime care să reprezinte variabilele dependente și o altă submulțime care să reprezinte variabile independente. În aceste situații, sunt utilizate metode de analiză specifice, cunoscute sub numele de *metode de analiză a interdependenței*.

Metodele de analiză a interdependenței sunt metodele statistiche utilizate în cazul în care nu există posibilitatea de a identifica în mulțimea variabilelor analizate variabile dependente și variabile independente și care au scopul de a stabili *din ce cauză și în ce măsură* variabilele analizate sunt legate între ele. Utilizarea acestor metode are o mare varietate de scopuri, dintre care, printre cele mai importante, amintim: *măsurarea gradului de interdependentă, identificarea variabilelor cu semnificație relevantă, identificarea unor categorii sau clase de variabile*.

1.5.1 Metode de analiză a dependenței

Metodele de analiză a dependenței pot fi clasificate în funcție de mai multe criterii, dintre care cele mai importante sunt următoarele:

- numărul de variabile dependente: *o singură variabilă dependentă sau mai multe variabile dependente*;
- numărul de variabile independente: *o singură variabilă independentă sau mai multe variabile independente*;
- tipul scalei pe care sunt măsurate variabilele dependente: *scală non-metrică sau scală metrică*;
- tipul scalei pe care sunt măsurate variabilele independente: *scală non-metrică sau scală metrică*;

Metodele de analiză a dependenței pot fi clasificate în funcție de numărul de variabile dependente și de numărul de variabile independente în două grupe:

- *metode unidimensionale de analiză a dependenței*;
- *metode multidimensionale de analiză a dependenței*.

În categoria metodelor unidimensionale se încadrează analiza regresiei simple, analiza discriminantului, procedurile unidimensionale de verificare a unor ipoteze etc.

Metodele de analiză multidimensională a dependenței se împart, în funcție de numărul variabilelor dependente, în două categorii:

- *metode de analiză multidimensională cu o singură variabilă dependentă*;
- *metode de analiză multidimensională cu mai multe variabile dependente*.

În rândul metodelor de analiză multidimensională cu o singură variabilă dependentă pot fi menționate: analiza regresiei multiple, analiza discriminantului, analiza varianței etc.

Ca metode de analiză multidimensională cu mai multe variabile dependente putem menționa: analiza corelațiilor canonice, analiza multidimensională a varianței etc. O clasificare sintetică a metodelor de analiză a dependenței, în funcție de criteriile de mai sus, este prezentată în tabelul următor.

Clasificarea metodelor de analiză a dependenței

Număr variabile independente		Număr variabile dependente			
		O variabilă		Mai multe variabile	
O variabilă	Scală non-metrică	Scală non-metrică	Scală metrică	Scală non-metrică	Scală metrică
	Scală metrică	• Analiza discriminantului (discretă)	• testul <i>t</i>	• Analiza multigrup a discriminantului (discretă)	• Analiza multidimensională a varianței (MANOVA)
Mai multe variabile	Scală non-metrică	• Analiza discriminantului • Regresie logistică	• Regresie simplă	• Analiza multigrup a discriminantului	• Analiza corelațiilor canonice
	Scală metrică	• Analiza discriminantului • Analiza simultană (MONANOVA)	• Analiza ANOVA	• Analiza multigrup a discriminantului (discretă)	• Analiza multidimensională a varianței (MANOVA)
	Scală metrică	• Analiza discriminantului • Regresie logistică	• Regresie multiplă	• Analiza multigrup a discriminantului	• Analiza corelațiilor canonice

1.5.2 Metode de analiză a interdependenței

Analiza interdependenței are ca scop să identifice și să evidențieze situațiile în care variabilele sunt corelate între ele și să explice modul în care are loc corelarea variabilelor supuse analizei. În funcție de numărul de variabile analizate, metodele de

analiză a interdependenței se împart în două categorii:

- *metode de analiză a interdependenței dintre două variabile;*
- *metode de analiză a interdependenței dintre mai multe variabile.*

În cadrul metodelor de analiză a interdependenței dintre două variabile putem menționa: analiza corelației simple, analiza bazată pe tabele de contingență etc.

Dintre metodele de analiză a interdependenței dintre mai multe variabile putem menționa: analiza componentelor principale, analiza factorială, analiza corespondențelor, analiza cluster etc. În tabelul următor este prezentată sintetic clasificarea metodelor și tehniciilor de analiză a interdependenței.

Clasificarea metodelor de analiză a interdependentelor

Număr variabile	Tipul scalei	
	Scală non-metrică	Scală metrică
<i>Două variabile</i>	<ul style="list-style-type: none"> • Tabele de contingență (bidimensionale) • Modele log-liniare 	<ul style="list-style-type: none"> • Analiza corelației simple
<i>Mai multe variabile</i>	<ul style="list-style-type: none"> • Analiza corespondențelor • Modele log-liniare • Tabele de contingență (multidimensionale) 	<ul style="list-style-type: none"> • Analiza componentelor principale • Analiza factorială

1.6 Analiza preliminară a datelor

Una din cele mai importante etape ale procesului de analiză a datelor din domeniul economico-social este cea a **analizei preliminare**, cunoscută și sub numele de **analiză exploratorie** a datelor. Analiza preliminară este o activitate anterioară, pregătitoare, a analizei propriu-zise a datelor, care are ca scop *initializarea* procesului de analiză. În cadrul acestei etape, informațiile primare disponibile sunt supuse unui proces de prelucrare în cadrul căruia are loc o filtrare a informațiilor din punct de vedere al semnificației și utilității pe care le au acestea în raport cu scopurile urmărite. Activitatea de analiză preliminară a datelor presupune utilizarea unei game variate de metode și tehnici statistico-matematice în scopul obținerii unei sugestive caracterizări statistică a acestor informații.

Preponderentă ca utilizare în faza de analiză preliminară a datelor și cu o utilitate de necontestat pentru activitatea științifică din această etapă, este **analiza grafică**. Utilizarea tehniciilor de analiză grafică în faza preliminară a analizei datelor vizează, în principal, următoarele categorii de probleme:

- *identificarea principalelor tendințe manifestate la nivelul observațiilor disponibile;*
- *depistarea principalelor legături existente între variabilele supuse analizei;*
- *detectarea valorilor extreme, izolate, a căror apariție în multimea datelor analizate nu se justifică din punct de vedere statistic.*

Există în prezent numeroase instrumente software, cu ajutorul cărora pot fi efectuate, într-o manieră comodă și eficientă, cele mai complexe reprezentări grafice.

1.7 Suportul software în analiza datelor

Apariția calculatoarelor bazate pe microprocesoare și larga accesibilitate a acestora din punct de vedere al prețului și software-ului aplicativ au făcut ca, în prezent, să nu existe domeniu de cercetare în care calculatorul electronic să nu fie instrumentul cel mai frecvent folosit pentru rezolvarea celor mai diverse probleme. Prelucrarea unui volum uriaș de informații, pe baza unor algoritmi de o complexitate deosebită și în condiții de precizie ridicată, nu se poate realiza decât cu ajutorul calculatorului electronic, chiar a unui calculator electronic performant.

În domeniul analizei datelor, există zeci și zeci de produse software, de dată mai mult sau mai puțin recentă. O inventariere, pe care nu o putem pretinde a fi exhaustivă, dar care poate fi considerată ca fiind minimală, pe care am făcut-o în anul 2005, cu privire la instrumentele software existente în lume la acest moment, ne-a condus la identificarea a peste 100 de astfel de instrumente software specializate. Numărul lor este cu atât mai semnificativ cu cât aceste produse software sunt de notorietate mondială, având o largă utilizare în domeniul analizei datelor.

Printre cele mai noi și mai performante instrumente software existente în prezent, destinate, total sau parțial, activităților de analiză a datelor, putem menționa:

- **SPSS 10.0**, SPSS Inc., Chicago, IL, USA, 1999;
- **STATISTICA 6.0**, StatSoft Inc., Tulsa, OK, USA, 2001;
- **S-PLUS 2000**, MathSoft Inc., Seattle, Washington, USA, 1999;
- **SAS 8.2**, SAS Institute Inc., Cary, NC, USA, 2001;
- **SYSTAT 9.0**, SPSS Inc., Chicago, IL, USA, 1999;
- **MATHEMATICA 4.0**, Wolfram Research, Champaign, IL, USA, 1999;
- **EVIEWS 3.0**, Quantitative Micro Software, Irvine, CA, USA, 2000.

În afară de existența unei mari varietăți de instrumente software destinate analizei datelor, în prezent se poate identifica o tendință de dezvoltare explozivă a preocupărilor din acest domeniu, tendință determinată de necesitățile informaționale tot mai mari și mai rafinate ale indivizilor și entităților economice și sociale.

Tema 2. Concepțe și operații specifice econometriei și analizei datelor

2.1 Concepțe fundamentale ale analizei datelor

Teoria și practica analizei datelor se bazează pe o serie de concepțe fundamentale, a căror definire este deosebit de importantă pentru înțelegerea demersului științific presupus de acest tip de analiză, pentru definirea și înțelegerea procedurilor și instrumentelor specifice ale acestei discipline, pentru desfășurarea eficientă a organizării și proiectării activităților de analiză și pentru interpretarea rezultatelor obținute din analiză.

Concepțele utilizate în analiza datelor și definirea riguroasă a acestora prezintă importanță nu numai pentru a facilita definirea și înțelegerea procedurilor, metodelor și tehnicielor de analiză a datelor, ci și pentru asigurarea unei modalități coerente și sugestive de interpretare și prezentare a concluziilor rezultate din analiză. Pe lângă concepțele proprii, specifice, în analiza datelor sunt utilizate numeroase concepțe care țin de domenii înrudite cu analiza datelor, cum ar fi: teoria probabilităților și statistică, econometria, teoria economică, informatică etc. Din acest motiv, mulțimea conceptelor cu care se operează în domeniul analizei datelor este extrem de cuprinzătoare și extrem de variată.

2.1.1 Data

Conceptul cel mai important și cel mai frecvent întâlnit în analiza datelor, care, de fapt, intră și în alcătuirea numelui acestui tip de activitate științifică, este reprezentat de **dată**. Importanța acestui concept pentru domeniul analizei datelor este, cu adevărat, covârșitoare, deoarece el este cel care definește atât intrările oricărui proces de analiză a datelor, *materia primă* supusă prelucrării, cât și, într-un sens general, ieșirile sale, rezultatele și concluziile obținute.

Datele pot fi privite ca reprezentând *semnale și mesaje* provenite din realitatea înconjurătoare, pe baza cărora receptorul își poate forma o anumită imagine despre respectiva realitate, poate obține un anumit grad de cunoaștere a acelei realități. Imaginea formată este cu atât mai fidelă în raport cu realitatea, cu cât cantitatea semnalelor și mesajelor este mai mare, respectiv, cu cât acestea sunt mai puțin afectate de perturbații și de distorsiuni. De regulă, datele nu sunt recepționate în mod pasiv de beneficiarul lor, fără nici un efort din partea acestuia. Aproape fără excepție, obținerea datelor necesare pentru orice activitate de analiză constituie un proces costisitor și laborios.

În analiza datelor, datele reprezintă expresia *cantitativă* sau *calitativă* a unor fapte reale, care sunt manifestări ale fenomenelor și proceselor investigate. Eterogenitatea fenomenelor și a manifestării concrete a acestora face ca datele referitoare la ele să fie extrem de variate.

Definiție: Datele reprezintă expresii cantitative și calitative ale unor fenomene și procese din realitatea înconjurătoare.

Datele pot să difere în funcție de mai mulți factori: de *sursa* care le-a generat, de *tipul* și de *natura* lor. Indiferent de varietatea lor, datele pot fi grupate în trei categorii fundamentale: date *cantitative*, date *calitative* și date *mixte*. Toate cele trei tipuri de date pot fi, însă, exprimate sub formă cantitativă.

2.1.2 Populația și eșantionul

Unul din concepțele fundamentale ale analizei datelor, de care este legată definirea multora dintre concepțele uzuale ale acestei discipline este cel de **populație statistică**.

În raport cu acest concept fundamental sunt definite majoritatea celorlalte concepțe utilizate în analiza datelor: eșantion, caracteristici, variabile, observații, parametri, grade de libertate.

Definiție: Populația sau *colectivitatea generală* este reprezentată de mulțimea tuturor măsurătorilor efective sau conceptuale care prezintă interes pentru cercetător sau experimentator.

În general, se poate spune că *populația statistică* reprezintă, obiectul de studiu global al analizei datelor. Aceasta deoarece majoritatea tehniciilor și instrumentelor de analiză a datelor au ca scop deducerea unor legități care guvernează populația statistică, obținerea unor estimări pentru o serie de mărimi specifice aceleiași populații statisticice, efectuarea de predicții referitoare la structurarea pe tipologii sau la evoluția populației statisticice.

Deși populația reprezintă obiectul investigației științifice, totuși, analiza datelor vizează în mod direct, aproape exclusiv, observațiile de la nivelul eșantionului.

Populația statistică poate fi definită ca reprezentând *totalitatea observațiilor posibile dintr-un studiu*. Generic, o unitate componentă a unei populații statisticice se numește **unitate elementară, element, individ, subiect, obiect, profil, formă, articol sau caz**. Ca exemple de unități elementare ale unei populații statisticice putem menționa: cumpărătorul, firma, locuitorul unei țări sau al unui oraș, produsul, familia etc.

În funcție de numărul, finit sau infinit, al elementelor din care este alcătuită o populație statistică, aceasta poate fi de două tipuri: **populație finită și populație infinită**.

Theoretic, într-o problemă de analiză a datelor pot fi studiate, fie toate observațiile posibile, adică întreaga populație, fie o parte, mai mare sau mai mică a acestora, numită **eșantion**.

Definiție: Eșantionul reprezintă o submulțime de măsurători selectate dintr-o populație, o submulțime a populației statisticice supusă investigației științifice.

Eșantionul are o importanță fundamentală în analiza datelor deoarece acesta, și nu populația totală, reprezintă, de fapt, baza informațională utilizată în procesele de analiză a datelor. Informațiile primare manipulate în activitatea de analiză a datelor sunt

de fapt rezultatele măsurătorilor efectuate la nivel de eșantion.

O modalitate de vizualizare a relației, a raportului în care se găsește eșantionul față de populația statistică este prezentată în figura următoare.

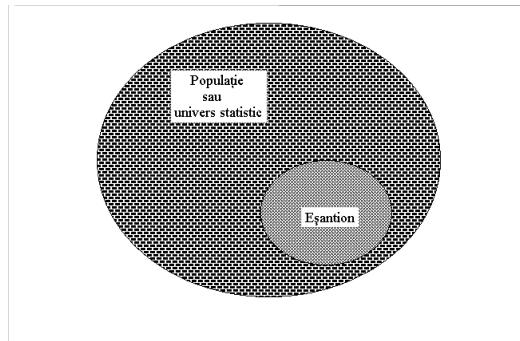


Figura 2.1: Ilustrarea relației dintre eșantion și populație

O foarte mare importanță principală pentru analizele cantitative bazate pe studiul eșantioanelor, o are postulatul statistic în conformitate cu care *un eșantion exprimă informațional într-o oarecare măsură populația din care a fost extras, proprietățile și structura populației fiind cu atât mai fidel exprimate de eșantion cu cât volumul acestuia este mai mare.*

2.1.3 Caracteristici și variabile

Din punct de vedere al informației statistice, o populație statistică prezintă interes nu în raport cu unitățile, ca atare, din care este alcătuită populația, ci în raport cu *trăsăturile, proprietățile* acestor unități.

Proprietățile unităților elementare aparținând unei populații statistic sunt numite în analiza datelor **caracteristici** sau **attribute**. Fiecare unitate elementară a populației investigate poate avea *o singură* caracteristică sau *mai multe* caracteristici. În cel de-al doilea caz, în studierea populației statisticе apare necesitatea abordării multidimensionale, necesitatea utilizării unor instrumente și tehnici de analiză specifice caracterului multidimensional.

În funcție de natura lor, caracteristicile unităților unei populații pot fi de două tipuri: **caracteristici calitative** și **caracteristici cantitative**.

Caracteristicile sau attributele unităților elementare ce alcătuiesc o anumită populație sunt elemente ale unei realități date, de natură empirică. De regulă, în activitatea științifică nu se operează cu elementele realității ca atare, ci cu **simboluri** care sunt *rezentări abstracte ale realității*.

Măsurarea caracteristicilor unităților unei populații este echivalentă cu atribuirea de simboluri, numerice sau nenumerice, acelor caracteristici. În general, simbolurile care pot să ia o varietate de valori se numesc **variabile**.

În cadrul demersurilor științifice care au ca scop investigarea fenomenelor și proceselor economice sau sociale, caracteristicile unităților unei populații sunt reflectate prin intermediul conceptului de **variabilă**, tocmai pentru a sugera natura schimbătoare a acestora, variabilitatea lor în timp și spațiu. Variabila este, poate, cel mai important concept vehiculat în cadrul oricărui proces de analiză a datelor, în raport cu care se definește întreaga succesiune de operații de prelucrare specifice acestui proces.

Informațiile care sunt elemente de intrare în procesele de analiză a datelor, reprezintă, aproape fără excepție, măsurători asupra unor caracteristici, măsurători care definesc valori ale variabilelor ce simbolizează caracteristicile populației analizate.

Definiție: *Variabila* reprezintă o abstractizare a mulțimii de valori posibile pe care le poate înregistra o caracteristică a unui anumit fenomen.

Varietatea fenomenelor economico-sociale și modalitățile diferite de exprimare a caracteristicilor acestora fac ca variabilele prin intermediul cărora sunt descrise aceste caracteristici să aibă natură diferită. Ca și caracteristicile populaților, după natura pe care o au, variabilele pot fi de două tipuri: **variabile calitative** și **variabile cantitative**.

În analiza datelor apare necesitatea tratării diferențiate a datelor de tip calitativ și cantitativ deoarece există diferențe substanțiale între aceste tipuri de date atât din punct de vedere al modalităților de abordare și interpretare, cât și din punct de vedere al metodelor și tehniciilor utilizate în analiză. Din aceste motive, se face o deosebire netă între variabilele de tip calitativ și variabilele de tip cantitativ.

Definiție: *Variabilele calitative* sunt variabile ce *diferă prin tip*, se referă la proprietăți nenumerice ale unităților elementare aparținând unei populații și nu pot fi exprimate numeric.

În cazul în care, în mod convențional, valorile lor sunt codificate prin numere, *această exprimare nu este relevantă numeric*. Variabile calitative sunt: sexul, opțiunea cumpărătorului, opțiunea alegătorului, profesia, starea civilă etc.

Definiție: *Variabilele cantitative* sunt variabile care *diferă prin mărime*, se referă la proprietăți numerice ale unităților elementare dintr-o populație și sunt exprimate în unități numerice: de lungime, de greutate, valorice etc.

Variabile cantitative sunt: prețul unui produs, cheltuielile lunare ale unei familii, salariul mediu lunar, venitul național, volumul fizic al producției etc.

În funcție de natura valorilor pe care le iau, variabilele se împart în două categorii: **variabile de tip discret** și **variabile de tip continuu**.

Definiție: *Variabilele de tip discret* sunt variabile care pot lua o mulțime limitată, finită de valori și care se mai numesc și **variabile categoriale**.

Valorile luate de variabilele discrete se numesc **alternative, categorii, variante** sau **modalități**. De regulă, variabilele

calitative sunt variabile de tip discret. Variabile de tip discret pot fi însă și unele variabile cantitative.

Definiție: *Variabilele de tip continuu* sunt variabile care pot lua valori aparținând unui interval continuu.

Practic, mulțimea valorilor posibile ale variabilelor de tip continuu este o mulțime infinită. De regulă, variabilele calitative nu sunt variabile de tip continuu.

2.1.4 Observații

Strâns legat de conceptul de variabilă, este un alt concept, la fel de important și frecvent utilizat în analiza datelor, și anume, conceptul de *observație*.

Definiție: *Observația* este reprezentată de *valoarea* sau *setul de valori* înregistrate pentru o anumită unitate elementară a populației, la una sau mai multe caracteristici ale acesteia.

De fapt, observațiile sunt valori pe care le iau variabilele supuse analizei, valori care sunt rezultate din măsurările efectuate asupra caracteristicilor unităților aparținând populației investigate.

Observația reprezintă în procesul de analiză a datelor unitatea elementară de informație utilizată în procesele de prelucrare, mulțimea observațiilor constituind *baza informatională* a analizei datelor. Practic, mulțimea de observații supuse procesului de analiză este echivalentă cu eșantionul, care, la rândul său, nu este altceva decât tot o mulțime de observații.

De cele mai multe ori, prin observație se înțelege chiar entitatea elementară care intră în alcătuirea populației analizate și de la care se obțin informații. În acest sens, observația este sinonimă cu *cazul, obiectul, individul, subiectul, articolul*.

2.2 Tipuri de date primare

În funcție de *modul* în care sunt obținute, datele primare pot fi clasificate în mai multe categorii. Vom prezenta, în continuare, două dintre clasificările cele mai importante ale datelor primare: clasificarea în funcție de *natura contextului* în care acestea sunt obținute și clasificarea în funcție de *modul de obținere* a acestora.

2.2.1 Date experimentale și date non-experimentale

Din punct de vedere al *naturii contextului* în care datele sunt obținute, al modului în care investigatorul controlează sau nu, în procesul obținerii datelor, populația supusă studiului, datele pot fi grupate în două categorii:

- *date experimentale*;
- *date non-experimentale* sau *date observaționale*.

Diferențele dintre datele experimentale și non-experimentale țin nu numai de natura contextului în care ele sunt obținute, ci și de modul în care ele sunt utilizate în analiza datelor, de metodele, instrumentele și procedurile utilizate pentru prelucrarea acestora.

2.2.1.1 Date experimentale

Datele experimentale sunt datele obținute prin organizarea unor *experimente de tip controlat*, desfășurate în condiții clare și prestabiliite. Contextul obținerii datelor de tip experimental este restricționat, prin impunerea unor reguli specifice.

În general, organizarea unui experiment controlat presupune, în primul rând, *izolarea fenomenelor și proceselor studiate*, precum și *eliminarea*, în cât mai mare măsură, a influențelor externe, care nu prezintă interes pentru analiză. O altă etapă importantă a experimentului controlat este aceea a *stabilității factorilor și cauzelor* importante, cu semnificație, care își exercită influența asupra mărimilor exogene, precum și a *alegerii nivelelor* la care influența acestora va fi urmărită. Cea de-a treia etapă a procesului de obținere a datelor experimentale constă în desfășurarea experimentului propriu-zis, etapă în care sunt *“dozați”*, în condiții specifice, factorii de influență. În această etapă, se fixează *nivele concrete* pentru fiecare din factorii de influență, se *crează condițiile necesare* pentru exercitarea influențelor factorilor și se *măsoară efectul* influenței factorilor asupra mărimii sau mărimilor de ieșire.

Datele obținute în urma unui astfel de proces, adică datele experimentale, sunt reprezentate de două categorii de informații: *valorile fixate pentru nivelurile factorilor* de influență și *valorile rezultate pentru variabila sau variabilele dependente*, în urma exercitării acestor influențe.

Definiție: *Datele experimentale* reprezintă informații obținute prin organizarea unor *experimente controlate*, în care influențele factorilor asupra efectului sunt controlate în mod direct, prin fixarea unor combinații precise de influențe.

Datele experimentale sunt caracteristice doar unor domenii de cercetare, și anume acelor domenii în care pot fi organizate experimente specifice, necesare obținerii acestor date. Experimentarea este posibilă doar în anumite domenii ale cunoașterii, cum ar fi, de exemplu, domeniul *științelor naturale*: fizică, chimie, biologie etc.

Într-o altă modalitate de exprimare, se poate spune că datele experimentale sunt *date de laborator*, prin *“laborator”* înțelegând aici o serie de condiții speciale, care se referă atât la o serie de restricții și instrumente specifice de măsurare, cât la modalitatea de desfășurare a unor procese cauzale specifice.

Spre deosebire de aceste domenii, în domeniul economico-social experimentarea este fie total imposibilă, fie posibilă, dar numai foarte rar și în condiții foarte restrictive și costisitoare.

2.2.1.2 Date non-experimentale

Datele non-experimentale, care se mai numesc și *date observaționale*, sunt datele obținute prin *“observarea”* fenomenelor și proceselor în mișcarea lor *naturală, liberă*, fără impunerea unor restricții, fără a se exercita un control de un anumit fel asupra fenomenelor și proceselor investigate.

Definiție: Datele non-experimentale reprezintă informații obținute prin observarea liberă a mișcării fenomenelor și proceselor studiate, fără intervenția directă a investigatorului asupra condițiilor în care se desfășoară această mișcare.

Obținerea datelor de tip non-experimental reprezintă rezultatul *observării pasive, constatării*. Intervenția observatorului, a celui care face măsurătorile, este de tip *ex-post*, are loc după ce desfășurarea fenomenelor și proceselor reale a avut loc.

Datele de tip non-experimental sunt datele specifice domeniului economico-social, domeniu în care organizarea de experimente este fie dificilă, fie imposibilă. Mai mult decât atât, complexitatea influențelor din domeniul economico-social, multitudinea interacțiunilor din acest domeniu, determină o relevanță foarte scăzută pentru eventualele date de natură experimentală.

2.2.2 Date de tip profil, serii de timp și date panel

Din punct de vedere *cronologic*, observarea unei populații de fenomene sau procese, în scopul obținerii informațiilor necesare, poate avea loc sub două forme diferite: *static* și *dinamic*. Din acest punct de vedere, datele primare pot să constituie fie o *imagină de tip static* a populației, în care aceste date să reprezinte informații legate de *starea unităților populației la un moment dat*, fie o *imagină de tip dinamic, evolutiv*, în care datele să reprezinte informații legate de *evoluția în timp* a unei sau unor unități ale populației.

Având în vedere aceste două modalități de a observa unitățile unei populații, datele primare pot fi grupate în trei categorii:

- *date de tip profil*;
- *date de tip serii de timp*;
- *date de tip panel*;

Ca și în cazul datelor experimentale, analiza datelor diferențiază semnificativ modalitățile de abordare, în raport cu fiecare din aceste categorii de date. Din acest motiv, considerăm că este necesară o scurtă prezentare a acestor trei tipuri de date.

2.2.2.1 Date de tip profil

Datele de tip *profil* reprezintă rezultate ale unor măsurători efectuate, *la un moment dat*, asupra uneia sau mai multor caracteristici, *de-a lungul* unităților populației, adică pe mulțimea unităților sau obiectelor care sunt supuse studiului.

Datele de tip profil se mai numesc *date de tip secvență* sau *date de tip secțiune* și reprezintă *"tăieturi informaționale"* efectuate într-o anumită populație la un moment dat, "tăieturi" care sunt de tip *transversal*, în raport cu axa timpului.

Definiție: Datele de tip *profil* reprezintă informații obținute prin măsurători de natură *statică*, efectuate asupra caracteristicilor unor unități ale unei populații, *la același moment de timp*.

O *observație* în contextul datelor de tip profil este reprezentată de valoarea sau de valorile unei singure entități, ale unei singure unități din populație. Numărul de observații coincide, în cazul datelor de tip profil, cu numărul de unități observate și înregistrate. Datele de tip profil nu încorporează în semnificația pe care acestea o poartă, influența timpului asupra formării caracteristicilor la nivelul populației și sensul surgerii timpului, nici în mod explicit și nici în mod implicit.

Ca exemple de date de tip profil, putem menționa: datele referitoare la salariul individual dintr-o lună al lucrătorilor unei firme; datele referitoare la populația medie a statelor lumii într-un anumit an; datele referitoare la rata inflației înregistrată de țările lumii într-o anumită perioadă; sexul cumpărătorilor ce cumpără un anumit bun într-o anumită perioadă; numărul mediu înregistrat de populația județelor unei țări într-un anumit an; volumul anual al vânzărilor unor mărci de autoturisme, numărul voturilor înregistrate de partidele înscrise într-o campanie electorală etc.

De regulă, datele de tip profil se referă la *starea* pe care o au la un anumit moment indivizii aparținând unor anumite colectivități, gospodăriile, firmele, ramurile, unitățile administrativ-teritoriale, țările lumii etc.

2.2.2.2 Date de tip serii de timp

Datele de tip *serii de timp*, numite și *serii cronologice* sau, pur și simplu, *serii de timp*, reprezintă rezultate ale unor măsurători efectuate asupra caracteristicilor unei unități a populației studiate, *de-a lungul timpului, la momente succesive ale evoluției* acesteia, *la anumite intervale de timp*.

Intervalele de timp pentru care se fac măsurătorile pot fi reprezentate de: ore sau fracțiuni de ore, zile, săptămâni, decade, luni, trimestre, semestre, ani. Deoarece intervalele sunt egale și reprezintă surgereea timpului, observațiile rezultate în urma acestor măsurători sunt *successive* și, de regulă, *echidistante* în timp.

Definiție: Datele de tip *serii de timp* sau *seriile cronologice* reprezintă informații obținute prin măsurători de natură *dinamică*, efectuate asupra caracteristicilor unei unități a unei populații *la momente sau în intervale succesive de timp*.

Datele reprezentate de seriile de timp se referă la *evoluția în timp* a stării unui individ, gospodării, zone geografice, țări etc. Datele de acest tip pot fi date de tip *interval* sau date de tip *moment*.

Datele de tip *interval* sunt datele care se referă la caracteristici care sunt mărimi de tip *stoc*, în timp ce *datele de tip moment* sunt date care se referă la caracteristici care sunt mărimi de tip *flux*. Și în acest caz, datele de tipul seriilor de timp pot fi privite ca reprezentând *"secțiuni informaționale"*, însă aceste secțiuni sunt *de-a lungul axei timpului, de-a lungul evoluției*, adică sunt *secțiuni longitudinale* în raport cu axa timpului.

2.2.2.3 Date de tip panel

Datele de tip *panel* sunt date care reprezintă *combinății, mixturi* ale datelor de tip profil și datelor de tipul seriilor de timp. Ele sunt rezultate ale măsurătorilor efectuate asupra caracteristicilor unor unități individuale, atât *de-a lungul unuiașilor individuale*, cât și *de-a lungul timpului*.

Definiție: Datele de tip *panel* reprezintă informații obținute prin măsurători mixte, de natură *statică* și de natură *dinamică*, efectuate asupra caracteristicilor *acelorăși* unități ale unei populații *la momente sau în intervale succesive de timp*.

Datele de tip panel pot fi imaginatate ca reprezentând *"tăieturi informaționale mixte"*, transversale și longitudinale, în raport

cu axa timpului. În cazul datelor de tip panel, observarea se face într-o notă de *simultaneitate*: atât asupra mai multor unități ale populației, cât și asupra evoluției în timp a acestor unități. Exemplul cel mai sugestiv pentru datele de tip panel este cel al *bugetelor de familie*, în contextul cărora se fac înregistrări pe perioade de mai mulți ani a veniturilor și cheltuielilor tuturor familiilor care alcătuesc eșantionul respectiv.

2.2.3 Scala de măsurare a variabilelor

Așa cum am menționat anterior, "materia primă" în analiza datelor este reprezentată de măsurătorile efectuate asupra unor caracteristici ale populației statisticice. În analiza datelor, aceste măsurători sunt considerate a reprezenta *valori* ale unor variabile definite în raport cu caracteristicile analizate. Indiferent de tipul ei, valoarea înregistrată de o caracteristică la nivelul unui obiect este totdeauna rezultatul exprimării univoce a valorii respectivei caracteristici în funcție de "gradățile" unei anumite *scaale*.

2.2.3.1 Definirea scalei

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

Atribuirea de numere sau simboluri pentru caracteristicile sau proprietățile unor obiecte se face pe baza *respectării unor reguli de stabilire și prin utilizarea unor proceduri specifice*. De exemplu, dacă obiectul studiului este reprezentat de indivizi care sunt potențiali cumpărători ai unui anumit produs, atunci caracteristicile cărora este necesar să li se atribu numere sau simboluri pot fi: vârstă, venitul, sexul, profesia etc.

Măsurarea caracteristicilor sau proprietăților unor obiecte sau subiecți este totdeauna caracterizată de o anumită specificitate, determinată de natura caracteristicii măsurate, și presupune, cu necesitate, existența unor repere, a unor sisteme de referință, cunoscute sub numele de *scală*. Ca element fundamental al procesului de măsurare a caracteristicilor fenomenelor și proceselor economice, scala poate fi definită sub forma următoare.

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

- a stabili o mulțime de valori posibile ale variabilei, mulțime numită și spațiu de selecție;
- a preciza regulile după care sunt atribuite simboluri pentru elementele unei realități date, adică a defini o structură a spațiului de selecție.

În funcție de natura variabilelor exprimate cu ajutorul lor, există patru tipuri de scale, pe care le vom defini în cele ce urmează.

2.2.3.2 Tipuri de scale

Ca și procesul de măsurare ca atare, scala sau sistemul de referință 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* și *scala raport*. Primele două tipuri de scale sunt *scale de tip non-metric*, iar ultimele două sunt *scale de tip metric*.

2.2.3.2.1 Scala nominală

Scala nominală este o scală *non-metrică*, pe baza căreia *valorile variabilelor sunt definite prin intermediul simbolurilor nenumeric*. 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, totuși, pur convenționale.

Definiție: *Scala nominală* este o scală *non-metrică*, prin intermediul căreia valorilor posibile ale caracteristicilor măsurate li se atribuie simboluri fără relevanță numerică, în funcție 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* și sunt variabile a căror formă de exprimare este de tip atrutiv și care pot fi folosite numai pentru stabilirea apartenenței la o anumită clasă a entității 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* și pot fi utilizate numai în scopuri de clasificare de tip calitativ, natura nenumerică a acestor variabile făcând imposibilă utilizarea lor pentru comparații, 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 evidențiate 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 mulțimea acestor simboluri. De exemplu, mulțimile:

- {"masculin", "feminin"},
- {"industria", "agricultură", "construcții", ...},
- {"muncitor", "țăran", "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 subiecții studiați nu pot fi comparați 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 puțin, "în ce măsură" un subiect este situat mai bine decât altul.

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

Pentru caracteristicile măsurate pe scara nominală, poate fi calculat un număr limitat de indicatori statistici, care reprezintă, de fapt, *contorizări* ale simbolurilor apărute pe scara nominală. Acești indicatori sunt *modulul* și *frecvența*. În cazul caracteristicilor măsurate pe scara nominală poate fi evidențiată și *distribuția de frecvență*.

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 invariant a scalei nominale este reprezentată de operația de *recodificare*, această operație neafectând apartenența la o anumită clasă a valorilor măsurate pe această scală.

2.2.3.2.2 Scala ordinală

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 preferințelor consumatorilor*.

Scala ordinală permite clasificarea valorilor unei variabile în funcție de rangul acestora, însă *diferențele între ranguri nu sunt relevante și nu au sens*. Acest tip de scală nu dă posibilitatea stabilirii gradului în care caracteristicile a două entități distincte diferă între ele (mai mult, mai puțin).

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

Variabilele măsurate pe această scală se numesc *variabile ordinate*, sunt variabile calitative de tip discret și nu pot fi exprimate sub o formă numerică reală. Ca exemple de variabile ordinate putem menționa: *categoria de venit* (mic, mediu, mare), *nivelul studiilor* (elementare, medii, superioare), *preferința 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 subiecților supuși analizei determină o diferențiere a subiecților din punct de vedere al *poziției pe care fiecare dintre aceștia 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 scara ordinală sunt *valori ordinate* sau *note*, cunoscute și sub numele de *ranguri*. Aceste valori li se atribuie fie *numere de ordine*, fie *simboluri* care evidențiază o anumită ordine a valorilor caracteristicii.

Pe scara ordinală, două valori diferite ale unei caracteristici sunt evidențiate prin intermediul a două *ranguri* diferite, adică prin intermediul a două poziții diferite în cadrul ierarhiei. Elementele scalei ordinate, "diviziunile" acesteia, sunt reprezentate de *numerele* sau de *simbolurile* folosite pentru reprezentarea rangurilor, respectiv de pozițiile posibile în respectiva ordonare. Scara nominală este reprezentată chiar de multimea acestor numere sau simboluri.

Cu toate că valorile caracteristicilor de tip ordinal nu sunt numere propriu-zise, ele diferențiază, totuși, poziția unui subiect în raport cu un alt subiect, "spun ceva" despre această poziție. Valorile unei caracteristici măsurate pe scara ordinală permit doar *ordonarea* subiecților din punct de vedere al acestei caracteristici, determinând o ierarhizare a subiecților sau obiectelor.

Prin intermediul valorilor pe care le pot lua caracteristicile măsurate pe scara ordinală, indivizii se diferențiază 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 ordinate pentru măsurarea caracteristicilor măsurabile pe această scală generează *ierarhii, ordonări ale subiecților*.

Măsurarea pe scara ordinală *permite comparații* între subiecți din punct de vedere al caracteristicii măsurate, dar aceste comparații se referă numai la modul în care un subiect "este situat" în raport cu altul, fără a se putea spune și "în ce măsură" subiecții diferă între ei după caracteristica respectivă. Diferențele dintre două valori succese de pe scara ordinală nu pot fi considerate ca fiind egale, ele nedeterminând o distanță 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 scara ordinală, pot fi calculați o serie de indicatori statistici cum ar fi: *modulul, mediana, coeficientul de corelație a rangurilor, frecvența*. De asemenea, pentru caracteristicile de tip ordinal se poate evidenția și *distribuția de frecvență*. Este important să se facă, în acest context, precizarea că *media și diferențele valorilor variabilelor ordinate sunt nerelevante*, nu au sens informațional și nici sens logic.

Singura transformare invariantă a scalei ordinate este *translația*, adică transformarea care păstrează ordinea valorilor unei variabile. Analitic, acest tip de transformare invariantă a scalei ordinate poate fi definit astfel:

$$y=a+x$$

unde a este o constantă, pozitivă sau negativă, care dă sensul și mărimea translației valorilor scalei ordinate, valori reprezentate de x.

2.2.3.2.3 Scala interval

Este o scală *quasi-metrică* pe care se poate defini un punct de referință, dar acest punct nu este o origine "zero" reală, ci convențională, arbitrară. Valoarea "zero" pe acest tip de scală nu indică niciodată absența fenomenului măsurat. Deci, pentru scara de tip interval, *originea scalei este arbitrară*, având importanță doar *scalarea valorilor în interiorul intervalului*.

Această scăă permite scalarea valorilor unei variabile în cadrul unui interval de valori reale, fără a avea însă o origine precisă a acestor valori. Cu valorile măsurate pe această scăă pot fi efectuate diferențe, raporturi ale diferențelor, dar nu are sens determinarea raportului a două valori.

Definiție: *Scala interval* este o scală quasi-metrică, prin intermediul căreia valorilor posibile ale caracteristicilor măsurate li se atribuie valori numerice, fără ca pentru aceste valori numerice să existe o origine prestabilită.

Variabilele măsurate pe scara interval se numesc *variabile tip interval* și sunt variabile cantitative. Ele pot fi utilizate în comparații al căror rezultat permite o exprimare numerică.

În cazul măsurării pe scara interval, diferența dintre două valori succesive ale scalei are o semnificație numerică sigură, permitând măsurarea modului în care subiecții se distanțează din acest punct de vedere. O astfel de scală este, de exemplu, cea în contextul căreia măsurarea caracteristicilor constă în acordarea unui *număr de puncte* sau în acordarea unei *note*, în funcție de importanța pe care o are caracteristica respectivă la nivelul unui subiect, în funcție de *magnitudinea* sa.

O caracteristică a scalei de tip interval este aceea că evaluarea caracteristicii măsurate nu este afectată dacă scara este *translatată* sau dacă scara este *multiplicată* cu o anumită constantă. Translatarea este echivalentă cu o schimbare a originii, care, oricum, este o origine arbitrară. Multiplicarea este echivalentă cu o mărire proporțională a distanțelor dintre valorile scalei, adică o mărire care conservă proporțiile între aceste distanțe. În consecință, se poate spune că transformarea până la care scara interval rămâne invariantă, este transformarea de tip liniar următoare:

$$y = a + bx$$

unde a și b reprezintă două constante reale, iar x și y reprezintă valorile scalei originale, respectiv cele ale scalei transformate.

Operațiile care pot fi efectuate cu valorile măsurate pe scara de tip interval sunt mai numeroase decât cele care sunt posibile pe scara nominală și ordinală. În plus față de operațiile permise pe primele două scale, scara interval mai permite: *calculul mediei*, *calculul abaterii standard*, *calculul momentelor*, *calculul coeficienților de corelație Pearson*. Ca exemplu de variabilă tip interval, putem menționa variabila reprezentată de *durata programului de lucru*, timpul, ca variabilă specifică seriilor cronologice etc.

2.2.3.2.4 Scala raport

Scara tip raport este scara care are toate proprietățile scalei de tip interval, însă, în plus față de aceasta, are o origine naturală, neconvențională, care nu poate fi schimbată. Este o scă *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 absența proprietății, caracteristicii. În plus față de scalele precedente, pe această scă este *definit și raportul valorilor*, adică se poate compara de câte ori o valoare este mai mare decât alta.

Definiție: *Scala raport* este o scă metrică, prin intermediul căreia valorilor posibile pe care le pot lua caracteristicile măsurate li se atribuie numere definite în raport cu o origine prestabilită.

Scara raport este invariantă până la o transformare proporțională pozitivă, adică până la transformarea:

$$y = ax$$

Variabilele măsurate pe scara raport se numesc *variabile tip raport* și sunt variabile cantitative. Cu aceste variabile sunt permise toate operațiile definite pentru variabilele numerice.

Ca exemple de variabile tip raport putem menționa: *prețul, venitul, vârstă, salariul, profitul, volumul vânzărilor, numărul cumpărătorilor* etc.

2.3 Moduri de reprezentare a datelor

Pentru a se asigura o manipulare mai convenabilă și 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 *structuri simple și clare* a datelor, cât și avantajul de a oferi posibilitatea generalizării conceptului de mulțime de date.

În cele mai multe ipostaze din analiza datelor, matricea este entitatea care definește și, în același timp, conține totalitatea informațiilor, totalitatea datelor, supuse procesului de analiză.

În principiu, datele primare sunt reprezentate în analiza de date sub trei forme matriciale principale: *matrici de observații, matrici sau tabele de contingență și matrici sau tabele de proximitate*.

2.3.1 Matrici de observații

O *matrice de observații* 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 și numele generic de *scoruri*. Matricile de observații se mai numesc și 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 observații are forma următoare:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots \\ x_{T1} & x_{T2} & \dots & x_{Tn} \end{pmatrix}$$

unde un element x_{ij} reprezintă valoarea înregistrată pentru cea de-a j -a caracteristică a obiectului i . O linie i a matricii de observații X definește un obiect O_i și reprezintă valorile înregistrate de acest obiect la cele n caracteristici pe care le posedă. O coloană j a matricii de observații X reprezintă valorile înregistrate de caracteristica j pe mulțimea tuturor celor T obiecte supuse analizei. De regulă, în analiza de date, fiecare linie a matricii de observații X este numită *observație* și fiecare coloană a acestei matrici este numită *variabilă*.

În multe situații, nu pot fi obținute informații despre toate caracteristicile tuturor obiectelor supuse analizei. În cazul în care datele ce definesc obiectele nu sunt complete, matricea de observații definită mai sus poartă numele de ***matrice de observații cu valori omise***.

2.3.2 Matrici de contingență

Sunt tablouri rectangulare de dimensiune $m \times n$, utilizate pentru reprezentarea datelor referitoare la frecvențele relative sau absolute înregistrate pe o mulțime 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 contingență reprezintă valorile posibile ale primei variabile discrete, iar coloanele acestei matrici reprezintă valorile posibile ale celei de-a două variabile discrete. În analiza datelor, matricile de contingență se mai numesc și matrici de tip "***modalități × modalități***".

Un element x_{ij} reprezintă frecvența, absolută sau relativă, a obiectelor pentru care prima variabilă ia valoarea u_i și cea de-a două variabilă ia valoarea v_j . Acest element arată la câte obiecte cele două variabile analizate au simultan valorile u_i și v_j .

2.3.3 Matrici de proximitate

Sunt matrici pătratice de dimensiune $n \times n$, 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 coeficienți de similaritate, coeficienți de nesimilaritate sau distanțe. Un element x_{ij} din această matrice măsoară gradul de proximitate dintre obiectul i și obiectul j .

Matricile de proximitate se mai numesc și matrici de tip "***obiecte × obiecte***" și sunt utilizate în problemele de clasificare cu ajutorul tehniciilor de tip cluster și în problemele de scalare multidimensională.

Tema 3. Transformarea și summarizarea datelor primare

3.1 Transformarea datelor primare

Sub forma lor inițială, rezultată din procesul de măsurare pe o scală corespunzătoare și nesupuse nici unui proces de transformare sau de prelucrare, datele sunt cunoscute sub numele de *date primare*, *date brute* sau *date originale*. În majoritatea cazurilor, în analizele propriu-zise datele nu sunt folosite sub forma lor primară, brută. De obicei, înainte de a fi utilizate, datele brute sunt supuse la două categorii de operații preliminare: *operații de rafinare* și *operații de transformare*.

3.1.1 Rafinarea datelor

Sunt situații în care, din diferite motive, este imposibilă utilizarea datelor sub forma lor brută, primară. Pentru a putea fi utilizate, datele primare trebuie să fie supuse mai întâi unui proces de *purificare*, *de rafinare*, care să le asigure consistență, relevanță și comparabilitate.

Necesitatea rafinării datelor este determinată de numeroși factori, însă cei mai importanți dintre aceștia sunt cei legați de existența *datelor omise* și a *datelor necomparabile*.

Un exemplu care poate să sugereze necesitatea și utilitatea operațiilor de rafinare este cel reprezentat de necesitatea de a asigura comparabilitatea unor date privind indicatorii macroeconomici.

Deoarece majoritatea indicatorilor macroeconomici sunt exprimati sub formă valorică, mărimea acestora este artificial și puternic influențată de evoluția prețurilor. Această influență face ca valorile din diferite perioade ale acestor indicatori să nu fie comparabile între ele, deoarece influența perturbatorie a evoluției prețurilor nu este uniformă de la o perioadă la alta. De aceea, seriile de timp referitoare la evoluția indicatorilor macroeconomici trebuie supuse unor operații de *curățire*, *de rafinare*. Rafinarea datelor include o serie de operații specifice, dintre care mai frecvent utilizate sunt cele de *interpolare*, *de extrapolare* și de *ajustare*.

3.1.1.1 Interpolarea datelor

Necesitatea interpolării datelor apare în legătură cu seriile de timp, în situațiile în care, în datele analizate, există observații omise. Interpolarea reprezintă una dintre metodele cele mai frecvent utilizate pentru *completarea* datelor omise.

Prin metoda interpolării, completarea unei valori omise se face pe baza utilizării celor două valori care o încadrează, adică a celor două valori vecine. Pentru a deduce din cele două valori cunoscute valoarea pentru observația omisă, se pot folosi mai multe modalități de interpolare: *liniară*, *exponențială* etc.

În cazul interpolării liniare valoarea care va înlocui observația omisă se determină ca medie aritmetică a celor două valori care încadrează valoarea omisă. Această metodă poate fi aplicată numai în cazul în care observațiile sunt egal distanțate în timp.

O altă modalitate de interpolare a valorilor omise este aceea a folosirii analizei de regresie. Pentru aceasta se construiește un model de regresie adecvat, în care variabila dependentă este reprezentată de variabila ale cărei observații conțin valori omise, iar variabila independentă este *tempul*.

Forma modelului se alege în funcție de tendința manifestată de observațiile variabilei dependente. Pe baza observațiilor existente cu privire la variabila dependentă, mai puțin cele omise, se estimează parametrii modelului, după care, cu ajutorul modelului și a valorilor cunoscute pentru variabila *temp*, se determină valorile care vor înlocui observațiile omise.

3.1.1.2 Extrapolarea datelor

Este o altă metodă de rafinare a datelor, care se folosește tot pentru completarea unor valori omise, însă în cazuri mai speciale. Spre deosebire de cazul precedent, valorile omise sunt la una din extremitățile seriilor de timp, astfel încât aceste valori au doar un singur vecin, ceea ce face imposibilă obținerea valorilor pentru observațiile omise prin interpolare, adică prin "medierea" după o anumită regulă a valorilor vecine.

Completarea valorilor omise poate fi făcută în acest caz aplicând aceeași tehnică și aceleași reguli ca mai sus, adică utilizând analiza de regresie. Cu ajutorul modelului rezultat în urma estimării, pot fi produse valori care să înlocuiască observațiile omise atât de la începutul, cât și de la finalul seriei de timp, prin simpla introducere în model a valorii momentului de timp ce corespunde observațiilor omise.

3.1.1.3 Ajustarea datelor

Operațiile de ajustare a datelor se aplică tot în cazul datelor de tipul seriilor de timp, și anume în situațiile în care datele conțin o serie de perturbații cunoscute sub numele de *zgomote*. Cele mai frecvent întâlnite perturbații care influențează datele sunt *perturbațiile aleatoare*, cunoscute și sub numele de *zgomot alb*.

Operațiile de ajustare au ca scop *netezirea* seriilor de date, prin eliminarea perturbațiilor incorporate în datele seriei de timp, perturbații care pot fi reprezentate componente accidentale sau ciclice ale evoluției fenomenelor sau proceselor studiate.

Operațiile de ajustare a seriilor de timp mai sunt cunoscute și sub numele de *operații de filtrare* a datelor, iar seriile de date rezultate în urma acestor operații sunt cunoscute sub numele de *trend*.

3.1.2 Prelucrarea preliminară a datelor

Așa cum am arătat anterior, înainte de utilizarea lor în analiza datelor, datele originale sunt supuse unui proces de transformare, de prelucrare preliminară.

Două dintre cele mai caracteristice operații pentru acest proces sunt reprezentate de *operația de centrare a datelor originale* și de *operația de standardizare a datelor originale*.

3.1.2.1 Centrarea observațiilor

Operația de *centrare* a datelor constă în substituirea valorii fiecărei observații aparținând unei variabile cu o nouă valoare, reprezentând abaterea valorii originale față de media calculată prin luarea în considerare a observațiilor inițiale.

Dacă analiza presupune existența unui număr de n variabile și a unui număr de T observații, atunci operația de centrare a observațiilor variabilei \mathbf{x}_i constă în calculul noilor observații, adică al valorilor centrate, după relația:

$$\mathbf{x}_{ti}^c = \mathbf{x}_{ti} - \bar{\mathbf{x}}_i,$$

unde $\bar{\mathbf{x}}_i$ reprezintă media celei de-a i -a variabile.

Datorită faptului că suma abaterilor valorilor originale ale observațiilor față de medie este totdeauna nulă, adică:

$$\sum_{t=1}^T \mathbf{x}_{ti}^c = \sum_{t=1}^T \mathbf{x}_{ti} - \sum_{t=1}^T \bar{\mathbf{x}}_i = T \left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_{ti} \right) - \sum_{t=1}^T \bar{\mathbf{x}}_i = T \bar{\mathbf{x}}_i - T \bar{\mathbf{x}}_i = 0.$$

Operația de centrare a valorilor observațiilor efectuate asupra unei caracteristici va face ca variabilele centrate să aibă *media nulă*:

$$\bar{\mathbf{x}}_i^c = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_{ti}^c = 0.$$

În cazul în care variabilele originale sunt centrate, ca urmare a faptului că aceste variabile sunt de medie nulă, varianța unei variabile este proporțională cu pătratul lungimii vectorului reprezentat de observațiile respectivei variabile, iar abaterea standard este proporțională cu lungimea același vector.

Dacă v este o variabilă centrală, atunci cele T observații ale acesteia, $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_T$, definesc un punct sau un vector în spațiul T -dimensional al observațiilor. Varianța variabilei centrate v este, în acest caz:

$$s_v^2 = \frac{1}{T-1} \sum_{t=1}^T (\mathbf{v}_t - \bar{\mathbf{v}})^2 = \frac{1}{T-1} \sum_{t=1}^T v_t^2.$$

Aceasta înseamnă că varianța variabilei centrate v poate fi scrisă în funcție de lungimea vectorului v , astfel:

$$s_v^2 = \frac{1}{T-1} \|v\|^2,$$

unde $\|v\|$ reprezintă lungimea vectorului v :

$$\|v\| = \sqrt{v_1^2 + v_2^2 + \dots + v_T^2}.$$

În mod similar, abaterea standard a variabilei centrate v poate fi scrisă în funcție de lungimea vectorului v astfel:

$$s_v = \frac{1}{\sqrt{T-1}} \|v\|.$$

Dacă v și w sunt două variabile centrate, atunci covarianța dintre aceste variabile poate fi exprimată în funcție de produsul scalar al vectorilor v și w care reprezintă observațiile celor două variabile. Covarianța dintre variabilele centrate v și w este dată de relația:

$$s_{vw} = \frac{1}{T-1} \sum_{t=1}^T (\mathbf{v}_t - \bar{\mathbf{v}})(\mathbf{w}_t - \bar{\mathbf{w}}) = \frac{1}{T-1} \sum_{t=1}^T v_t w_t.$$

Rezultă că, în cazul variabilelor centrate v și w , covarianța este proporțională cu produsul scalar al vectorilor v și w care reprezintă observațiile celor două variabile:

$$s_{vw} = \frac{1}{T-1} v^t w,$$

unde $v^t w$ reprezintă produsul scalar al vectorilor v și w .

Coefficienții de corelație de tip Pearson pentru variabilele centrate pot fi și ei exprimați în aceeași manieră. Coeficientul de corelație dintre variabilele centrate v și w este dat de relația:

$$r_{vw} = \frac{s_{vw}}{s_v s_w} = \frac{\frac{1}{T-1} v^t w}{\frac{1}{\sqrt{T-1}} \|v\| \cdot \frac{1}{\sqrt{T-1}} \|w\|}.$$

Rezultă că, în cazul variabilelor centrate, coeficientul de corelație dintre două variabile este raportul dintre produsul scalar al vectorilor ce reprezintă observațiile asupra variabilelor și produsul lungimilor acestor vectori:

$$r_{vw} = \frac{v^t w}{\|v\| \cdot \|w\|}.$$

Deoarece raportul dintre produsul scalar a doi vectori și produsul lungimilor acestor doi vectori este egal cu cosinusul unghiului dintre cei doi vectori, rezultă că:

$$r_{vw} = \cos \phi,$$

unde ϕ reprezintă unghiul format de cei doi vectori v și w .

3.1.2.2 Standardizarea observațiilor

Operația de standardizare a valorilor unei variabile constă în substituirea valorilor fiecărei observații cu o nouă valoare reprezentând raportul dintre valoarea centrală a respectivei operații și abaterea standard a respectivei variabile. În condițiile notațiilor utilizate mai înainte, operația de standardizare a valorilor variabilei x_i presupune calculul noilor valori după relația:

$$x_{ti}^s = \frac{x_{ti}^c}{s_i} = \frac{x_{ti} - \bar{x}_i}{s_i}, \quad t=1,2,\dots,T,$$

unde \bar{x}_i reprezintă media celei de-a i-a variabile, iar s_i reprezintă abaterea standard a variabilei x_i , adică rădăcina pătrată a varianței, calculată cu ajutorul relațiilor:

- pentru cazul deplasat:

$$s_i^2 = \frac{1}{T} \sum_{t=1}^T (x_{ti} - \bar{x}_i)^2;$$

- pentru cazul nedeplasat:

$$s_i^2 = \frac{1}{T-1} \sum_{t=1}^T (x_{ti} - \bar{x}_i)^2.$$

În mod similar cu cazul variabilelor centrate, variabilele standardizate sunt variabile care *au media aritmetică nulă*:

$$\bar{x}_i^s = \frac{1}{T} \sum_{t=1}^T x_{ti}^s = \frac{1}{T} \sum_{t=1}^T \frac{x_{ti} - \bar{x}_i}{s_i} = \frac{1}{T} \frac{1}{s_i} \sum_{t=1}^T x_{ti}^c = 0.$$

În plus față de aceasta, variabilele standardizate au proprietatea că *varianța lor este egală cu unitatea*:

$$\langle s_i^s \rangle^2 = \frac{1}{T-1} \sum_{t=1}^T (x_{ti}^s - \bar{x}_i^s)^2 = \frac{1}{T-1} \sum_{t=1}^T \left(\frac{x_{ti} - \bar{x}_i}{s_i} - \bar{x}_i^s \right)^2 = \frac{1}{s_i^2} \left[\frac{1}{T-1} \sum_{t=1}^T (x_{ti} - \bar{x}_i)^2 \right] = \frac{1}{s_i^2} s_i^2 = 1.$$

De asemenea, variabilele standardizate au proprietatea că *au covarianțele scalate în intervalul [-1, 1]*:

$$-1 \leq s_{ij}^s \leq 1.$$

În cazul în care covarianța are valoarea egală cu 1, se consideră că există o *perfectă asociere liniară directă* între cele două variabile, iar în cazul în care covarianța are valoarea egală cu -1 se consideră că între cele două variabile există o *perfectă asociere liniară indirectă*. De asemenea, dacă valoarea covarianței este *nulă*, se consideră că nu există asociere de tip liniar între cele două variabile. O consecință importantă a acestei ultime proprietăți este reprezentată de faptul că, în cazul variabilelor standardizate, covarianțele sunt chiar coeficienți de corelație Pearson.

Dacă z este o variabilă standardizată, atunci cele T observații ale acesteia, z_1, z_2, \dots, z_T , definesc un punct sau un vector z în spațiul T-dimensional al observațiilor. Varianța variabilei standardizate z este, în acest caz:

$$s_z^2 = \frac{1}{T-1} \sum_{t=1}^T (z_t - \bar{z})^2 = \frac{1}{T-1} \sum_{t=1}^T z_t^2.$$

În aceste condiții, varianța variabilei standardizate z poate fi scrisă în funcție de lungimea vectorului z astfel:

$$s_z^2 = \frac{1}{T-1} \|z\|^2,$$

unde $\|z\|$ reprezintă lungimea vectorului z :

$$\|z\| = \sqrt{z_1^2 + z_2^2 + \dots + z_T^2}.$$

În mod similar, abaterea standard a variabilei standardizate z poate fi scrisă în funcție de lungimea vectorului z astfel:

$$s_z = \frac{1}{\sqrt{T-1}} \|z\|.$$

Proprietatea variabilelor standardizate de a avea varianță și, implicit, abaterea standard egale cu unitatea, evidențiază proprietatea conform căreia, lungimea vectorului ce reprezintă observațiile unei variabile standardizate este egală cu $\sqrt{T-1}$, adică:

$$\|z\| = \sqrt{T-1}.$$

Cele de mai sus arată că pentru a *normalize* vectorii observațiilor standardizate este suficient să împărți fiecare componentă a acestora cu mărimea $\sqrt{T-1}$, adică:

$$z_{norm} = \frac{1}{\sqrt{T-1}} \cdot z; \quad \|z_{norm}\| = \frac{1}{\sqrt{T-1}} \|z\| = 1.$$

Tot în condițiile stabilite anterior, covarianța dintre două variabile standardizate z și w poate fi exprimată în funcție de vectorii z și w care reprezintă observațiile celor două variabile. Covarianța dintre variabilele standardizate z și w este dată de relația:

$$s_{zw} = \frac{1}{T-1} \sum_{t=1}^T (z_t - \bar{z})(w_t - \bar{w}) = \frac{1}{T-1} \sum_{t=1}^T z_t w_t$$

Rezultă că, în cazul variabilelor standardizate z și w , covarianța este proporțională cu produsul scalar al vectorilor z și w , care reprezintă observațiile celor două variabile:

$$s_{zw} = \frac{1}{T-1} z^t w,$$

unde $z^t w$ reprezintă produsul scalar al vectorilor z și w .

Și în cazul variabilelor standardizate, coeficienții de corelație de tip Pearson pot fi exprimați prin intermediul produsului scalar și lungimilor vectorilor corespunzători. Astfel, coeficientul de corelație dintre variabilele standardizate z și w este dat de relația:

$$r_{zw} = \frac{s_{zw}}{s_z s_w} = \frac{1}{T-1} z^t w.$$

Rezultă că, în cazul variabilelor standardizate, coeficientul de corelație dintre două variabile este identic cu covarianța și este proporțional cu produsul scalar al vectorilor ce reprezintă observațiile asupra variabilelor:

$$r_{zw} = \frac{1}{T-1} z^t w.$$

3.2 Sumarizarea datelor primare

De cele mai multe ori, în analiza datelor nu sunt utilizate datele primare, brute, ci *date derivate* obținute din transformarea celor primare. Analiza datelor presupune că datele de intrare au o anumită *formă specifică*, în funcție de natura metodelor, instrumentelor și procedurilor de analiză folosite.

În scopul obținerii informațiilor sub forma care este necesară pentru o analiză a datelor, pe baza datelor primare disponibile pentru analiză, se calculează o serie de mărimi statistice cum ar fi: *media*, *suma de pătrate*, *varianța*, *covarianța*, *corelația* etc. Aceste mărimi reprezintă *esențializări* particulare ale datelor primare, caracterizează anumite aspecte ale datelor primare, reprezentând, în același timp, *elemente de intrare* pentru orice analiză de date. Pe de altă parte, prin intermediul acestor mărimi, o foarte mare varietate de date, nedecelabilă și dificilă din punct de vedere al înțelegerii și interpretării semnificației pe care ea o reprezintă, poate fi *sintetizată informational*, *rezumată numeric*, sintetizare care este importantă și utilă atât pentru modul de prezentare a unor date extrem de variate, cât și pentru modul de înțelegere a semnificației reprezentată de datele respective.

În continuare, vom face o scurtă prezentare a principalelor măsuri specifice sintetizării numerice și a principalelor modalități de exprimare a acestora prin intermediul unor mărimi specifice. De asemenea, vom face și o scurtă descriere și prezentare a conținutului acestor mărimi și a modului de calcul al acestora.

3.2.1 Măsura tendinței centrale

Una dintre măsurile cele mai importante și mai relevante pentru descrierea valorilor unei caracteristici este cea reprezentată de *tendința centrală*.

Măsurarea tendinței centrale are ca scop principal determinarea unei mărimi care să sintetizeze, să rezume, multitudinea de valori reprezentate de observațiile efectuate asupra unor variabile, din punct de vedere al magnitudinii acestora.

Este evident că, pentru a fi relevantă, mărimea utilizată pentru măsurarea tendinței centrale trebuie să fie un fel de “*centru de greutate*” al observațiilor disponibile, valorile observațiilor fiind repartizate în jurul acestei mărimi.

În figura următoare este evidențiată poziția posibilă a mărimii care măsoară tendința centrală, mărime notată cu c .

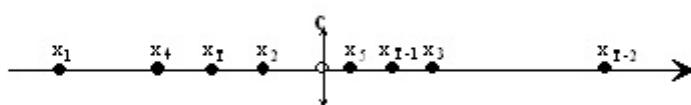


Figura 3.1: Poziționarea indicatorului ce măsoară tendința centrală

Din punct de vedere geometric, determinarea unei măsuri pentru exprimarea tendinței centrale este echivalentă cu a găsi un vector care să aibă *același sens și aceeași direcție* cu vectorul ale cărui componente sunt egale cu unitatea și care să fie *cât mai apropiat* de vectorul observațiilor. În acest sens, se poate spune că, în cazul metricii euclidiene, mărimea care exprimă în mod *optimal* tendința centrală este *media aritmetică*.

Tendința centrală poate fi evidențiată prin intermediul unor indicatori statistici, între care cei mai importanți sunt: *media*, *mediana* și *modulul*. Fiecare dintre acești indicatori exprimă, într-un fel sau altul, mai mult sau mai puțin sugestiv, nivelul caracteristicii analizate de-a lungul obiectelor.

3.2.2 Măsura variabilității

O altă măsură importantă pentru sintetizarea valorilor unei caracteristici este aceea a *variabilității* ce caracterizează observațiile variabilei, a *împrăștierii*, a *dispersiei* acestor valori. Un indicator sintetic, utilizat pentru măsurarea și exprimarea variabilității valorilor unei caracteristici, este **varianța**.

Variabilitatea care caracterizează mulțimea observațiilor efectuate asupra unei anumite caracteristici este evidențiată prin diferențele care există între valorile pe care le înregistrează caracteristica pe mulțimea subiecților, prin mărimea variațiilor valorilor caracteristicii de la un subiect la altul.

Variabilitatea este importantă atât din punct de vedere informațional, cât și ca mărime în contextul căreia poate fi judecată relevanța mediei. Cu cât variabilitatea unei mulțimi de observații este mai mică, cu atât media constituie o sintetizare, o rezumare mai potrivită și mai relevantă pentru mulțimea de observații.

Pe de altă parte, cu cât variabilitatea este mai mare, cu atât mai puțin media poate fi considerată o expresie sintetică relevantă a valorilor observate. Prin urmare, se poate spune că încrederea mai mare sau mai mică pe care o putem acorda mediei ca mărime ce sintetizează valorile observate depinde de mărimea variabilității acestor valori. Aceasta înseamnă că pentru a avea o măsură a relevanței mediei este necesar să se stabilească o măsură a variabilității.

În principiu, o măsură a variabilității valorilor unei caracteristici s-ar putea deduce prin luarea în considerare a variațiilor succesive, de la un individ la altul, înregistrate de valorile acestei caracteristici. O astfel de construcție nu ar fi însă consistentă și măsura rezultată în urma acestei construcții nu ar fi relevantă, din cauza faptului că variațiile succesive ale valorilor caracteristicii pe mulțimea indivizilor analizați nu ar avea comparabilitate, ele fiind determinate, de fiecare dată, în raport cu un reper variabil.

Varianța este *direct proporțională* cu *mărimea variației* valorilor caracteristicii măsurate sau cu *mărimea informației* care este conținută de observațiile disponibile pentru analiza de date. În condițiile notațiilor anterioare, varianța variabilei x_i , notată cu s_i^2 , se determină cu ajutorul formulei următoare:

$$s_i^2 = \frac{1}{T-1} \sum_{t=1}^T (x_{ti} - \bar{x}_i)^2.$$

În mod concret, varianța reprezintă suma pătratelor abaterilor valorilor individuale în raport cu media ce revine, *în medie*, pe fiecare valoare individuală, adică pe fiecare observație efectuată asupra variabilei.

Ca rezultat al faptului că variabilitatea poate exista sau nu poate exista, varianța, ca măsură a acestei variabilități, este totdeauna o mărime *nenegativă*. Aceasta este și unul din motivele pentru care varianța poate fi considerată ca o măsură informațională, ca o măsură a cantității de informație conținută în observațiile disponibile.

Pornind de la modul în care varianța măsoară variabilitatea și de la importanța pe care o are această variabilitate în analiza datelor, se poate face afirmația că, într-un anumit sens, varianța reprezintă o măsură a informației conținute în datele analizate. Această proprietate remarcabilă a varianței poate fi foarte simplu intuită dacă ne gândim că o mulțime de date cu variabilitate nulă, pentru care, implicit, varianța este egală cu zero, nu spune nimic din punct de vedere statistic, nu explică nimic din ceea ce se întâmplă cu fenomenul la care se referă. De fapt, în acest caz, deoarece *toate observațiile sunt egale*, există o redundanță informațională maximă, toate observațiile reprezentând, în fond, aceeași informație.

Pe de altă parte, o mare variabilitate a datelor este semnul faptului că fiecare observație este purtătoarea unei informații specifice, diferită de informația conținută în celelalte observații. Cu cât variabilitatea este mai mare, cu atât observațiile diferă mai mult între ele și fiecare din ele evidențiază o informație cu relevanță mai mare, explicând într-o măsură din ce în ce mai mare natura fenomenului analizat și modul de mișcare a acestuia.

O deficiență majoră a varianței, ca indicator de măsurare a variabilității, a cantității de informație conținută în datele primare, este legată de faptul că varianțele a două caracteristici sau a două variabile exprimate în unități de măsură diferite nu pot fi comparate. Compararea varianțelor este, totuși, posibilă numai în cazul în care măsurătorile caracteristicilor sunt exprimate în aceleași unități de măsură.

Tot în acest sens, există și o altă deficiență importantă a varianței: aceea că ea este o *mărime nescalată*. Cu toate că mărimea varianței este limitată inferior, ea având o *margine inferioară* reprezentată de valoarea zero și evidențierind lipsa variabilității sau constanță, ea nu este limitată superior, nu are o *margine superioară*:

$$0 \leq s_i^2 < \infty.$$

Din acest motiv, apar dificultăți legate de interpretarea magnitudinii varianței și de utilizarea acesteia pentru efectuare de comparații.

O altă problemă dificilă, care apare în legătură cu varianța, este aceea că unitățile de măsură în care aceasta este exprimată sunt diferite de unitățile de măsură ale caracteristicii a cărei variabilitate o măsoară.

De fapt, varianța este măsurată în unități de măsură care reprezintă *pătrate* ale unităților de măsură ale observațiilor efectuate asupra caracteristicii considerate. Această trăsătură a varianței crează o serie de dificultăți legate de interpretarea concretă a mărimii acestui indicator al varianței.

Datorită lipsei de semnificație a unităților de măsură ale varianței, pentru măsurarea varianței se utilizează și un alt indicator, derivat din varianță și reprezentat de *rădăcina pătrată* a varianței. Acest indicator este cunoscut sub numele de **abatere standard** și se calculează cu ajutorul relației:

$$s_i = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (x_{ti} - \bar{x}_i)^2}.$$

Spre deosebire de varianță, exprimată în unități de măsură nefirești, nenaturale, abaterea standard este exprimată în aceleași unități de măsură ca și observațiile efectuate asupra caracteristicii.

3.2.3 Măsura legăturii de tip liniar

Intensitatea și sensul legăturii sau asocierii de tip liniar dintre două caracteristici ale unor obiecte sau indivizi reprezintă o altă măsură importantă utilizabilă în sintetizarea numerică a datelor.

Măsura asocierii de tip liniar poate fi exprimată prin intermediul corelării *variațiilor simultane* sau *covariațiilor* a două caracteristici pe o mulțime de obiecte sau indivizi. Această măsură evidențiază cum se corelează, cum se asociază valorile a două caracteristici la nivelul unei mulțimi de indivizi care posedă aceste caracteristici. Mărimea de bază utilizată pentru exprimarea variațiilor simultane a două caracteristici este reprezentată de indicatorul cunoscut sub numele de *covarianță*. Pentru cazul a două variabile x_i și x_j , covarianța acestora se calculează cu ajutorul formulei:

$$s_{ij} = \frac{1}{T-1} \sum_{t=1}^T (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j),$$

care, în cazul în care cele două variabile coincid, adică $x_i = x_j$, covarianța coincide cu varianța:

$$s_{ii} = \frac{1}{T-1} \sum_{t=1}^T (x_{ti} - \bar{x}_i)(x_{ti} - \bar{x}_i) = s_i^2.$$

Covarianța este o măsură a variației simultane a două variabile, ea fiind, în valoare absolută, cu atât mai mare cu cât valorile absolute ale variațiilor celor două variabile în jurul mediei sunt mai apropiate ca magnitudine, evidențierind o anumită proporționalitate pe mulțimea subiecțiilor studiați. Covarianța este considerată a fi o expresie numerică a gradului de asociere a două caracteristici ca urmare a faptului că, în toate cazurile în care două variabile sunt semnificativ legate între ele, o variație într-un sens a uneia dintre ele va determina o variație proporțională de același sens (în cazul legăturii *directe*) sau de sens contrar (în cazul legăturii *inverse*) a celeilalte variabile.

În mod similar cu varianța, și în cazul exprimării covarianței apare problema unor unități de măsură nefirești, nenaturale. După modul în care este definită, covarianța este exprimată în unități de măsură care sunt de fapt *produs* al unităților de măsură ale caracteristicilor considerate. Ca și în cazul varianței, există o dificultate și mai mare în legătură cu măsura numită covarianță. Aceasta constă în faptul că ea este o *mărime nescalată*. Deși, în valoare absolută, covarianța *are o margine inferioară*, reprezentată de valoarea zero și care evidențiază lipsa asocierii de tip liniar, ea nu este limitată superior, *nu are o margine superioară*:

$$0 \leq |s_{ij}| < \infty.$$

Ca urmare a acestei proprietăți, apar dificultăți legate de interpretarea magnitudinii covarianței și de utilizarea acesteia pentru efectuare de comparații.

O măsură scalată a gradului de asociere liniară între două variabile, care elimină unele deficiențe ale covarianței ca indicator de măsurare a asocierii de tip liniar, o reprezintă *coeficientul de corelație Pearson*. Pentru cazul a T observații existente cu privire la două variabile x_i și x_j , coeficientul de corelație Pearson este dat de relația:

$$r_{ij} = \frac{s_{ij}}{s_i s_j} = \frac{\frac{1}{T-1} \sum_{t=1}^T (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)}{\sqrt{\frac{1}{T-1} \sum_{t=1}^T (x_{ti} - \bar{x}_i)^2} \sqrt{\frac{1}{T-1} \sum_{t=1}^T (x_{tj} - \bar{x}_j)^2}}.$$

Spre deosebire de covarianță, coeficientul de corelație este o mărime scalată în intervalul închis $[-1, 1]$:

$$-1 \leq r_{ij} \leq 1.$$

O valoare nulă a coeficientului de corelație evidențiază absența legăturii de tip liniar între cele două variabile, după cum o valoare absolută egală cu unitatea evidențiază o legătură liniară perfectă, legătură care este *directă* dacă valoarea este egală cu 1 și *inversă* dacă valoarea este egală cu -1.

3.2.4 Măsuri generalizate ale variabilității

Așa cum am văzut mai înainte, în accepțiunea sa comună, varianța reprezintă o măsură a variabilității individuale, la nivelul fiecărei caracteristici. Fiecare din aceste varianțe individuale reprezintă o măsură a unei părți din variabilitatea ce caracterizează observațiile variabilelor analizate, oferind doar o imagine parțială a variabilității conținute în aceste observații.

În mod corespunzător, mulțimea valorilor varianțelor tuturor variabilelor supuse analizei, constituie o imagine mai cuprinzătoare a variabilității conținută în observațiile respectivelor variabile. Din nefericire însă, în acest caz, exprimarea variabilității nu este sintetizată, cum ar fi de dorit, prin intermediul unui singur indicator, ci prin intermediul unei multimi de indicatori.

Una dintre posibilitățile de a da un răspuns corespunzător problemei rezultate din necesitatea de a exprima cât mai adecvat

și mai sintetic variabilitatea conținută în observațiile variabilelor analizate constă în definirea altor doi indicatori ai varianței: *varianța totală* și *varianța generalizată*.

3.2.4.1 Varianța totală

Am arătat anterior că măsurarea variabilității este o problemă dificilă și că utilizarea varianței simple pentru sintetizarea acesteia nu este satisfăcătoare. O modalitate de a elimina acest neajuns o reprezintă deducerea unei măsuri globale, unice, pentru variabilitatea ce caracterizează observațiile variabilelor studiate.

O astfel de măsură a variabilității este *varianța totală*, care este unul dintre indicatorii importanți în analiza datelor, utilizat în numeroase proceduri de analiză a datelor.

Definiție: *Varianța totală* măsoară variabilitatea ce caracterizează observațiile unei mulțimi de variabile și se definește ca sumă a varianțelor individuale ale variabilelor:

$$V_T = \sum_{i=1}^n s_i^2$$

Cu toate că varianța totală oferă o imagine cuprindătoare asupra variabilității globale ce caracterizează observațiile variabilelor analizate, ea măsoară această variabilitate doar în sens individual, neluând în considerare variabilitatea comună, simultană a observațiilor, adică variabilitatea interacțiunilor.

O măsură interesantă a variabilității totale, care ține seama atât de variabilitatea individuală, cât și de variabilitatea rezultată din interacțiuni, este reprezentată de varianța generalizată.

3.2.4.2 Varianța generalizată

O extindere importantă a conceptului de măsură a variabilității o reprezintă *varianța generalizată* care măsoară variabilitatea ce caracterizează observațiile mulțimii de variabile, atât din punct de vedere individual, cât și din punct de vedere al simultaneității, al interactivității informaționale ce caracterizează variabilele.

Pentru a da o interpretare intuitivă varianței generalizate, vom porni de la o construcție geometrică. În acest scop, vom considera că variabilele \mathbf{x}_1 și \mathbf{x}_2 reprezintă doi vectori în spațiul observațiilor.

Există o strânsă legătură între mărimea unghiului format de cei doi vectori și corelația dintre cele două variabile. Aceasta constă în faptul că, de fapt, *coeficientul de corelație este cosinusul unghiului* dintre vectorii ce reprezintă cele două variabile. Într-adevăr, dacă unghiul dintre cei doi vectori este zero, adică vectorii se suprapun, legătura perfectă existentă în această situație este evidențiată atât printr-o valoare a coeficientului de corelație egală cu unitatea, cât și prin valoarea unitară a cosinusului unghiului respectiv. Invers, dacă unghiul dintre vectori este de 90 de grade, adică vectorii sunt ortogonali, inexistența legăturii specifice acestei situații este evidențiată prin faptul că atât coeficientul de corelație, cât și cosinusul unghiului respectiv sunt egale cu zero. Cele trei situații de corelare posibilă a două variabile \mathbf{x}_1 și \mathbf{x}_2 , ale căror observații sunt reprezentate prin intermediul vectorilor \mathbf{x}^1 și \mathbf{x}^2 , sunt evidențiate în graficele din figura 3.2.

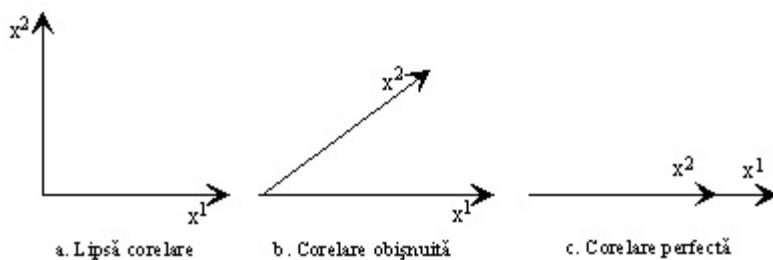


Figura 3.2: Situații posibile de corelare a două variabile reprezentate prin intermediul vectorilor \mathbf{x}^1 și \mathbf{x}^2

Vom presupune că unghiul format de cei doi vectori este ϕ și că cei doi vectori sunt scalări prin înmulțirea cu mărimea $1/\sqrt{T-1}$, adică cei doi vectori scalări au componente de formă:

$$\mathbf{z}_i^k = \frac{1}{\sqrt{T-1}} \mathbf{x}_i^k, \quad k=1,2; i=1,2,\dots,T.$$

Lungimea unui astfel de vector va fi:

$$\|\mathbf{z}\| = \sqrt{\frac{\mathbf{x}_1^2}{T-1} + \frac{\mathbf{x}_2^2}{T-1} + \dots + \frac{\mathbf{x}_T^2}{T-1}} = \frac{1}{\sqrt{T-1}} \sqrt{\mathbf{x}_1^2 + \mathbf{x}_2^2 + \dots + \mathbf{x}_T^2} = \frac{1}{\sqrt{T-1}} \|\mathbf{x}\|,$$

unde \mathbf{x}_t reprezintă cea de-a t -a observație efectuată asupra variabilei x .

Dacă variabilele \mathbf{x}_1 și \mathbf{x}_2 sunt variabile centrate, adică de medie nulă, atunci pătratul lungimii vectorilor \mathbf{z}^1 și \mathbf{z}^2 reprezintă chiar varianțele celor două variabile:

$$\|z\|^2 = \frac{1}{T-1} \sum_{t=1}^T (x_i - \bar{x})^2 = \frac{1}{T-1} \sum_{t=1}^T x_i^2 = s_i^2.$$

În cazul *lipsei de corelație*, evidențiată prin ortogonalitatea celor doi vectori, aria paralelogramului este maximă. Aceasta corespunde unei situații în care redundanța informațională aferentă observațiilor efectuate asupra celor două variabile este nulă. În cazul în care *corelația este perfectă*, adică cei doi vectori sunt *coliniari*, aria paralelogramului este minimă. În această situație redundanța informațională corespunzătoare observațiilor efectuate asupra celor două variabile, este maximă. În figura 3.3, este reprezentată aria paralelogramului având ca laturi vectorii ce definesc cele două variabile analizate.

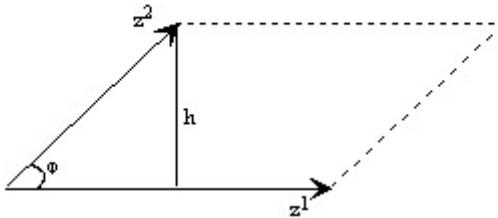


Figura 3.3: Interpretarea redundanței informaționale prin intermediul ariei paralelogramului

Din punct de vedere al analizei datelor, situația de redundanță minimă este ideală, aceasta evidențind faptul că între cele două variabile menționate nu există nici o suprapunere informațională. În această situație, variabilitatea indusă de cele două variabile este maximă, ceea ce din punct de vedere geometric este echivalent cu faptul că vectorii sunt ortogonali, respectiv că aria paralelogramului este maximă. Pe de altă parte, situația de redundanță maximă este cea mai puțin dorită, aceasta însemnând că cele două variabile reprezintă unul și același lucru din punct de vedere informațional. În acest caz, variabilitatea corespunzătoare celor două variabile este minimă și este evidențiată de coliniaritatea vectorilor ce reprezintă cele două variabile, adică de faptul că aria paralelogramului este nulă.

În afară de poziția pe care o au cei doi vectori unul față de altul, aria paralelogramului depinde și de lungimea fiecărui dintre vectori, fiind cu atât mai mare, cu cât lungimea celor doi vectori este mai mare.

Deoarece pătratul lungimii fiecărui din cei doi vectori z^1 și z^2 este chiar varianța corespunzătoare variabilei pe care acesta o reprezintă, este evident că aria paralelogramului este și măsură a varianței variabilelor standardizate.

Cele menționate anterior evidențiază un fapt de o însemnatate excepțională pentru problematica măsurării variabilității individuale și comune ce caracterizează observațiile unei mulțimi de variabile: *aria paralelogramului poate fi folosită ca măsură comună atât pentru variabilitatea individuală, exprimată prin intermediul varianțelor variabilelor, cât și pentru variabilitatea comună, exprimată prin intermediul covarianțelor dintre aceste variabile.*

Cele două situații menționate evidențiază faptul că aria paralelogramului determinat de cei doi vectori poate fi utilizată pentru *determinarea unei măsuri a redundanței informaționale și a variabilității generale* ce caracterizează observațiile variabilelor. O astfel de măsură este reprezentată de *pătratul ariei paralelogramului* ce corespunde celor doi vectori și este cunoscută sub numele de *varianță generalizată*.

Deoarece baza paralelogramului este reprezentată de lungimea vectorului z^1 , adică de mărimea $\|z^1\| = \frac{\|x^1\|}{\sqrt{T-1}}$, iar înălțimea paralelogramului este dată de relația:

$$h = \frac{1}{\sqrt{T-1}} \|x^2\| \cdot \sin \phi,$$

aria paralelogramului va fi:

$$A_p = \|z^1\| \cdot h = \frac{1}{T-1} \|x^1\| \cdot \|x^2\| \cdot \sin \phi.$$

În cazul în care există un număr de n variabile, varianța generalizată corespunzătoare acestora este chiar *pătratul volumului hiperparalelipipedului* format de cei n vectori în spațiul observațiilor.

Din cele arătate mai sus rezultă că, în sens geometric, varianța generalizată poate fi definită sub forma următoare:

Definiție: *Varianța generalizată* corespunzătoare spațiului observațiilor celor două variabile considerate este dată de relația:

$$V_g = \left(\frac{1}{T-1} \|x^1\| \cdot \|x^2\| \cdot \sin \phi \right)^2$$

Se poate arăta că varianța generalizată este reprezentată de *determinantul matricii de covarianță* ce corespunde variabilelor supuse studiului, respectiv:

$$V_g = |S|.$$

Varianța generalizată este o măsură extrem de importantă a variabilității totale, formată atât ca urmare a variabilității

individuale ce caracterizează variabilele, cât și ca urmare a variabilității comune ce caracterizează interacțiunea variabilelor.

3.3 Matrici utilizate în analiza multidimensională a datelor

În urma efectuării unor operații preliminare asupra datelor primare, reprezentate prin intermediul matricii de observații X :

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots \\ x_{T1} & x_{T2} & \dots & x_{Tn} \end{pmatrix},$$

rezultă următoarele trei tipuri de matrici foarte importante în analiza datelor:

- matricea observațiilor centrate;
- matricea observațiilor standardizate;
- matricea produselor încrucișate,

pe care le vom defini în continuare.

3.3.1 Matricea observațiilor centrate

Matricea observațiilor **centrate** poate fi obținută ca diferență între matricea de observații și matricea \bar{X} ale cărei coloane sunt mediile celor n variabile:

$$X_c = X - \bar{X} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1n} - \bar{x}_n \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2n} - \bar{x}_n \\ \dots \\ x_{T1} - \bar{x}_1 & x_{T2} - \bar{x}_2 & \dots & x_{Tn} - \bar{x}_n \end{pmatrix}, \text{ unde } \bar{X} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_n \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_n \\ \dots \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_n \end{pmatrix}.$$

3.3.2 Matricea observațiilor standardizate

Matricea observațiilor **standardizate** poate fi obținută ca produs între matricea variabilelor centrate și inversa matricii diagonale V , ale cărei elemente sunt abaterile standard ale celor n variabile:

$$Z = X_c V^{-1} = \begin{pmatrix} \frac{x_{11} - \bar{x}_1}{s_1} & \frac{x_{12} - \bar{x}_2}{s_2} & \dots & \frac{x_{1n} - \bar{x}_n}{s_n} \\ \frac{x_{21} - \bar{x}_1}{s_1} & \frac{x_{22} - \bar{x}_2}{s_2} & \dots & \frac{x_{2n} - \bar{x}_n}{s_n} \\ \dots \\ \frac{x_{T1} - \bar{x}_1}{s_1} & \frac{x_{T2} - \bar{x}_2}{s_2} & \dots & \frac{x_{Tn} - \bar{x}_n}{s_n} \end{pmatrix}, \text{ unde } V = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \dots \\ 0 & 0 & \dots & s_n \end{pmatrix}.$$

3.3.3 Matricea produselor încrucișate

Matricea produselor **încrucișate** poate fi determinată atât pentru variabilele originale, cât și pentru variabilele centrate și standardizate. Pentru cazul variabilelor *originale*, matricea produselor încrucișate se obține ca produs între transpusa matricii X și matricea X :

$$C = X^t X = \begin{pmatrix} \sum_{t=1}^T x_{t1}^2 & \sum_{t=1}^T x_{t1} x_{t2} & \dots & \sum_{t=1}^T x_{t1} x_{tn} \\ \sum_{t=1}^T x_{t2} x_{t1} & \sum_{t=1}^T x_{t2}^2 & \dots & \sum_{t=1}^T x_{t2} x_{tn} \\ \dots \\ \sum_{t=1}^T x_{tn} x_{t1} & \sum_{t=1}^T x_{tn} x_{t2} & \dots & \sum_{t=1}^T x_{tn}^2 \end{pmatrix}.$$

Utilizând scrierea bazată pe lungimile vectorilor de observații și pe produsele scalare ale acestora, matricea produselor încrucișate pentru situația în care variabilele sunt sub forma originală poate fi scrisă sub formă:

$$C = \begin{pmatrix} \|x^1\|^2 & (x^1)^t \cdot x^2 & \dots & (x^1)^t \cdot x^n \\ (x^2)^t \cdot x^1 & \|x^2\|^2 & \dots & (x^2)^t \cdot x^n \\ \dots \\ (x^n)^t \cdot x^1 & (x^n)^t \cdot x^2 & \dots & \|x^n\|^2 \end{pmatrix},$$

unde x^i este vectorul observațiilor variabilei x_i .

În cazul în care variabilele sunt *centrate*, matricea produselor încrucișate poate fi determinată astfel:

$$C_c = X_c^t \cdot X_c = \begin{pmatrix} \sum_{t=1}^T (x_{t1} - \bar{x}_1)^2 & \sum_{t=1}^T (x_{t1} - \bar{x}_1)(x_{t2} - \bar{x}_2) & \dots & \sum_{t=1}^T (x_{t1} - \bar{x}_1)(x_{tn} - \bar{x}_n) \\ \sum_{t=1}^T (x_{t2} - \bar{x}_2)(x_{t1} - \bar{x}_1) & \sum_{t=1}^T (x_{t2} - \bar{x}_2)^2 & \dots & \sum_{t=1}^T (x_{t2} - \bar{x}_2)(x_{tn} - \bar{x}_n) \\ \dots & \dots & \dots & \dots \\ \sum_{t=1}^T (x_{tn} - \bar{x}_n)(x_{t1} - \bar{x}_1) & \sum_{t=1}^T (x_{tn} - \bar{x}_n)(x_{t2} - \bar{x}_2) & \dots & \sum_{t=1}^T (x_{tn} - \bar{x}_n)^2 \end{pmatrix}.$$

Folosind lungimile vectorilor de observații centrate și produsele scalare ale acestora, matricea produselor încrucișate pentru situația în care variabilele sunt centrate poate fi scrisă sub forma:

$$C_c = \begin{pmatrix} \|x^1 - \bar{x}^1\|^2 & (x^1 - \bar{x}^1)^t \cdot (x^2 - \bar{x}^2) & \dots & (x^1 - \bar{x}^1)^t \cdot (x^n - \bar{x}^n) \\ (x^2 - \bar{x}^2)^t \cdot (x^1 - \bar{x}^1) & \|x^2 - \bar{x}^2\|^2 & \dots & (x^2 - \bar{x}^2)^t \cdot (x^n - \bar{x}^n) \\ \dots & \dots & \dots & \dots \\ (x^n - \bar{x}^n)^t \cdot (x^1 - \bar{x}^1) & (x^n - \bar{x}^n)^t \cdot (x^2 - \bar{x}^2) & \dots & \|x^n - \bar{x}^n\|^2 \end{pmatrix}.$$

3.3.4 Matricea de covarianță

Matricea de covarianță constituie una dintre cele mai frecvent utilizate matrici în analiza datelor, majoritatea tehniciilor de analiză a datelor presupunând calculul acestei matrici. Pentru situația în care numărul de variabile analizate este egal cu n , covarianțele dintre orice două variabile pot fi aranjate sub forma unei matrici pătrate și simetrice, de dimensiune $n \times n$, numită **matrice de covarianță**:

$$S = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1n} \\ s_{21} & s_2^2 & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ s_{n1} & s_{n2} & \dots & s_n^2 \end{pmatrix}, \text{ unde } s_i = \frac{1}{T-1} \|x^i - \bar{x}^i\|^2 \text{ și } s_{ij} = \frac{1}{T-1} (x^i - \bar{x}^i)^t \cdot (x^j - \bar{x}^j).$$

În condițiile notațiilor anterioare, matricea de covarianță pentru variabilele originale poate fi scrisă cu ajutorul matricii produselor încrucișate pentru cazul variabilelor centrate, sub forma:

$$S = \frac{1}{T-1} \cdot C_c = \begin{pmatrix} \frac{1}{T-1} \|x^1 - \bar{x}^1\|^2 & \frac{1}{T-1} (x^1 - \bar{x}^1)^t \cdot (x^2 - \bar{x}^2) & \dots & \frac{1}{T-1} (x^1 - \bar{x}^1)^t \cdot (x^n - \bar{x}^n) \\ \frac{1}{T-1} (x^2 - \bar{x}^2)^t \cdot (x^1 - \bar{x}^1) & \frac{1}{T-1} \|x^2 - \bar{x}^2\|^2 & \dots & \frac{1}{T-1} (x^2 - \bar{x}^2)^t \cdot (x^n - \bar{x}^n) \\ \dots & \dots & \dots & \dots \\ \frac{1}{T-1} (x^n - \bar{x}^n)^t \cdot (x^1 - \bar{x}^1) & \frac{1}{T-1} (x^n - \bar{x}^n)^t \cdot (x^2 - \bar{x}^2) & \dots & \frac{1}{T-1} \|x^n - \bar{x}^n\|^2 \end{pmatrix}.$$

3.3.5 Matricea de corelație

Matricea de corelație este o altă matrice importantă în contextul multor metode și tehnici de analiză a datelor. Matricea de corelație este o matrice importantă în analiza datelor, în primul rând, pentru faptul că o serie de metode și tehnici ale analizei datelor își bazează procedurile pe analiza spectrală a acestei matrici.

În mod similar cu matricea de covarianță, se definește **matricea de corelație** corespunzătoare celor n variabile originale, care este o matrice simetrică având următoarea formă:

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}, \text{ unde } r_{ij} = \frac{(x^i - \bar{x}^i)^t \cdot (x^j - \bar{x}^j)}{\|x^i - \bar{x}^i\| \cdot \|x^j - \bar{x}^j\|}.$$

Matricea de corelație a variabilelor originale poate fi scrisă cu ajutorul matricii produselor încrucișate pentru cazul variabilelor standardizate, astfel:

$$R = \frac{1}{T-1} Z^t \cdot Z = \begin{pmatrix} \frac{1}{T-1} \|z^1\|^2 & \frac{1}{T-1} (z^1)^t \cdot z^2 & \dots & \frac{1}{T-1} (z^1)^t \cdot z^n \\ \frac{1}{T-1} (z^2)^t \cdot z^1 & \frac{1}{T-1} \|z^2\|^2 & \dots & \frac{1}{T-1} (z^2)^t \cdot z^n \\ \dots & \dots & \dots & \dots \\ \frac{1}{T-1} (z^n)^t \cdot z^1 & \frac{1}{T-1} (z^n)^t \cdot z^2 & \dots & \frac{1}{T-1} \|z^n\|^2 \end{pmatrix}$$

Tema 4. Spații și distanțe utilizate în analiza datelor

Interpretarea geometrică a obiectelor și variabilelor ca puncte din spații cu mai multe dimensiuni și, în mod implicit, ca vectori din aceste spații este foarte utilă pentru înțelegerea unor metode și tehnici de analiză a datelor și pentru simplificarea interpretării unor rezultate obținute în urma analizei.

4.1 Reprezentarea geometrică a observațiilor și variabilelor

În funcție de elementele care fac obiectul reprezentării, obiecte sau observații, respectiv variabile, în analiza datelor sunt definite două tipuri de spații: *spațiu variabilelor* și *spațiu observațiilor*. Aceste două spații au un rol fundamental în numeroase raționamente și în definirea multor tehnici de analiză multidimensională a datelor. În fiecare din cele două spații sunt definite, în principal, două operații importante: determinarea *gradului de apropiere* dintre punctele spațiului și determinarea *gradului de asociere* dintre punctele spațiului.

4.1.1 Spațiu variabilelor

Obiectele sau observațiile disponibile într-o analiză de date pot fi privite ca *puncte* sau ca *vectori* dintr-un spațiu ale căruia dimensiuni sunt reprezentate de variabilele ce caracterizează obiectele. Într-o astfel de reprezentare, axele spațiului corespund variabilelor, iar valorile înregistrate de obiecte la fiecare din variabilele analizate sunt proiecții ale punctelor reprezentate de observații pe axele spațiului.

Spațiul în care sunt reprezentate observațiile sau obiectele analizate este cunoscut în analiza datelor sub numele de *spațiu variabilelor*. Dacă numărul de observații dintr-o analiză de date este T , iar numărul de variabile este n , cele T observații pot fi privite ca puncte din spațiul real n -dimensional.

Definiție: Se numește *spațiu variabilelor*, spațiu real n -dimensional \mathbb{R}^n în care sunt reprezentate obiectele supuse analizei și ale cărui axe sunt reprezentate de variabilele analizate.

În spațiul variabilelor, un obiect, de exemplu cel de-al i -lea, este reprezentat prin intermediul vectorului n -dimensional:

$$\mathbf{o}^i = \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \vdots \\ \mathbf{x}_{in} \end{pmatrix}$$

Într-un astfel de spațiu, pot fi determinate distanțele dintre obiecte, poate fi evidențiată și măsurată variabilitatea obiectelor de-a lungul axelor, pot fi determinate și măsurate eventuale legături între obiecte etc.

Reprezentarea obiectelor în spațiul variabilelor este foarte utilă și sugestivă pentru înțelegerea tehnicii de analiză a discriminantului și de analiză cluster, metodelor și tehnicii de analiză a legăturilor etc.

Figura următoare conține reprezentarea grafică a celor două obiecte în spațiul variabilelor, ale cărui axe sunt reprezentate de caracteristicile obiectelor.

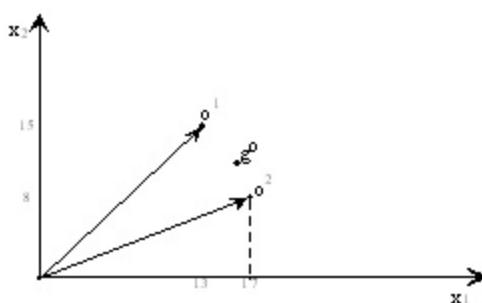


Figura 4.1: Reprezentarea obiectelor \mathbf{o}^1 și \mathbf{o}^2 în spațiul variabilelor

În spațiul variabilelor, obiectele analizate sunt reprezentate sub formă unui *nor* de puncte, centrul de greutate al norului de puncte fiind chiar punctul care reprezintă mediile caracteristicilor obiectelor, punct care se numește *centroid* al obiectelor.

Definiție: Se numește *centroid* în spațiul variabilelor punctul ale cărui coordonate sunt reprezentate de mediile celor n variabile analizate, adică punctul $\mathbf{g}^o = (g_1^o, g_2^o, \dots, g_n^o) \in \mathbb{R}^n$.

Cele n coordonate ale centrului de greutate sau ale centroidului obiectelor în spațiul variabilelor sunt date de relațiile următoare:

$$g_j^o = \frac{1}{T} \cdot \sum_{t=1}^T x_{jt}; \quad j=1,2,\dots,n,$$

unde x_{jt} reprezintă coordonata celui de-al t-lea obiect în raport cu cea de-a j-a axă a spațiului variabilelor.

Dacă obiectele din spațiul variabilelor sunt privite ca vectori din acest spațiu, atunci cosinusul unghiiului dintre doi vectori ce definesc două obiecte reprezintă măsura legăturii liniare ce există între cele două obiecte.

Ceea ce are importanță pentru caracterizarea obiectelor este *poziția* pe care fiecare obiect o are în spațiul variabilelor, poziție atât în raport cu axele spațiului, cât și în raport cu alte obiecte.

Cea mai importantă operație în spațiul variabilelor este aceea a determinării *gradului de apropiere sau de depărtare dintre obiecte*, astfel încât, mărimea cea mai relevantă pentru caracterizarea obiectelor reprezentate în spațiul variabilelor este *distanța*.

Definirea și evaluarea distanței în spațiul variabilelor sunt posibile numai în condițiile în care pe spațiul respectiv este definită o anumită *metrică*.

Conceptul care stă la baza definirii unei metriki într-un anumit spațiu este *produsul scalar*. În cazul spațiului real n-dimensional \mathbb{R}^n , produsul scalar dintre doi vectori x și y este numărul real definit astfel:

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \dots + x_n y_n.$$

Cu ajutorul produsului scalar poate fi definită *lungimea unui vector* din spațiul n-dimensional \mathbb{R}^n :

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2},$$

și *distanța euclidiană dintre doi vectori* din spațiul n-dimensional \mathbb{R}^n , respectiv:

$$d(x, y) = \|x - y\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}.$$

Metrica în spațiul variabilelor este introdusă prin intermediu *distanței euclidiene*, în conformitate cu care, distanța dintre două obiecte o^k și o^l este dată de rădăcina pătrată a sumei pătratelor diferențelor coordonatelor celor două obiecte:

$$d_E(o^k, o^l) = \sqrt{(x_{k1} - x_{l1})^2 + (x_{k2} - x_{l2})^2 + \dots + (x_{kn} - x_{ln})^2}.$$

Analiza varianței obiectelor de-a lungul axelor este extrem de importantă deoarece, în cazul în care de-a lungul unei axe, varianța este nesemnificativă în raport cu varianțele obiectelor în raport cu celelalte axe, se poate renunța la această axă, considerându-se că variabila ce reprezintă respectiva axă nu are o semnificație relevantă în definirea obiectelor. Aceasta este de fapt ideea centrală care stă la baza tehniciilor de simplificare și de reducere a dependențelor cauzale.

4.1.2 Spațiul observațiilor

În numeroase situații concrete, apare necesitatea analizei variabilelor ce definesc obiectele, analizei legăturilor care există între acestea sau a modului în care variabilele se asociază, astfel încât, analizele vizează nu obiectele ca atare, ci caracteristicile acestora, variabilele. Analizele de acest fel se află într-un anumit raport de dualitate cu analizele care se fac în spațiul variabilelor și presupun definirea unui spațiu adecvat acestei situații: spațiul observațiilor.

În mod similar cu cazul obiectelor, variabilele implicate într-o analiză de date pot fi reprezentate ca *puncte* sau ca *vectori* într-un spațiu ale cărui dimensiuni sunt reprezentate de obiectele supuse analizei. Axele spațiului în care sunt reprezentate variabilele corespund observațiilor sau obiectelor analizate.

Coordonatele variabilelor în spațiul observațiilor sunt valorile înregistrate de variabile la nivelul fiecărui obiect, adică proiecții ale punctelor reprezentate de variabile pe axele acestui spațiu.

Spațiul în care sunt reprezentate variabilele analizate este cunoscut în analiza datelor sub numele de *spațiu observațiilor*. În cazul în care numărul de variabile supuse analizei este n , iar numărul de observații din această analiză este T , cele n variabile pot fi privite ca puncte sau vectori din spațiu real T -dimensional.

Definiție: Se numește *spațiu observațiilor*, spațiu real T -dimensional \mathbb{R}^T , în care sunt reprezentate variabilele supuse analizei și ale cărui axe sunt reprezentate de observațiile sau obiectele analizate.

În spațiul observațiilor, o variabilă, de exemplu cea de-a j-a, este reprezentată prin intermediu vectorului T -dimensional:

$$x^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{Tj} \end{pmatrix}.$$

În spațiul observațiilor, pot fi determinate și măsurate legăturile de tip liniar dintre variabile, poate fi stabilit modul în care variabilele sau caracteristicile obiectelor se asociază, pot fi stabilite distanțe între variabile etc. Ca și în cazul reprezentării obiectelor în spațiul variabilelor, reprezentarea variabilelor în spațiul observațiilor este utilă în numeroase analize de tip multidimensional, între cele două tipuri de reprezentare fiind o legătură de tip *dual*.

Variabilele analizate reprezintă în spațiul observațiilor un *nor de puncte*, al cărui centru de greutate este numit *centroid* al variabilelor.

Definiție: Se numește *centroid* în spațiul observațiilor punctul ale căruia coordonate sunt reprezentate de mediile celor T observații analizate, adică punctul $\mathbf{g}^v = \left(\mathbf{g}_1^v, \mathbf{g}_2^v, \dots, \mathbf{g}_T^v \right) \in \mathfrak{R}^T$.

Cele T coordonate ale centroidului variabilelor în spațiul observațiilor sunt date de relațiile:

$$\mathbf{g}_t^v = \frac{1}{n} \cdot \sum_{j=1}^n \mathbf{x}_{tj}; \quad t=1,2,\dots,T,$$

unde \mathbf{g}_t^v reprezintă valoarea medie înregistrată de cele n variabile la cel de-al t -lea obiect. Reprezentarea variabilelor în spațiul observațiilor și a centroidului acestora sunt evidențiate în figura următoare.

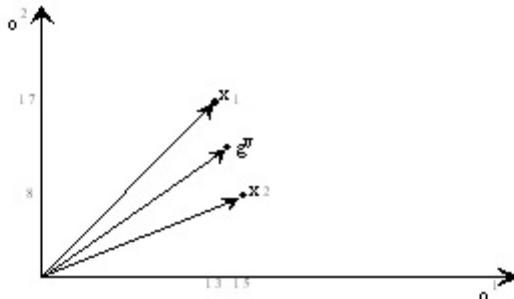


Figura 4.2: Reprezentarea variabilelor \mathbf{x}_1 și \mathbf{x}_2 în spațiul observațiilor

Ca și în cazul spațiului variabilelor, definirea produsului scalar în spațiul observațiilor permite definirea conceptului de distanță euclidiană între două variabile. Deși această distanță nu are relevanță pe care o are în cazul obiectelor, există situații în care măsurarea distanței dintre variabile are o importanță practică. Distanța euclidiană dintre variabilele \mathbf{x}_i și \mathbf{x}_j în spațiul observațiilor este:

$$d_E(\mathbf{x}_i^v, \mathbf{x}_j^v) = \sqrt{(\mathbf{x}_{1i} - \mathbf{x}_{1j})^2 + (\mathbf{x}_{2i} - \mathbf{x}_{2j})^2 + \dots + (\mathbf{x}_{Ti} - \mathbf{x}_{Tj})^2}.$$

4.2 Distanță în analiza datelor

În orice proces de analiză a datelor apare o foarte importantă problemă de natură metodologică: *măsurarea distanței dintre două obiecte sau indivizi*. Alegerea modalității de exprimare a distanței dintre obiecte este anteroară analizelor efective și influențează în mod direct și sensibil calitatea rezultatelor obținute. Din aceste motive, considerăm că este foarte utilă definirea conceptului de *distanță* și evidențierea modalităților care sunt cele mai potrivite pentru exprimarea acesteia.

Distanța reprezintă unul dintre cele mai importante și mai frecvent utilizate concepte din domeniul analizei datelor. În același timp, distanța constituie una dintre cele mai relevante modalități de sumarizare a informațiilor manipulate în analiza datelor, mai ales în situațiile în care sunt investigate interdependențele dintre fenomene și procese. Ca mărime, distanța se calculează pentru a evalua apropierea sau depărtarea dintre obiectele sau caracteristicile care se supun studiului, pentru a măsura gradul de *similitudine* sau *nesimilitudine* dintre acestea, din punct de vedere al caracteristicilor studiate.

Definirea și interpretarea conceptului de distanță presupune, în mod implicit, existența unui spațiu în raport cu care are loc nu numai definirea, ci și evaluarea numerică a distanței. Spațiul în care este posibilă o distanță se numește *spațiu metric* și poate fi spațiul variabilelor sau spațiul observațiilor.

Corespunzător celor două modalități de reprezentare, în spațiul variabilelor și în spațiul observațiilor, distanța poate fi utilizată pentru a evalua apropierea sau depărtarea dintre puncte ale unui spațiu multidimensional, puncte ce pot reprezenta atât obiecte, cât și caracteristici.

Definiție: Funcția reală $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ se numește *distanță* dacă, fiind date punctele $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3 \in \mathfrak{R}^n$, verifică următoarele proprietăți:

- a. este *nenegativă*: $d(\mathbf{x}^1, \mathbf{x}^2) \geq 0$ și $d(\mathbf{x}^1, \mathbf{x}^2) = 0$ numai dacă $\mathbf{x}^1 = \mathbf{x}^2$
- b. este *simetrică*: $d(\mathbf{x}^1, \mathbf{x}^2) = d(\mathbf{x}^2, \mathbf{x}^1)$
- c. verifică *inegalitatea triunghiului*: $d(\mathbf{x}^1, \mathbf{x}^2) \leq d(\mathbf{x}^1, \mathbf{x}^3) + d(\mathbf{x}^3, \mathbf{x}^2)$

În funcție de modul în care distanța este evaluată, adică în funcție de modul în care se evaluatează gradul de depărtare sau apropiere dintre două obiecte, există mai multe tipuri importante de distanțe: *distanța euclidiană*, *distanța statistică*, *distanța standardizată*, *distanța Mahalanobis* etc.

4.2.1 Distanța euclidiană

Cea mai cunoscută distanță utilizată pentru a măsura depărtarea sau apropierea unor puncte dintr-un spațiu multidimensional este *distanța în linie dreaptă*, cunoscută sub numele de *distanță euclidiană*.

Definiție: Se numește *distanță euclidiană* între două puncte x și y din spațiul n-dimensional \mathbb{R}^n , funcția reală definită astfel:

$$d_E(x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2},$$

unde x_i și y_i reprezintă cea de-a i-a coordonată a punctelor x , respectiv y .

Pentru a ilustra geometric distanța euclidiană, vom considera cazul unui spațiu bidimensional \mathbb{R}^2 , presupunând că cele două puncte x și y sunt reprezentate de vectorii x și y din figura următoare.

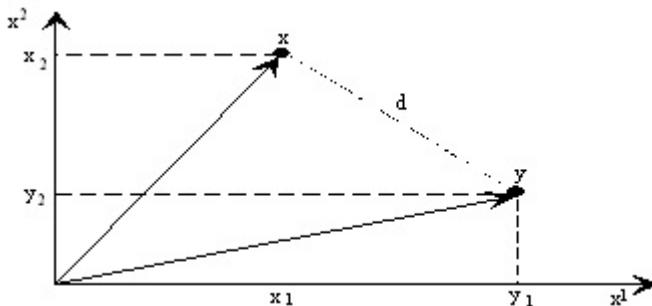


Figura 4.3: Distanța euclidiană dintre două puncte x și y

După cum se poate observa, în conformitate cu teorema lui Pitagora, lungimea segmentului ce unește punctele x și y , adică distanța dintre cele două puncte este:

$$d_E(x,y) = \sqrt{(y_1 - x_1)^2 + (x_2 - y_2)^2}.$$

Rezultă că *distanța euclidiană* dintre două puncte este numărul nenegativ reprezentat de *rădăcina pătrată* a sumei pătratelor diferențelor coordonatelor omoloage ale respectivelor puncte. Altfel spus, pătratul distanței euclidiene dintre două puncte este egal cu suma pătratelor diferențelor dintre coordonatele omoloage ale celor două puncte.

4.2.2 Distanța euclidiană ponderată

Din cauza specificității domeniilor în care este utilizată analiza datelor, folosirea distanței euclidiene pentru a măsura gradul de apropiere sau de depărtare dintre obiecte ridică două importante dificultăți. Prima dificultate se referă la faptul că, în analiza datelor, dimensiunile spațiului în care se face evaluarea distanței nu sunt de aceeași natură, nu sunt omogene. În domeniul fizic, măsurarea distanțelor prin intermediul distanței euclidiene are consistență necesară, deoarece toate dimensiunile spațiului în care se face măsurarea sunt omogene, au aceeași natură, adică sunt exprimate în unități de măsură comparabile.

Spre deosebire de fizică, în domeniul economic, punctele ce reprezintă obiectele studiate au de obicei caracteristici *eterogene*, cu importanță diferită și exprimate în unități de măsură diferite, astfel încât, distanța euclidiană, care presupune însumarea pătratelor unor diferențe de natură diferită, își pierde relevanță.

O modalitate simplă de asigurare a comparabilității caracteristicilor obiectelor studiate, din punct de vedere al importanței pe care acestea o dețin în *caracterizarea obiectelor* constă în definirea unor ponderi sau unor coeficienți de importanță pentru caracteristicile studiate și utilizarea acestora în calculul unei *distanțe euclidiene ponderate*.

Definiție: Considerând că p_1, p_2, \dots, p_n sunt ponderile atribuite celor n caracteristici ale obiectelor, *distanța euclidiană ponderată* este definită de relația:

$$d_{E_p}(x,y) = \sqrt{p_1(x_1 - y_1)^2 + p_2(x_2 - y_2)^2 + \dots + p_n(x_n - y_n)^2}$$

Calculul distanței euclidiene *ponderate* este echivalent cu calculul distanței euclidiene, dar pentru cazul în care observațiile fiecărei a i-a caracteristici au fost multiplicate cu mărimea $\sqrt{p_i}$.

4.2.3 Distanța standardizată

Două obiecte pot fi considerate a fi mai apropiate între ele dacă ele provin dintr-o populație cu o împrăștiere mai mare, și mai depărtate dacă provin dintr-o populație cu o împrăștiere mai mică. Această constatare determină necesitatea ca distanța euclidiană să fie supusă unei transformări, unei adaptări, pentru a putea surprinde cele menționate mai sus. O astfel de transformare este cea reprezentată de distanța standardizată.

O măsură de natură statistică a distanței între obiecte, care ține seama de împrăștierea populațiilor din care provin obiectele este *distanța standardizată*.

În cazul multidimensional, când obiectele reprezintă puncte din spațiul real n-dimensional, distanța standardizată între obiectele x^i și x^j se calculează astfel:

$$d_{std}(x^i, x^j) = \sum_{k=1}^n \left(\frac{x_{ik} - x_{jk}}{s_k} \right)^2.$$

4.2.4 Distanța Mahalanobis

Distanța standardizată ia în considerare numai variabilitatea individuală ce caracterizează observațiile variabilelor, ceea ce echivalează, în mod implicit, cu faptul că, în calculul acestei distanțe variabilele sunt presupuse a fi necorelate.

O generalizare a distanței standardizate, care, spre deosebire de distanța standardizată, ia în considerare și variabilitatea interacțiunii dintre variabile, o reprezintă **distanța Mahalanobis**.

Distanța Mahalanobis ia în considerare atât variabilitatea individuală conținută în observațiile efectuate asupra variabilelor, cât și variabilitatea comună conținută în respectivele observații.

Pentru a fi senzitivă în raport cu variabilitatea individuală, în construcția distanței Mahalanobis sunt implicate varianțele variabilelor, iar pentru a fi senzitivă în raport cu variabilitatea comună, în construcția distanței Mahalanobis sunt implicate covarianțele și coeficienții de corelație.

Definiție: În cazul bidimensional, în care se consideră obiecte având câte două caracteristici, x_k și x_l , **distanța Mahalanobis** dintre două obiecte o^i și o^j este dată de relația:

$$d_{Mah}(o^i, o^j) = \frac{1}{1-r^2} \left(\frac{(x_k^i - x_k^j)^2}{s_k^2} - 2r \cdot \frac{(x_k^i - x_k^j)(x_l^i - x_l^j)}{s_k s_l} + \frac{(x_l^i - x_l^j)^2}{s_l^2} \right),$$

unde r reprezintă coeficientul de corelație dintre cele două variabile ce reprezintă caracteristicile obiectelor, s_k^2 și s_l^2 reprezintă varianțele, iar s_k și s_l reprezintă abaterile standard ale celor două variabile.

Este important să observăm că distanța standardizată și distanța euclidiană sunt cazuri particulare ale distanței Mahalanobis. Într-adevăr, dacă cele două variabile ce caracterizează obiectele sunt necorelate, adică $r=0$, distanța Mahalanobis coincide cu distanța standardizată. Pe de altă parte, dacă varianțele variabilelor sunt egale cu unitatea și variabilele sunt necorelate, distanța Mahalanobis coincide cu distanța euclidiană.

Definiție: În cazul obiectelor *multidimensionale*, adică al obiectelor caracterizate prin intermediul a n variabile, **distanța Mahalanobis** este definită de mărimea:

$$d_{Mah}(o^i, o^j) = (x^i - x^j)^t S^{-1} (x^i - x^j)$$

unde x^i și x^j sunt vectori n -dimensionali ale căror componente sunt reprezentate de valorile caracteristicilor obiectelor o^i și o^j , iar S este *matricea de covarianta*.

Dacă cele n variabile ce caracterizează obiectele sunt necorelate, matricea de covarianta S este o matrice diagonală, elementele diagonale ale acesteia reprezentând varianțele variabilelor. În cazul în care variabilele sunt standardizate și necorelate, matricea de covarianta S este matricea unitate, ceea ce înseamnă că distanța Mahalanobis se reduce la distanța euclidiană.

Tema 5. Analiza intragrupală și analiza intergrupală

În analiza datelor apar foarte multe situații în care este necesară *analiza comparativă a datelor provenind din populații diferite sau analiza caracteristicilor care determină diferențierea unor grupe* de obiecte din mulțimea obiectelor analizate.

În anumite situații care apar în diferite analize de date, obiectele supuse analizei sunt caracterizate de o mare eterogenitate după anumite caracteristici, astfel încât, pentru obținerea unor rezultate relevante este necesar ca analiza să se facă în mod diferențiat, în funcție de grupurile care se conturează pe mulțimea obiectelor în contextul respectivei eterogenități.

O altă conjunctură în care analiza datelor conduce la investigarea grupelor de obiecte și a interrelațiilor dintre acestea este aceea în care apare necesitatea clasificării unor obiecte în grupe cunoscute aprioric sau, mai general, necesitatea generării unei clasificări sau partajări a obiectelor supuse analizei.

În general, analiza caracteristicilor de grupare sau analiza grupelor are ca scop să verifice sau să determine *gradul de omogenitate din interiorul grupelor și gradul de eterogenitate dintre grupe*. În plus față de aceasta, în analiza caracteristicilor de grupare se urmărește și modul în care variabilele supuse analizei contribuie la diferențierea grupelor de obiecte, la discriminarea observațiilor sau obiectelor. Necesitatea de a analiza observațiile sau obiectele diferențiate pe grupe apare în contextul a numeroase probleme din domeniul analizei datelor, în cazul unora din ele având o importanță de excepție.

În cele ce urmează, vom face o scurtă prezentare a modului în care datele primare pot fi supuse unui proces de prelucrare preliminară, care să implice analiza relațiilor intragrupale și intergrupale.

5.1 Analiza de tip intragrupal

Analiza intragrupală are ca scop principal evidențierea gradului de omogenitate a obiectelor din fiecare grupă, determinarea unei *măsuri comune a similarității* obiectelor din fiecare grupă. O astfel de măsură este data de elementele matricii comune de covarianță, care măsoară similaritatea obiectelor în raport cu fiecare din variabilele care definesc obiectele analizate.

În general, dacă numărul de grupe este G , numărul de observații din cele G grupe este n_1, n_2, \dots, n_G , iar matricile produselor încrucișate ale grupelor sunt C_1, C_2, \dots, C_G , atunci *matricea comună a produselor încrucișate* este definită astfel:

$$C_w = \sum_{i=1}^G C_i.$$

Matricea comună de covarianță este definită ca fiind matricea rezultată din împărțirea elementelor matricii comune a produselor încrucișate la numărul gradelor de libertate:

$$df = n_1 + n_2 + \dots + n_G - G,$$

adică matricea:

$$S_w = \begin{pmatrix} \frac{c_{11}^w}{df} & \frac{c_{12}^w}{df} & \dots & \frac{c_{1n}^w}{df} \\ \frac{c_{21}^w}{df} & \frac{c_{22}^w}{df} & \dots & \frac{c_{2n}^w}{df} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{c_{n1}^w}{df} & \frac{c_{n2}^w}{df} & \dots & \frac{c_{nn}^w}{df} \end{pmatrix},$$

unde c_{ij}^w este element al matricii comune a produselor încrucișate.

Matricea comună de covarianță exprimă cantitatea de variație prezentă în observațiile fiecărei din cele G grupe de observații. Matricea comună de covarianță este foarte importantă în analiza datelor grupate, deoarece ea furnizează informația cu privire la omogenitatea sau similaritatea din interiorul grupelor, adică *omogenitatea intragrupală*.

De exemplu, dacă obiectele reprezentate de observațiile fiecărei grupe sunt identice din punct de vedere al tuturor variabilelor, adică toate observațiile unei variabile coincid cu media, atunci elementele matricii comune de covarianță vor fi *nule*, ceea ce evidențiază *omogenitatea perfectă* în interiorul grupelor. Valori mai mari decât zero ale elementelor matricii comune de covarianță evidențiază un anumit grad de eterogenitate a observațiilor în cadrul grupelor, eterogenitate care este cu atât mai mare cu cât valorile elementelor matricii comune de covarianță sunt mai mari.

5.2 Analiza de tip intergrupal

Analiza intergrupală se bazează pe studierea abaterilor înregistrate de mediile grupelor față de media generală, adică media calculată prin luarea în considerare a tuturor observațiilor, fără a ține seama de apartenența acestora la grupele existente. Aceste medii sunt definite în raport cu fiecare din variabilele analizate.

Elementul cheie în analiza intergrupală este reprezentat de *suma pătratelor abaterilor dintre grupe*. Această mărime se

definește și se calculează pentru fiecare din variabilele analizate. Pentru cea de-a i-a variabilă, suma pătratelor abaterilor dintre grupe este dată de relația:

$$SP_i = \sum_{j=1}^G n_j (\bar{x}_{ij} - \bar{x}_i)^2, \quad i=1,2,\dots,n,$$

unde G reprezintă numărul grupelor, n_j reprezintă numărul de observații din cea de-a j-a grupă, \bar{x}_{ij} reprezintă media înregistrată de cea de-a i-a variabilă la nivelul celei de-a j-a grupe, iar \bar{x}_i reprezintă media celei de-a i-a variabile la nivelul tuturor observațiilor din cele G grupe.

O matrice importantă utilizată în analiza intergrupală este **matricea produselor încrucișate intergrupale**, care se definește astfel:

$$C_b = \begin{pmatrix} \sum_{j=1}^G n_j (\bar{x}_{1j} - \bar{x}_1)^2 & \sum_{j=1}^G n_j (\bar{x}_{1j} - \bar{x}_1)(\bar{x}_{2j} - \bar{x}_2) & \dots & \sum_{j=1}^G n_j (\bar{x}_{1j} - \bar{x}_1)(\bar{x}_{nj} - \bar{x}_n) \\ \sum_{j=1}^G n_j (\bar{x}_{2j} - \bar{x}_2)(\bar{x}_{1j} - \bar{x}_1) & \sum_{j=1}^G n_j (\bar{x}_{2j} - \bar{x}_2)^2 & \dots & \sum_{j=1}^G n_j (\bar{x}_{2j} - \bar{x}_2)(\bar{x}_{nj} - \bar{x}_n) \\ \dots & & & \\ \sum_{j=1}^G n_j (\bar{x}_{nj} - \bar{x}_n)(\bar{x}_{1j} - \bar{x}_1) & \sum_{j=1}^G n_j (\bar{x}_{nj} - \bar{x}_n)(\bar{x}_{2j} - \bar{x}_2) & \dots & \sum_{j=1}^G n_j (\bar{x}_{nj} - \bar{x}_n)^2 \end{pmatrix}.$$

Matricea produselor încrucișate intergrupală este extrem de importantă în analiza grupelor sau a caracteristicilor de grupare, deoarece ea furnizează informația cu privire la *eterogenitatea grupelor* de observații.

Ea reprezintă o măsură a diferențelor care există între grupele de observații, arătând care este rolul variabilelor în diferențierea grupelor de observații.

Cu ajutorul matricii produselor încrucișate intergrupale poate fi calculată **matricea de covarianță intergrupală**:

$$S_b = \frac{1}{G-1} \cdot \begin{pmatrix} \sum_{j=1}^G n_j (\bar{x}_{1j} - \bar{x}_1)^2 & \sum_{j=1}^G n_j (\bar{x}_{1j} - \bar{x}_1)(\bar{x}_{2j} - \bar{x}_2) & \dots & \sum_{j=1}^G n_j (\bar{x}_{1j} - \bar{x}_1)(\bar{x}_{nj} - \bar{x}_n) \\ \sum_{j=1}^G n_j (\bar{x}_{2j} - \bar{x}_2)(\bar{x}_{1j} - \bar{x}_1) & \sum_{j=1}^G n_j (\bar{x}_{2j} - \bar{x}_2)^2 & \dots & \sum_{j=1}^G n_j (\bar{x}_{2j} - \bar{x}_2)(\bar{x}_{nj} - \bar{x}_n) \\ \dots & & & \\ \sum_{j=1}^G n_j (\bar{x}_{nj} - \bar{x}_n)(\bar{x}_{1j} - \bar{x}_1) & \sum_{j=1}^G n_j (\bar{x}_{nj} - \bar{x}_n)(\bar{x}_{2j} - \bar{x}_2) & \dots & \sum_{j=1}^G n_j (\bar{x}_{nj} - \bar{x}_n)^2 \end{pmatrix}.$$

Matricea de covarianță intergrupală exprimă informația referitoare la *eterogenitatea intergrupală*, arătând cât de mult se diferențiază grupele după fiecare din variabilele analizate.

Considerând că numărul total de observații este T și notând cu C_t matricea produselor încrucișate pentru totalitatea observațiilor centrate, indiferent de împărțirea lor pe grupe, adică matricea:

$$C_t = \begin{pmatrix} \sum_{t=1}^T (x_{t1} - \bar{x}_1)^2 & \sum_{t=1}^T (x_{t1} - \bar{x}_1)(x_{t2} - \bar{x}_2) & \dots & \sum_{t=1}^T (x_{t1} - \bar{x}_1)(x_{tn} - \bar{x}_n) \\ \sum_{t=1}^T (x_{t2} - \bar{x}_2)(x_{t1} - \bar{x}_1) & \sum_{t=1}^T (x_{t2} - \bar{x}_2)^2 & \dots & \sum_{t=1}^T (x_{t2} - \bar{x}_2)(x_{tn} - \bar{x}_n) \\ \dots & & & \\ \sum_{t=1}^T (x_{tn} - \bar{x}_n)(x_{t1} - \bar{x}_1) & \sum_{t=1}^T (x_{tn} - \bar{x}_n)(x_{t2} - \bar{x}_2) & \dots & \sum_{t=1}^T (x_{tn} - \bar{x}_n)^2 \end{pmatrix},$$

este verificată relația fundamentală:

$$C_t = C_w + C_b.$$

Această relație, foarte importantă în analiza datelor, evidențiază faptul că **matricea generală sau totală a produselor încrucișate** (C_t) poate fi descompusă sub formă a două matrici: **matricea comună sau cumulată a produselor încrucișate** (C_w) și **matricea produselor încrucișate intergrupale** (C_b).

În termeni informaționali, relația precedentă evidențiază faptul că *informația totală*, exprimată de variabilitatea ce caracterizează toate observațiile, poate fi împărțită pe două componente:

- *informația reprezentată de variabilitatea existentă în interiorul grupelor* și care este măsurată prin **matricea comună a produselor încrucișate** (C_w);

- informația reprezentată de *variabilitatea existentă între grupe și care este măsurată prin intermediul matricii produselor încrucișate intergrupale \mathbf{C}_b* :

Așa cum am mai menționat, cu ajutorul *matricii comune de covarianță* se exprimă similaritatea, *omogenitatea intragrupală*, iar cu ajutorul *matricii de covarianță intergrupală* se exprimă *eterogenitatea intergrupală*.

Din punct de vedere al unei singure variabile, se poate spune că, există diferențe cu atât mai mari între grupe cu cât suma pătratelor abaterilor intergrupale este mai mare în comparație cu suma pătratelor abaterilor intragrupale.

Această afirmație furnizează implicit criteriul teoretic ideal pentru construirea celui mai bun algoritm de clasificare: *maximizarea sumei pătratelor abaterilor intergrupale și minimizarea sumei pătratelor abaterilor intragrupale*.

În mod similar cu descompunerea matricii totale a produselor încrucișate pe cele două componente, se poate defini și o relație de descompunere a gradelor de libertate ce corespund totalității observațiilor, notate cu df , astfel:

$$df = df_w + df_b,$$

unde df_w reprezintă gradele de libertate corespunzătoare variabilității intragrupale, adică:

$$df_w = n_1 + n_2 + \dots + n_G - G,$$

iar df_b reprezintă gradele de libertate corespunzătoare variabilității intergrupale, adică:

$$df_b = G - 1.$$

După cum se poate observa, într-adevăr suma $df_w + df_b$ reprezintă numărul de grade de libertate ce caracterizează cele T observații:

$$df = (n_1 + n_2 + \dots + n_G - G) + (G - 1) = n_1 + n_2 + \dots + n_G - 1 = T - 1.$$

Analiza caracteristicilor de grupare și matricile definite în legătură cu aceasta sunt utilizate frecvent în probleme de clasificare și discriminare. În capitolul destinat prezentării tehnicilor de recunoaștere a formelor, vom discuta, detaliat, modul în care pot fi folosite procedurile de analiză a grupelor.

Tema 6. Analiza componentelor principale

6.1 Necessitatea și importanța simplificării spațiului cauzal

Demersul științific implicat în cadrul oricărei probleme de analiză și predicție din domeniul economico-social include, ca o etapă importantă a sa, o detaliată investigare a legăturilor funcționale existente între variabilele explicative. Variabilele explicative care intervin într-un demers științific de acest fel, variabile cunoscute și sub numele de variabile independente, reprezentă simboluri care exprimă diferențe cantitative sau calitative ale unor fenomene care au calitatea de *factori de influență* sau de *cauze* pentru alte fenomene sau procese.

În domeniul analizei datelor se consideră că ansamblul variabilelor explicative implicate într-o analiză multidimensională particulară definește un anumit spațiu numeric numit *spațiu inițial de cauzalitate*. Dimensiunea acestui spațiu este dată de numărul de variabile cauzale implicate în analiză, ceea ce înseamnă că numărul și natura axelor acestui spațiu sunt determinate de numărul și natura variabilelor analizate. Fiecare variabilă cauzală este reprezentată pe către o axă a acestui spațiu, un punct de pe o anumită axă a spațiului reprezentând o valoare posibilă pe care o poate lua variabila asociată cu axa respectivă. Punctele spațiului cauzal sunt reprezentate de obiectele supuse investigației, iar proiecțiile acestor obiecte pe axele spațiului sunt valorile înregistrate de obiecte la caracteristicile asociate cu axele.

Remarcă: Analiza componentelor principale asigură o descompunere exprimată printr-un *număr redus de componente și neredundanță* a variabilității totale din spațiul cauzal inițial.

Din punct de vedere al analizei datelor, cea mai importantă caracteristică pe care o poate evidenția spațiul de cauzalitate, caracteristică fundamentală, de fapt, pentru toate tehniciile de analiză multidimensională a datelor, o reprezintă *variabilitatea* acestui spațiu de cauzalitate.

Ca expresie directă și suficient de cuprindătoare a cantității de informație semnificativă pe care o conțin datele supuse analizei, variabilitatea spațiului cauzal, pe de o parte, precum și modalitățile prin intermediul căror aceasta poate fi exprimată sub cea mai simplificată și sugestivă formă, pe de altă parte, definesc esența și conținutul celor mai multe metode de analiză multidimensională a datelor.

Așa cum o să vedem în continuare, variabilitatea conținută în spațiul cauzal determinat de variabilele explicative poate fi exprimată sub mai multe forme posibile, forme mai mult sau mai puțin eficiente din punct de vedere al posibilităților acestora de a evidenția relațiile de cauzalitate care au relevanță pe mulțimea variabilelor explicative. Din acest punct de vedere, analiza componentelor principale poate fi privită ca o tehnică de descompunere a variabilității totale din spațiul cauzal inițial pe un număr mai redus de componente și fără nici o suprapunere a contribuțiilor individuale la formarea acestei variabilități.

Remarcă: Analiza componentelor principale este o tehnică de analiză multidimensională care are ca scop descompunerea variabilității totale din spațiul cauzal inițial sub forma unui *număr redus de componente* și fără ca această descompunere să conțină redundanțe informative.

În cazul specific al investigațiilor din domeniul economico-social, ca rezultat direct al numeroaselor și variatelor interdependențe manifestate *pe mai multe* palieri între variabilele explicative, spațiul de cauzalitate este caracterizat de o complexitate deosebită. Acest grad ridicat de complexitate a manifestării legăturilor cauzale dintre variabilele explicative determină serioase dificultăți legate de exprimarea și interpretarea facilă a raporturilor de cauzalitate, astfel încât apare necesitatea unei *reduceri*, unei *simplificări* a spațiului cauzal. Această reducere sau simplificare se referă la reducerea dimensionalității spațiului cauzal, la reducerea numărului de axe în funcție de care se face reprezentarea punctelor în acest spațiu.

Modalitatea cea mai adecvată pentru realizarea acestui obiectiv de mare importanță și utilitate, este aceea a utilizării tehniciilor de analiză multidimensională cunoscute sub numele de *analiza componentelor principale*.

Remarcă: Analiza componentelor principale este o tehnică de analiză multidimensională care are ca scop *reducerea dimensionalității spațiului cauzal inițial*, în condițiile unei pierderi informative minime.

Într-o exprimare extrem de sintetică, se poate spune că analiza componentelor principale este o metodă de analiză multidimensională care are ca scop găsirea unei anumite *modalități de transformare* a caracteristicilor inițiale ale unor *obiecte* sau *forme*, astfel încât, pe baza acestei transformări, să se asigure o *repräsentare optimală* a acestora, în condițiile folosirii unui număr mult mai redus de caracteristici.

Noile caracteristici rezultate în urma transformării caracteristicilor inițiale, al căror număr este semnificativ mai redus, sunt așa-numitele *componente principale*.

Caracterul de *optimalitate* pe care transformarea menționată trebuie să-l asigure noii modalități de reprezentare a obiectelor se referă la pierderea de informație referitoare la obiecte, care are loc odată cu trecerea de la vechile la noile caracteristici.

Se consideră că noile caracteristici asigură o reprezentare optimală a obiectelor dacă și numai dacă trecerea de la vechea reprezentare la noua reprezentare se face sub restricția *minimizării pierderii de informație*. Aceasta înseamnă că transformarea caracteristicilor se determină în așa fel încât pierderea de informație antrenată de reducerea dimensionalității caracteristicilor

să fie minimă. Realizarea acestei cerințe impuse transformării caracteristicilor presupune definirea unei performanțe, unei funcții obiectiv specifice, astfel încât transformarea să asigure, după caz, fie minimizarea, fie maximizarea acestei funcții obiectiv.

În cazul concret al analizei componentelor principale, performanța este reprezentată de *maximizarea varianței* caracteristicilor obiectelor, iar transformarea este o *transformare de tip liniar*.

Sintetizând cele menționate anterior, putem spune că analiza componentelor principale este utilizabilă pentru rezolvarea a două categorii generale de probleme: *simplificarea structurii dependenței cauzale și reducerea dimensionalității spațiului cauzal*.

6.1.1 Simplificarea structurii dependenței cauzale

Legăturile funcționale dintre variabilele explicative pot fi evidențiate prin intermediul definirii unei *structuri a dependenței* acestor variabile și măsurate, de obicei, cu ajutorul coeficienților de covarianță și corelație.

Remarcă: Sub forma sa cea mai simplă, *structura dependenței* este reprezentată de *mulțimea variabilelor cauzale supuse analizei*.

Necesitatea evidențierii structurii dependenței cauzale și a interpretării corecte a acesteia apare în orice analiză care are ca scop investigarea cauzalităților și este fundamentală pentru orice proces de cunoaștere științifică.

Pentru a facilita posibilitatea unei interpretări sugestive a cauzalității este necesar ca structura dependenței specifice unei mulțimi de variabile cauzale să fie cât *mai simplu și mai clar* exprimată, fără manifestarea perturbatorie a unor redundanțe informaționale. Evidențierea unei structuri a dependenței cât mai simple și mai clare reprezintă, de fapt, unul dintre principalele scopuri urmărite în analiza datelor. Pe baza unei astfel de structuri pot fi formulate concluzii mai clare cu privire la relațiile de cauzalitate manifestate la nivelul ansamblului de variabile explicative.

De cele mai multe ori însă, datorită complexității legăturilor cauzale și naturii specifice a informațiilor ce caracterizează variabilele analizate, o exprimare simplă și clară, sintetică și nereductată a cauzalităților este deosebit de dificilă și nu poate fi obținută în mod direct. Realizarea acestui deziderat poate fi obținută însă indirect, prin utilizarea unor instrumente specifice, cu ajutorul cărora să se poată face o anumită *simplificare a spațiului cauzal* determinat de variabilele explicative.

Remarcă: Prin *simplificarea spațiului cauzal* se înțelege reducerea dimensionalității acestuia, astfel încât să se obțină un spațiu cauzal de dimensiune mai mică (spațiul redus) și care să permită o reprezentare mai simplă și mai sugestivă a obiectelor.

În spațiul cauzal rezultat în urma acestei simplificări, spațiu a cărui cardinalitate este mult mai redusă în comparație cu spațiul cauzal inițial, este mult mai ușor să se evidențieze morfologia legăturilor cauzale și să se exprime o structură adecvată a dependențelor.

6.1.1.1 Eliminarea redundanțelor informaționale

În situația în care investigația științifică are ca obiect direct spațiu cauzal inițial, este foarte dificil să se deducă și să se exprime o dependență structurală care să evidențieze cu claritate *contribuțiile nete* ale variabilelor analizate la formarea variabilității întregului spațiu cauzal, mai ales atunci când aceste variabile sunt intercorelate.

Corelarea variabilelor cauzale determină o structură complicată și amalgamată a dependenței, o structură *redundantă*, care include anumite *suprapunerি informaționale* ale influențelor variabilelor cauzale. Faptul că structura cauzală inițială este complicată și include numeroase suprapunerি informaționale generează o serie de dificultăți privind înțelegerea clară a raporturilor de cauzalitate și formularea unor concluzii pertinente privind structurile de cauzalitate analizate.

O altă problemă importantă a analizei datelor, complementară intr-un anumit sens redundanței informaționale, constă în aceea că *variabilitatea conținută în spațiul inițial nu este la fel de semnificativă în toate direcțiile, de-a lungul tuturor axelor spațiului*, existând situații în care, după o anumită direcție, variabilitatea să fie neglijabilă din punct de vedere al magnitudinii.

În situații de acest fel se poate renunța la includerea în analiză a informației corespunzătoare acestor direcții, această renunțare fiind echivalentă cu a considera în analiză un spațiu de cauzalitate de dimensiune mai redusă.

6.1.2 Reducerea dimensionalității

În viziunea specifică analizei componentelor principale, unitățile elementare ale populației, obiectele supuse studiului, sunt considerate a fi puncte dintr-un spațiu ale cărui axe reprezintă caracteristicile posedate de respectivele obiecte. Aceasta înseamnă că în contextul analizei componentelor principale valorile caracteristicilor, adică observațiile, sunt coordonate ale punctelor ce definesc elementele populației analizate.

La baza analizei componentelor principale stă ideea că reprezentarea unităților în sistemul inițial de coordonate, adică în sistemul pe ale cărui axe sunt măsurate caracteristicile originale ale unităților, nu este totdeauna cea mai potrivită, considerându-se că *poate exista o altă modalitate de reprezentare mai relevanță, mai eficientă din punct de vedere informațional*.

Această modalitate de reprezentare, mai avantajoasă din punct de vedere informațional, poate fi obținută considerând un nou spațiu de reprezentare, spațiu care definește prin axe sale, în mod implicit, noi caracteristici ale obiectelor. Coordonatele obiectelor în acest nou spațiu sunt valorile înregistrate de obiecte la aceste noi caracteristici.

În contextul simbolizării cu ajutorul variabilelor, noile caracteristici sunt numite *componente principale*, iar valorile înregistrate de obiecte la aceste noi caracteristici sunt numite *scoruri*.

Având în vedere logica pe care se bazează determinarea lor, se consideră că noile caracteristici sunt mai relevante și mai adecvate pentru evaluarea informațională a obiectelor. Este evident că problema realizării noii reprezentări își pierde sensul dacă

noile caracteristici nu preiau, într-un fel sau altul, conținutul informațional al caracteristicilor inițiale. Mai mult decât atât, noua reprezentare se construiește în aşa fel încât conservarea informațională să fie *maximă*, acceptându-se, totuși, o *pierdere informațională minimală*.

Din punct de vedere geometric, analiza componentelor principale este echivalentă cu o “*rescriere*” a unităților unei populații într-un nou sistem de axe, cu o reprezentare mai adecvată din punct de vedere informațional a acestor unități. Noul sistem de coordonate rezultat din analiza componentelor principale este caracterizat prin trei trăsături fundamentale:

- are o *dimensiune redusă*;
- axele sale sunt *ortogonale*;
- coordonatele în acest sistem sunt *maximizatoare* de varianță.

Problema reprezentării într-un spațiu mai redus este cunoscută sub numele de problemă a *reducerii dimensionalității*. Din acest motiv, analiza componentelor principale este cunoscută și ca *tehnica de reducere a dimensionalității*.

Dacă vom considera că există n variabile originale, reprezentate de elementele mulțimii $\{x_1, x_2, \dots, x_n\}$, esența analizei componentelor principale poate fi reprezentată în mod simplificat prin intermediul transformării următoare:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \xrightarrow{\Psi} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix},$$

unde $\{w_1, w_2, \dots, w_k\}$ reprezintă mulțimea componentelor principale, iar transformarea Ψ este astfel încât asigură, prin intermediul componentelor principale, conservarea variabilității din spațiul cauzal inițial *într-o proporție maximă posibilă*.

Pentru a ilustra restricțiile sub care poate fi făcută reducerea dimensionalității, vom considera reprezentările grafice din figura următoare, referitoare la un număr de 10 obiecte care posedă câte două caracteistică. Obiectele au fost alese astfel să sugereze atât conținutul procesului de reducere a dimensionalității, cât și necesitatea acestui proces.

În această figură sunt evidențiate două modalități de reprezentare a celor 10 obiecte: prima este cea originală, adică aceea în care obiectele sunt reprezentate în coordonatele inițiale, netransformate, iar cea de-a doua este cea rezultată în urma reducerii dimensionalității.

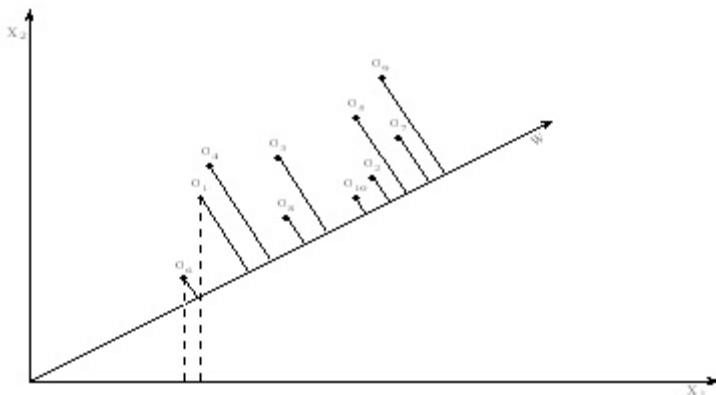


Figura 6.1: Exemplificarea reducerii dimensionalității de la două axe - \mathbf{X}_1 și \mathbf{X}_2 , la o singură axă - W

În cadrul primei reprezentări, obiectele sunt considerate a fi puncte din planul x_1Ox_2 , cele două coordonate ale fiecăruiu dintre aceste puncte, măsurate pe axele reprezentate de \mathbf{X}_1 și \mathbf{X}_2 , fiind evaluări ale două presupuse caracteisticăi. Ceea ce este deranjant în contextul acestei reprezentări inițiale a obiectelor este faptul că *rezoluția* acestor obiecte de-a lungul axei \mathbf{X}_1 este foarte mică. Obiectele pot fi distinse cu dificultate unele de altele, unele fiind chiar confundabile. În termeni statistici, rezoluția scăzută a obiectelor în sistemul de coordonate inițial este echivalentă cu o variabilitate scăzută a primei caracteisticăi, de-a lungul obiectelor. Mai departe, această variabilitate scăzută poate fi interpretată ca însemnând o semnificație redusă a primei caracteisticăi, din punct de vedere ale puterii ei de a diferenția obiectele analizate.

Reducerea dimensionalității constă, în acest caz, în trecerea de la două dimensiuni la o singură dimensiune. Ca urmare a reducerii dimensionalității, se trece de la reprezentarea obiectelor prin intermediul a două coordonate la reprezentarea obiectelor prin intermediul unei singure coordonate. Aceasta înseamnă trecerea de la reprezentarea în plan la reprezentarea pe o dreaptă.

Noua axă rezultată în urma reducerii dimensionalității, notată în grafic cu W , poate fi interpretată ca fiind expresia unei caracteisticăi noi, rezultată din combinarea într-o anumită formă, a celor două caracteisticăi originale.

Se poate observa că reprezentarea rezultată în urma reducerii dimensionalității crește rezoluția obiectelor. Aceasta înseamnă că noua caracteistica, simbolizată prin intermediul variabilei W , este caracterizată de o mai mare variabilitate de-a lungul obiectelor, comparativ cu caracteistica simbolizată prin intermediul variabilei \mathbf{X}_1 . Noua variabilă W poate fi interpretată ca

fiind o componentă principală.

În urma reducerii dimensionalității s-a obținut o nouă entitate informațională, care poate fi interpretată ca reprezentând o nouă caracteristică a obiectelor. Informația conținută în această nouă entitate este mai relevantă decât informația conținută în observațiile făcute asupra primei caracteristici a obiectelor.

6.1.3 Necessitatea reducerii dimensionalității

Necesitatea de a simplifica spațiul cauzal inițial, de a exprima cauzalitatea sub o formă mai simplă, prin intermediul unui număr mai redus de variabile, apare în numeroase probleme de analiză a datelor și, în plus, condiționează posibilitățile de aplicare a multor proceduri specifice analizei datelor.

Problema *reducerii dimensionalității* definește în mod sintetic atât esența tehniciilor de analiză a componentelor principale, cât și scopul majorității modalităților de utilizare a acestor tehnici.

Anterior am evidențiat faptul că în orice activitate de cunoaștere științifică apare necesitatea identificării unei structuri simple și clare a dependenței, arătând că, pentru realizarea acestui scop, trebuie să se facă o reducere, o simplificare a cauzalității. Modalitatea de abordare cea mai adekvată pentru realizarea acestui scop este cea bazată pe tehniciile oferite de analiza componentelor principale.

Separat de această importantă și generală modalitate de utilizare, reducerea dimensionalității spațiului de cauzalitate este deosebit de utilă și pentru rezolvarea unor probleme concrete cum ar fi cele legate de *selectarea variabilelor de influență, simplificarea modelelor matematice, eliminarea redundanțelor informative, vizualizarea relațiilor de cauzalitate complexe, compresia și restaurarea datelor în informatică* etc.

În continuare, vom face o scurtă prezentare a modului în care analiza componentelor principale poate fi folosită pentru soluționarea fiecarei dintre problemele menționate.

6.1.3.1 Selectarea variabilelor de influență

Caracteristica principală a celor mai multe probleme de analiză și predicție din domeniul economico-social este dată de faptul că dimensiunea *spațiului cauzal* investigat, dimensiune determinată de numărul de variabile ce pot fi identificate drept *cauze* care determină evoluția unui fenomen *efect*, este foarte mare.

Deoarece nu toate variabilele de influență au aceeași importanță în formarea caracteristicilor fenomenului *efect*, este necesar ca aceste variabile să fie supuse unui *proces de filtrare*, prin intermediul căruia unele variabile să fie eliminate, iar altele reținute, în funcție de semnificația fiecarei dintre acestea.

Selectarea variabilelor independente, în funcție de semnificația influenței pe care o au acestea asupra caracteristicilor fenomenului *efect*, poate fi făcută și cu ajutorul tehniciilor de reducere a dimensionalității.

În afara procesului de filtrare a variabilelor independente în funcție de importanța acestora, în analiza datelor apare frecvent necesitatea de a *grupa sau categorisi* variabilele independente în funcție de natura influenței pe care acestea o exercită.

Clasificarea variabilelor explicative este foarte importantă mai ales în domeniul economico-financiar, unde apare în mod frecvent necesitatea identificării unor grupe specifice de variabile, care evidențiază anumite caracteristici globale ale structurilor economice, cum ar fi: dezvoltarea economică, dezvoltarea socială, forța financiară etc.

Deși există numerose metode și tehnici care pot fi utilizate scopul grupării variabilelor de influență, totuși, analiza componentelor principale este printre metodele cele mai simple și mai eficiente în acest sens.

6.1.3.2 Simplificarea modelelor matematice

În condițiile în care specificitatea investigațiilor din domeniul economico-social este reprezentată de complexitate, apare necesitatea ca în procesul de modelare a fenomenelor și proceselor economico-sociale să se ia în considerare o mulțime foarte mare de variabile cauzale. Aceasta determină însă obținerea unor modele prea complicate, care includ un număr mult prea mare de variabile cauzale și care crează numeroase dificultăți de ordin metodologic și tehnic.

Există multe motive care fac dificilă și incomodă reținerea în cadrul unui model matematic de analiză sau predicție a unui număr foarte mare de variabile și care conduc la necesitatea unei anumite simplificări a modelului din acest punct de vedere.

În primul rând, *semnificația fiecarei variabile într-un model care include un număr prea mare de variabile este foarte mult diminuată*. De exemplu, dacă în construcția unui model au fost reținute peste 100 de variabile și presupunând pentru simplificare că, din punct de vedere al scopului urmărit, semnificația fiecarei variabile este comparabilă cu a celorlalte, ar rezulta pentru fiecare variabilă o "contribuție" mai mică decât un procent. În acest fel, semnificația fiecarei variabile apare mult diminuată, devine aproape neglijabilă, ea fiind mai redusă chiar decât cea a erorilor de observare sau a factorilor cu influență accidentală.

În al doilea rând, *obținerea informațiilor necesare estimării unui model care conține un număr foarte mare de variabile ar implica un efort și un cost prohibitive*, în anumite situații fiind greu sau chiar imposibil de obținut informații semnificative pentru toate variabilele; în cazul unui astfel de model, pentru a păstra un număr de grade de libertate suficient de ridicat în vederea asigurării unei semnificații rezonabile pentru estimarea parametrilor modelului, ar fi necesar un număr foarte mare de observații efectuate asupra fiecarei variabile.

În al treilea rând, *este posibil ca în cazul unui număr mare de variabile să avem o probabilitate ridicată de existență a unor variabile foarte puternic intercorelate*, incluzându-aceste variabile în model însenmând, în afara introducerii unui anumit grad de redundanță informațională, și dificultăți serioase de estimare a modelului ca urmare a apariției fenomenului de *colinearitate* sau *multicolinearitate*.

În al patrulea rând, *un număr foarte mare de variabile reținute într-un model ar ridica serioase probleme de complexitate*

a calculului chiar în raport cu ridicatele performanțe ale configurațiilor hardware actuale, presupunând implicit costuri foarte ridicate de prelucrare.

Având în vedere cele menționate, rezultă că procesul de construire a modelului presupune necesitatea efectuării unor simplificări și unor transformări corespunzătoare asupra mulțimii de variabile independente.

În toate situațiile de acest fel, analiza componentelor principale poate servi ca un puternic instrument complementar, atât în faza de construire a modelelor, cât și în faza de estimare a parametrilor acestor modele.

6.1.3.3 Compresia și restaurarea datelor

O problemă importantă din domeniul informaticii, legată în special de eficiența cu care are loc transmiterea și stocarea informației, este aceea a *compresiei și restaurării* datelor.

Analiza componentelor principale este una dintre tehnicele cele mai frecvent utilizate pentru soluționarea unei astfel de probleme, pentru activitatea de dezvoltare a algoritmilor de comprimare și restaurare a datelor. Eliminarea redundanțelor informaționale, asigurată prin intermediul utilizării analizei componentelor principale, oferă un foarte mare avantaj pentru orice proces de stocare și transmitere de informație din sistemele informaționale.

Dacă noua reprezentare a obiectelor asigurată prin aplicarea tehnicii de analiză a componentelor principale este însotită de o pierdere informațională neglijabilă, atunci atât în stocarea, cât și în transmiterea de informație, pot fi utilizate ca purtătoare de informație noile caracteristici ale obiectelor. Cum numărul acestora este mult mai redus în comparație cu numărul caracteristicilor inițiale, rezultă că manipularea informației se poate face cu consum mai mic de resurse, cu costuri mai reduse.

6.1.3.4 Vizualizarea unor relații de cauzalitate complexe

În situația în care numărul de variabile explicative este mai mare decât trei, nu mai există posibilitatea de a vizualiza poziționarea obiectelor ale căror caracteristici sunt decrete prin intermediul acestor variabile, depășindu-se limita maximă posibilă a reprezentărilor grafice, limită dată de contextul tridimensional.

Chiar și în cazul în care numărul de variabile explicative este egal cu trei, reprezentarea grafică tridimensională a obiectelor devine greoai și, de cele mai multe ori, este lipsită de relevanță.

Prin intermediul analizei componentelor principale pot fi create condițiile necesare pentru ca să poată fi reprezentate grafic chiar obiecte ce sunt caracterizate de un număr foarte mare de variabile.

Modalitatea concretă de reprezentare grafică a obiectelor constă în determinarea primelor două componente principale, corespunzătoare variabilelor originale și reprezentarea obiectelor într-un spațiu ortogonal ale cărui axe sunt reprezentate de cele două componente principale.

Calitatea reprezentării grafice realizată prin intermediul acestei modalități depinde de măsura în care cele două componente principale exprimă conținutul informațional al variabilelor originale, respectiv de proporția pe care componentele principale o preiau din varianța spațiului cauzal inițial.

Situatiile menționate mai sus conturează câteva dintre cele mai importante argumente justificative pentru necesitatea de a efectua o *simplificare*, o *reducere* a complexității cauzalității, fără însă ca această reducere să determine o pierdere importantă din informația semnificativă, relevantă.

O astfel de simplificare este echivalentă cu o reducere, de dorit cât mai semnificativă, a dimensiunii spațiului cauzal, o trecere de la un spațiu cu un număr foarte mare de dimensiuni la un spațiu cu un număr mai redus de dimensiuni, asigurând prin această trecere un procent cât mai ridicat de conservare a informației conținută în spațiul cauzal inițial.

Orientarea axelor spațiului redus trebuie să fie astfel încât să permită o *descompunere neredundantă a variabilității spațiului cauzal inițial pe factori de influență*, o exprimare a structurii dependenței fără nici o suprapunere a contribuțiilor individuale ale factorilor cauzali.

Remarcă: Prin *descompunere neredundantă a variabilității spațiului cauzal inițial* se înțelege exprimarea varianței totale care caracterizează acest spațiu ca sumă a varianțelor unui număr mai mic de noi variabile, în condițiile în care acestea sunt necorelate între ele.

Este evident că obținerea unei astfel de simplificări nu poate avea loc printr-o reducere a dimensiunii spațiului cauzal inițial făcută în mod simplist, bazată pe renunțarea arbitrară la unele din variabilele cauzale și reținerea variabilelor cauzale rămase, mai ales atunci când importanța apriorică a variabilelor pare a fi aproximativ egală.

Singura posibilitate de a face o astfel de simplificare, sub restricțiile menționate mai sus, este cea a *definirii unor noi variabile*, mai puține la număr decât variabilele inițiale, care să exprime într-o manieră consistentă semnificația informațională a variabilelor originale și care să asigure într-o măsură suficient de mare "conservarea" informației conținute în variabilele inițiale.

6.1.4 Domenii ale utilizării analizei componentelor principale

Există numeroase situații concrete în care se dorește obținerea unor informații cu caracter mai special, care să evidențieze profunzimea și subtilitatea interdependențelor existente la nivelul unei realități oarecare. Aceste situații conduc, în mod inevitabil, la necesitatea utilizării tehnicii specifice analizei componentelor principale.

Pentru a ilustra natura situațiilor în care apare necesitatea utilizării tehnicii de analiză a componentelor principale, menționăm următoarele exemple:

- Într-o cercetare intreprinsă la nivelul unui număr de firme dintr-un anumit domeniu, în scopul determinării *forței financiare* a acestora, a fost identificat un *număr foarte mare de indicatori* economico-financiari, astfel încât este foarte dificilă deducerea unei ierarhii financiare pe mulțimea firmelor analizate; pentru operaționalizarea informațiilor repre-

zentate de acești indicatori și pentru creștrea relevanței acestora este necesară utilizarea analizei componentelor principale;

- o investigație științifică în domeniul social are ca scop identificarea unor tipologii socio-culturale, specifice unor zone geografice; informațiile de natură socială și culturală disponibile pot fi utilizate pentru construirea acestor tipologii numai în condițiile existenței unor metode și tehnici adecvate, în rândul căror analiza componentelor principale ocupă cel mai important loc;

- în activitatea de control al calității producției se dorește ca, pe baza a numeroase informații privitoare la desfășurarea procesului de fabricație, să se definească un număr mic de *indicatori relevanți* pentru a aprecia dacă procesul se desfășoară în parametrii calitativi corespunzători; acești indicatori pot fi obținuți prin utilizarea tehnicii specifice analizei componentelor principale;

- într-o cercetare din domeniul economico-financiar a fost identificat un model în care variabilele independente sunt afectate de *fenomenul de colinearitate*; în aceste condiții este posibil ca erorile standard ale estimărilor parametrilor să fie foarte mari, astfel încât calitatea modelului să fie negativ afectată; pentru a putea obține estimări corespunzătoare este necesar ca variabilele originale să fie *substituite* cu alte variabile noi, necorelate, cum ar fi componentele principale.

Analiza datelor, indiferent dacă respectivele date sunt de natură economică, socială, medicală, biologică sau tehnică, reprezintă domeniul predilect al utilizării analizei componentelor principale. Utilizarea analizei componentelor principale în analiza datelor are loc atât în sens individual, ca tehnică independentă de analiză a datelor, cât și împreună, în complementaritate, cu alte metode și tehnici de analiză.

Analiza componentelor principale este folosită în probleme de analiză a datelor atât în faza inițială a acestora, ca *tehnica de analiză preliminară*, cât și în fazele ulterioare ale acestor analize, în special în faza de *interpretare a rezultatelor*.

În cele ce urmează, vom preciza câteva dintre cele mai importante domenii și activități ale analizei datelor, în care utilizarea analizei componentelor principale este nu numai posibilă, ci și strict necesară.

- analiza preliminară a datelor;
- construirea modelelor matematice;
- soluționarea problemelor de analiză factorială;
- scalarea multidimensională;
- recunoașterea formelor;
- analiza grafică;
- prezentarea și interpretarea rezultatelor.

Anterior, am evidențiat necesitatea simplificării spațiului cauzal și am menționat unele din situațiile în care această simplificare se impune. Tehnica specifică folosită pentru reducerea dimensiunii spațiului cauzal inițial, în sensul prezentat anterior, poartă numele de **analiza componentelor principale**, iar noile variabile care definesc spațiul redus de cauzalitate se numesc **componente principale**. În cadrul paragrafelor următoare, vom face o definire a analizei componentelor principale, precum și a noilor variabile construite în contextul acestei analize, respectiv a componentelor principale.

6.2 Definirea analizei componentelor principale

În mod concret și într-o viziune simplificată, tehnica pe care se bazează analiza componentelor principale constă în *calculul proiecțiilor fiecărui punct din spațiul inițial*, determinat de variabilele originale supuse analizei, *pe axele unui nou spațiu, a cărui dimensiune este semnificativ mai redusă*. În sens riguros, dar totuși foarte general, analiza componentelor principale poate fi definită sub următoarea formă:

Definiție: *Analiza componentelor principale* este o metodă de analiză multidimensională care are ca scop determinarea unor noi variabile, numite **componente principale** și exprimate sub forma *combinațiilor liniare de variabilele originale*, astfel încât aceste variabile noi să fie caracterizate de o variabilitate maximă.

În mod firesc, numărul de combinații liniare posibil a fi formate cu variabilele originale este extrem de mare. Deoarece, din punct de vedere al principiilor pe care se bazează activitatea de analiză a datelor, prezintă interes deosebit numai acele combinații liniare semnificative din punct de vedere informațional, caracterizate de o mare variabilitate, este necesară o triere, o selectare a acestor combinații liniare. Efectuarea acestei selectări presupune definirea unui *criteriu* care să stea la baza deciziei de reținere sau de eliminare a unei anumite combinații liniare.

În cadrul analizei componentelor principale acest criteriu este bazat pe magnitudinea varianței fiecărei combinații liniare și poate fi formulat astfel: *se elimină combinațiile liniare cu varianță mică, nesemnificativă și se rețin pentru studiu acele combinații liniare care au o varianță maximă*.

Refinarea în analiză doar a celor combinații liniare care au varianță maximă are ca scop final realizarea unui eventual context în care variabilele originale să poată fi înlocuite cu un număr mult mai mic de astfel de combinații liniare, în condițiile în care prin intermediul combinațiilor liniare reținute se preia o parte cât mai mare din variabilitatea conținută în observațiile variabilelor originale.

6.3 Definirea componentelor principale

După cum am mai arătat, componentele principale sunt noi variabile al căror conținut informațional, preluat în exclusivitate de la variabilele originale, este definit, mai ales, în raport cu legăturile care există între variabilele originale.

Sintetizând cele menționate până acum în legătură cu componentele principale, putem da următoarea definiție a acestora:

Definiție: Componentele principale sunt variabile vectoriale abstrakte, definite sub forma unor *combinații liniare de variabilele originale* și care au următoarele două proprietăți fundamentale:

- sunt *necorelate* două câte două și suma pătratelor coeficienților care definesc combinația liniară ce corespunde unei componente principale este egală cu unitatea;
- prima componentă principală este o *combinație liniară normalizată a cărei varianță este maximă*, cea de-a două componentă principală este o combinație liniară necorelată cu prima componentă principală și care are o varianță cât mai mare posibilă, însă mai mică decât cea a primei componente etc.

Verificarea primei proprietăți de către coeficienții combinațiilor liniare ce definesc componentele principale face ca acești coeficienți, priviți sub formă vectorială, să alcătuiască un *sistem ortonormal*.

Componentele principale sunt vectori ortogonali care preiau cât mai mult din varianța variabilelor vector originale astfel: prima componentă principală preia maximul posibil din varianța variabilelor originale, a doua componentă principală preia maximul din varianța rămasă după ce este eliminată varianța preluată de prima componentă și.a.m.d.

Verificarea primei proprietăți de către coeficienții combinațiilor liniare ce definesc componentele principale face ca acești coeficienți, priviți sub formă vectorială, să alcătuiască un *sistem ortonormal*.

Componentele principale sunt vectori ortogonali care preiau cât mai mult din varianța variabilelor vector originale astfel: prima componentă principală preia maximul posibil din varianța variabilelor originale, a doua componentă principală preia maximul din varianța rămasă după ce este eliminată varianța preluată de prima componentă și.a.m.d.

6.3.1 Caracteristici ale componentelor principale

În calitatea lor de construcții abstrakte, rezultate din utilizarea unor tehnici specifice, componentele principale reprezintă noi variabile care, prin proprietățile interesante pe care acestea le au, oferă noi și subtile posibilități de analiză și interpretare informațională a datelor originale.

Din punct de vedere geometric, variabilele numite componente principale definesc un nou spațiu al obiectelor, în contextul căruia sunt verificate următoarele proprietăți relevante pentru definirea analizei componentelor principale:

- axele noului spațiu sunt ortogonale două câte două și definesc noile variabile numite *componente principale*;
- coordonatele obiectelor în noul spațiu, adică proiecțiile obiectelor pe axele acestuia, sunt evaluări obiectelor în raport cu noile variabile și se numesc *scoruri ale componentelor principale* sau *scoruri principale*.
- din punct de vedere teoretic, *numărul de componente principale este egal cu numărul de variabile originale*; nu toate componentele principale au însă o semnificație informațională considerabilă, astfel cele mai puțin semnificative din punct de vedere informațional sunt eliminate;
- componentele principale sunt *combinații liniare de varianță maximală* ale variabilelor originale;
- componentele principale sunt *scalate* în funcție de magnitudinea varianței acestora, prima fiind componenta principală cu varianță maximă, iar ultima fiind componenta principală cu varianță minimă;
- componentele principale sunt *necorelate două câte două*;
- *suma varianțelor componentelor principale coincide cu suma varianțelor variabilelor originale*, astfel încât componentele principale preiau în totalitate variabilitatea conținută în variabilele originale.

Sintetizând cele arătate mai sus, putem spune că analiza componentelor principale este o metodă de *reexprimare* a variabilelor originale sub forma unui număr mai mic de noi variabile, numite componente principale, care sunt combinații liniare de varianță maximă ale variabilelor originale.

Cu ajutorul componentelor principale se poate defini o structură a dependenței dintre variabilele originale mai simplă și mai clară, deci mai ușor de interpretat. În cazul în care numărul de componente principale este egal cu numărul variabilelor originale putem privi analiza componentelor principale ca pe o metodă de rotație, de regulă ortogonală, a axelor spațiului inițial, semnificația fiecărei noi axe fiind măsurată prin varianța asociată unei componente principale.

6.4 Logica analizei componentelor principale

Cele mai interesante și mai utile aspecte ale analizei componentelor principale sunt în primul rând legate, nu de aparatul matematic pe care această analiză se bazează, ci de multiplele și nuanțatele interpretări posibile pe care aceasta le oferă.

Pentru a da o ilustrare intuitivă clară, bazată pe o interpretare geometrică simplificată, raționamentului primar care stă la baza analizei componentelor principale, vom dedica această parte, în exclusivitate, interpretărilor și exemplificărilor numerice. În acest sens, vom considera contextul numeric oferit de exemplul următor, context care va servi ca referință pentru multe din interpretările și exemplificările ulterioare.

Exemplul 6.1

Vom considera cazul unui număr de 10 obiecte sau observații, referitoare la două variabile, x_1 și x_2 . Tabelul următor conține observațiile inițiale disponibile pentru cele două variabile, precum și valorile centrate ce corespund acestor observații.

Valorile observațiilor inițiale și centrate

Tabelul 6.1

Observația	Valori inițiale		Valori centrate	
	x_1	x_2	x_1^c	x_2^c
O_1	7,0	10,0	0,6	-0,5
O_2	5,0	11,0	-1,4	0,5
O_3	10,0	15,0	3,6	4,5
O_4	2,0	5,0	-4,4	-5,5
O_5	5,0	10,0	-1,4	-0,5
O_6	6,0	13,0	-0,4	2,5
O_7	7,0	12,0	0,6	1,5
O_8	9,0	11,0	2,6	0,5
O_9	7,0	8,0	0,6	-2,5
O_{10}	6,0	10,0	-0,4	-0,5
Media	6,4	10,5	0	0
Varianța	4,933	7,389	4,933	7,389

Varianța individuală pentru fiecare din cele două variabile este 4,933, respectiv 7,389, iar varianța totală, corespunzătoare celor două variabile, x_1 și x_2 , este 12,322:

$$s_{11} = 4,933; \quad s_{22} = 7,389; \quad V_T = 12,322.$$

În aceste condiții, se poate spune că *rolul informațional* al celor două variabile este aproximativ *același*, că cele două variabile au aproximativ aceeași contribuție la formarea variabilității totale ce caracterizează spațiul cauzal inițial. Prima variabilă are o contribuție la formarea varianței totale de 46,45%, iar cea de-a doua variabilă contribuie cu 53,55% la formarea varianței totale:

$$\frac{s_{11}}{V_T} = 46,45\%, \quad \frac{s_{22}}{V_T} = 53,55\%.$$

Pentru observațiile din tabelul anterior, matricea produselor încrucișate, matricea de covarianță și matricea de corelație, corespunzătoare celor două variabile x_1 și x_2 , sunt următoarele:

$$C = \begin{pmatrix} 454,0 & 712,0 \\ 712,0 & 1169,0 \end{pmatrix} \quad S = \begin{pmatrix} 4,933 & 4,444 \\ 4,444 & 7,389 \end{pmatrix} \quad R = \begin{pmatrix} 1,000 & 0,736 \\ 0,736 & 1,000 \end{pmatrix}.$$

În cazul observațiilor centrate, matricea produselor încrucișate, matricea de covarianță și matricea de corelație sunt următoarele:

$$C = \begin{pmatrix} 44,4 & 40,0 \\ 40,0 & 66,5 \end{pmatrix} \quad S = \begin{pmatrix} 4,933 & 4,444 \\ 4,444 & 7,389 \end{pmatrix} \quad R = \begin{pmatrix} 1,000 & 0,736 \\ 0,736 & 1,000 \end{pmatrix}.$$

După cum se poate observa, în urma operației de centrare se modifică doar matricea produselor încrucișate, matricea de covarianță și matricea de corelație rămânând neschimbate. Matricea de corelație evidențiază faptul că cele două variabile sunt *corelate*, la nivelul unui coeficient de corelație de 0,736, adică:

$$r_{12} = r_{21} = 0,736.$$

Având în vedere intensitatea relativ ridicată a legăturii dintre cele două variabile originale, este de așteptat ca aceste variabile să poată fi sintetizate prin intermediul unei singure componente principale, în condițiile unei pierderi informative minime.

6.4.1 Rotația axelor și maximizarea varianței

Logica analizei componentelor principale se bazează pe ideea fundamentală că *se pot face anumite transformări asupra observațiilor inițiale*, care să determine *maximizarea* varianței individuale pentru anumite variabile și *minimizarea* varianței pentru alte variabile. Pe această cale, se accentuează semnificația logico-informațională a unor variabile și se diminuează cea a altor variabile.

Maximizarea varianței unor variabile, în detrimentul varianței celorlalte variabile, conduce la creșterea contribuției variabilelor a căror varianță este maximizată la formarea varianței totale. Maximizarea varianței unor variabile și, pe această bază, accentuarea semnificației informative a acestor variabile în raport cu celelalte este cu atât mai relevantă cu cât între variabilele originale există legături mai puternice, respectiv sunt *mai puternic corelate*. Așa cum o să arătăm în continuare, transformările care asigură maximizarea varianțelor individuale ale unor variabile sunt reprezentate de transformările de coordonate corespunzătoare efectuării unor rotații ortogonale ale axelor originale.

Pentru a evidenția modul în care pot fi deduse componente principale, adică noile variabile care au proprietatea de a conserva variabilitatea ce caracterizează spațiul cauzal inițial și care sunt necorelate, vom proceda la rotații successive ale celor două axe inițiale, măsurând varianța ce caracterizează cele două variabile pentru fiecare poziție modificată prin rotație a sistemului de axe. Ca urmare a faptului că sistemul de axe este rotit cu un anumit număr de grade, coordonatele celor două variabile se modifică în mod corespunzător, astfel încât, reprezentarea observațiilor în noile coordonate este diferită.

Dacă vom considera că în sistemul inițial de axe punctul reprezentat de cea de-a t-a observație are coordonatele (x_{1t}, x_{2t}) ,

atunci rotația axelor cu un anumit unghi va determina o modificare a coordonatelor acestui punct, respectiv a valorilor celor două observații.

În cazul în care unghiul de rotație este θ , noile coordonate ale punctului (x_{t1}, x_{t2}) , noteate x'_{t1} și x'_{t2} , sunt date de relațiile următoare:

$$\begin{aligned} x'_{t1} &= \cos\theta \cdot x_{t1} + \sin\theta \cdot x_{t2} \\ x'_{t2} &= -\sin\theta \cdot x_{t1} + \cos\theta \cdot x_{t2} \end{aligned}$$

Prin intermediul exemplului următor se evidențiază modul în care se schimbă varianța fiecărei variabile odată cu efectuarea unei rotații a axelor originale.

Exemplul 6.2

Considerând datele din Exemplul 6.1, pentru o rotație a axelor cu un unghi de 10 grade, coordonatele primei observații centrate, respectiv coordonatele punctului $(0,6; -0,5)$, devin $0,504$ și $-0,597$:

$$\begin{aligned} 0,504 &= \cos 10 \cdot (0,6) + \sin 10 \cdot (-0,5) \\ -0,597 &= -\sin 10 \cdot (0,6) + \cos 10 \cdot (-0,5) \end{aligned}$$

În tabelul următor sunt prezentate coordonatele celor două variabile într-un sistem de axe în care axele sunt rotite cu 10, 30, 45, 60 și 90 grade. Penultima linie a tabelului conține varianțele celor două variabile, calculate pentru fiecare poziție obținută din rotația axelor cu un număr de grade.

După cum se poate observa, varianțele celor două variabile sunt diferite pentru diferitele poziții ale axelor, deși suma acestor varianțe, adică varianța totală, rămâne neschimbată prin rotația axelor. Aceasta înseamnă că variabilitatea conținută în observațiile corespunzătoare sistemului de axe inițial este integral conservată odată cu rotația axelor.

Coordonatele observațiilor variabilelor în condițiile rotației axelor

Tabelul 6.2

Observația	Coordonatele variabilelor pentru diferite unghiuri de rotație									
	10 grade		30 grade		45 grade		60 grade		90 grade	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
O_1	0,504	-0,597	0,269	-0,733	0,071	-0,778	-0,133	-0,769	-0,500	-0,60
O_2	-1,292	0,736	-0,962	1,133	-0,636	1,343	-0,267	1,462	0,500	1,400
O_3	4,327	3,806	5,368	2,097	5,728	0,636	5,697	-0,868	4,500	-3,60
O_4	-5,288	-4,652	-6,561	-2,563	-7,000	-0,778	-6,963	1,060	-5,500	4,400
O_5	-1,466	-0,249	-1,462	0,267	-1,343	0,636	-1,133	0,962	-0,500	1,400
O_6	0,040	2,531	0,904	2,365	1,485	2,051	1,965	1,596	2,500	0,400
O_7	0,851	1,373	1,269	0,999	1,485	0,636	1,599	0,230	1,500	-0,60
O_8	2,647	0,041	2,502	-0,867	2,192	-1,485	1,733	-2,002	0,500	-2,60
O_9	0,157	-2,566	-0,730	-2,465	-1,343	-2,192	-1,865	-1,769	-2,500	-0,60
O_{10}	-0,481	-0,423	-0,596	-0,233	-0,636	-0,071	-0,633	0,096	-0,500	0,400
Varianța	6,527	5,795	9,396	2,926	10,606	1,716	10,624	1,698	7,389	4,933
Varianța totală	12,322		12,322		12,322		12,322		12,322	

Este evident că rotația axelor cu 90 de grade, determină interschimbarea valorilor observațiilor celor două variabile. În mod corespunzător, are loc și interschimbarea valorilor varianțelor celor două variabile.

Pentru a ilustra modificarea valorilor pe care le iau varianțele, în tabelul următor sunt prezentate varianțele individuale ale celor două variabile, varianța totală și ponderile varianțelor individuale în varianța totală, pentru coordonatele calculate corespunzător unor rotații ale axelor inițiale din 5 în 5 grade.

Schimbarea varianțelor individuale pentru diferite unghiuri de rotație a axelor

Tabelul 6.3

Unghi de rotație θ	Varianța individuală		Varianță totală	Procent față de varianță totală (%)	
	x_1	x_2		x_1	x_2
0	4,9333	7,3888	12,3221	46,45	53,55
5	5,7237	6,5984	12,3221	52,97	47,03
10	6,5274	5,7947	12,3221	59,41	40,59

Unghi de rotație θ	Varianța individuală		Varianță totală	Procent față de varianță totală (%)	
	x_1	x_2		x_1	x_2
15	7,3200	5,0021	12,3221	65,55	34,45
20	8,0774	4,2448	12,3221	71,23	28,77
25	8,7765	3,5456	12,3221	76,25	23,75
30	9,3962	2,9259	12,3221	80,49	19,51
35	9,9175	2,4046	12,3221	83,79	16,21
40	10,3248	1,9973	12,3221	86,07	13,93
45	10,6055	1,7166	12,3221	87,25	12,75
50	10,7512	1,5709	12,3221	87,30	12,70
52,7214	10,7720	1,5501	12,3221	87,42	12,58
55	10,7574	1,5647	12,3221	86,22	13,78
60	10,6240	1,6982	12,3221	84,03	15,97
65	10,3549	1,9672	12,3221	80,82	19,18
70	9,9584	2,3637	12,3221	76,66	23,34
75	9,4466	2,8756	12,3221	71,70	28,30
80	8,8349	3,4872	12,3221	66,08	33,92
85	8,1420	4,1802	12,3221	59,96	40,04
90	7,3888	4,9333	12,3221	81,59	18,41

În figura următoare este reprezentată varianța variabilei pentru fiecare rotație cu corespunzătoare rotației axelor din 5 în 5 grade.

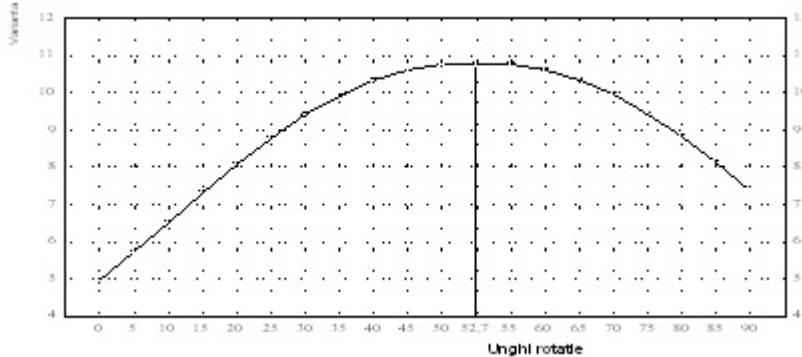


Figura 6.2: Reprezentarea grafică a varianței primei variabile în funcție de mărimea unghiului de rotație a axelor

După cum se poate observa, pe măsură ce unghiul de rotație crește, varianța crește, atingând un maxim pentru un unghi de rotație de 52,7 grade, după care varianța începe să se reducă. Pe de altă parte, pe măsură ce varianța primei variabile crește, varianța celei de-a două variabile scade, astfel încât suma celor două varianțe sau varianța totală rămâne constantă. Similar, scăderea varianței primei variabile este însoțită de creșterea varianței celei de-a două variabile. Aceasta înseamnă că atunci când varianța variabilei x_1 este maximă, varianța variabilei este minimă. În final, se poate spune că, rotația axelor inițiale cu un unghi de 52,7 grade maximizează varianța variabilei x_1 și minimizează varianța variabilei x_2 . Această rotație a axelor cu un unghi de 52,7 grade este chiar transformarea de care este nevoie pentru a maximiza relevanța observațiilor primei variabile.

Exemplul precedent evidențiază faptul că prin rotația axelor cu un anumit număr de grade se poate obține o diferențiere a semnificației variabilelor originale, din punct de vedere al proporției pe care acestea o explică din varianța totală. În aceste condiții, se pune problema de a găsi o rotație optimală a axelor, astfel încât în noul sistem de axe semnificațiile informaționale ale variabilelor să fie cât mai accentuate, problemă care definește în mod sintetic și sugestiv esența logicii componentelor principale.

6.4.2 Transformarea coordonatelor obiectelor prin rotația axelor

Ca urmare a rotației axelor cu un număr oarecare de grade are loc, în mod implicit, o modificare a coordonatelor inițiale ale obiectelor. Pentru ilustrarea modului în care are loc modificarea coordonatelor obiectelor, ca urmare a rotației axelor, vom considera exemplul următor.

Exemplul 6.3

Presupunând contextul informațional din Exemplul 6.1, vom calcula coordonatele ce corespund obiectelor în condițiile în care axele

sunt rotite în aşa fel încât să asigure maximizarea varianței primei variabile. Coordonatele noilor punctelor, reprezentate de observațiile celor două variabile în sistemul de axe rotit cu un unghi de 52,7 grade, sunt prezentate în tabelul următor.

Coordonatele observațiilor în sistemul de axe rotit cu 52,7 grade
Tabelul 6.4

Observația	Coordonate transformate	
	x_1	x_2
O_1	-0,344	-7,803
O_2	-4,502	14,168
O_3	57,611	-1,388
O_4	-70,414	1,696
O_5	-12,458	8,111
O_6	17,469	18,325
O_7	15,569	4,312
O_8	19,727	-17,659
O_9	-16,258	-19,917
O_{10}	-6,401	0,154
Media	0	0
Varianța	10,7720	1,5501

Dacă vom presupune că cele două serii de observații din tabel, rezultate în urma transformării reprezentate de rotația axelor cu un unghi de 52,7 grade, sunt observațiile corespunzătoare unor noi variabile, notate cu w_1 și w_2 , vom putea spune că *am definit două noi variabile, care sunt necorelate și care au varianțele în aşa fel încât varianța primei variabile este maximală*. Aceste noi variabile sunt chiar *componentele principale*, iar observațiile corespunzătoare acestora și definite de relațiile:

$$\begin{aligned} w_{t1}' &= \cos 52,7^\circ \cdot x_{t1} + \sin 52,7^\circ \cdot x_{t2}, \\ w_{t2}' &= -\sin 52,7^\circ \cdot x_{t1} + \cos 52,7^\circ \cdot x_{t2}, \end{aligned}$$

reprezintă scorurile componentelor principale.

Prima componentă principală are varianța 10,772, iar cea de-a doua componentă principală are varianța 1,5501. În acest fel, prima componentă principală preia 87,42% din varianța totală și deci sintetizează, explică într-o proporție suficient de mare cele două variabile originale. Matricea produselor încrucișate, matricea de covarianță și matricea de corelație pentru cele două componente principale sunt:

$$C_w = \begin{pmatrix} 96,9482 & 0,0000 \\ 0,0000 & 13,9518 \end{pmatrix} \quad S_w = \begin{pmatrix} 10,772 & 0,000 \\ 0,000 & 1,550 \end{pmatrix} \quad R_w = \begin{pmatrix} 1,000 & 0,000 \\ 0,000 & 1,000 \end{pmatrix}.$$

În consecință, în locul observațiilor ce corespund celor două variabile originale, în analiză pot fi utilizate observațiile corespunzătoare primei componente principale, adică scorurile acesteia, în condițiile unei pierderi informaționale de 12,78%.

Axele noului spațiu, rezultate din rotația axelor inițiale cu 52,7 grade, sunt definite, în mod similar, de următorii doi vectori:

$$\begin{aligned} j_1 &= \cos 52,7^\circ \cdot i_1 + \sin 52,7^\circ \cdot i_2, & \text{respectiv: } j_1 = \begin{pmatrix} 0,605691 \\ 0,795700 \end{pmatrix}; \\ j_2 &= -\sin 52,7^\circ \cdot i_1 + \cos 52,7^\circ \cdot i_2, & j_2 = \begin{pmatrix} -0,795700 \\ 0,605691 \end{pmatrix}. \end{aligned}$$

Cei doi vectori, j_1 și j_2 , care definesc noile axe, sunt vectori de lungime unitară și sunt ortogonali, adică verifică următoarele relații:

$$\begin{aligned} j_1^t \cdot j_1 &= 0,605691 \cdot 0,605691 + 0,795700 \cdot 0,795700 = 1,000000 \\ j_2^t \cdot j_2 &= (-0,795700) \cdot (-0,795700) + 0,605691 \cdot 0,605691 = 1,000000. \\ j_1^t \cdot j_2 &= 0,605691 \cdot (-0,795700) + 0,795700 \cdot 0,605691 = 0,000000 \end{aligned}$$

Unghiul dintre vectorul j_1 , reprezentând prima axă nouă, și vectorul i_1 , reprezentând prima din axele inițiale, reprezintă chiar unghiul de rotație a axelor, respectiv 52,72 grade. În virtutea ortogonalității, aceeași valoare o are și unghiul dintre vectorul j_2 și vectorul i_2 . Cosinusurile unghiurilor dintre vectorii fiecareia dintre cele cele două perechi vor fi:

$$\begin{aligned} \cos \theta &= \frac{j_1^t \cdot i_1}{\|j_1\| \cdot \|i_1\|} = (0,605691 \quad 0,795700) \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 0,605691; \\ &= \frac{j_2^t \cdot i_2}{\|j_2\| \cdot \|i_2\|} = (-0,795700 \quad 0,605691) \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0,605691, \end{aligned}$$

iar acestora le vor corespunde același unghi, de 52,72 grade:

$$\arccos(0,605691) = 52,72 \text{ grade.}$$

În graficul din figura următoare este să se reprezinte atât axele inițiale, cât și noile axe, adică axele rotite cu 52,72 grade. Coordonatele observațiilor în noile axe, adică valorile din tabelul anterior, reprezintă proiecțiile punctelor pe noile axe ale spațiului.

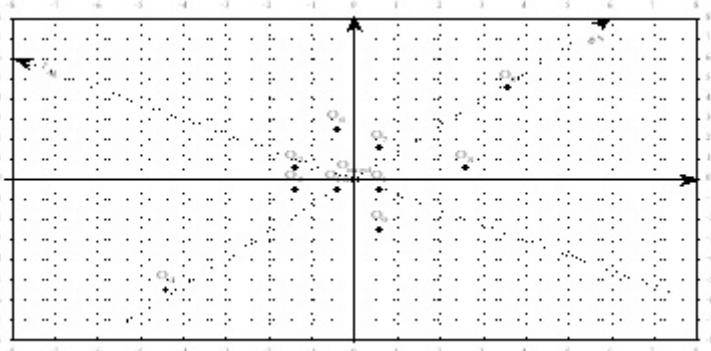


Figura 6.3: Reprezentarea grafică a observațiilor în sistemul de axe rotite cu 52,72 grade

Referitor la exemplificarea anterioară cu privire la logica determinării componentelor principale, putem face următoarele observații finale:

- rotația axelor inițiale cu un anumit unghi, în vederea maximizării varianței de-a lungul unei axe, nu modifică poziția sau configurația punctelor reprezentate de observațiile originale; ceea ce se schimbă sunt doar coordonatele acestora în raport cu noile axe;
- noile axe rezultate în urma rotației care maximizează varianța după prima axă definesc două noi variabile, numite **componente principale** și care au proprietatea că *au media nulă*;
- cele două variabilele numite **componente principale** reprezintă combinații liniare de variabilele originale și sunt **necorelate** între ele;
- coordonatele noilor variabile sunt proiecțiile punctelor reprezentate de observațiile inițiale și se numesc **scoruri ale componentelor principale**;
- cele două componente principale *conservează integral varianța totală* corespunzătoare variabilelor originale, adică *suma varianțelor celor două componente principale este egală cu suma varianțelor celor două variabile originale*;
- prima componentă principală are **varianță maximală**, preluând maximum posibil din varianța totală ce caracterizează variabilele originale.

6.4.3 Corelarea variabilelor și componentelor principale

Eficiența exprimării variabilelor originale prin intermediul componentelor principale este strâns legată de *gradul de corelare* a variabilelor originale și, mai ales, de *felul în care sunt structurate* aceste variabile din punct de vedere al corelării.

În legătură cu gradul de corelare a variabilelor originale, se poate face o observație extrem de interesantă din punct de vedere teoretic și foarte utilă din punct de vedere practic. Această observație se referă la faptul că *există o puternică legătură între gradul de corelare a variabilelor originale și numărul de componente principale* cu ajutorul cărora pot fi reexprimate, în mod eficient, variabilele originale.

Dacă pe mulțimea variabilelor originale se evidențiază cu claritate existența unor *submulțimi* formate din variabile care au proprietatea că sunt *foarte puternic corelate între ele*, pe de o parte, și *foarte slab corelate cu variabile aparținând altor submulțimi*, pe de altă parte, atunci se poate face afirmația că *variabilele originale pot fi reexprimate suficient de bine printr-un număr de componente principale egal cu numărul de submulțimi* de acest fel.

Remarcă: Se poate spune că, în general, numărul de componente principale este egal cu numărul grupelor de variabile care sunt foarte puternic corelate între ele.

Cu cât gradul de corelare a variabilelor din interiorul fiecărei submulțimi este mai ridicat și al variabilelor din mulțimi diferite este mai scăzut, cu atât este mai mare puterea de reprezentare a informației inițiale prin intermediul unui număr de componente principale egal cu numărul submulțimilor.

De pildă, dacă gradul de corelare a oricărora două variabile din mulțimea de variabile originale este foarte scăzut, există puține șanse de reuși să se exprime convenabil variabilele originale printr-un număr componentă principale mai mic decât al variabilelor originale. Dacă, dimpotrivă, corelarea oricărora două variabile este foarte puternică, atunci variabilele originale pot fi exprimate suficient de bine chiar prin intermediul unei singure componente principale.

Vom ilustra legătura directă care există între gradul de corelare al variabilelor și numărul de componente principale cu ajutorul următorului exemplu.

Exemplul 6.4

În cadrul acestui exemplu vom arăta cum, în funcție de diferite moduri în care se coreleză variabilele originale, rezultă diferite posibilități de reprezentare a acestora prin intermediul componentelor principale. Pentru început, vom considera cazul unui număr de 5 variabile originale, care au proprietatea că sunt foarte slab corelate între ele. Matricea de corelație pentru aceste variabile, matrice care evidențiază necorelarea variabilelor, este următoarea:

$$R = \begin{pmatrix} 1,000000 & -0,128307 & -0,023963 & -0,083163 & -0,018717 \\ -0,128307 & 1,000000 & 0,081026 & 0,124651 & -0,052883 \\ -0,023963 & 0,081026 & 1,000000 & 0,012432 & -0,024491 \\ -0,083163 & 0,124652 & 0,012432 & 1,000000 & 0,076912 \\ -0,018717 & -0,052884 & -0,024491 & 0,076912 & 1,000000 \end{pmatrix}.$$

Valorile proprii pentru această matrice de corelație și semnificațiile lor relative se găsesc în tabelul următor.

Informații cu privire la valorile proprii ale matricii de corelație

Tabelul 6.5

Nr. Crt.	Valori proprii	Pondere (%)	Pondere cumulată (%)
1	1,292867	25,857	25,857
2	1,112724	22,255	48,112
3	0,965477	19,309	67,421
4	0,893816	17,877	85,298
5	0,735115	14,702	100,000

Analiza valorilor proprii prezentate în tabel arată că exprimarea prin intermediul componentelor principale a celor 5 variabile originale poate fi satisfăcătoare numai dacă se folosesc 4 sau chiar 5 componente principale. Acest lucru se întâmplă deoarece chiar contribuția ultimelor componente principale la explicarea varianței totale este relativ ridicată, astfel încât renunțarea la ele este echivalentă cu o pierdere semnificativă de informație.

Cazul limită pentru situația de necorelare a variabilelor originale este cel în care matricea de corelație este egală cu matricea unitate. În acest caz, toate valorile proprii sunt egale cu unitatea, iar contribuția fiecărei la formarea varianței totale este de 20%, adică un procent relativ ridicat. Rezultă că pentru a asigura conservarea varianței totale în proporție de peste 80% este necesară luarea în considerare a tuturor componentelor principale.

Vom considera acum o a doua situație, opusă celei precedente, și anume aceea în care toate variabilele originale sunt puternic corelate. Matricea de corelație presupusă pentru acest caz este:

$$R = \begin{pmatrix} 1,00 & 0,95 & 0,98 & 0,95 & 0,97 \\ 0,95 & 1,00 & 0,96 & 0,98 & 0,95 \\ 0,98 & 0,96 & 1,00 & 0,94 & 0,93 \\ 0,95 & 0,98 & 0,94 & 1,00 & 0,96 \\ 0,97 & 0,95 & 0,93 & 0,96 & 1,00 \end{pmatrix}.$$

Tabelul următor conține situația celor 5 valori proprii corespunzătoare matricii de corelație considerate.

Informații cu privire la valorile proprii ale matricii de corelație

Tabelul 6.6

Nr. Crt.	Valori proprii	Pondere (%)	Pondere cumulată (%)
1	4,828053	96,561	96,561
2	0,082870	1,657	98,218
3	0,067152	1,343	99,561
4	0,017526	0,351	99,912
5	0,004399	0,088	100,000

În acest caz, o singură componentă principală, și anume prima, poate fi folosită pentru a sintetiza din punct de vedere informațional toate cele 5 variabile originale. Prin intermediul acestei componente principale se asigură conservarea a peste 96% din variabilitatea conținută în spațiul cauzal inițial, determinat de cele 5 variabile.

Și în această situație există un caz limită, anume acela în care matricea de corelație are toate elementele egale cu unitatea, expresie a unei corelații perfecte a oricărora două variabile originale. În acest caz, prima valoare proprie este egală cu 5,0, iar celelalte patru valori proprii sunt nule. Aceasta înseamnă că prima componentă principală asigură captarea întregii variabilități individuale, varianța ei fiind egală cu varianța totală din spațiul cauzal inițial.

Rezultă că prima componentă principală exprimă din punct de vedere informațional toate cele 5 variabile originale, fără pierdere de informație. Acest lucru este natural deoarece corelarea perfectă a oricărora două variabile originale nu înseamnă altceva decât că aceste variabile sunt identice.

În sfârșit, o a treia situație pe care vrem să o luăm în considerare este aceea în care pe mulțimea variabilelor originale se evidențiază grupe de variabile puternic corelate. Vom considera cazul a 6 variabile originale, pe mulțimea căror se evidențiază trei grupe de variabile puternic corelate. Matricea de corelație care exprimă o astfel de situație este următoarea:

$$R = \begin{pmatrix} 1,000 & 0,040 & 0,895 & 0,110 & 0,005 & 0,150 \\ 0,040 & 1,000 & 0,100 & 0,045 & 0,155 & 0,975 \\ 0,895 & 0,100 & 1,000 & 0,120 & 0,100 & 0,070 \\ 0,110 & 0,045 & 0,120 & 1,000 & 0,950 & 0,160 \\ 0,005 & 0,155 & 0,100 & 0,950 & 1,000 & 0,035 \\ 0,150 & 0,975 & 0,070 & 0,160 & 0,035 & 1,000 \end{pmatrix}.$$

Din analiza matricii de corelație se poate observa că există trei grupe de variabile corelate puternic: x_1 și x_3 , x_2 și x_6 , respectiv x_4 și x_5 . Pe de altă parte, între cele trei grupe există legături foarte slabe. Aceasta înseamnă că pentru a exprima cele 6 variabile sunt suficiente, practic, doar trei componente principale, care vor acoperi variabilitatea din spațiul celor 6 variabile într-o proporție foarte mare. În tabelul următor sunt prezentate cele 6 valori proprii ale acestei matrice de corelație și informațiile privind ponderea fiecărei valori proprii.

Informații cu privire la valorile proprii ale matricii de corelație

Tabelul 6.7

Nr. Crt.	Valori proprii	Pondere (%)	Pondere cumulată (%)
1	2,308133	38,469	38,469
2	1,766032	29,434	67,903
3	1,750709	29,178	97,071
4	0,132365	2,206	99,287
5	0,042761	0,713	100,000
6	0,000000	100,000	100,000

Într-adevăr valorile proprii și ponderea lor în varianța totală arată că primele trei componente principale preiau din variabilitatea spațiului inițial 97,071%. Fiecare dintre aceste trei componente principale exprimă variabilele originale din fiecare cele trei grupe de variabile corelate. Astfel, se justifică ideea că *numărul de componente principale este egal cu numărul de submulțimi de variabile originale puternic corelate*.

6.5 Modelul matematic al componentelor principale

În scopul formulării modelului matematic care să la baza analizei componentelor principale, vom considera că spațiul cauzal inițial supus investigării este determinat de un număr de n variabile explicative noteate x_1, x_2, \dots, x_n . Aceste variabile simbolizează caracteristici ale obiectelor supuse analizei, ceea ce înseamnă că fiecare obiect este presupus a fi caracterizat de n variabile.

Activitatea de determinare a componentelor principale poate fi descrisă prin intermediul unei transformări de tipul următor:

$$\Psi : \mathbb{R}^n \rightarrow \mathbb{R}^k,$$

unde \mathbb{R}^n și \mathbb{R}^k sunt două spații vectoriale reale, iar dimensiunea celui de-al doilea spațiu este mult mai mică decât dimensiunea primului spațiu, respectiv $k < n$.

Prin intermediul transformării Ψ , un anumit obiect x , aparținând spațiului n -dimensional \mathbb{R}^n , este transformat într-un obiect w , aparținând spațiului k -dimensional \mathbb{R}^k . Transformarea vizează atât modificarea coordonatelor obiectului, cât și reducerea numărului acestor coordonate.

Vom arăta în continuare că, dacă $x \in \mathbb{R}^n$ și $w \in \mathbb{R}^k$, atunci transformarea Ψ este o aplicație liniară de tipul următor:

$$w = A^t \cdot x,$$

unde A este o matrice de numere reale, de dimensiune $n \times k$.

Rezolvarea problemei constă în determinarea matricii A , astfel încât un obiect w să constituie o reprezentare *cât mai bună* pentru obiectul x .

6.5.1 Relațiile de definire a componentelor principale

Am anticipat anterior că, de fapt, componentele principale sunt combinații liniare de variabilele originale, combinații care au însă o serie de proprietăți. În condițiile ipotezelor și precizărilor anterioare, putem privi cele n componente principale, corespunzătoare spațiului cauzal analizat, sub forma unui vector n -dimensional, notat cu w :

$$w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}.$$

Fiecare coordonată w_i a acestui vector reprezintă o componentă principală definită în raport cu variabilele originale, cu ajutorul combinației liniare următoare:

$$w_i = \alpha_1^{(i)} \cdot x_1 + \alpha_2^{(i)} \cdot x_2 + \dots + \alpha_n^{(i)} \cdot x_n \quad i=1,2,\dots,n \quad (6.1)$$

Este evident că pentru determinarea componentei principale w_i este necesară determinarea coeficienților $\alpha_j^{(i)}$, care definesc combinația liniară corespunzătoare acestei componente principale.

Așa cum o să vedem în continuare, coeficienții $\alpha_j^{(i)}$ sunt chiar *coordonatele vectorilor proprii* corespunzători matricii de covarianță a variabilelor originale x_1, x_2, \dots, x_n , iar varianțele componentelor principale sunt chiar *valorile proprii* ale acestei matrici.

Conform definiției componentelor principale, determinarea acestor coeficienți trebuie să se facă în așa fel încât componenta principală w_i să aibă varianță maximă.

Considerând că cei n coeficienți $\alpha_1^{(i)}, \alpha_2^{(i)}, \dots, \alpha_n^{(i)}$ ai combinației liniare de mai sus sunt coordinatele vectorului n-dimensional $\alpha^{(i)}$, respectiv:

$$\alpha^{(i)} = \begin{pmatrix} \alpha_1^{(i)} \\ \alpha_2^{(i)} \\ \vdots \\ \alpha_n^{(i)} \end{pmatrix} \quad i=1,2,\dots,n,$$

putem defini componenta principală w_i sub forma următoare:

$$w_i = (\alpha^{(i)})^t \cdot x \quad i=1,2,\dots,n \quad (6.2)$$

unde coordinatele vectorului $\alpha^{(i)}$ sunt alese astfel încât să se asigure *maximizarea varianței* componentei principale w_i .

Deoarece ceea de-a i-a componentă principală, w_i , este în realitate o *transformare liniară* a elementelor vectorului x , presupus a fi repartizat normal, de medie μ și matrice de covarianță Σ , rezultă că această componentă principală este ea însăși, de asemenea, o variabilă aleatoare, repartizată normal.

Pe baza relației (6.2), care definește componenta principală w_i , pot fi deduse media și varianța acestei componente principale astfel:

$$E(w_i) = E[(\alpha^{(i)})^t \cdot x] = (\alpha^{(i)})^t \cdot \mu \quad \text{Var}(w_i) = (\alpha^{(i)})^t \cdot \Sigma \cdot \alpha^{(i)}.$$

Rezultă că:

$$w_i \sim N[(\alpha^{(i)})^t \cdot \mu, (\alpha^{(i)})^t \cdot \Sigma \cdot \alpha^{(i)}] \quad i=1,2,\dots,n.$$

Având făcute precizările de mai sus, vom trece în continuare la descrierea modelului matematic pe care se bazează analiza componentelor principale.

6.5.2 Formularea modelului matematic

Am arătat anterior că soluționarea problemei de analiză a componentelor principale din punct de vedere matematic este echivalentă cu rezolvarea problemei de extrem următoare:

$$\begin{cases} \text{opt} & \phi(x, w) \\ A \in M_{n \times k} \\ \text{SR: } w = A^t \cdot x \end{cases},$$

unde criteriul de optim poate fi *maxim* sau *minim*, în funcție de natura funcției ϕ . Dacă funcția ϕ este o funcție de tip distanță, atunci criteriul de optim va fi reprezentat de *minimizarea* funcției ϕ . În cazul în care funcția ϕ este o măsură a cantității de informație adusă de noua modalitate de reprezentare a obiectelor, criteriul de optim va fi reprezentat de *maximizarea* funcției ϕ . O astfel de situație este specifică variantei standard de soluționare a problemei componentelor principale, în care se urmărește maximizarea varianței componentelor principale, ca măsură a cantității de informație exprimată de fiecare dintre acestea.

În scopul definirii modelului matematic al analizei componentelor principale, vom considera că vectorii $\alpha^{(i)}$ reprezintă coloanele unei matrici A de dimensiune $n \times n$ de forma:

$$A = \begin{pmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} & \dots & \alpha_1^{(n)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} & \dots & \alpha_2^{(n)} \\ \vdots & & & \\ \alpha_n^{(1)} & \alpha_n^{(2)} & \dots & \alpha_n^{(n)} \end{pmatrix}.$$

De asemenea, vom presupune că x este vectorul ale cărui coordonate sunt variabilele originale x_1, x_2, \dots, x_n și că w este

vectorul ale cărui coordonate sunt componentele principale w_1, w_2, \dots, w_n . În aceste condiții, combinațiile liniare care definesc componentele principale pot fi scrise sub forma:

$$\begin{cases} w_1 = \alpha_1^{(1)}x_1 + \alpha_2^{(1)}x_2 + \dots + \alpha_n^{(1)}x_n \\ w_2 = \alpha_1^{(2)}x_1 + \alpha_2^{(2)}x_2 + \dots + \alpha_n^{(2)}x_n \\ \dots \\ w_n = \alpha_1^{(n)}x_1 + \alpha_2^{(n)}x_2 + \dots + \alpha_n^{(n)}x_n \end{cases},$$

sau, în scriere matricială, sub forma:

$$\begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{pmatrix} = \begin{pmatrix} \alpha_1^{(1)} & \alpha_2^{(1)} & \dots & \alpha_n^{(1)} \\ \alpha_1^{(2)} & \alpha_2^{(2)} & \dots & \alpha_n^{(2)} \\ \dots \\ \alpha_1^{(n)} & \alpha_2^{(n)} & \dots & \alpha_n^{(n)} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}.$$

Pe baza acestor notații, *modelul matematic al analizei componentelor principale* poate fi definit astfel:

$$\begin{cases} \max_{A \in M_{n \times n}} \text{Var}(w) \\ w = A^t \cdot x \end{cases} \quad (6.3)$$

Așa cum vom vedea în continuare, cele n coloane ale matricii A reprezintă de fapt *vectorii proprii normalizați* ai matricii de covarianță Σ , iar varianța fiecărei componente principale w_i , care este o varianță maximală în raport cu varianțele componentelor principale anterioare, este reprezentată chiar de valoarea proprie λ_i a aceleiași matrici de covarianță. Această modalitate de determinare a elementelor matricii A este echivalentă cu calculul proiecțiilor obiectelor de tip $x \in \mathbb{R}^n$ pe subspațiul liniar generat de vectorii coloanelor matricii A.

Am văzut anterior că cele n componente principale ale spațiului cauzal determinat de variabilele originale x_1, x_2, \dots, x_n , sunt definite de combinațiile liniare:

$$w_i = \alpha_1^{(i)} \cdot x_1 + \alpha_2^{(i)} \cdot x_2 + \dots + \alpha_n^{(i)} \cdot x_n \quad i=1,2,\dots,n,$$

ale căror ponderi $\alpha_j^{(i)}$ se determină în aşa fel încât să maximizeze varianța componentelor principale w_i .

În scopul simplificării notațiilor, vom renunța, temporar, la unii dintre indicii care apar în relații. Astfel, vom considera în continuare că w este notația generică pentru o anumită componentă principală, iar α este notația generică pentru vectorul coeficienților ce definesc combinația liniară pentru această componentă principală.

În acest sens, vom avea grija să specificăm explicit, la fiecare apariție a notației w, dacă este vorba de vectorul w sau de componentă principală w și să menționăm explicit indicele componentei principale atunci când o privim ca pe un element al vectorului componentelor principale w.

6.5.3 Formularea problemei de maximizare a varianței

Din punct de vedere teoretic, problema determinării componentelor principale poate fi formulată ca o problemă de maximizare cu restricții, iar elementele implicate în definirea componentelor principale sunt soluțiile acestei probleme. Criteriul de optim atașat acestei probleme de extrem este reprezentat de maximizarea varianței componentelor principale, astfel încât acestea să preia o proporție cât mai mare din variabilitatea spațiului cauzal inițial. În condițiile folosirii notației simplificate, menționată anterior, varianța componentei principale w poate fi scrisă astfel:

$$\text{Var}(w) = \alpha^t \cdot \Sigma \cdot \alpha.$$

Principala problemă care se pune în legătură cu definirea componentelor principale este aceea a determinării necunoscutelor reprezentate de coordonatele vectorului α astfel încât varianța componentei principale w, adică $\text{Var}(w)$, să fie maximă. Din nefericire, problema determinării acestor necunoscute este o problemă nedeterminată. Astfel, admitând că vectorul α este o soluție a problemei, rezultă că și produsul $\alpha\alpha$, unde α este o constantă arbitrară, este o soluție a aceleiași probleme, iar varianța componentei principale w este cu atât mai mare cu cât luăm o valoare mai mare pentru constanta α .

Pentru a avea asigurarea că problema enunțată este bine determinată, va trebui să impunem o restricție asupra vectorului α necunoscut α , cea mai naturală restricție fiind dată, în acest caz, de condiția ca lungimea vectorului α să fie unitară:

$$\alpha^t \cdot \alpha = 1.$$

Verificarea acestei restricții în definirea componentelor principale este echivalentă cu *fixarea scalei coeficienților* și oferă asigurarea că efectuarea oricărei operații de scalare asupra coeficienților combinațiilor liniare care definesc componentele principale, nu determină modificarea varianțelor acestora.

În aceste condiții, problema determinării componentelor vectorului α ce definesc combinația liniară reprezentând componenta principală w se reduce la rezolvarea următoarei probleme de extrem cu legături:

$$\begin{cases} \max_{\alpha} \alpha^t \cdot \Sigma \cdot \alpha \\ \text{SR: } \alpha^t \cdot \alpha = 1 \end{cases} \quad (6.4)$$

unde variabilele de alegere sau necunoscutele problemei sunt reprezentate de componentelete vectorului α . Vectorul $\tilde{\alpha}$, care este soluție a acestei probleme, definește o componentă principală de varianță maximală.

Rezolvarea problemei de extrem condiționat precedente poate fi făcută cu ajutorul metodei multiplicatorilor lui Lagrange, metodă care reduce problema de extrem condiționat la o problemă de extrem liber. Funcția Lagrangean asociată problemei de extrem (6.4), folosită pentru rezolvarea indirectă a acestei probleme de extrem, are forma următoare:

$$L(\alpha, \lambda) = \alpha^t \cdot \Sigma \cdot \alpha - \lambda(\alpha^t \cdot \alpha - 1) \quad (6.5)$$

Condițiile necesare de extrem pentru funcția Lagrangean definită de relația (6.5) sunt date de anularea derivatelor parțiale în raport cu componentelete vectorului α și cu multiplicatorul λ , respectiv:

$$\begin{cases} \frac{\partial L(\alpha, \lambda)}{\partial \alpha} = 0 \\ \frac{\partial L(\alpha, \lambda)}{\partial \lambda} = 0 \end{cases} \quad \text{adică:} \quad \begin{cases} 2 \cdot \Sigma \cdot \alpha - 2 \cdot \lambda \cdot \alpha = 0 \\ \alpha^t \cdot \alpha - 1 = 0 \end{cases} \quad (6.6)$$

Din prima condiție necesară de extrem, care mai poate fi scrisă și sub forma următoare:

$$\Sigma \cdot \alpha = \lambda \cdot \alpha.$$

rezultă că soluția $\tilde{\alpha}$ a problemei de extrem (6.4) este chiar unul din vectorii proprii ai matricii de covarianță Σ , anume cel asociat valorii proprii $\tilde{\lambda}$ a aceleiași matrici. Mai mult, se observă că valoarea maximă a formei pătratice $\alpha^t \cdot \Sigma \cdot \alpha$ este, în punctul de extrem $\tilde{\alpha}$, egală cu $\tilde{\lambda}$, respectiv:

$$\text{Var}(\tilde{w}) = \tilde{\alpha}^t \cdot \Sigma \cdot \tilde{\alpha} = \tilde{\lambda}.$$

Această ultimă relație evidențiază faptul că varianța unei componente principale este egală cu o valoare proprie a matricii de covarianță.

6.5.4 Deducerea componentelor principale

Rezultatele obținute mai sus arată că determinarea coeficienților combinației liniare ce definește componenta principală w , în condițiile maximizării varianței acestei componente principale, este echivalentă cu a *alege* dintre cele n valori proprii ale matricii de covarianță Σ pe cea mai mare și a *determina* componentelete vectorului de ponderi α ce definește respectiva componentă principală prin calculul vectorului propriu al matricii Σ asociat cu acea valoare proprie.

Așa cum vom vedea în continuare, pentru fiecare valoare proprie λ_i din cele n valori proprii ale matricii de covarianță Σ , vom avea câte o soluție a problemei de maxim de mai sus, adică câte un vector $\alpha^{(i)}$ și deci câte o componentă principală w_i .

Presupunând că cele n valori proprii ale matricii de covarianță Σ sunt ordonate în aşa fel încât:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n,$$

prima componentă principală w_1 , care va avea varianță maximă λ_1 , este dată de combinația liniară:

$$w_1 = (\alpha^{(1)})^t \cdot x.$$

Vectorul $\alpha^{(1)}$ este acel vector propriu al matricii de covarianță Σ căruia îi corespunde valoarea proprie cea mai mare, λ_1 , adică este vectorul care verifică restricțiile de mai jos:

$$\begin{cases} (\Sigma - \lambda_1 I) \cdot \alpha^{(1)} = 0 \\ (\alpha^{(1)})^t \cdot \alpha^{(1)} = 1 \end{cases}.$$

Valoarea proprie λ_1 este rădăcină a ecuației caracteristice:

$$|\Sigma - \lambda \cdot I| = 0,$$

iar I este notația pentru matricea unitate.

Determinarea în acest fel a componentei principale w_1 , face ca aceasta să aibă proprietăți ilustrate prin relațiile următoare:

$$\bullet E(w_1) = (\alpha^{(1)})^t \cdot \mu \quad \bullet \text{Var}(w_1) = \lambda_1.$$

După determinarea primei componente principale w_1 , urmează determinarea celei de-a doua componente principale w , componentă care trebuie să fie caracterizată, la rândul său, de următoarele proprietăți: să aibă varianță maximală și să fie necorelată cu prima componentă principală w_1 .

Exemplul 6.5

Pentru a ilustra modul de calcul implicat de analiza componentelor principale, vom considera cazul unui număr de 10 obiecte, fiecare obiect având un număr de 5 caracteristici exprimate prin intermediul variabilelor x_1, x_2, x_3, x_4 și x_5 . Măsurările efectuate asupra caracteristicilor celor 10 obiecte sunt presupuse a fi cele din tabelul următor.

Observații inițiale

Tabelul 6.8

Obiecte	Caracteristici				
	x_1	x_2	x_3	x_4	x_5
O_1	3,31	3,02	4,27	5,31	4,24
O_2	2,76	3,94	4,14	8,07	11,08
O_3	10,86	15,16	15,19	6,42	6,55
O_4	9,29	9,71	10,94	8,15	11,34
O_5	8,36	9,12	11,91	7,13	8,33
O_6	12,07	11,12	13,69	8,05	11,01
O_7	4,73	2,99	4,76	18,12	28,21
O_8	8,77	14,13	13,56	6,17	5,99
O_9	18,10	21,00	13,67	11,16	23,39
O_{10}	4,58	8,83	7,43	14,84	16,97
Medie	8,283	9,902	9,957	9,342	12,711
Abatere standard	4,716	5,798	4,378	4,150	7,845

Estimăția pentru matricea de covarianță corespunzătoare celor 5 variabile originale este următoarea:

$$\Sigma = \begin{pmatrix} 22,24 & 24,95 & 17,24 & -2,44 & 6,42 \\ 24,95 & 33,62 & 22,14 & -4,38 & 1,73 \\ 17,24 & 22,14 & 19,17 & -6,61 & -7,42 \\ -2,44 & -4,38 & -6,61 & 17,22 & 29,77 \\ 6,42 & 1,73 & -7,42 & 29,77 & 61,54 \end{pmatrix},$$

iar valorile proprii ale matricii de covarianță sunt:

$$\lambda_1 = 78,625; \quad \lambda_2 = 69,362; \quad \lambda_3 = 3,197; \quad \lambda_4 = 2,385; \quad \lambda_5 = 0,249.$$

Suma celor 5 valori proprii, reprezentând varianțele celor 5 componente principale, este egală cu suma varianțelor variabilelor originale:

$$\sum_{i=1}^5 s_i^2 = 153,79 \approx \sum_{i=1}^5 \lambda_i = 153,82.$$

Vectorii proprii ai matricii de covarianță Σ , corespunzători valorilor proprii menționate anterior sunt prezentate sub forma liniilor matricii următoare:

$$U^t = \begin{pmatrix} -0,070 & -0,167 & -0,237 & 0,448 & 0,842 \\ 0,545 & 0,659 & 0,432 & 0,026 & 0,284 \\ 0,527 & -0,155 & -0,518 & -0,624 & 0,199 \\ 0,357 & -0,714 & 0,599 & -0,001 & 0,057 \\ 0,542 & -0,063 & -0,356 & 0,639 & -0,408 \end{pmatrix}.$$

Cei 5 vectori proprii sunt de lungime egală cu unitatea și sunt ortogonali doi către doi, astfel încât matricea U este *ortogonală*. Deoarece suma varianțelor primelor două componente principale reprezintă 96,623% din varianța celor 5 variabile originale, adică:

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^5 s_i^2} = \frac{147,987}{153,79} = 96,623\%,$$

se poate considera că cele 5 variabile originale pot fi reexprimate suficient de bine din punct de vedere informațional - în limita unei pierderi de 3,4% - prin intermediul primelor două componente principale ale căror ecuații sunt:

$$\begin{aligned} w_1 &= -0,0701 \cdot x_1 - 0,167 \cdot x_2 - 0,237 \cdot x_3 + 0,448 \cdot x_4 + 0,842 \cdot x_5 \\ w_2 &= 0,5448 \cdot x_1 + 0,659 \cdot x_2 + 0,432 \cdot x_3 + 0,026 \cdot x_4 + 0,284 \cdot x_5. \end{aligned}$$

Coordonatele celor 10 obiecte în spațiul redus, ale cărui axe sunt reprezentate de primii doi vectori ai matricii de covarianță Σ , sunt prezentate în tabelul următor:

Scorurile principale ale obiectelor

Tabelul 6.9

Obiectele	Coordonatele obiectelor	
	w_1	w_2
O_1	-54,037	-101,762

Obiectele	Coordonatele obiectelor	
	w_1	w_2
O_2	7,284	-82,893
O_3	-78,067	44,211
O_4	-17,385	3,534
O_5	-44,532	-7,746
O_6	-29,873	33,017
O_7	174,078	-34,196
O_8	-76,982	21,818
O_9	56,611	144,532
O_{10}	62,904	-20,515

În analiza componentelor principale coordonatele obiectelor în spațiul redus se mai numesc și **scoruri principale ale obiectelor**.

Dacă vom presupune că au fost reținute p componente principale și dacă vom nota cu \tilde{U} matricea de dimensiune $n \times p$, ale cărei coloane sunt cei p vectori proprii care definesc cele p componente principale, atunci matricea scorurilor poate fi determinată astfel:

$$W_{p \times T} = \tilde{A}_{p \times n}^T \cdot X_{n \times T}.$$

Liniile matricii W reprezintă *scorurile* corespunzătoare noilor variabile sau *observațiile* celor p componente principale. O dată determinate, scorurile principale pot fi folosite în analiză ca substitut al observațiilor originale, simplificând, în acest fel, baza informațională inițială. În legătură cu această problemă, considerăm că este extrem de important să facem precizarea că *scorurile principale sunt mai potrivite* pentru a fi folosite în analize deoarece sunt mai puțin afectate de erori, în comparație cu măsurările originale. Faptul că scorurile principale sunt mai *robuste* în raport cu perturbațiile introduse de erori, că au o anumită invarianță în raport cu erorile, le face să devină mai importante din punct de vedere informațional decât observațiile originale. Deoarece nou spațiu redus are numai două axe, cele 10 obiecte pot fi reprezentate grafic în acest spațiu. Reprezentarea grafică din figura următoare arată poziționarea celor 10 obiecte în raport cu axele noului spațiu.

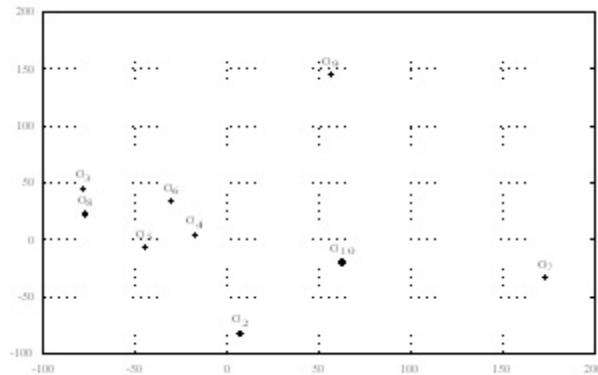


Figura 4.6: Reprezentarea obiectelor în spațiul redus

În condițiile în care obiectele reprezintă entități reale, un astfel de grafic poate servi ca bază eficientă pentru o analiză a mulțimii de obiecte. Poziționarea obiectelor în raport cu cele două axe oferă o primă imagine cu privire la legăturile dintre obiecte, evidențierind cu claritate similaritățile sau nesimilaritățile dintre acestea.

6.6 Proprietățile componentelor principale

Componentele principale au o serie de proprietăți extrem de interesante, care decurg chiar din modul lor de definire și care sunt importante pentru a înțelege natura și conținutul acestor construcții abstrakte.

Una dintre proprietățile menționate se referă la faptul că *varianța fiecărei componente principale este maximă și este egală cu o valoare proprie a matricii de covarianță*. Altă proprietate a componentelor principale este aceea că ele sunt *necorelate două către două*, această proprietate fiind echivalentă, în cazul în care componentele principale sunt distribuite după legea de probabilitate normală, cu proprietatea de independență.

În afara acestor proprietăți, implicate de însăși modul lor de definire, componentele principale au o altă serie de proprietăți deosebit de importante pentru modelarea matematică, în general, și pentru analiza economică, în special. În continuare, vom prezenta pe scurt fiecare dintre aceste proprietăți.

6.6.1 Distribuirea după legea normală

În condițiile în care variabilele originale sunt repartizate normal, *vectorul componentelor principale w este repartizat*

normal cu media $\mathbf{A}^t \boldsymbol{\mu}$ și matricea de covarianță Λ , adică:

$$\mathbf{w} \sim N(\mathbf{A}^t \boldsymbol{\mu}, \Lambda),$$

unde Λ este matricea diagonală ale cărei elemente sunt valorile proprii $\lambda_1, \lambda_2, \dots, \lambda_n$ ale matricii de covarianță Σ .

Normalitatea celor n variabile reprezentând componentele principale rezultă din faptul că acestea sunt combinații liniare de cele n variabile originale, care, prin ipoteză, sunt variabile normale. Pentru a arăta că matricea de covarianță a vectorului w este matricea Λ este suficient să arătăm că dacă:

$$\mathbf{w} = \mathbf{A}^t \mathbf{x},$$

x fiind repartizat normal, cu matricea de covarianță Σ , atunci matricea de covarianță a transformării liniare w este:

$$\text{Cov}(\mathbf{w}) = \mathbf{A}^t \Sigma \mathbf{A}.$$

6.6.2 Conservarea varianței totale

Componentele principale au o proprietate care le face să fie adecvate din punct de vedere informațional pentru a substitui variabilele originale. Această proprietate se referă la faptul că prin intermediul componentelor principale se asigură *conservarea* variabilității din spațiul cauzal inițial.

Componentele principale $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ asigură *conservarea integrală a varianței totale a variabilelor originale* $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, ceea ce înseamnă că:

$$\sum_{i=1}^n \text{Var}(\mathbf{x}_i) = \sum_{i=1}^n \text{Var}(\mathbf{w}_i) \quad \text{deci:} \quad \sum_{i=1}^n \sigma_{ii} = \sum_{i=1}^n \lambda_i.$$

Exemplul 6.6

Vom presupune că estimarea pentru o matrice de covarianță corespunzătoare observațiilor efectuate asupra a 3 variabile este de forma următoare:

$$\hat{\Sigma} = \begin{pmatrix} 54,131 & 13,304 & 3,106 \\ 13,304 & 7,496 & 0,690 \\ 3,106 & 0,690 & 0,238 \end{pmatrix}.$$

Elementele diagonale ale acestei matrici reprezintă varianțele corespunzătoare celor trei variabile originale, respectiv:

$$\sigma_1^2 = 54,131; \sigma_2^2 = 7,496; \sigma_3^2 = 0,238.$$

Cele trei valori proprii corespunzătoare acestei matrici de covarianță au valorile următoare:

$$\lambda_1 = 57,835; \lambda_2 = 3,972; \lambda_3 = 0,058.$$

Așa cum se poate observa imediat, este verificată proprietatea menționată anterior, respectiv:

$$\text{tr}(\hat{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 = 61,865 = \lambda_1 + \lambda_2 + \lambda_3 = \text{tr}(\hat{\Lambda}).$$

6.6.3 Conservarea varianței generalizate

Componentele componente principale $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ asigură *conservarea integrală a varianței generalizate* a variabilelor originale $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. Aceasta înseamnă că:

$$\mathbf{V}_G(\mathbf{x}) = \mathbf{V}_G(\mathbf{w}).$$

Acastă proprietatea evidențiază calitatea informațională pe care o au componentele principale de a reprezenta o reexprimare a variabilelor originale.

Exemplul 6.7

Considerând matricea de covarianță din exemplul precedent, se poate observa că determinantul acestei matrici este egal cu produsul celor trei valori proprii, respectiv este egal cu determinantul matricii de covarianță ce coreponde celor trei componente principale:

$$|\Sigma| = 1,34424396 = 57,835 \cdot 3,972 \cdot 0,058 = |\Lambda|.$$

6.6.4 Dependența de unitățile de măsură

Componentele principale $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ și varianțele acestora *depind de unitățile de măsură* în care sunt măsurate variabilele originale $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. Aceasta înseamnă că, odată cu schimbarea unităților de măsură ale variabilelor originale se schimbă atât componentele principale, cât și varianțele acestora.

Cunoașterea proprietăților pe care le au componentele principale este deosebit de importantă în procesul de analiză a datelor, permitând stabilirea modificărilor induse asupra componentelor principale și asupra mărimilor asociate acestora de către transformările aplicate asupra observațiilor variabilelor originale.

6.7 Matricea factor

O matrice importantă utilizată în contextul analizei componentelor principale, ale cărei elemente oferă premize pentru interpretări interesante, este matricea factor, pe care o vom defini în continuare.

În acest scop, vom presupune că cele n componente principale sunt reprezentate prin intermediul vectorului w, iar matricea de covarianță a componentelor principale este matricea diagonală Λ . De asemenea, vom considera legătura dintre vectorul variabilelor originale și vectorul componentelor principale ca fiind dată de relația:

$$\mathbf{x} = \mathbf{A} \cdot \mathbf{w},$$

unde A este matricea vectorilor proprii ai matricii de covarianță Σ . Atunci matricea de covarianță dintre vectorul x al variabilelor originale și vectorul w al componentelor principale poate fi definită sub forma:

$$\text{Cov}(\mathbf{x}, \mathbf{w}) = E(\mathbf{x} \cdot \mathbf{w}^t) = E(\mathbf{A} \mathbf{w} \cdot \mathbf{w}^t) = \mathbf{A} \cdot E(\mathbf{w} \mathbf{w}^t) = \mathbf{A} \cdot \Lambda,$$

matricea de covarianță a componentelor principale Λ fiind matricea diagonală formată din valorile proprii ale matricii de covarianță Σ . Pe baza acestui rezultat, matricea de corelație dintre vectorii n-dimensionali x și w poate fi definită sub forma:

$$\text{Corr}(\mathbf{x}, \mathbf{w}) = [\text{Var}(\mathbf{x})]^{-\frac{1}{2}} \cdot \text{Cov}(\mathbf{x}, \mathbf{w}) \cdot [\text{Var}(\mathbf{w})]^{-\frac{1}{2}},$$

unde $\text{Var}(\mathbf{x})$ este matricea diagonală ale cărei elemente sunt reprezentate de varianțele variabilelor originale, iar $\text{Var}(\mathbf{w})$ este matricea diagonală ale cărei elemente sunt varianțele componentelor principale. Deci matricea $\text{Var}(\mathbf{x})$ are forma:

$$\text{Var}(\mathbf{x}) = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix},$$

iar matricea $\text{Var}(\mathbf{w})$ este chiar matricea Λ .

Tinând seama de exprimarea anterioară a covarianței dintre x și w, matricea de corelație dintre x și w devine:

$$\text{Corr}(\mathbf{x}, \mathbf{w}) = [\text{Var}(\mathbf{x})]^{-\frac{1}{2}} \cdot \mathbf{A} \cdot \Lambda \cdot \Lambda^{-\frac{1}{2}} = [\text{Var}(\mathbf{x})]^{-\frac{1}{2}} \cdot \mathbf{A} \cdot \Lambda^{\frac{1}{2}} = \Omega.$$

Matricea Ω este o matrice foarte importantă pentru analiza componentelor principale și este cunoscută sub numele de **matrice factor**. Modalitatea detaliată în care această matrice poate fi calculată este definită de relația:

$$\Omega = \begin{pmatrix} \sigma_1^{-1} & 0 & \dots & 0 \\ 0 & \sigma_2^{-1} & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & \sigma_n^{-1} \end{pmatrix} \cdot \begin{pmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} & \dots & \alpha_1^{(n)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} & \dots & \alpha_2^{(n)} \\ \dots & & & \\ \alpha_n^{(1)} & \alpha_n^{(2)} & \dots & \alpha_n^{(n)} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & \sqrt{\lambda_n} \end{pmatrix}.$$

După efectuarea produselor matriciale în relația de mai sus, matricea Ω capătă forma următoare:

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & \dots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \dots & \omega_{2n} \\ \dots & & & \\ \omega_{n1} & \omega_{n2} & \dots & \omega_{nn} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{\lambda_1} \cdot \alpha_1^{(1)}}{\sigma_1} & \frac{\sqrt{\lambda_2} \cdot \alpha_1^{(2)}}{\sigma_1} & \dots & \frac{\sqrt{\lambda_n} \cdot \alpha_1^{(n)}}{\sigma_1} \\ \frac{\sqrt{\lambda_1} \cdot \alpha_2^{(1)}}{\sigma_2} & \frac{\sqrt{\lambda_2} \cdot \alpha_2^{(2)}}{\sigma_2} & \dots & \frac{\sqrt{\lambda_n} \cdot \alpha_2^{(n)}}{\sigma_2} \\ \dots & & & \\ \frac{\sqrt{\lambda_1} \cdot \alpha_n^{(1)}}{\sigma_n} & \frac{\sqrt{\lambda_2} \cdot \alpha_n^{(2)}}{\sigma_n} & \dots & \frac{\sqrt{\lambda_n} \cdot \alpha_n^{(n)}}{\sigma_n} \end{pmatrix},$$

un element generic ω_{ij} din matricea factor Ω fiind determinat de relația:

$$\omega_{ij} = \frac{\sqrt{\lambda_j}}{\sigma_i} \alpha_i^{(j)}, \quad i=1,2,\dots,n; j=1,2,\dots,n.$$

Elementele matricii factor Ω se numes **intensitățile factorilor** și au o interpretare deosebit de interesantă din punct de vedere al legăturii dintre variabilele originale x_1, x_2, \dots, x_n și componentele principale w_1, w_2, \dots, w_n . Astfel, elementul care se

găsește la intersecția liniei i cu coloana j în matricea factor Ω , adică elementul $\omega_{ij} = \frac{\sqrt{\lambda_j}}{\sigma_i} \alpha_i^{(j)}$, reprezintă **coeficientul de corelație** dintre cea de-a i-a variabilă standardizată x_i și cea de-a j-a componentă principală w_j .

Intensitățile factorilor sunt indicatori ai măsurii în care variabilele originale participă la formarea componentelor principale sau, mai corect, ai măsurii în care componentele principale sintetizează informația conținută în variabilele originale. Cu cât este

mai mare valoarea coeficientului de corelație dintr-o variabilă originală și o componentă principală, cu atât este mai adecvată și mai completă exprimarea informațională a variabilei originale prin intermediul componentei principale respective.

Matricea factor este foarte importantă deoarece, pe baza analizei valorilor elementelor ei, pot fi identificate o serie de *partiții* sau *cluster-e* pe mulțimea variabilelor, partiții sau clustere care, asociate cu anumite componente principale, pot conduce la stabilirea unor semnificații intuitive pentru acele componente. Aceasta înseamnă că analiza elementelor matricii factor Ω poate permite identificarea celor variabile originale care sunt reprezentate prin intermediul unei anumite componente principale și, pe această bază, crearea posibilității de *atribuire a unei semnificații concrete* pentru fiecare componentă principală.

În cazul în care variabilele care intră în componentă vectorului x sunt standardizate, varianțele acestora sunt egale cu unitatea, ceea ce înseamnă că matricea $\text{Var}(x)$ este egală cu matricea unitate. Rezultă că:

$$\text{Cor}(x_s, w) = V\Lambda \cdot \Lambda^{-\frac{1}{2}} = V\Lambda^{\frac{1}{2}} = \Omega_s.$$

În acest caz, *coeficientul de corelație* dintre ce-a de-a i-a variabilă originală și cea de-a j-a componentă principală este definit sub forma:

$$\omega_{ij} = \sqrt{\lambda_j} \cdot \alpha_i^{(j)}.$$

În această variantă, matricea factor are o proprietate importantă care constă în aceea că suma pătratelor elementelor din fiecare coloană să coincide cu varianța componentei principale care se asociază cu respectiva coloană, respectiv:

$$\sum_{i=1}^n \omega_{ij}^2 = \sum_{i=1}^n \lambda_j \cdot (\alpha_i^{(j)})^2 = \lambda_j \cdot \sum_{i=1}^n (\alpha_i^{(j)})^2 = \lambda_j.$$

Ca rezultat al acestei proprietăți, pătratul unui coeficient de corelație din matricea factor poate fi interpretat ca măsură a contribuției pe care o are fiecare variabilă originală la formarea varianței componentei principale.

Exemplul 6.8

Vom presupune că pentru cazul a patru variabile originale matricea de covarianță este de forma următoare:

$$\hat{\Sigma}_v = \begin{pmatrix} 7,603292 & 2,767139 & -1,042365 & -0,686163 \\ 2,767139 & 1,028130 & -0,256115 & -0,197692 \\ -1,042365 & -0,256115 & 1,672133 & 0,824653 \\ -0,686163 & -0,197692 & 0,824653 & 0,429222 \end{pmatrix}.$$

Valorile proprii ale acestei matrici sunt $\hat{\lambda}_1=8,855058$, $\hat{\lambda}_2=1,852882$, $\hat{\lambda}_3=0,01900945$ și $\hat{\lambda}_4=0,00582685$. Vectorii proprii corespunzător acestor valori proprii sunt reprezentați sub forma coloanelor matricii \hat{V} :

$$\hat{V} = \begin{pmatrix} 0,9240069 & 0,1498512 & 0,2157849 & -0,2778358 \\ 0,3343096 & 0,1266964 & -0,5253742 & 0,7721186 \\ -0,1573165 & 0,8847755 & -0,3229776 & -0,2968321 \\ -0,0984877 & 0,4226878 & 0,7570365 & 0,4883963 \end{pmatrix}.$$

Matricea factor, obținută din înmulțirea coloanelor matricii \hat{V} cu rădăcina pătrată a valorii proprii corespunzătoare componentelor principale și din împărțirea liniilor cu abaterea standard corespunzătoare variabilelor originale, este exprimată prin intermediul următorului tabelou:

	w_1	w_2	w_3	w_n
x_1 :	0,99717	0,07397	0,01079	-0,00769
x_2 :	0,98112	0,17008	-0,07144	0,05813
x_3 :	-0,36202	0,93137	-0,03444	-0,01752
x_4 :	-0,44734	0,87822	0,15932	0,05690

Legătura foarte puternică exprimată de primii doi coeficienți de corelație din prima coloană evidențiază faptul că prima componentă principală exprimă conținutul informațional al variabilelor originale x_1 și x_2 . În mod similar, corelația foarte puternică exprimată de ultimii doi coeficienți din cea de-a doua coloană arată că cea de-a două componentă principală sintetizează informațional variabilele originale x_1 și x_2 . Din faptul că primele două valori proprii reprezintă 99,77% din suma tuturor valorilor proprii rezultă că cele patru variabile originale pot fi exprimate prin intermediul primelor două componente principale, cu pierdere neglijabilă de informație.

Tema 7. Analiza factorială

7.1 Scopul analizei factoriale

Încă de la începuturile utilizării sale în activitatea de cercetare științifică, analiza factorială a fost folosită, aproape exclusiv, pentru a fundamenta o serie de teorii psihologice referitoare la comportamentul și abilitatea umană. Utilizarea analizei factoriale în acest scop se bazează pe o facilitatea extraordinară pe care o oferă acest instrument, facilitate care constă în existența posibilității de măsurare indirectă, de cuantificare a unor factori neobservabili.

În psihologie, precum și în alte științe din domeniul economico-social, apare frecvent necesitatea identificării și studierii unor entități informaționale complexe, care *nu pot fi observable* în sens statistic, ceea ce înseamnă că, în mod implicit, aceste entități *nu pot fi direct măsurabile*. Aceste entități sunt cunoscute în analiza factorială sub numele de **factori comuni și factori specifici**. Ca exemple de astfel de entități informaționale de acest tip putem menționa *statusul social, abilitatea managerială, inteligența, profilul psihologic* etc.

Există însă posibilitatea de utiliza alte entități informaționale, care au proprietatea că sunt *măsurabile* și care *reflectă* entitățile informaționale neobservabile menționate anterior, astfel încât prin intermediul acestora putem face o *evaluare indirectă* a entităților neobservabile. Entitățile informaționale utilizate pentru această evaluare indirectă sunt cunoscute în analiza factorială sub numele de **indicatori**.

În acest sens, analiza factorială poate fi privită ca o modalitate de a explica și interpreta legăturile dintre anumite variabile observabile (variabile indicator) în termeni de mărimi care nu pot fi direct observabile (factori). De exemplu, rezultatele înregistrate la anumite teste de performanță, care au natură observabilă și direct măsurabilă, pot fi explicate și interpretate în termenii unui factor abstract și neobservabil în mod direct, care este *nivelul de inteligență* al celor testați.

7.2 Importanța și necesitatea sintetizării cauzalității

O problemă importantă în cadrul procesului de studiere și descriere a relațiilor de cauzalitate dintre fenomenele și procesele economice, cu o pondere foarte mare în analiza și predicția din domeniul economico-social, este cea legată de necesitatea de a *sintetiza relațiile de cauzalitate*. Această problemă este direct și strâns legată de o altă problemă de importanță teoretică fundamentală, a cărei rezolvare apare cu necesitate în orice investigație de natură economico-socială, și anume problema *măsurării indirecte*, a măsurării factorilor de *natură neobservabilă*.

Necesitatea sintetizării relațiilor de cauzalitate, a detectării unor relații de cauzalitate ascunse și a evaluării unor factori de natură neobservabilă apare în mod frecvent, în contextul soluționării unor probleme de măsurare și cuantificare dintr-o mare varietate de domenii: economic, social, politic, psihologic etc. În general, domeniile în care apar probleme de acest fel sunt domenii caracterizate de complexitate și de imposibilitatea efectuării unor măsurători experimentale.

7.2.1 Măsurarea factorilor neobservabili

În orice investigare sau activitate de cercetare științifică din domeniul economico-social intervine, aproape în mod inevitabil, o problemă extrem de dificilă, a cărei rezolvare depășește cadrul standard, presupunând utilizarea unor raționamente și instrumente de mare subtilitate și rafinament științific. O astfel de problemă apare, cu exclusivitate, în cîmpul măsurării și cuantificării economico-sociale, și este legată de existența factorilor *latenți*, a factorilor *ascunși*, a factorilor de *natură neobservabilă*. Acești factori sunt cunoscuți în analiza factorială sub numele de **factori comuni**.

Exemplul 7.1

Este evident, de exemplu, că fiecare țară din lumea contemporană are un anumit *nivel de dezvoltare economică*, dispune de o anumită *forță economică*. Această caracteristică este *comună* tuturor țărilor lumii, însă are *valori diferite* de la o țară la alta.

Forța economică de care dispune o anumită țară, la un moment dat, este o apreciere de natură eminentă *sintetică, abstractă*, care poate fi considerată ca rezultantă a compunerii și agregării unui număr foarte mare de fenomene și procese economice, de influențe și interdependențe. Este evident că este foarte dificil, chiar imposibil, ca o astfel caracteristică extrem de complexă să poată fi apreciată în mod direct, doar prin intermediul unei singure mărimi, pe baza unui singur indicator, indiferent dacă acest indicator este ritmul creșterii economice, volumul produsului intern brut, volumul schimburilor economice externe, rata inflației, rata șomajului sau oricare alt indicator macroeconomic. În aceste condiții, se poate afirma cu certitudine că nu pot exista nici metodologia, nici unitățile de măsură și nici instrumentele cu ajutorul căror să se poată calcula, în mod direct și la un nivel de rigoare acceptabil, nivelul de dezvoltare a unei țări.

Prin urmare, nivelul de dezvoltare economică a unei țări constituie un exemplu relevant de ceea ce înseamnă un factor neobservabil. Cu toate acestea, prin utilizarea unor tehnici de analiză specifice, pot fi deduse *măsuri numerice unice, indicatori agregați*, care să reflecte gradul de dezvoltare economică a unei țări în toată complexitatea sa.

În obținerea măsurilor aggregate de acest fel sunt utilizati indicatori macroeconomici parțiali, din categoria celor enumerate anterior. Informațiile purtate de acești indicatori sunt considerate a fi *semnale generate de factorul latent numit nivel de dezvoltare economică*.

Raționalul teoretic pe care se fundamentează evaluările de acest fel constă în presupunerea că factorul latent induce variații, mai mult sau mai puțin semnificative, în magnitudinea unor indicatori de natură observabilă. Cu cât variații induse sunt mai puternice, cu atât

se consideră că respectivii indicatori au o capabilitate mai ridicată de a servi la exprimarea, indirectă, a factorului latent. Astfel, pentru un exemplu de tipul celui menționat anterior, se poate considera că volumul produsului intern brut, ritmul creșterii economice, productivitatea socială a muncii, nivelul de instruire a populației, volumul schimburilor economice externe etc., sunt rezultate ale unui anumit nivel de dezvoltare, sunt expresii ale acestui nivel de dezvoltare și, în consecință, pot servi la evaluarea acestuia.

În ceea ce privește legătura dintre valoarea unei variabile indicator și factorul comun, facem precizarea că mărimea înregistrată de valoarea unui indicator nu este determinată în mod *exclusiv* de factorul comun, ea depinzând, în afară de acesta, și de influența altor factori, cunoscuți sub numele de *factori specifici*. Influенța acestor factori este diferită de la un indicator la altul și nu este comparabilă pe multimea acestor indicatori. Dată fiind natura lor, din rândul factorilor specifici fac parte și erorile sau reziduurile.

De exemplu, produsul intern brut și volumul schimburilor economice externe au o determinare *comună*, dată de nivelul de dezvoltare a economiei, dar au și o determinare *specifică*, dată de mărimea țării, în cazul primului indicator, respectiv de poziția geografică a țării, în cazul celui de-al doilea indicator. Mărimea țării este factor specific pentru produsul intern brut, iar poziția geografică a țării este indicator specific pentru volumul schimburilor externe.

Schematic, legăturile dintre indicatori, pe de o parte, și factorul comun și specifici, pe de altă parte poate fi prezentată schematic sub forma din figura următoare.

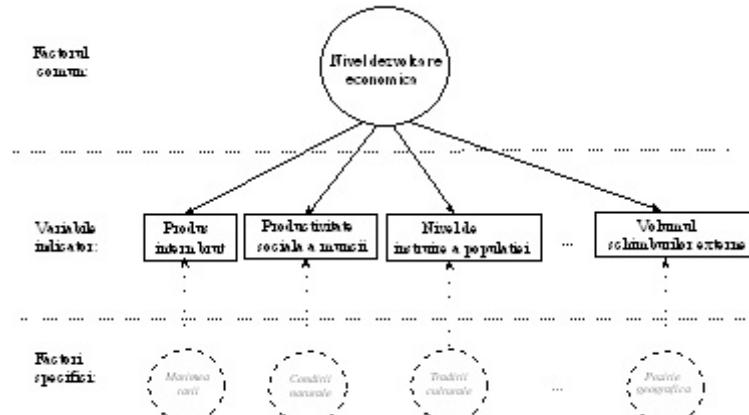


Figura 7.1: Exemplificarea unor legături factoriale posibile

Măsurarea influenței specificului local asupra volumului vânzărilor dintr-un produs, evaluarea aptitudinilor manageriale ale unei persoane, canticarea imaginii pe care o firmă o are pe piață, măsurarea forței financiare a unei firme, aprecierea gradului de dezvoltare economică a țării, evaluarea gradului de profitabilitate a unei firme etc., reprezintă probleme care conduc la necesitatea efectuării unor evaluări de tip *indirect*, pe bază de *intermediere*, făcându-se apel la o serie de entități observabile, cunoscute sub numele de *teste* sau *indicatori*.

Pe baza informațiilor colaterale și cu ajutorul unor instrumente specifice, pot fi obținute dimensionări de natură cantitativă pentru factorii neobservabili și pot fi construite scale de măsurare pe care aceștia să poată fi evaluați.

Un rol unic în acest sens, rol bine determinat și extrem de important în activitatea de evaluare și canticarea a factorilor de natură neobservabilă și de construire a unor scale de măsurare pentru acești factori, revine tehniciilor de analiză cunoscute sub numele generic de *analiză factorială*.

Deducerea, pe cale indirectă, a informațiilor referitoare la factorii neobservabili se bazează pe identificarea unor *variabile indicator* sau pe construirea unor *teste*, care să furnizeze informațiile necesare aplicării tehniciilor de analiză factorială.

Pe baza informațiilor referitoare la variabilele de tip indicator, analiza factorială își propune de deducă substanța informațională conținută în observațiile acestor variabile indicator, substanță comună tuturor acestor variabile.

7.2.2 Detectarea substanței informaționale comune

Ideea fundamentală care stă la baza oricărui demers ce vizează sintetizarea relațiilor de cauzalitate este aceea că, în general, influențele exercitate de variabilele explicative ce intervin într-o relație de cauzalitate complexă reprezintă o *intermediere* a influențelor unor alți factori, comuni unor submulțimi ale acestor variabile și aflați în spatele acestora.

Cu alte cuvinte, se consideră că influențele exercitate de variabilele explicative, variabile care, în aparență, par a fi cauze directe ce determină un anumit efect, nu sunt altceva decât forme particulare de manifestare indirectă, intermediară, ale unor alte influențe de natură primordială și sintetică, influențe specifice unor factori neobservabili, cu caracter latent.

Acceptând ipoteza că influențele fiecărui din acești factori comuni, cu natură latentă, se manifestă prin intermediul uneia sau a mai multor variabile explicative, putem separa aceste variabile în *submulțimi* de variabile ce au un *element comun*, iar prin intermediul acestor submulțimi să definim o corectă și sugestivă *structură a dependenței*.

În sensul ideii exprimate anterior, problema sintetizării relațiilor de cauzalitate devine echivalentă cu problema detectării și canticării unei anumite *substanțe comune*, de natură pur informațională, conținută de o submulțime a variabilelor explicative supuse studiului.

Această substanță poate fi interpretată ca fiind *fundamentul informațional comun* al variabilelor explicative ce alcătuiesc respectiva submulțime și care există, sub formă mascată și latentă, dincolo de aparența relevată de valorile obținute pe baza

măsurătorilor efectuate asupra acestor variabile.

Din punct de vedere teoretic, se consideră că *mărimea cantității din această substanță* regăsită în fiecare variabilă este cea care determină, în cea mai mare parte, nivelul și evoluția respectivelor variabile, constituind în același timp fundamentalul invizibil care generează și dimensionează relațiile de interdependență dintre variabile.

În funcție de cantitatea din respectiva substanță comună, conținută în fiecare variabilă explicativă, se poate determina o structură riguroasă a dependenței acestora, se pot construi "cluster-e" de variabile explicative și se pot deduce informații a căror natură să nu mai fie parazitată de modalitățile particulare de manifestare a formei fenomenelor descrise de respectivele variabile.

7.3 Domenii ale utilizării analizei factoriale

Printre domeniile a căror investigare presupune cu preponderență existența unui instrument de analiză care să permită depistarea substanței comune conținută în variabilele urmărite putem menționa pe cel sociologic și pe cel psihologic.

Instrumentul a cărui operaționalitate și eficiență sunt unanim recunoscute în rezolvarea problemelor care au ca scop sintetizarea relațiilor de cauzalitate este **analiza factorială**.

În domeniul economic aplicațiile analizei factoriale sunt întâlnite în majoritatea analizelor și predicțiilor care vizează *prospectarea pieței, fundamentarea strategiilor și deciziilor manageriale* la nivelul agenților economici, *comportamentul consumatorilor sau procesul decizional din domeniul finanțier-bancar*.

Practic, nu există analiză de date sau predicție din domeniul economic care să nu apeleze, cel puțin în faza preliminară, la utilizarea unui instrument cum este cel oferit de analiza factorială.

Printre cele mai frecvent întâlnite aplicații ale analizei factoriale în domeniul economic putem menționa pe cele care au ca scop *determinarea unor indici agregați* pentru diferite fenomene economice complexe sau pe cele care își propun *clasificarea și ierarhizarea unor opțiuni sau entități economice* în funcție de mai multe criterii economice și sociale.

Din ce în ce mai mult în ultimul timp folosirea analizei factoriale este întâlnită și în domeniul *politologiei*. Studiile riguroase de fundamentare a strategiilor politice folosesc analiza factorială în special pentru determinarea opțiunilor politice ale electoratului înaintea organizării scrutinurilor și pentru cunoașterea motivațiilor concrete care determină aceste opțiuni.

7.4 Definirea analizei factoriale și a conceptelor acesteia

Analiza factorială este unul dintre cele mai eficiente și mai frecvent utilizate instrumente utilizate în modelarea statistică-matematică a problemelor economic-sociale. Posibilitățile largi pe care acest instrument de lucru le oferă în probleme de analiză științifică și de interpretare a datelor sunt ilustrate prin existența unui număr foarte mare de aplicații și prin bogata literatură de specialitate dedicată problemelor teoretice și practice specifice domeniului.

Prin intermediul tehniciilor de acest fel se încearcă aplicarea unor transformări și aproximări, astfel încât să se obțină reprezentări în spații cu mai puține dimensiuni, respectiv în spații unidimensionale, bidimensionale sau, cel mult, tridimensionale. Aceste transformări și aproximări trebuie să fie aplicate în aşa fel încât să se verifice condiția ca distanțele obținute în *noul spațiu* între obiecte sau variabile să reflecte cât mai bine distanțele existente între acestea în spațiu original. Noul spațiu, rezultat din efectuarea analizei factoriale, se numește *spațiu factor, spațiu redus sau spațiu factorial*.

Utilizarea analizei factoriale în soluționarea problemelor specifice analizei datelor se face sub forma a două modalități de abordare: *modalitatea exploratorie* și *modalitatea confirmatorie*.

7.4.1 Definirea conceptelor fundamentale

În analiza factorială sunt manipulate o serie de concepte importante, al căror conținut și interpretare se caracterizează, de cele mai multe ori, prin multiple și subtile aspecte informaționale. Conceptul fundamental al analizei factoriale este cel de *factor comun* sau de *factor latent*, celelalte concepte utilizate fiind definite, într-o măsură mai mare sau mai mică, în raport cu acest concept de referință.

Factorul comun este o construcție abstractă, care încearcă să exprime sub o formă numerică o entitate informațională neobservabilă, ascunsă în spatele unei mari varietăți de manifestări eterogene, dar a cărei semnificație este extrem de importantă, atât din punct de vedere teoretic, cât și practic. Factorul comun are o natură similară cu cea a unui factor cauzal cu semnificație stabilă și consistentă, a căruia existență și manifestare influențează valorile unei întregi mulțimi de variabile numite *indicatori*.

Definiție: *Factorul comun* sau *factorul latent* reprezintă o entitate informațională de natură generală, care exprimă o caracteristică esențială a unui fenomen sau proces din realitate, își exercită influența asupra tuturor elementelor unei mulțimi de indicatori și care nu poate fi supusă unei procese direct de observare și măsurare.

După cum rezultă și din definiție, caracteristica principală a factorilor comuni constă în aceea că ei sunt *mărimi neobservabile*. Cu toate că factorii comuni nu pot fi măsurăți direct, pot fi definite și supuse procesului de observare și măsurare o serie de alte entități informaționale, care au proprietatea că *reflectă*, într-o măsură mai mare sau mai mică, factorii comuni.

Aceste entități informaționale sunt *variabilele indicator*, care sunt considerate a fi forme de manifestare cantitativă indirectă a factorilor comuni și care prin valorile lor *indică* atât prezența factorilor comuni, cât și intensitatea cu care se manifestă aceștia.

Factorii comuni pot fi priviți ca variabile aleatoare, care au o anumită distribuție de probabilitate. De regulă, se consideră că factorii comuni sunt repartizați după legea de probabilitate normală. Vom nota numărul de factori comuni cu p , iar pentru cei p factori comuni vom folosi notațiile f_1, f_2, \dots, f_p .

Obținerea de informații necesare pentru deducerea unor aproximări cantitative pentru factorii comuni este bazată pe existența unor variabile speciale denumite *indicatori, teste sau măsuri*.

Definiție: *Indicatorul sau testul* este o variabilă ale cărei observații, cunoscute sub numele de *scoruri*, sunt utilizate în cadrul analizei factoriale în scopul de a produce evaluări numerice pentru factorul sau factorii comuni.

În analiza factorială se presupune că legătura dintre variabilele indicator și factorii comuni poate fi exprimată sub o formă liniară, exprimare care conduce la necesitatea estimării coeficienților care intervin în definirea respectivei forme.

Prin raportare la contextul terminologiei utilizate în cadrul analizei componentelor principale, se poate spune că indicatorii sau testele sunt reprezentate de variabilele originale. Vom presupune în continuare că în analiză există n indicatori sau n teste și vom simboliza acești indicatori cu x_1, x_2, \dots, x_n .

Observațiile existente cu privire la cei n indicatori, observații obținute prin extragerea unui eșantion aleator de volum T din populația ale cărei unități sunt caracterizate de cei n indicatori, reprezintă *scorurile indicatorilor* sau *scorurile testelor*.

O ipoteză importantă a analizei factoriale constă în presupunerea conform căreia nivelul unei variabile indicator se formează ca urmare a unor influențe conjugate, exercitată atât de factorul sau factorii comuni, cât și de un *factor unic*. În afara acestor influențe cu natură semnificativă, asupra nivelului unei variabile indicator se mai exercită și influența erorilor de măsurare, influență considerată a fi neglijabilă.

Departate de a avea o semnificație măcar comparabilă cu aceea a factorilor comuni, factorul unic are, totuși, o natură similară cu cea a acestor factori: influențează nivelul unei variabile indicator și are natură neobservabilă. Spre deosebire de factorul comun, a cărui influență se manifestă la nivelul tuturor variabilelor indicator, factorul unic este caracterizat prin aceea că influența sa are o natură *particulară, unilaterală*, considerată a se exercita sau exprima numai la nivelul unei singure variabile indicator. Din acest motiv, numărul de factori unici coincide cu numărul de indicatori sau de teste.

Vom folosi pentru notarea celor n factori unici, care sunt asociații celor n indicatori, simbolurile v_1, v_2, \dots, v_n . Factorul unic poate fi definit sub forma următoare:

Definiție: *Factorul unic* reprezintă o entitate informațională de natură *particulară*, care își exercită influența în mod *unilateral*, asupra unei singure variabile indicator, și care nu poate fi supusă unei procese directe de observare și măsurare.

În analiza factorială variabilele indicator sunt considerate a fi dependente de factorii comuni și de factorul unic, în mod similar cu dependența descrisă de modelele de regresie, în care *variabila dependentă* este reprezentată de indicator sau test, *variabilele independente* sunt reprezentate de factorii comuni, iar *termenul eroare* este reprezentat de factorul unic și de factorul rezidual. Considerând cazul celei de-a i-a variabile indicator, un astfel de model de regresie are forma următoare:

$$x_j = a_{j1} \cdot f_1 + a_{j2} \cdot f_2 + \dots + a_{jp} \cdot f_p + \varepsilon_j.$$

Deosebirea acestui tip de model față de un model de regresie autentic, constă în aceea că variabilele sale independente, f_1, f_2, \dots, f_p , sunt variabile aleatoare *neobservable*.

Vom ilustra și vom concretiza conținutul celor trei mărimi fundamentale definite anterior, respectiv indicator, factor comun și factor unic, prin intermediul următorului exemplu.

Exemplul 7.2

În scopul evaluării nivelului de inteligență generală și a capacitatei de memorare ce caracterizează fiecare dintre cei 10 studenți ai unei grupe ipotetice, vom presupune că dispunem de notele obținute de aceștia la 5 examene: Matematică, Informatică, Economie, Istorie, Engleză. Rezultatele se găsesc în tabelul următor.

Situarea notelor obținute de studenții unei grupe

Tabelul 7.1

Student	Note obținute					Medie student
	Matematică	Informatică	Economie	Istorie (h)	Engleză (g)	
S ₁	6,25	8,25	7,25	8,00	8,50	7,65
S ₂	5,50	7,50	7,50	6,75	7,00	6,85
S ₃	9,25	9,75	9,25	8,50	8,25	9,00
S ₄	8,00	7,75	9,00	8,50	9,25	8,50
S ₅	5,75	6,00	7,25	10,00	8,25	7,45
S ₆	7,50	8,00	6,75	8,75	9,00	8,00
S ₇	10,00	9,25	9,00	9,00	9,25	9,30
S ₈	9,50	9,75	8,50	10,00	9,00	9,35
S ₉	6,50	8,25	9,25	9,75	10,00	8,75
S ₁₀	5,25	6,75	7,00	8,25	8,50	7,15
Media	7,350	8,125	8,075	8,750	8,700	
Varianța	3,1139	1,5035	1,0285	1,0139	0,6500	

În limbajul analizei factoriale, cele cinci examene reprezintă variabilele indicator sau teste. Factorii comuni sunt, în acest caz, *inteligența și memoria*. Notele obținute de studenți la examene reprezintă *scorurile testelor*.

Ipoteza raționamentului specific analizei factoriale este aceea că rezultatele obținute de studenți sunt *intercorelate*, ca urmare a faptului că ele sunt influențate de două caracteristici care sunt comune tuturor studenților: inteligența și capacitatea de memorare.

În tabelul următor sunt prezentate coeficienții de corelație dintre cele cinci variabile indicator. Faptul că notele obținute la cele cinci discipline sunt corelate, între ele, într-o măsură mai mare sau mai mică, constituie suportul pentru a considera că în obținerea acestor note se manifestă ceva care este comun tuturor disciplinelor, indiferent de natura acestora. În aceste condiții, este natural a presupune că elementele comune, care determină corelarea rezultatelor, țin de inteligență nativă a indivizilor și de capacitatea lor de memorare.

Matricea de corelație a variabilelor indicator

Tabelul 7.2

Discipline	Matematică	Informatică	Economie	Istorie	Engleză
Matematică	1,00	0,86	0,77	0,15	0,30
Informatică	0,86	1,00	0,73	0,09	0,24
Economie	0,77	0,73	1,00	0,23	0,39
Istorie	0,15	0,09	0,23	1,00	0,73
Engleză	0,30	0,24	0,39	0,73	1,00

Inteligența și capacitatea de memorare reprezintă doi factori care influențează notele obținute de fiecare student la fiecare dintre examenele susținute, cu diferențe de la student la student, în funcție de inteligență și memoria proprii fiecărui dintre studenți.

În afara celor doi factori comuni, nota obținută de studenți la fiecare dintre examene este influențată și de un *factor unic*, factor care reprezintă aptitudinile studentului pentru domeniul de care aparține disciplina respectivă. De exemplu, separat de nivelul de inteligență și de capacitatea de memorare, un student poate avea aptitudini speciale pentru domeniul Informaticii. Influența acestor aptitudini asupra notei obținute la informatică se va concretiza prin intermediul factorului unic asociat cu acest indicator, adică prin intermediul factorului unic v_i .

Rezultă că performanțele obținute de studenți la fiecare examen pot fi descrise cu ajutorul unor ecuații de regresie de forma:

$$N_m = a_{mI} \cdot I + a_{mM} \cdot M + v_m; \quad N_i = a_{iI} \cdot I + a_{iM} \cdot M + v_i; \quad N_e = a_{eI} \cdot I + a_{eM} \cdot M + v_e$$

$$N_h = a_{hI} \cdot I + a_{hM} \cdot M + v_h; \quad N_g = a_{gI} \cdot I + a_{gM} \cdot M + v_g$$

unde N_m, N_i, N_e, N_h și N_g definesc variabilele indicator, care reprezintă notele la cele cinci examene, I și M sunt cei doi factori comuni care influențează notele, iar v_m, v_i, v_e, v_h și v_g reprezintă factorii unici celor cinci discipline considerate. Mărimile a_I și a_M reprezintă coeficienții corespunzători celor doi factori comuni.

Efectuând analiza factorială pe datele conținute în tabelul de mai sus, am obținut următoarele rezultate:

$$N_m = 0,891 \cdot I + 0,323 \cdot M + v_m; \quad N_i = 0,832 \cdot I + 0,366 \cdot M + v_i; \quad N_e = 0,813 \cdot I + 0,143 \cdot M + v_e$$

$$N_h = 0,410 \cdot I - 0,667 \cdot M + v_h; \quad N_g = 0,603 \cdot I - 0,730 \cdot M + v_g$$

Cei doi factori comuni, inteligența și capacitatea de memorare, explică o proporție semnificativ de mare din varianța totală a variabilelor indicator, respectiv un procent de 78,30%, din care primul factor deține 53,58%, iar cel de-al doilea 24,72%.

Cu toate acestea, rezultatele manifestă o inadvertență logică legată de interpretarea celui de-al doilea factor comun. Această inadvertență constă în faptul că, în cazul variabilelor indicator Istorie și Engleză, coeficienții corespunzători factorului Memorie au valori negative.

Existența acestor coeficienți cu valori negative ridicătoare conduce la o interpretare aberantă, interpretare conform căreia capacitatea de memorare ar influența foarte puternic performanțele la Istorie și Engleză, însă în sens negativ.

Deoarece există soluția obținută nu este unică, poate fi căutată altă soluție, care să fie compatibilă cu o interpretare naturală și corectă a factorilor. Această soluție poate fi obținută printr-o procedură de transformare a coeficienților factorilor, procedură cunoscută sub numele de *rotația structurii factor*. În urma aplicării acestei proceduri, poate obține o interpretare mai corectă a factorilor și o creștere a semnificației acesteia, în condițiile în care contribuția totală a factorilor comuni la formarea variabilității variabilelor indicator rămâne neschimbăță.

În cazul exemplului nostru, prin aplicarea procedurii de rotație a structurii factor au fost obținute rezultate care conduc la următoarea formă a modelului factorial:

$$N_m = 0,938 \cdot I + 0,132 \cdot M + v_m; \quad N_i = 0,907 \cdot I + 0,066 \cdot M + v_i; \quad N_e = 0,785 \cdot I + 0,255 \cdot M + v_e$$

$$N_h = 0,050 \cdot I + 0,782 \cdot M + v_h; \quad N_g = 0,191 \cdot I + 0,927 \cdot M + v_g$$

Se poate observa că inadvertența legată de interpretarea celui de-al doilea factor a dispărut, în condițiile în care proporția explicată de cei doi factori din varianța totală a variabilelor indicator a rămas tot la nivelul de 78,30%. Ceea ce s-a modificat în urma rotației este structura acestei proporții pe cei doi factori, structură conform căreia, din procentul de 78,30%, primul factor deține 47,17%, iar cel de-al doilea 31,13%.

Rezultatele obținute evidențiază că indicatorii reprezentând notele la Matematică, Informatică și Economie exprimă foarte bine nivelul de inteligență al studenților, după cum indicatorii care reprezintă notele la Istorie și Engleză reflectă foarte bine capacitatea de memorare a studenților. Contribuțiiile factorilor unici sunt relativ reduse, cu excepția Economiei, unde se înregistrează o contribuție ceva mai ridicată a factorului unic.

În graficul din figura următoare sunt prezentate legăturile dintre cele trei categorii de mărimi și sunt evidențiate intensitățile cu care aceste legături se manifestă.

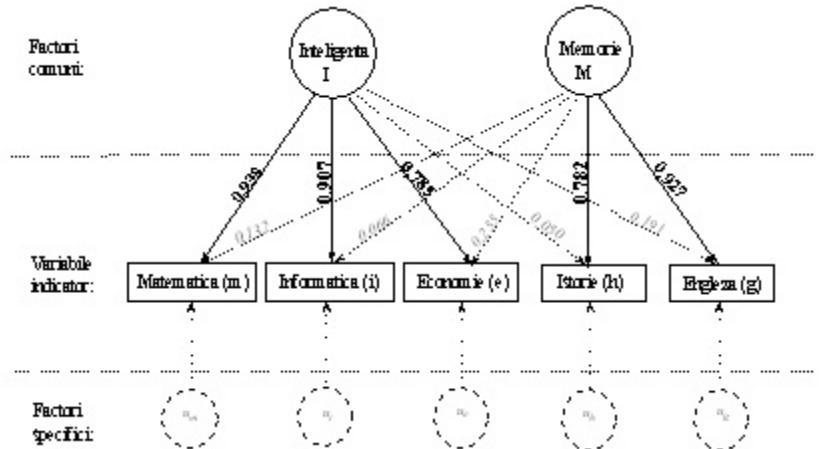


Figura 7.2: Exemplificarea legăturilor dintre notele la examene, nivelul de inteligență și capacitatea de memorare

7.4.2 Definirea analizei factoriale

Analiza factorială este unul dintre cele mai puternice, mai subtile și mai utile instrumente de analiză a datelor, eficiența utilizării ei fiind de necontestat în aproape toate fazele importante ale unui astfel de demers, inclusiv în faza exploratorie a analizei. Scopul principal al analizei factoriale este acela al deducerii unor construcții ipotetice, al identificării unor factori esențiali, prin intermediul cărora să poată fi explicate interdependențele existente între elementele unei mulțimi de variabile indicator.

Spre deosebire de analiza componentelor principale, care are ca scop *reexprimarea unei proporții cât mai mari din variabilitatea spațiului cauzal initial*, analiza factorială își propune să găsească cea mai bună modalitate de reproducere a corelațiilor manifestate între elementele unei mulțimi de indicatori, ceea ce este o modalitate de *reexprimare a interdependențelor* dintre variabilele indicator. Din punct de vedere al acestui scop, analiza factorială poate fi definită sub forma următoare:

Definiție: *Analiza factorială* este o analiză multivariată, care are ca scop să explice corelațiile manifestate între o serie de variabile, numite *indicatori* sau *teste*, prin intermediul unui *număr mai mic* de factori ordonați și necorelați, numiți *factori comuni*.

Proprietatea de necorelare a factorilor, care apare în definiția anterioară, se referă la definirea și determinarea acestora sub restricția inexistenței redundanței informaționale. În mod similar, ordonarea factorilor se referă la ierarhizarea acestora într-o manieră descrescătoare, în funcție de mărimea varianței fiecărui factor.

Ca metodă de analiză multidimensională, analiza factorială este folosită, în principal, în rezolvarea problemelor al căror scop este legat de:

- studierea *nivelelor diferite de manifestare a interdependențelor* dintre variabilele explicative, în special atunci când numărul acestora este foarte mare;
- detectarea unei structuri simplificate și clare a relațiilor de interdependență existente între variabilele explicative;
- obținerea unei "cluster-izări", unei clasificări a variabilelor explicative prin intermediul unor entități numite *factori*, astfel încât variabilele aparținând unui anumit factor să fie puternic intercorelate;
- obținerea unor informații specifice, sub forma așa-numiților *factori*, pe baza cărora să se poată face o interpretare sintetică a relațiilor de cauzalitate;
- verificarea unor ipoteze cu privire la existența unei structuri factoriale particulare sau cu privire la existența unui anumit număr de factori comuni;
- *sintetizarea potențialului cauzal comun* al mai multor variabile explicative sub forma unui număr cât mai redus de factori.

Privită în sens restrâns, ca mulțime de proceduri logice și numerice executate pe date de un anumit tip, analiza factorială poate fi definită ca reprezentând un proces a cărui desfășurare include următoarele etape esențiale:

- determinarea numărului *minimal* de factori comuni cu ajutorul căruia pot fi explicate în mod *optimal* corelațiile existente între variabilele indicator;
- efectuarea unor *rotații* ale factorilor, în scopul determinării *soluției factor* sub cea mai simplă și mai clară formă;
- estimarea *intensităților factorilor*, *structurii legăturilor*, *comunalităților* și *varianțelor factorilor unici*;
- deducerea unor *interpretări* adecvate pentru factorii comuni;
- estimarea *scorurilor factorilor*.

Dintre toate activitățile implicate de analiza factorială, problema care ridică cele mai dificultăți în executarea acestei analize este aceea a *estimării comunalităților*, sau, ceea ce este același lucru, estimarea intensității factorilor comuni.

7.4.3 Tipurile analizei factoriale

În funcție de modalitatea în care este implicată în studiu și de scopul concret în care este folosită, analiza factorială poate fi considerată ca fiind de două tipuri: *analiză factorială exploratorie* și *analiză factorială confirmatorie*.

Definiție: *Analiza factorială exploratorie* reprezintă acea modalitate de utilizare a analizei factoriale care are ca scop detectarea unei structuri a dependenței și generarea unor construcții teoretice, cunoscute sub numele de *factori comuni*.

În varianta sa de analiză exploratorie, analiza factorială reprezintă o tehnică de identificare a structurii dependenței, de generare a unor construcții teoretice. Analiza factorială de tip exploratoriu nu presupune cunoașterea apriorică a structurii dependenței cauzale sau cunoașterea apriorică a factorilor.

Definiție: *Analiza factorială confirmatorie* reprezintă acea modalitate de utilizare a analizei factoriale, care are ca scop *confirmarea unor ipoteze și teorii* privind structura unei dependențe cauzale.

În analiza factorială confirmatorie, structura dependenței cauzale sau construcția cauzală teoretică sunt presupuse a fi cunoscute, adică date prin ipoteză. Prin intermediul teoriei de acest tip se urmărește *confirmarea* unei anumite teorii, verificarea acestei teorii pe cale empirică.

7.5 Structura generală a modelului factorial

În formularea sa cea mai generală, formulare extrem de necesară pentru precizări cu caracter terminologic și pentru formulareră unor ipoteze de natură teoretică, modelul analizei factoriale are la bază două ipoteze fundamentale.

Prima ipoteză se referă la presupunerea că că *nivelul* sau *valorile* unui ansamblu de variabile aleatoare $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ se formează ca rezultat exclusiv al influenței a trei categorii de factori:

- o mulțime formată din *p factori comuni*, $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_p$, a căror influență se consideră a se exercita asupra *fiecărei* dintre cele *n* variabile considerate;
- o mulțime formată din *n factori unici*, $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$, a căror influență se consideră a se exercita în mod *individual*, fiecare factor unic influențând *una și numai una* dintre variabilele considerate;
- o mulțime de *n factori reziduali*, $\mathbf{\epsilon}_1, \mathbf{\epsilon}_2, \dots, \mathbf{\epsilon}_n$, a căror influență se consideră a fi exercitată tot în mod *individual*, fiecare factor rezidual influențând *câte o singură variabilă*.

Din punct de vedere statistic, se consideră că influențele semnificative, care trebuie reținute în analiză, sunt cele exercitate de factorii comuni și unici, în timp ce influențele factorilor reziduali, se consideră a avea caracter accidental, nesemnificativ. La nivelul fiecărei variabile, influența factorului rezidual corespunzător poate fi considerată a fi neglijabilă și este asimilabilă erorilor de măsurare. Din acest motiv, factorii reziduali se mai numesc și *erori*.

În ceea ce privește factorii comuni, există posibilitatea ca în cazul anumitor variabile influența lor asupra acestor variabile să fie neglijabilă sau chiar nulă, ceea ce înseamnă că factorii respectivi pot fi eliberați din lista factorilor pentru variabila respectivă. În aceste condiții, este posibil ca schema de influență pentru anumite variabile să conțină mai mulți factori comuni, iar pentru alte variabile mai puțini. Numărul de factori comuni cu influență semnificativă asupra variabilei indicator determină *complexitatea* variabilei indicator respective.

Faptul că influențele considerate sunt structurate pe cele trei categorii de factori, determină o anumită structură a modelului factorial general, structură evidențiată de reațiile următoare:

$$\mathbf{x}_1 = \varphi_1(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_p; \mathbf{u}_1) + \mathbf{\epsilon}_1$$

$$\mathbf{x}_2 = \varphi_2(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_p; \mathbf{u}_2) + \mathbf{\epsilon}_2$$

...

$$\mathbf{x}_n = \varphi_n(\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_p; \mathbf{u}_n) + \mathbf{\epsilon}_n$$

unde funcțiile φ_i sunt funcții reale de $p+1$ variabile, funcții care pot fi liniare sau neliniare în raport cu variabilele.

Cu toate că funcțiile φ_i pot fi, în principiu, liniare sau neliniare, aproape în toate cazurile, atât în cele legate de teoria analizei factoriale, cât și în cele legate de utilizarea acesteia în aplicații, este considerată varianta *liniară*, ceea ce înseamnă că relațiile care definesc modelul au forma următoare:

$$\mathbf{x}_1 = a_{11} \cdot \mathbf{f}_1 + a_{12} \cdot \mathbf{f}_2 + \dots + a_{1p} \cdot \mathbf{f}_p + a_1 \cdot \mathbf{u}_1 + \mathbf{\epsilon}_1$$

$$\mathbf{x}_2 = a_{21} \cdot \mathbf{f}_1 + a_{22} \cdot \mathbf{f}_2 + \dots + a_{2p} \cdot \mathbf{f}_p + a_2 \cdot \mathbf{u}_2 + \mathbf{\epsilon}_2$$

...

$$\mathbf{x}_n = a_{n1} \cdot \mathbf{f}_1 + a_{n2} \cdot \mathbf{f}_2 + \dots + a_{np} \cdot \mathbf{f}_p + a_n \cdot \mathbf{u}_n + \mathbf{\epsilon}_n$$

Coeficienții factorilor sunt cunoscuți sub numele de *intensitățile* factorilor. Prin *magnitudinea* se coeficientul măsoară *intensitatea* influenței exercitate de factorul corespunzător asupra nivelului variabilei indicator, iar prin *semnul* său măsoară *sensul* influenței exercitate.

Definiție: Se numește *intensitate* a unui factor comun \mathbf{f}_j în raport cu o variabilă indicator \mathbf{x}_i mărimea a_{ij} , care arată cu câte unități se modifică nivelul variabilei indicator \mathbf{x}_i , atunci când nivelul factorului \mathbf{f}_j crește cu o unitate.

Cea de-a doua ipoteză pe care se fundamentează analiza factorială este aceea că în conținutul informațional al variabilelor aleatoare $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ se regăsesc informații cu privire la factorii comuni și unici, ceea ce înseamnă că ele pot fi folosite ca indicatori ai acestor factori, ca semnale informative generate de acești factori.

Având în vedere că la nivelul unei variabile indicator, nu se poate face, sub nici o formă, o distincție clară între factorul unic și factorul rezidual, din motive legate de simplificare și de crearea posibilităților de soluționare efectivă a problemei de analiză factorială, factorul rezidual este neglijat sau, ceea ce înseamnă același lucru, este unificat cu factorul unic. În consecință, modelul factorial capătă forma următoare:

$$\begin{aligned} x_1 &= a_{11} \cdot f_1 + a_{12} \cdot f_2 + \dots + a_{1p} \cdot f_p + a_1 \cdot u_1 \\ x_2 &= a_{21} \cdot f_1 + a_{22} \cdot f_2 + \dots + a_{2p} \cdot f_p + a_2 \cdot u_2 \\ &\dots \\ x_n &= a_{n1} \cdot f_1 + a_{n2} \cdot f_2 + \dots + a_{np} \cdot f_p + a_n \cdot u_n \end{aligned}$$

Dacă vom face notațiile următoare:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}, \quad f = \begin{pmatrix} f_1 \\ f_2 \\ \dots \\ f_p \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}, \quad F = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \dots & & & \\ a_{n1} & a_{n2} & \dots & a_{np} \end{pmatrix}, \quad D = \begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & a_n \end{pmatrix},$$

atunci modelul factorial poate fi scris sub forma următoare:

$$x = F \cdot f + D \cdot u.$$

În raport cu această ultimă formă a modelului factorial se definește conceptul de *configurație factorială*, concept care este folosit și într-un sens mai larg, cu referire la întregul set de ecuații care definește modelul.

În continuarea prezentării, vom nota matricea de corelație a variabilelor indicator cu **R**, respectiv:

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \dots & & & \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{pmatrix},$$

unde elementele diagonale sunt egale cu unitatea, respectiv $r_{ii} = 1$.

Deoarece, de regulă variabilele indicator sunt considerate a fi centrate, matricea de corelație a acestora coincide cu matricea de covarianță. În ceea ce privește cei n factori comuni, vom nota matricea de covarianță a acestora sub forma următoare:

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1p} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2p} \\ \dots & & & \\ \lambda_{p1} & \lambda_{p2} & \dots & \lambda_{pp} \end{pmatrix},$$

unde elementul diagonal λ_{kk} din această matrice reprezintă varianța factorului comun f_k , iar elementul nediagonal λ_{ij} reprezintă covarianța dintre factorul f_i și f_j . În anumite situații, pentru varianța factorului comun f_k se folosește notația λ_k , adică $\lambda_k = \lambda_{kk}$. Dacă factorii comuni sunt mărimi standardizate, atunci matricea de corelație a factorilor comuni coincide cu matricea de covarianță a acestor factori. În această situație, elementele diagonale sunt egale cu unitatea, respectiv $\lambda_k = \lambda_{kk} = 1$.

7.6 Descompunerea variabilității spațiului inițial

În mod similar cu analiza componentelor principale, analiza factorială își propune să reexprime variabilitatea conținută în spațiu cauzal inițial, într-o manieră diferențiată, în funcție de rolul pe care îl au în formarea acesteia factorii comuni, pe de o parte, și factorii unici, pe de altă parte.

În cadrul acestui paragraf, vom trata modul în care varianța unei variabile aleatoare poate fi descompusă în componente relevante din punct de vedere al interpretărilor interdependențelor cauzale.

7.6.1 Spațiul factor și exprimarea conținutului său informational

Prin utilizarea tehniciile de analiză multidimensională care au ca scop reducerea dimensionalității, variabilitatea spațiului cauzal n-dimensional, determinat de mulțimea de variabile indicator x_1, x_2, \dots, x_n , este conservată într-o proporție, mai mare sau mai mică, prin intermediul variabilității induse de un număr mai redus de factori abstracți f_1, f_2, \dots, f_p ($p < n$), care sunt factorii comuni. Împreună cu *factorul unic*, acești factori determină un spațiu $(p+1)$ -dimensional numit *spațiu test* sau *spațiu factor*.

Definiție: *Spațiu test* sau *spațiu factor* este un spațiu real, de dimensiune $(p+1)$, ale cărui axe sunt *ortogonale* două câte două și sunt reprezentate de *factorii comuni* f_1, f_2, \dots, f_p și de *factorul unic* u .

Variabilitatea ce caracterizează celor două spații implicate în analiză, *spațiu original* și *spațiu test*, este măsurată prin intermediul *varianței* sau *dispersiei*.

În analiza datelor se consideră că, pentru oricare dintre variabilele care definesc spațiul cauzal original, varianța este o măsură a informației conținută în observațiile efectuate asupra respectivei variabile.

Între gradul de variabilitate specific unei variabile și semnificația informațional-statistică produsă de observațiile făcute asupra respectivei variabile există un stabil raport de directă proporționalitate, considerându-se că o variabilă este cu atât mai semnificativă cu cât variabilitatea sa este mai mare.

7.6.2 Componentele varianței

După cum am mai evidențiat, spre deosebire de analiza componentelor principale, în contextul căreia variabilitatea ce caracterizează spațiu cauzal inițial este privită nediferențiat, în analiza factorială variabilitatea spațiului cauzal inițial este considerată a fi o compunere de variabilități, care se formează sub influența factorilor considerați.

Corespunzător celor trei categorii generale de factori considerați a influență variabila indicator x_j , respectiv *comuni*, *unici* și *reziduali*, în analiza factorială se postulează ipoteza conformă căreia varianța σ_j^2 , corespunzătoare variabilei x_j , poate fi împărțită în trei componente importante:

- componenta h_j^2 , numită *comunalitate*, care este acea parte a varianței totale σ_j^2 ce exprimă informația *comună* tuturor variabilelor care definesc spațiu cauzal n-dimensional și care se formează sub influența factorilor comuni f_1, f_2, \dots, f_p .
- componenta a_j^2 , numită *unicitate*, care este acea parte a varianței totale σ_j^2 ce exprimă informație *semnificativă* de natură *specifică*, caracteristică variabilei particulare x_j și care se formează sub influența factorului unic u_j ;
- componenta e_j^2 , numită *rezidualitate* sau *eroare*, care este acea parte a varianței totale σ_j^2 formată sub influența factorului rezidual asociat cu variabila respectivă și exprimă informație *nesemnificativă* cu caracter *specific* variabilei x_j .

Pe baza celor trei tipuri de influențe menționate anterior, varianța variabilei indicator x_j poate fi descompusă sub forma:

$$\text{Var}(x_j) = \text{Comunalitate} + \text{Unicitate} + \text{Rezidualitate}$$

sau, utilizând notațiile corespunzătoare, sub forma:

$$\sigma_j^2 = h_j^2 + a_j^2 + e_j^2 \quad j=1,2,\dots,n,$$

unde mărimile h_j^2 , a_j^2 și e_j^2 reprezintă, aşa cum o să arătăm în continuare, tot varianțe. Această relație definește *descompunerea varianței unei variabile indicator în funcție de varianțele celor trei categorii de factori* care influențează variabila respectivă. Deși cele trei componente ale descompunerii au natură de varianțe, ele nu trebuie identificate ca reprezentând varianțe ale factorilor, deoarece descompunerea este făcută doar în funcție de varianțele factorilor. Cu excepția celei de-a treia componente a descompunerii, care este chiar varianța factorului rezidual, primelor două componente nu li se poate atribui calitatea de varianțe ale factorilor.

Primele două componente sunt determinate de coeficienții care ponderează varianțele factorilor, ceea ce înseamnă că ele reprezintă *contribuții* ale varianțelor factorilor la formarea varianței variabilei indicator. De fapt, în condițiile unor ipoteze privind necorelarea factorilor, forma *completă* a relației de descompunere a varianței variabilei indicator x_j poate fi scrisă astfel:

$$\begin{aligned} \sigma_j^2 &= a_{j1}^2 \cdot \text{Var}(f_1) + a_{j2}^2 \cdot \text{Var}(f_2) + \dots + a_{jp}^2 \cdot \text{Var}(f_p) + a_j^2 \cdot \text{Var}(u_j) + \text{Var}(e_j) \\ &= a_{j1}^2 \cdot \lambda_1 + a_{j2}^2 \cdot \lambda_2 + \dots + a_{jp}^2 \cdot \lambda_p + a_j^2 \cdot v_j^2 + e_j^2 \end{aligned}$$

Este evident că pătratele coeficienților care ponderează varianțele factorilor reprezintă *contribuții* ale factorilor la formarea varianței variabilei indicator. Forma simplificată a descompunerii rezultă din impunerea ipotezei că varianțele factorilor comuni și unic sunt egale cu unitatea și din definirea *comunalității* ca sumă a contribuților factorilor comuni la formarea varianței indicator, respectiv:

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + \dots + a_{jp}^2.$$

Componenta cea mai importantă a varianței variabila indicator x_j este *comunalitatea*, care exprimă contribuția factorilor comuni la formarea varianței variabilei indicator și care poate fi privită, la rândul său, ca o varianță.

Definiție: *Comunalitatea* este acea parte a varianței unei variabile indicator, care exprimă variabilitatea indusă de influența factorilor comuni f_1, f_2, \dots, f_p .

Având în vedere că la formarea variabilității unei variabile indicator contribuie, în principiu, *toti* cei n factori comuni, *comunalitatea* poate fi descompusă, la rândul său, în raport cu cele n influențe.

Cea de-a două componentă a varianței variabilei indicator, numită *unicitate*, exprimă contribuția factorului unic la formarea varianței variabilei indicator și are, de asemenea, natură de varianță. Ea se mai numește și varianță unică, și se definește astfel:

Definiție: *Unicitatea* este acea parte a varianței unei variabile indicator, care exprimă variabilitatea indusă de influența factorului unic u și care nu poate fi explicată prin intermediul factorilor comuni.

Ultima componentă a varianței variabilei indicator, respectiv cea datorată factorului rezidual, reflectă influența erorilor de măsurare asupra formării variabilității variabilei indicator și se mai numește și **varianță reziduală** sau **varianță eroare**. De cele mai multe ori, această influență este considerată a fi neglijabilă în raport cu celelalte influențe.

Cu excepția varianței reziduale, care exprimă influența factorilor cu natură accidentală, a cauzelor *aleatorii, imprevizibile și nesemnificate*, celelalte două componente, *comunalitatea și unicitatea*, exprimă influențe de natură *sistemerică* asupra formării variabilității variabilei indicator, influențe cu caracter *permanent și stabil*. Din acest punct de vedere, varianța variabilei indicator poate fi privită ca fiind suma a două tipuri de varianțe: **varianța sistemerică și varianța reziduală**. Dacă vom nota varianța sistemerică cu g_j^2 , adică:

$$g_j^2 = h_j^2 + a_j^2,$$

atunci varianța variabilei indicator poate fi scrisă sub forma:

$$\sigma_j^2 = g_j^2 + e_j^2.$$

Din această reformulare rezultă că varianța variabilei indicator poate fi descompusă ca sumă a două componente numite *sistemericitate și rezidualitate*, respectiv:

$$\text{Var}(x_i) = \text{Sistemericitate} + \text{Rezidualitate}.$$

Pe de altă parte, influența factorului unic u_j și influența factorului rezidual e_j se caracterizează prin faptul că au o direcționalitate *specifică*, bine determinată, fiind raportate numai la o anumită variabilă indicator. Din acest punct de vedere, contribuția celor doi factori la formarea varianței variabilei indicator este numită **specificitate**, care se definește sub forma:

$$s_j^2 = a_j^2 + e_j^2.$$

Pe baza acestei noi redefiniri a componentelor, varianța variabilei indicator poate fi rescrisă sub forma următoare:

$$\sigma_j^2 = h_j^2 + s_j^2,$$

ceea ce înseamnă că:

$$\text{Var}(x_i) = \text{Comunalitate} + \text{Specificitate}.$$

Exemplul 7.3

Pentru a evidenția componentele prin intermediul cărora poate fi descompusă varianța variabilelor indicator, vom considera datele din exemplul 7.1 și vom efectua analiza factorială pe matricea de corelație următoare:

	Matematică	Informatică	Economie	Istorie	Engleză	
$R =$	1,00	0,86	0,77	0,15	0,30	Matematică
	0,86	1,00	0,73	0,09	0,24	Informatică
	0,77	0,73	1,00	0,23	0,39	Economie
	0,15	0,09	0,23	1,00	0,73	Istorie
	0,30	0,24	0,39	0,73	1,00	Engleză

Valorile proprii corespunzătoare matriciei de corelație evidențiază faptul că există doi factori comuni importanți, care explică un procent de 78,30% din varianța totală a variabilelor indicator. În tabelul următor sunt prezentate informațiile referitoare la contribuția varianțelor factorilor la formarea varianței totale a variabilelor indicator.

Contribuțiiile inițiale ale varianțelor factorilor la formarea varianței totale a variabilelor indicator

Tabelul 7.3

Factor comun	Valoare proprie	Varianță explicată		Procent varianță explicată	
		Individual	Cumulat	Individual	Cumulat
f_1	2,679102	2,358245	2,679102	53,58204	53,58204
f_2	1,235815	1,556673	3,914917	24,71630	78,29835
Diferență	-	1,085083	5,000000	21,70164	100,00000

După rotația structurii factor, efectuată în scopul asigurării unei interpretabilități mai corecte pentru cei doi factori, varianțele factorilor și contribuțiiile acestora la formarea varianței variabilelor indicator se modifică astfel:

Contribuțiiile modificate ale varianțelor factorilor la formarea varianței totale a variabilelor indicator

Tabelul 7.4

Factor comun	Valoare proprie	Varianță explicată		Procent varianță explicată	
		Individual	Cumulat	Individual	Cumulat
f_1	2,358245	2,358245	2,679102	47,16490	47,16490
f_2	1,556673	1,556673	3,914917	31,13346	78,29836
Diferență	-	1,085083	5,000000	21,70164	100,00000

În ceea ce privește descompunerea varianței variabilelor indicator pe componente, vom considera că cele două componente sunt reprezentate de comunalitate și specificitate, ceea ce înseamnă că presupunem o combinație a factorului unic cu factorul rezidual. Rezultatele

descompunerii varianței sunt prezentate în tabelul următor.

Descompunerea varianței variabilelor indicator

Tabelul 7.5

Variabila indicator	Intensități		Tipuri de varianțe		
	Inteligentă	Memorie	Totală	Comunalitate	Specificitate
Matematică	0,93797	0,13158	1,00000	0,89710	0,10290
Informatică	0,90689	0,06630	1,00000	0,82685	0,17315
Economie	0,78539	0,25461	1,00000	0,68167	0,31833
Istorie	0,05008	0,78158	1,00000	0,61338	0,38662
Engleză	0,19147	0,92696	1,00000	0,89592	0,10408
Varianță	2,358245	1,556673	5,00000	3,91492	1,08508

Pentru fiecare variabilă indicator, prima componentă a varianței, communalitatea, reprezintă suma pătratelor intensităților celor doi factori. De exemplu, pentru prima variabilă indicator, communalitatea se obține astfel:

$$h_1^2 = a_{11}^2 + a_{12}^2 = 0,93797^2 + 0,13158^2 = 0,89710.$$

Informațiile referitoare la specificitate, adică datele din ultima coloană, au fost determinate prin diferență între varianța fiecărei variabile și communalitatea corespunzătoare celor doi factori. De exemplu, pentru cazul primei variabile indicator, specificitatea se determină astfel:

$$s_1^2 = \sigma_1^2 - h_1^2 = 1,00000 - 0,89710 = 0,10290.$$

Facem precizarea că această componentă a varianței include influența combinată a *factorului unic* și *a factorului rezidual*. De asemenea, se poate verifica faptul că suma pătratelor intensităților care apar în coloana fiecărui factor comun reprezintă varianța factorului comun respectiv. În cazul primului factor, vom avea:

$$\begin{aligned} c_{f_1}^2 &= a_{11}^2 + a_{21}^2 + a_{31}^2 + a_{41}^2 + a_{51}^2 \\ &= 0,93797^2 + 0,90689^2 + 0,78539^2 + 0,05008^2 + 0,19147^2 = 2,358245 \end{aligned}$$

7.7 Configurația factor și structura factor

În analiza factorială se definesc două concepte fundamentale, care sintetizează o serie de mărimi importante, specifice analizei factoriale. Aceste mărimi sunt reprezentate de coeficienții factorilor și de coeficienții de corelație dintre variabilele indicator și factori, iar sintetizarea este făcută prin intermediul conceptelor numite *configurație factor* și *structură factor*.

7.7.1 Definirea configurației factor

Conceptul de *configurație factor* sau *configurație factorială* se referă la *intensitățile* factorilor comuni și poate fi definit astfel:

Definiție: Se numește *configurație factor* mulțimea intensităților corespunzătoare factorilor *comuni* ce apar într-un model factorial.

Dacă se ținea seama și de intervenția într-un model de analiză factorială a factorului specific, atunci poate fi definit conceptul de *configurație factor extinsă*.

Configurația factor este foarte importantă în cadrul analizei factoriale din mai multe puncte de vedere. În primul rând, elementele care intră în alcătuirea configurației factor caracterizează intensitatea și sensul influențelor exercitate de factori asupra formării nivelului variabilelor indicator. De asemenea, configurația factor, în varianta sa extinsă, este importantă deoarece ea se constituie ca o reprezentare simplificată și completă a modelului factorial.

Un alt element de importanță pentru configurația factorială rezultă din faptul că elementele acesteia stau la baza determinării elementelor structurii factor și pot fi utilizate în reproducerea, pe diferite nivele, a corelațiilor dintre variabilele indicator. În sfârșit, configurația factor este utilă în efectuarea unor comparații între diferite sisteme de factori, care sunt exprimabile prin intermediul aceleiași combinații de indicatori.

7.7.2 Definirea structurii factor

O problemă importantă a analizei factoriale, a cărei soluționare ține de esență analizei factoriale, este aceea a determinării corelațiilor dintre variabilele indicator și factorii comuni, pe de o parte, și dintre variabilele indicator și factorii specifici, pe de altă parte.

Pe baza conceptului de corelație între variabilele indicator și factori, poate fi definit cel de-al doilea concept, *structura factor* sau *structura factorială*. Structura factor mai este cunoscută în teoria destinată analizei factoriale și sub numele de *matrice factor*. Matricea factor are aceeași interpretare cu cea dată în cazul componentelor principale.

Definiție: Se numește *structură factor* sau *structură factorială* mulțimea coeficienților de corelație dintre variabilele indicator ale unui model factorial și factorii comuni incluși în acest model.

Importanța structurii factor în cadrul analizei factoriale constă în aceea că ea evidențiază corelațiile existente între

variabilele indicator și factorii comuni și facilitează interpretarea esenței factorilor comuni, prin prisma naturii pe care o au variabilele indicator. Spre deosebire de configurația factorială care sintetizează exprimarea variabilelor indicatori sub formă de combinații liniare de factorii comuni și specifici, structura factorială evidențiază schema legăturilor existente între indicatori și factori.

Structura factorială este considerată în cadrul analizei factoriale ca fiind una dintre numeroasele soluții posibile pentru o anumită problemă concretă. De altfel, structura factorială mai este cunoscută în literatura de specialitate și sub numele de **soluție factor sau soluție factorială**.

În numeroase situații practice, soluția factorială obținută inițial nu întrunește toate condițiile necesare pentru a permite o interpretare facilă a factorilor. De multe ori, se poate întâmpla, ca elementele structurii factor să sugereze interpretări ale factorilor care să vină în contradicție cu posibila esență ce poate fi atribuită, în mod logic, factorilor.

În aceste situații este necesară găsirea altor soluții factor, care să permită o mai bună interpretare și o mai consistentă interpretare a factorilor. În scopul creșterii posibilităților de interpretare și a consistenței acestor interpretări, structura factor poate fi supusă unei procese de rotație, în urma căruia rezultă *structura factor rotată*.

7.8 Calculul scorurilor factoriale

O anumită observație, corespunzătoare unui factor dat, este determinată sub forma unui scor corespunzător respectivului factor, scor format pe baza contribuției variabilelor originale. Exprimarea generică a scorurilor pentru un anumit factor în funcție de variabilele originale este dată de următoarea relație:

$$f_i = b_{i1}x_1 + b_{i2}x_2 + \dots + b_{in}x_n \quad i=1,2,\dots,p,$$

unde b_{ij} reprezintă coeficienții scorurilor factor și sunt elemente ale transpusenei matricii factor F. Sub formă matricială această relație poate fi scrisă astfel:

$$\mathbf{f} = \mathbf{F}^t \mathbf{x}.$$

În mod practic, exprimarea celor T observații efectuate asupra variabilelor originale sub forma **scorurilor factor**, respectiv calculul concret al scorurilor factor, se bazează pe următoarele relații:

$$z_{kj} = \sum_{i=1}^n b_{ki}x_{ij} \quad k=1,2,\dots,p; \quad j=1,2,\dots,T,$$

unde z_{kj} reprezintă scorurile factorilor, b_{ki} este elementul din linia k și coloana i a transpusenei matricii factor, iar x_{ij} este cea de-a j-a observație efectuată asupra celei de-a i-a variabile originale.

Dacă vom considera matricea de observații X, ale cărei linii reprezintă cele n variabile originale și ale cărei coloane reprezintă cele T observații făcute asupra acestor variabile, ca având forma următoare:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1T} \\ x_{21} & x_{22} & \dots & x_{2T} \\ \dots \\ x_{n1} & x_{n2} & \dots & x_{nT} \end{pmatrix},$$

și matricea factor F definită mai sus, atunci matricea Z de dimensiune $p \times T$ definită astfel:

$$\mathbf{Z} = \mathbf{F}^t \mathbf{X},$$

se numește **matricea scorurilor factor**. Scorurile factor determinate în acest fel pot fi folosite în analize ulterioare, în locul valorilor variabilelor originale.

Exemplul 7.4

În cadrul unei cercetări având ca obiectiv determinarea nivelului de dezvoltare economico-socială a 10 zone geografice din țară, a fost selectat un număr de 6 indicatori de natură economico-socială: *capitalul industrial*, *cifra de afaceri a firmelor*, *profiturile obținute*, *cheltuielile pentru învățământ*, *cheltuielile pentru cultură* și *cheltuielile pentru sănătate*. Valorile înregistrate, la nivelul unui an, de cele 10 zone geografice la cei șase indicatori sunt cuprinse în tabelul următor.

Indicatori economico-sociali ai zonelor geografice

Tabelul 7.6

Zona	Capital industrial	Cifra afaceri	Profituri	Cheltuieli învățământ	Cheltuieli cultură	Cheltuieli sănătate
	K _i	C _a	P _r	C _i	C _c	C _s
Z ₁	2148,900	1210,550	545,600	588,560	257,860	501,650
Z ₂	1056,550	1213,010	531,790	1045,790	424,770	631,150
Z ₃	2198,990	1088,890	565,130	956,020	418,230	698,470
Z ₄	2632,350	1820,490	670,400	925,580	292,300	679,330
Z ₅	1636,510	1067,480	459,570	635,510	263,940	451,970
Z ₆	2267,880	1647,610	554,970	737,490	255,550	520,530
Z ₇	1906,490	1853,010	855,100	656,820	290,810	628,740

Zona	Capital industrial	Cifra afaceri	Profituri	Cheltuieli învațământ	Cheltuieli cultură	Cheltuieli sănătate
	K _i	C _a	P _r	C _i	C _c	C _s
Z ₈	2756,080	1708,620	724,730	754,030	320,480	454,930
Z ₉	1299,630	1077,580	401,760	768,900	205,690	592,250
Z ₁₀	1160,220	1059,650	305,210	459,540	206,590	477,780
Media	1906,361	1374,690	561,427	752,824	293,622	563,680
Abatere std	600,378	338,329	158,985	180,565	76,285	93,416

Matricea de corelație corespunzătoare valorilor celor 6 variabile indicator din tabel, este următoarea:

$$R = \begin{pmatrix} K_i & C_a & P_r & C_i & C_c & C_s \\ 1,000 & 0,666 & 0,644 & 0,136 & 0,123 & 0,039 & K_i \\ 0,666 & 1,000 & 0,842 & 0,132 & 0,036 & 0,170 & C_a \\ 0,644 & 0,842 & 1,000 & 0,271 & 0,368 & 0,318 & P_r \\ 0,136 & 0,132 & 0,271 & 1,000 & 0,789 & 0,727 & C_i \\ 0,123 & 0,036 & 0,368 & 0,789 & 1,000 & 0,538 & C_c \\ 0,039 & 0,170 & 0,318 & 0,727 & 0,538 & 1,000 & C_s \end{pmatrix}.$$

Matricea de corelație are primele două valori proprii $\lambda_1=2,72029$ și $\lambda_2=1,66104$, ceea ce înseamnă că există doi factori comuni asociați acestor valori proprii, factori care justifică un procent de 73,02% din varianța variabilelor indicator.

În tabelul următor sunt prezentate informații cu privire la modul în care varianța totală a variabilelor indicator poate fi explicată prin intermediul a doi factori comuni.

Contribuțiiile inițiale ale varianțelor factorilor la formarea varianței totale a variabilelor indicator

Tabelul 7.7

Factor comun	Valoare proprie	Varianță explicată		Procent varianță explicată	
		Individual	Cumulat	Individual	Cumulat
f ₁	2,72029	2,72029	2,72029	45,3381	45,3381
f ₂	1,66104	1,66104	4,38132	27,6839	73,0220
Diferență	-	1,61868	6,00000	26,79797	100,00000

În tabelul următor sunt prezentate informații referitoare la o primă soluție obținută din aplicarea analizei factoriale. Tabelul conține informații cu privire la intensitățile factorilor, la descompunerea varianței variabilelor indicator între factorii comuni și factorul specific, dprecum și la coeficientul de corelație multiplă dintre fiecare variabilă indicator și cei doi factori comuni.

Coeficienții factorilor comuni și compoziția varianței variabilelor indicator

Tabelul 7.8

Indicatori	Coeficienții factorilor	Totală	Varianță			R ²		
			E	S	E + S			
K _i	0,60126	-0,37631	1,00000	0,36152	0,14161	0,50313	0,49687	0,50461
C _a	0,76731	-0,53782	1,00000	0,58876	0,28925	0,87801	0,12199	0,84882
P _r	0,85686	-0,34414	1,00000	0,73421	0,11843	0,85264	0,14736	0,87266
C _i	0,63448	0,69868	1,00000	0,40256	0,48815	0,89072	0,10928	0,81838
C _c	0,56859	0,61254	1,00000	0,32329	0,37521	0,69850	0,30150	0,80812
C _s	0,55673	0,49837	1,00000	0,30995	0,24837	0,55832	0,44168	0,60667
Total	-	-	6,00000	2,72029	1,66103	4,38132	1,61868	
Procent	-	-	100,0%	45,34%	27,68%	73,02%	26,98%	

Având în vedere natura variabilelor indicator utilizate și rezultatele obținute, se poate trage concluzia că cei doi factori comuni pot fi interpretați ca reprezentând *gradul de dezvoltare economică* (E) și *gradul de dezvoltare socială* (S). Configurația factorială și structura factor, corespunzătoare soluției inițiale, sunt prezentate în tabelul următor.

Descompunerea varianței variabilelor indicator

Tabelul 7.9

Variabila indicator	Configurația factor		Structura factor		Varianță comună	
	E	S	E	S	E	S
K _i	0,60126	-0,37631	0,60126	-0,37631	0,36152	0,14161
C _a	0,76731	-0,53782	0,76731	-0,53782	0,58876	0,28925
P _r	0,85686	-0,34414	0,85686	-0,34414	0,73421	0,11843

C_i	0,63448	0,69868	0,63448	0,69868	0,40256	0,48815
C_e	0,56859	0,61254	0,56859	0,61254	0,32329	0,37521
C_s	0,55673	0,49837	0,55673	0,49837	0,30995	0,24837
Varianță comună totală						2,72029
						1,66103

Deoarece elementele configurației factoriale rezultate din analiză, configurație reprezentată de valorile din coloanele 2 și 3 ale tabelului, au semn care vin în contradicție cu o interpretare convenabilă a celor doi factori, în termeni de grad de dezvoltare economică și grad de dezvoltare socială, este necesară obținerea unei alte soluții factor, care să permită o mai bună interpretare a factorilor. Această nouă soluție factor poate fi obținută prin efectuarea unei rotații a structurii factor. Rezultatele obținute în urma rotației structurii factor sunt prezentate în tabelul următor.

Coeficienții factorilor comuni și compoziția varianței variabilelor indicator după rotația structurii factor

Tabelul 7.10

Indicatori	Coeficienții factorilor	Totală	Varianță			Specifică	R^2
			E	S	E + S		
K_i	0,70639	0,06434	1,00000	0,49899	0,00414	0,50313	0,49687 0,50461
C_a	0,93632	0,03627	1,00000	0,87669	0,00132	0,87801	0,12199 0,84882
P_r	0,89039	0,24465	1,00000	0,79279	0,05985	0,85264	0,14736 0,87266
C_i	0,08221	0,94019	1,00000	0,00676	0,88396	0,89072	0,10928 0,81838
C_e	0,08190	0,83174	1,00000	0,00671	0,69179	0,69850	0,30150 0,80812
C_s	0,14155	0,73368	1,00000	0,02004	0,53829	0,55832	0,44168 0,60667
Total	-		6,00000	2,20198	2,17935	4,38190	1,61868
Procent	-		100,00	36,70%	36,32%	73,02%	26,98%

Analiza soluției obținute în urma rotației structurii factor, evidențiază că primul factor comun este puternic corelat cu primele trei variabile indicator, iar cel de-al doilea factor este puternic corelat cu ultimele trei variabile indicator. În ambele situații, coeficienții de corelație au valori pozitive și mai mari decât 0,70, justificând ideea de corelație puternică.

Faptul că primele trei variabile indicator, respectiv *capitalul industrial*, *cifra de afaceri și profiturile*, sunt de natură economică, iar ultimele trei variabile indicator, respectiv *cheltuielile pentru învățământ*, *cheltuielile pentru cultură și cheltuielile pentru sănătate*, sunt de natură socială, permite ca primului factor comun să i se atribuie semnificația de *factor economic*, iar celui de-al doilea factor comun să i se atribuie semnificația de *factor social*. Corelarea foarte puternică dintre cele două grupe de variabile indicator și cei doi factori comuni constituie un temei pentru a considera că, într-adevăr, primul factor comun reprezintă *gradul de dezvoltare economică*, iar cel de-al doilea factor comun reprezintă *gradul de dezvoltare socială*.

În afara celor doi factori comuni, nivelul variabilelor indicator se formează și sub influența factorilor specifici. Măsura în care factorii specifici influențează variabilele indicator este reflectată de mărimea varianței specifice. Modelul factorial, corespunzător informațiilor definite anterior, este reprezentat de următoarele ecuații factoriale:

$$K_i = a_{K_E} \cdot E + a_{K_S} \cdot S + a_{K_U} \cdot U_{K_i}; \quad C_a = a_{C_E} \cdot E + a_{C_S} \cdot S + a_{C_U} \cdot U_{C_a}; \quad P_r = a_{P_E} \cdot E + a_{P_S} \cdot S + a_{P_U} \cdot U_{P_r}$$

$$C_i = a_{C_E} \cdot E + a_{C_S} \cdot S + a_{C_U} \cdot U_{C_i}; \quad C_e = a_{C_E} \cdot E + a_{C_S} \cdot S + a_{C_U} \cdot U_{C_e}; \quad C_s = a_{C_E} \cdot E + a_{C_S} \cdot S + a_{C_U} \cdot U_{C_s}$$

Efectuând analiza factorială pe datele conținute în tabelul de mai sus, am obținut următoarea formă estimată a ecuațiilor modelului factorial:

$$K_i = 0,706 \cdot E + 0,064 \cdot S + 0,705 \cdot U_{K_i}; \quad C_a = 0,936 \cdot E + 0,036 \cdot S + 0,349 \cdot U_{C_a}; \quad P_r = 0,890 \cdot E + 0,245 \cdot S + 0,384 \cdot U_{P_r}$$

$$C_i = 0,082 \cdot E + 0,940 \cdot S + 0,331 \cdot U_{C_i}; \quad C_e = 0,082 \cdot E + 0,832 \cdot S + 0,549 \cdot U_{C_e}; \quad C_s = 0,142 \cdot E + 0,734 \cdot S + 0,665 \cdot U_{C_s}$$

În graficul din figura următoare sunt prezentate legăturile dintre variabilele indicator și factori și sunt evidențiate intensitățile cu care aceste legături se manifestă.

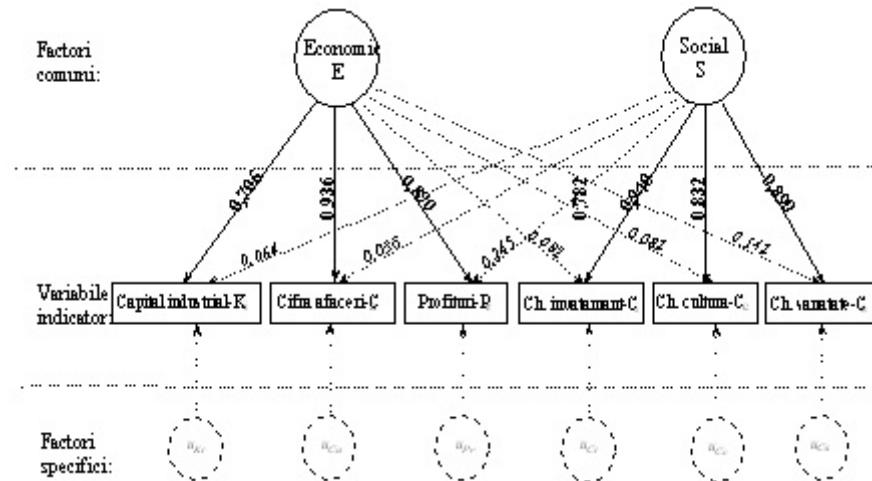


Figura 7.3: Exemplificarea legăturilor dintre indicatori și factori

Identificarea celor doi factori comuni, evaluarea influenței lor asupra variabilelor indicator și evaluarea corelațiilor existente între factorii comuni și variabilele indicator, permit reprezentarea grafică a variabilelor indicator în sistemul de axe factoriale. Această reprezentare este utilă deoarece ea oferă o imagine cu privire la modul în care variabilele indicator se asociază între ele, pe de o parte, și cu factorii comuni, pe de altă parte. În figura următoare se găsește reprezentarea grafică a variabilelor indicator, în sistemul de axe ce corespunde celor doi factori comuni. Reprezentarea grafică este suficient de sugestivă pentru a observa și a înțelege modul în care se asociază variabilele indicator, asociere determinată de legătura lor cu factorii comuni.

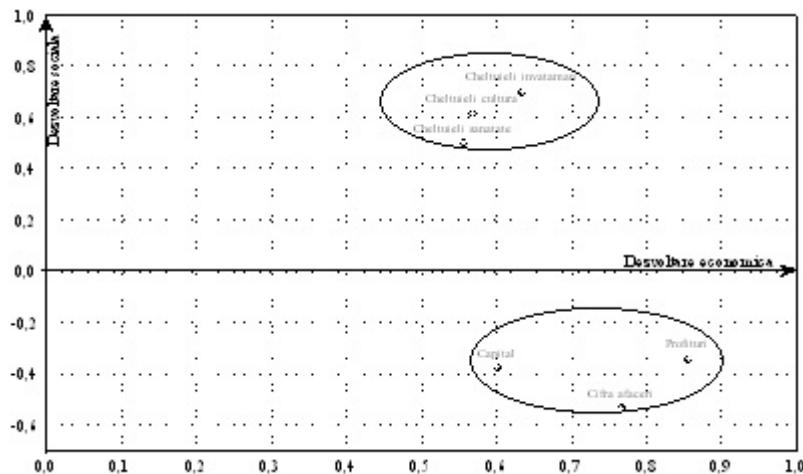


Figura 7.4: Exemplificarea legăturilor indicatori și factorii comuni și specifici

În figura următoare sunt vizualizate grafic pozițiile pe care le au cele 10 zone geografice în raport cu cele două axe factoriale, care au semnificația de *dezvoltare economică*, respectiv de *dezvoltare socială*.

O astfel de vizualizare a entităților de tip *obiect* este extrem de utilă deoarece ea evidențiază foarte clar valorile pe care le înregistrează obiectele la caracteristicile latente, neobservabile, caracteristici reprezentabile prin intermediul de factorilor comuni. Pe baza acestor valori, obiectele analizate pot fi supuse unor aprecieri de natură *globală* și sunt create premize pentru efectuarea unor *comparații multicriteriale*.

Pentru cazul exemplului considerat, din reprezentarea grafică rezultă că cea mai bine situată din punct de vedere economic este "Zona 4", că "Zona 2" și "Zona 3" stau foarte bine din punct de vedere social, după cum "Zona 10" este cea mai vitregită, atât din punct de vedere economic, cât și social.

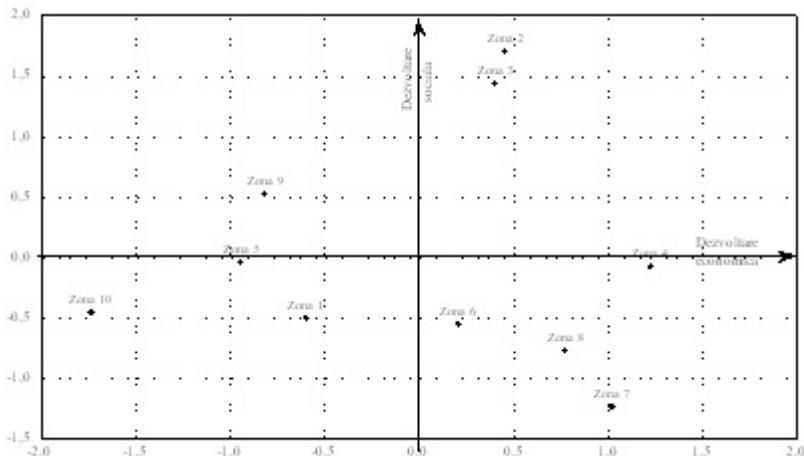


Figura 7.5: Exemplificarea legăturilor indicatori - factorii comuni și specifici

Potibilitățile pe care le oferă analiza factorială în acest sens sunt cu atât mai utile, cu cât efectuarea unor astfel de evaluări și de aprecieri nu putea fi făcută pe baza variabilelor indicator. În situația în care structura factor este supusă unei rotații care nu mai conservă ortogonalitatea axelor, se obțin factori comuni care sunt corelați, iar structura factor rezultată diferă de configurația factor.

Pentru datele din acest exemplu, efectuarea unei rotații oblice a axelor a condus la obținerea a doi factori comuni care sunt corelați la nivelul unui coeficient de 0,254037, adică:

$$\lambda_{12} = \lambda_{21} = 0,254037.$$

Aceasta înseamnă că axele factoriale nu mai sunt ortogonale, ele formând un unghi ascuțit Θ , a cărui valoare este:

$$\Theta = \arccos(0,254037) = 75,283^\circ.$$

În această situație, communalitatea fiecărei variabile indicator este definită de relația următoare:

$$h_i^2 = a_{i1}^2 + 2 \cdot a_{i1} \cdot a_{i2} \cdot \lambda_{12} + a_{i2}^2.$$

Rezultatele obținute în urma efectuării rotației de acest tip, adică a rotației bazate pe oblicitatea axelor, sunt prezentate în tabelul următor.

Coeficienții factorilor comuni și componența varianței
variabilelor indicator după *rotația neortogonală* a structurii factor

Tabelul 7.11

Indicatori	Coeficienții factorilor	Varianță						Specifică	
		Totală	Comună			Total			
			E	S	E&S				
K_i	0,61824	-0,02381	1,00000	0,38223	0,00057	-0,00748	0,37532	0,62468	
C_a	0,82509	-0,07497	1,00000	0,68077	0,00562	-0,03143	0,65497	0,34503	
P_r	0,76058	0,11483	1,00000	0,57848	0,01319	0,04438	0,63604	0,36396	
C_i	-0,03472	0,82326	1,00000	0,00120	0,67776	-0,01452	0,66445	0,33555	
C_e	-0,02259	0,72725	1,00000	0,00051	0,52889	-0,00835	0,52105	0,47895	
C_s	0,04146	0,63358	1,00000	0,00172	0,40142	0,01335	0,41649	0,58351	
Total	-	6,00000	1,64491	1,62745	-0,00406	3,26830	2,73170		
Procent	-	100,00	27,42%	27,12%	-0,068%	54,47%	45,53%		

Deoarece factorii comuni sunt corelați, structura factorială este diferită de configurația factorială. În tabelul următor sunt prezentate informațiile referitoare la configurația factor și la structura factor, corespunzătoare rezultatelor anterioare.

Configurația factor și structura factor

Tabelul 7.12

Indicatori	Configurația factor		Structura factor	
K_i	0,70639	0,06434	0,61824	-0,02381
C_a	0,93632	0,03627	0,82509	-0,07497
P_r	0,89039	0,24465	0,76058	0,11483

C_i	0,08221	0,94019	-0,03472	0,82326
C_e	0,08190	0,83174	-0,02259	0,72725
C_s	0,14155	0,73368	0,04146	0,63358

7.9 Criterii de alegere a numărului de factori

Utilizarea analizei factoriale pentru rezolvarea unor probleme specifice presupune și determinarea numărului de factori comuni ce vor fi reținuți în model. Deși decizia de a reține un anumit număr de factori este, în principiu, subiectivă, există o serie de criterii care pot să orienteze utilizatorul atunci când ia o astfel de decizie.

7.9.1 Criteriul procentului de acoperire

În general, alegerea numărului de factori care să fie incluși în modelul factorial depinde de proporția din variabilitatea comună conținută în spațiul cauzal inițial pe care utilizatorul dorește să o exprime prin intermediul unei succesiuni de factori comuni. O estimare aproximativă a acestei proporții, pentru cazul în care numărul de factori reținuți este egal cu k , poate fi obținută cu ajutorul formulei:

$$p_k = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i},$$

unde k reprezintă numărul de factori reținuți în model, n reprezintă numărul variabilelor originale, iar λ_i reprezintă valoarea proprie în raport cu care este definit factorul comun i .

Dezavantajul major al utilizării unei astfel de aproximății în problemele de analiză factorială este dat de faptul că mărimea p_k arată care este ponderea varianței primelor k componente principale în varianță totală și nu ponderea varianței explicate de primii k factori comuni în varianța spațiului test; acest lucru reprezintă un inconvenient deoarece între componentele principale și factorii comuni există, așa cum am mai arătat, o deosebire de esență.

7.9.2 Criteriul lui Kaiser

Acest criteriu poate fi folosit atunci când analiza factorială este efectuată pe o matrice de corelație, adică atunci când se presupune că variabilele originale sunt standardizate. În conformitate cu acest criteriu, *numărul de factori necesari a fi incluși într-un model de analiză factorială este egal cu numărul de valori proprii mai mari sau egale cu 1*.

Justificarea acestui criteriu este dată de faptul că, pentru analiză, prezintă importanță numai acei factori comuni a căror varianță este cel puțin egală cu varianța variabilelor originale, variabile care, fiind normalizate, au varianță unitară.

În afara faptului că un astfel de criteriu poate fi folosit numai în cazul în care se lucrează cu variabile normalize, dezavantajul principal al criteriului Kaiser este dat de faptul că aplicarea lui conduce la reținerea în model a unui număr prea mare de factori.

7.9.3 Criteriul "granulozității"

După acest criteriu, numărul de factori ce vor fi reținuți în modelul de analiză factorială se stabilește pe baza unei analize grafice a valorilor proprii. Graficul pe care se face analiza se construiește luând în abscisă numărul de ordine al valorilor proprii, iar în ordonată valorile acestor valori proprii.

Faptul că, așa cum știm, valorile proprii sunt ordonate după magnitudinea lor descrescătoare va face ca graficul să aibă forma aproximativă a unei curbe de tipul exponențialei negative.

Numărul de factori ce se vor reține în model este determinat de punctul de pe grafic în dreapta căruia panta curbei devine neglijabilă, numărul de ordine al valorii proprii corespunzătoare acestui punct determinând numărul de factori ce se vor reține.

Dezavantajul acestui criteriu constă în faptul că aplicarea sa conduce la reținerea în modelul analizei factoriale a unui număr prea mic de factori comuni.

În practică, alegerea unui anumit număr de factori comuni este puternic dependentă de natura problemei analizate și de scopul urmărit în utilizarea analizei factoriale. De multe ori, scopurile practice urmărite în analize de acest fel pot conduce la necesitatea de a obține un singur factor sau cel mult doi. Construirea unui model cu unul sau doi factori comuni are și avantajul, care nu este de neglijat, că facilitează reprezentarea grafică a mărimilor analizei factoriale, reprezentare care poate să fie deosebit de utilă în fază de interpretare a rezultatelor.

Tema 8. Metode și tehnici de recunoaștere a formelor

8.1 Importanța și necesitatea recunoașterii formelor

În cele mai multe dintre activitățile umane apare necesitatea de a încadra, de a diferenția, de a grupa sau de a clasifica anumite entități sau obiecte sub forma unor categorii sau clase, a căror delimitare trebuie să fie foarte clară și foarte naturală. Semnificația concretă a acestor categorii trebuie să aibă o echivalență corespunzătoare în realitatea studiată, să fie consistentă și relevantă pentru procesul de cunoaștere, să aibă un anumit grad de generalitate și să ofere o interpretabilitate simplă și naturală.

Diferențierea obiectelor pe categorii sau clase se face în funcție de proprietățile fundamentale ale obiectelor, iar criteriile de asociere a obiectelor sub formă de clase au la bază gradul de asemănare a proprietăților respectivelor obiecte, măsurat în funcție de magnitudinea valorilor acestor proprietăți.

Necesitatea de a grupa sau clasifica obiecte apare foarte frecvent și în domeniile foarte variate ale cunoașterii și activității umane, cum ar fi: analiza financiară, marketingul, asigurările, informatica, biologia, medicina, arheologia, meteorologia, criminalistica, psihologia, știința politică sau domeniul militar.

În domeniul informaticii, cerințele legate de creșterea performanțelor în utilizarea tehnicii de calcul au determinat necesitatea dezvoltării și implementării unor dispozitive hardware și a unor instrumente software pentru recunoașterea vocii umane și a scrierii de mâna. De asemenea, în domeniul economic, manifestarea comportamentului rațional al agenților economici face necesară existența posibilității de a identifica, de exemplu, activitățile rentabile, clienții solvabili sau piețele potențiale. Cele mai potrivite și cele mai eficiente instrumente utilizabile pentru soluționarea problemelor de acest fel s-au dovedit a fi metodele și tehnicele de clasificare sau de recunoaștere a formelor.

În general, oamenii dispun de o serie de simțuri naturale, cum ar fi auzul, văzul, miroslul, pipăitul etc., simțuri care le permit acestora să percepă anumite proprietăți ale obiectelor pe care le analizează și, pe această cale, să poată structura, clasifica sau ierarhiza aceste obiecte sub forma unor submulțimi specifice și distințe.

Într-o activitate de clasificare, oamenii se pot folosi, în afara simțurilor naturale pe care le posedă, și de cunoștințele pe care le au cu privire la obiectele pe care trebuie să le clasifice sau cu privire la proprietățile acestor obiecte. În plus față de acestea, oamenii mai pot apela și la o serie de modele, instrumente și tehnici speciale, de natură statistică-matematică, cu ajutorul cărora să poată clasifica mai ușor și mai corect obiectele analizate.

Poate că simțurile naturale pe care le au ca ființe umane și a cunoștințelor de care dispun, oamenii pot diferenția, clasifica sau grupa cu relativă ușurință cele mai variate categorii de obiecte. Există însă numeroase situații în care simțurile naturale ale indivizilor și informațiile de care aceștia dispun nu mai sunt suficiente pentru a putea discrimina corect între anumite obiecte sau pentru a face clasificări corecte ale acestor obiecte. Aceste situații sunt întâlnite în cazul obiectelor de tip multidimensional, adică în cazul obiectelor cu mai multe caracteristici, mai ales dacă aceste caracteristici sunt exprimabile sub o formă numerică, iar obiectele care trebuie clasificate sunt în număr foarte mare.

În cazul în care obiectele sunt de tip multidimensional, cu caracteristici numeroase și preponderent cantitative, diferențierea acestora pe categorii specifice nu se poate face numai pe cale intuitivă, exclusiv pe baza simțurilor naturale, fiind necesar să se apeleze la o serie de metode și tehnici specifice, de mare complexitate și cu un solid fundament statistică-matematică.

8.2 Definirea recunoașterii formelor

Activitățile științifice care au ca scop diferențierea și structurarea unor mulțimi de obiecte pe categorii sau clase specifice, în funcție de proprietățile fundamentale ale obiectelor, sunt cunoscute sub denumiri variate, cum ar fi: *clasificare, clusterizare, grupare sau discriminare*.

Aceste denumiri sunt folosite în literatura de specialitate a domeniului cu sens terminologic diferit, în funcție de tipul școlii de care aparține respectiva literatură. De exemplu, în literatura anglo-saxonă termenul de *clusterizare* este folosit pentru a descrie activitățile de clasificare necontrolată, iar termenul de *clasificare* este folosit pentru a defini activitățile specifice sistemelor de recunoaștere controlată. În același timp, în literatura franceză termenul de *clasificare* este folosit pentru a face referire la clasificarea de tip necontrolat, iar termenul de *discriminare* este folosit pentru descrierea activităților de tip controlat.

În cadrul acestei teme, vom utiliza termenul de *clasificare* în două accepțiuni. Prima accepțiune va fi folosită cu sens generic, acoperind atât conținutul conceptului de clasificare controlată, cât și conținutul conceptului de clasificare necontrolată. Cea de-a doua accepțiune va fi folosită cu sens particular, cu referire la clasificarea de tip necontrolat. Pentru clasificarea de tip necontrolat vom folosi însă și termenii de *clusterizare* și de *grupare*. De asemenea, pentru referire la clasificarea de tip controlat vom utiliza, aproape exclusiv, termenul de *discriminare*.

În general, putem spune că *discriminarea și clusterizarea* reprezintă activitățile de *aranjare* sau de *asociere* a unor obiecte, indivizi sau observații, sub forma unor grupe, categorii sau clase, în funcție de gradul de *asemănare* sau de *deosebire* dintre acestea.

Totalitatea tehniciilor de clasificare, adică de discriminare și de clusterizare, este cunoscută și sub numele generic de *teoria recunoașterii formelor*.

Metodele și tehnicele de clasificare, respectiv teoria recunoașterii formelor, reprezintă una dintre cele mai moderne și interesante probleme ale gândirii științifice contemporane, constituind un domeniu științific căruia i se acordă o importanță din ce în ce mai mare, iar aplicațiile acestora sunt din ce în ce mai numeroase și mai variate.

Domeniu bine conturat al inteligenței artificiale, cu tendință din ce în ce mai accentuată de autonomizare, recunoașterea formelor s-a dezvoltat în strânsă interdependență cu progresele înregistrate în domeniul tehniciei de calcul și informaticii, ultimele decenii fiind caracterizate printr-o dezvoltare explozivă a tehnicilor de acest fel.

Scopul general al teoriei recunoașterii formelor îl reprezintă *identificarea* la nivelul unor mulțimi complexe și eterogene de forme sau obiecte a unor *structuri, grupări, clase* sau *clustere* existente la nivel latent în cadrul acestor mulțimi și care se conturează în mod natural, în funcție de asemănările și deosebirile existente între elementele acestor mulțimi.

Dezvoltarea tehnicilor de recunoaștere a formelor poate fi privită atât ca proces independent, impus de necesitatea adâncirii cunoașterii din diverse domenii de activitate și stimulat de evoluția tehniciei de calcul, cât și ca un răspuns la necesitățile de simplificare și perfecționare a schimbului informațional om-mașină.

Având aplicații în cele mai diverse și surprinzătoare domenii de activitate, teoria recunoașterii formelor a cunoscut în ultimii ani progrese cu adevărat uimitoare, ajungându-se în prezent la un nivel de performanță tehnico-științifică care permite chiar construirea unor mașini specializate în recunoașterea anumitor tipuri de forme.

Aflată la intersecția unor domenii fundamentale cum sunt știința calculatoarelor, teoria informației, teoria deciziei, geometria, teoria probabilităților și statistica matematică, recunoașterea formelor cunoaște în prezent aplicații a căror paletă se întinde de la cercetarea antropologică și până la proiectarea hardware și software.

În domeniul economico-social teoria recunoașterii formelor își găsește o largă utilizare mai ales în procesul de *analiză a datelor* și în activitatea de *predicție*. Problema clasificării unei mulțimi de obiecte este o problemă standard, frecvent întâlnită în investigarea socio-economică, iar abordarea ei presupune utilizarea metodelor și tehnicilor specifice teoriei recunoașterii formelor.

Numerouse probleme din domeniul analizei datelor, începând cu cele legate de *identificarea caracteristicilor definitorii* pentru cele mai diverse categorii de fenomene și terminând cu cele legate de *delimitarea funcțională, ierarhizarea structurală* sau *sintetizarea informațională* a unor mulțimi de fenomene și procese economico-sociale, își găsesc rezolvarea prin intermediul aplicării unor concepte și instrumente a căror paternitate este, indiscutabil, legată de teoria recunoașterii formelor.

Metodele și tehnicele aparținând teoriei recunoașterii formelor sunt de neînlocuit în analizele care operează cu cantități mari de informație, unde necesitatea de a *esențializa* și *sintetiza* interdependențele implică un proces continuu de clasificare și structurare a informațiilor. Practic, tehnicele de recunoaștere a formelor reprezintă instrumente cu ajutorul cărora poate fi stăpânită și controlată marea complexitate informațională ce caracterizează fenomenele și procesele economico-sociale.

O utilizare și mai largă a teoriei recunoașterii formelor este întâlnită în domeniul predicțiilor. Activitatea de realizare a predicțiilor poate fi privită ca un proces ale cărui caracteristici sunt foarte apropiate, mergând chiar până la identificare, de caracteristicile specifice ale unui proces de recunoaștere a formelor.

Evaluarea stărilor pe care le poate avea în viitor un fenomen aparținând unei realități date reprezentă, de fapt, un proces de recunoaștere a celor forme de evoluție a fenomenului care au cea mai mare probabilitate de realizare. Mai mult, atât în activitatea de predicție, cât și în procesul de clasificare sau de recunoaștere a formelor, modalitățile de abordare au o natură preponderent probabilistică. Pe de altă parte, problema recunoașterii formelor este, ea însăși, o problemă de predicție în care, pornind de la anumite caracteristici ale obiectelor analizate, obiecte numite și *forme*, se fac predicții cu privire la apartenența acestor obiecte la anumite clase. De altfel, stabilirea apartenenței formelor la anumite clase reprezintă scopul principal al utilizării tehnicilor de recunoaștere a formelor.

Semnificativ pentru legătura dintre activitatea de predicție și teoria recunoașterii formelor este faptul că, în prezent, cele mai moderne metode și tehnici din domeniul predicției sunt cele bazate pe o nouă clasă de modele, specifice conturării unei noi modalități de abordare în domeniul teoriei recunoașterii formelor, numite *rețele neuronale*. Modalitățile de abordare științifică bazate pe rețele neuronale sunt mult mai concordante cu pronunțata complexitate și imprevizibilitate ce caracterizează comportamentul fenomenelor și proceselor economico-sociale și oferă o serie de avantaje importante, în comparație cu alte metode și tehnici utilizate în același scop.

Avantajele pe care le are modelarea de tip *rețea neuronală* sunt comparabile cu avantajele pe care le oferă modelarea de tip *fuzzy* în comparație cu modelarea clasică. Spre deosebire de modelele de tip tradițional, care operează în condiții de simplificare severă, justificabilă sau nu, a realității, modelele de analiză și predicție bazate pe conceptul de rețea neuronală au avantajul unei mai mari flexibilități și al unei mai mari compatibilități cu specificitatea realității modelate, simularea realității prin intermediul lor având un mai mare grad de naturalețe.

Pe de altă parte, modelele de tip *rețea neuronală* au avantajul unei continue *adaptabilități* în raport cu modificările intervenite în evoluția fenomenelor modelate și unei continue *autoperfecționări*. De asemenea, modelarea bazată pe rețele neuronale nu presupune dezvoltarea și utilizarea unui aparat matematic foarte sofisticat, ceea ce face ca implementarea și utilizarea tehnicilor de acest fel să fie accesibile unor categorii foarte largi de analiști și cercetători.

Rețelele neuronale reprezintă modele de mare generalitate și flexibilitate, a căror structură funcțională este continuu adaptabilă și configurabilă specificului evoluției unui anumit fenomen și care încearcă să simuleze activitățile de evaluare și decizie propriei creierului uman.

Tehnicile de recunoaștere 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 funcție de impactul acestora asupra diverselor comportamente ale vieții economico-sociale, detectarea unor perioade

cu caracter specific din evoluția unor sisteme economice, stabilirea politicilor de creditare în domeniul finanțier-bancar, evaluarea eficienței activităților de promovare a unor produse, determinarea perioadelor cele mai potrivite pentru vânzarea anumitor sortimente de mărfuri, identificarea celor mai profitabile domenii de afaceri, clasificarea și ierarhizarea unor entități economico-sociale etc.

Definiție: Teoria recunoașterii formelor poate fi definită ca reprezentând totalitatea *normelor, principiilor, metodelor și instrumentelor* de analiză și decizie utilizate în scopul de a identifica *apartenența* unor forme sau obiecte (unități, fenomene, evenimente, acțiuni, procese etc.) la anumite clase cu individualitate bine determinată.

Se poate spune că recunoașterea formelor însumează toate încercările de construire a celor modele care simulează modul în care omul cunoscător, analizează și interpretează și anticipatează comportamentul evolutiv al fenomenelor și proceselor.

Din punct de vedere al teoriei sistemelor, recunoașterea formelor poate fi privită ca un sistem general în care *intrările* reprezintă mulțimea caracteristicilor obiectelor ce urmează să fie clasificate, *ieșirile* reprezintă mulțimea claselor posibile din care pot face parte obiectele analizate, iar *funcția de transfer* exprimă mecanismul decizional prin care un anumit obiect este identificat ca făcând parte dintr-o anumită clasă.

8.3 Concepte fundamentale ale teoriei recunoașterii formelor

În teoria recunoașterii formelor se operează cu o mulțime de concepte care sunt specifice acestui domeniu și se utilizează o terminologie proprie. Dintre numeroasele concepte utilizate în teoria recunoașterii formelor, trei pot fi considerate ca fiind fundamentale și definitorii pentru esența și scopurile teoriei recunoașterii formelor: *forma, clasa și clasificatorul*.

Forma reprezintă expresia numerică a obiectului studiat în vederea clasificării lui într-o anumită clasă și este rezultatul cunoscătoriei principalelor caracteristici posedate de obiectul respectiv.

Deși utilizarea alternativă și cu același sens a termenilor *obiect* și *formă* nu ridică nici un fel de problemă legată de înțelegere, cei doi termeni se deosebesc, totuși, între ei. În timp ce obiectul este o entitate cu existență reală, forma este doar o reprezentare matematică a obiectului, definită sub forma unui vector n-dimensional, ale cărui componente definesc caracteristicile obiectului real.

Cu toate că există această deosebire de esență între obiect și formă, în cele mai multe din situații vom utiliza cele două concepte în mod intervertibil, ca având sens echivalent, nefăcând deosebire între obiect și formă decât în situațiile în care apare necesitatea unei stricte nuanțări a celor două concepte.

Definiție: *Forma* sau *obiectul* este o entitate informațională individuală, caracterizată prin intermediul unui vector n-dimensional, ale cărui componente definesc valorile caracteristicilor acesteia, și care face obiectul procesului de clasificare sau de predicție.

Formele implicate într-un proces de clasificare pot fi: cumpărători, clienți, salariați, votanți, produse, firme, zone geografice, țări, activități economice, titluri de valoare etc.

Una dintre ipotezele fundamentale pe care se bazează teoria recunoașterii formelor este aceea că obiectele analizate sunt caracterizate de un anumit grad de *eterogenitate*. Aceasta înseamnă că se asumă, în mod implicit, existența posibilității definirii unor clase distințe pe mulțimea obiectelor. Pe de altă parte, se mai presupune că anumite obiecte aparținând mulțimii analizate au ceva comun, sunt caracterizate printr-un anumit grad de omogenitate.

În virtutea acestei ultime presupunerii, variabilele explicative ce reprezintă caracteristicile obiectelor au o "substanță" comună pentru anumite submulțimi ale obiectelor. Cele două cerințe impuse mulțimii obiectelor analizate sunt cunoscute sub numele de *similaritate și disimilaritate*.

Clasa, grupa sau *clusterul* reprezintă o submulțime distincă de obiecte care verifică următoarele două proprietăți: obiectele care alcătuiesc o clasă sunt *omogene din punct de vedere al caracteristicilor lor* definitorii; două obiecte între care există diferențe semnificative din punct de vedere al caracteristicilor definitorii fac parte din clase diferite.

Definiție: *Clasa, grupa* sau *clusterul* reprezintă o entitate informațională distinctă și cu semnificație concretă, formată din totalitatea obiectelor ale căror caracteristici sunt identice sau diferă foarte puțin și care sunt semnificativ diferite de caracteristicile obiectelor din alte clase sau grupe.

De exemplu, în cazul în care scopul utilizării tehnicilor de clasificare vizează diferențierea firmelor din punct de vedere al riscului care afectează performanțele financiare ale acestora, clasele sau grupele pot fi următoarele: "firme sănătoase financiar", "firme cu probleme financiare temporare", "firme cu risc ridicat de faliment". Dacă tehniciile de clasificare sunt utilizate pentru a fundamenta deciziile de acordare a creditelor, atunci pot exista două clase: "clienți solvabili" și "clienți insolvenți".

În figurile următoare sunt ilustrate două situații, referitoare la două populații distințe: prima populație este caracterizată printr-un grad relativ ridicat de *omogenitate*, astfel încât ea nu se structurează în mod natural sub forma unor clase și deci nu prezintă interes din punct de vedere al tehnicilor de clasificare. Cea de-a doua populație are o natură *eterogenă*, fiind structurată, cu evidență, sub forma a două clase distințe.

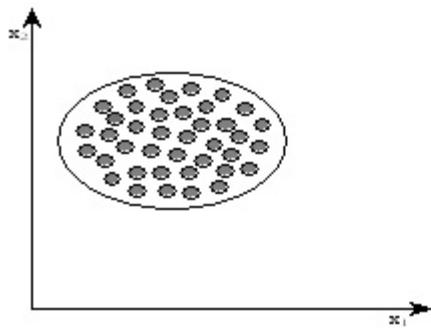


Figura 8.1: Mulțime relativ omogenă

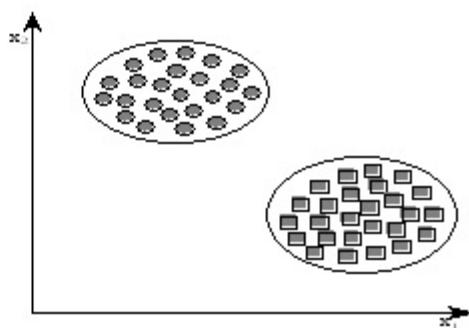


Figura 8.2: Mulțime structurată pe 2 clase

Numărul de clase care alcătuiesc mulțimea de ieșire a unui sistem de recunoaștere a formelor variază în funcție de specificul domeniului pentru care se folosește acest sistem și de scopurile urmărite.

Clasificatorul este un model statistic-matematic care, pe baza informațiilor referitoare la caracteristicile unui anumit obiect, determină decizia de clasificare a obiectului într-o anumită clasă. Clasificatorul poate fi privit ca fiind setul de principii, reguli sau criterii, în funcție de care obiectele analizate sunt atribuite unei clase sau alteia.

Definiție: *Clasificatorul sau criteriul de clasificare* reprezintă regula sau mulțimea de reguli pe baza cărora obiectele care aparțin mulțimii analizate sunt afectate sau atribuite unor clase sau grupe bine definite.

În funcție de natura regulilor utilizate în procesul de clasificare, există mai multe categorii de clasificatori: *clasificatori ie-rarhici*, *clasificatori de cost minim*, *clasificatori de distanță minimală*, *clasificatori de tip Bayes-ian*, *clasificatori euristic* etc.

8.4 Formularea problemei generale de clasificare

Sub cea mai generală formă a sa, problema de clasificare poate fi formulată 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 definită ca o problemă decizională. În acest scop, vom presupune existența unei populații de forme sau de obiecte, notată cu Ω și definită sub forma:

$$\Omega = \{o_1, o_2, \dots, o_M\},$$

unde M reprezintă numărul de unități ale populației analizate.

Fiecare obiect care alcătuiește populația Ω este definit prin intermediul unui număr de N caracteristici, pe care le vom nota cu x_1, x_2, \dots, x_N și care se numesc *variabile explicative*. În acest fel, un obiect din populația Ω poate fi reprezentat sub forma unui vector N -dimensional de forma:

$$x = (x_1 \ x_2 \ \dots \ x_N)^t.$$

Variabilele explicative, care definesc caracteristicile obiectelor analizate, sunt mărimile în funcție de care se stabilește apartenența unui obiect din populația Ω la una dintre clasele populației Ω , adică mărimile în funcție de care se poate face împărțirea acestei populații pe grupe sau clase. Variabilele explicative pot fi variabile de tip *calitativ* sau *cantitativ*. Ele pot fi măsurate pe una dintre cele patru scale cunoscute, respectiv scala *nominală*, scala *ordinală*, scala *interval* sau scala *raport*.

În cazul în care variabilele explicative sunt de tip cantitativ, mulțimea lor poate fi privită ca o submulțime a spațiului real N -dimensional, ceea ce înseamnă că $x \in \mathbb{R}^N$.

Dintre elementele care reprezintă variabilele explicative unele pot să aibă o *putere de discriminare mai redusă*, iar altele pot să aibă o *putere de discriminare mai mare*. Din acest punct de vedere, în construirea algoritmilor de clasificare trebuie să fie *selectate* acele variabile care au puterea de discriminare cea mai mare. De exemplu, în clasificarea firmelor în funcție de riscul posibil al evoluției viitoare a acestora, este plauzibil să considerăm că o variabilă cum ar fi “ponderea forței de muncă feminină” are o influență mai redusă în diferențierea firmelor pe clase de risc, în timp ce o variabilă cum ar fi “rata profitului” are o putere mult mai mare de discriminare a firmelor pe categorii de risc, în funcție de gradul de risc care poate afecta evoluția acestora.

Variabilele cu puterea de discriminare cea mai mare, definesc acele caracteristici ale obiectelor care permit o diferențiere mai puternică a claselor în care pot fi grupate respectivele obiecte și se numesc *variabile descriptor*. Pentru un anumit obiect, vectorul de valori ale variabilelor descriptor reprezintă chiar *forma* asociată respectivului obiect.

Vom presupune în continuare că variabilele cu putere mare de discriminare reprezintă o submulțime, pe care o vom nota cu $\{x_1, x_2, \dots, x_n\}$, a mulțimii $\{x_1, x_2, \dots, x_N\}$, respectiv:

$$\{x_1, x_2, \dots, x_n\} \subseteq \{x_1, x_2, \dots, x_N\}, \quad n \leq N.$$

Luând în considerare numai variabilele descriptor, orice obiect din populația Ω poate să fie reprezentat prin intermediul unui vector n -dimensional de forma:

$$x = (x_1 \ x_2 \ \dots \ x_n)^t,$$

vector cunoscut sub numele generic de *formă*.

Vom nota în continuare cu \mathbf{X} mulțimea formelor asociate tuturor obiectelor din populația Ω , mulțime cunoscută sub numele de *spațiul formelor*. Dacă cele n caracteristici ale obiectelor din populația Ω au valori numerice de tip continuu, atunci mulțimea \mathbf{X} este o submulțime a spațiului real n-dimensional, respectiv $\mathbf{X} \subseteq \mathbb{R}^n$.

Din punct de vedere concret, populația de obiecte Ω poate fi alcătuită din firme, bănci, clienți, cumpărători, țări, zone economice etc. În cazul în care obiectele din populația Ω sunt firme, variabilele descriptor pot fi reprezentate de o serie de indicatori economico-financiari, care caracterizează activitatea acestor firme, cum ar fi: cifra de afaceri, mărimea profitului, rata profitului, gradul de îndatorare, volumul investițiilor etc.

În raport cu o manifestare sau cu o acțiune viitoare, elementele populației Ω se pot găsi într-o din mai multe stări posibile, numite *stări ale naturii*. Stările naturii reprezintă conjuncturi fizice, economice sau sociale, în raport cu care mulțimea de obiecte analizate se structurează sub forma unor categorii bine individualizate.

Vom nota cu Θ mulțimea stărilor naturii și vom presupune existența apriorică a K stări posibile ale naturii, ceea ce înseamnă că mulțimea Θ este de forma:

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$$

Cele K stări posibile ale naturii se caracterizează prin *exhaustivitate* și prin *exclusivitate reciprocă*. Aceasta înseamnă că în afară celor K stări ale naturii nu mai poate exista nici o altă stare posibilă a naturii, respectiv că două stări diferite ale naturii nu se pot manifesta niciodată simultan. De exemplu, din punct de vedere al perspectivelor de evoluție în viitor, firmele dintr-o anumită țară se pot găsi, la un moment dat, în trei stări posibile: *firme performante*, *firme cu dificultăți temporare* și *firme cu risc ridicat de faliment*. Determinarea, dinainte, a stării posibile în care se va afla o firmă în viitor, prezintă o importanță maximă pentru orice decizie, client sau investitor.

Caracteristica principală a unei probleme de clasificare constă în faptul că deși stările posibile ale naturii sunt cunoscute aprioric, ca număr, ca natură și ca plauzibilitate a manifestării, iar fiecare element al populației Ω se găsește în mod sigur într-o singură stare, și nu într-o stare, de obicei nu se cunoaște, cu precizie și în mod aprioric, în care dintre stările naturii se găsește fiecare dintre unitățile populației.

Principala problemă care se pune în acest context constă în identificarea stării în care se află o anumită unitate din populația Ω , adică în stabilirea apartenenței acestei unități la o anumită categorie, clasă sau grupă. Cele K stări ale naturii, în care se pot găsi elementele mulțimii Θ , definesc o împărțire a populației Ω în K grupe sau clase, pe care le vom nota cu $\{\omega_1, \omega_2, \dots, \omega_K\}$.

Spre deosebire de stările naturii, care pot fi privite ca fiind realizări ale unei variabile aleatoare de tip discret, clasele sau grupurile reprezintă submulțimi de obiecte din populația Ω , toate obiectele dintr-o astfel de submulțime având proprietatea că se găsesc în aceeași stare a naturii.

În calitatea sa de submulțime a populației Ω , o clasă poate fi definită sub forma următoare $\omega_k = \{o_1^{(k)}, o_2^{(k)}, \dots, o_{M_k}^{(k)}\}$, unde M_k este numărul de obiecte din clasa k.

Ca o consecință directă a proprietăților pe care le au cele K stări ale naturii, clasele care trebuie identificate la nivelul populației Ω , verifică următoarele două proprietăți:

1. $\bigcup_{k=1}^K \omega_k = \Omega$
2. $\omega_i \cap \omega_j = \emptyset, \quad i,j=1,2,\dots,K; \quad i \neq j$

Prima proprietate implică faptul că orice obiect din populația Ω face parte, cu necesitate, dintr-o din cele K clase. Cea de-a doua proprietate implică faptul că un anumit obiect nu poate să fie afectat sau atribuit, în același timp, la două clase diferite. Mai mult decât atât, este verificată, în plus față de cele două condiții, și condiția:

$$M = M_1 + M_2 + \dots + M_K$$

Modul în care mulțimea stărilor naturii poate induce o structurare pe clase a populației Ω este ilustrat în tabelul următor.

Tabelul 8.1

Stări ale naturii	Clase în populația Ω	Variabile descriptor	Obiecte pe clase
θ_1	ω_1	x_1, x_2, \dots, x_n	$o_1^{(1)}, o_2^{(1)}, \dots, o_{M_1}^{(1)}$
θ_2	ω_2	x_1, x_2, \dots, x_n	$o_1^{(2)}, o_2^{(2)}, \dots, o_{M_2}^{(2)}$
...	...		
θ_K	ω_K	x_1, x_2, \dots, x_n	$o_1^{(K)}, o_2^{(K)}, \dots, o_{M_K}^{(K)}$

Scopul principal al metodelor și tehnicielor de clasificare este acela de a explica apartenența obiectelor mulțimii Ω la grupurile sau clasele $\omega_1, \omega_2, \dots, \omega_K$, utilizând în acest scop informațiile reprezentate de valorile variabilelor descriptor x_1, x_2, \dots, x_n .

Explicarea apartenenței obiectelor mulțimii Ω la cele K clase presupune, de fapt, deducerea sau identificarea unui criteriu de clasificare sau a unei reguli de clasificare, care să descrie modul de structurare a obiectelor populației pe clase. Criteriul de

clasificare mai este cunoscut și sub numele de *clasificator*.

Deducerea criteriului de clasificare se face pe baza informațiilor furnizate de un eșantion extras din populația Ω , eșantion format din obiecte a căror apartenență la clasele $\omega_1, \omega_2, \dots, \omega_K$ poate fi cunoscută sau necunoscută în mod *aprioric*.

Având în vedere cele menționate anterior, problema generală a clasificării poate fi formulată sub forma următoare:

Problema generală a clasificării: Fiind dată o mulțime de obiecte, se cere să se determine *criteriul* sau *regula* care să descrie *apartenența* obiectelor la clasele sub forma cărora se structurează respectivă mulțime de obiecte.

În funcție de cunoașterea sau necunoașterea apriorică a apartenenței la cele K clase a obiectelor care aparțin eșantionului extras din populația Ω , metodele de clasificare se împart în două mari categorii: de *clasificare controlată* și de *clasificare necontrolată*.

Odată ce criteriul de clasificare a fost stabilit, el poate fi folosit, în continuare, pentru efectuarea de *predicții* privind apartenența la o anumită clasă a unor noi obiecte, din afara eșantionului existent, obiecte a căror apartenență nu este cunoscută aprioric. După ce criteriul de clasificare a fost identificat, și cu condiția ca apartenența obiectelor aparținând eșantionului disponibil să fie cunoscută, el poate fi utilizat și pentru verificarea corectitudinii cu care acesta poate face clasificarea, adică pentru *testarea calității* clasificatorului. Calitatea criteriului de clasificare poate fi testată chiar pe obiectele din eșantionul pe care acest criteriu a fost identificat. În acest scop, fiecare obiect din eșantion, a cărui apartenență la o anumită clasă este cunoscută în mod efectiv, este reclasificat cu ajutorul respectivului criteriu, iar rezultatul noii clasificări este comparat cu clasificarea reală.

Testarea clasificatorului poate să conducă la o clasificare corectă a unor obiecte din eșantionul analizat și la o clasificare incorectă a altor obiecte din acest eșantion. Aceasta înseamnă că utilizarea clasificatorului respectiv poate să conducă la situația în care obiectele care aparțin în mod real unei anumite clase să fie clasificate fie în clasa corectă, fie incorect, în oricare din celelalte clase.

Modul în care un clasificator asigură clasificarea obiectelor cu apartenență cunoscută poate fi descris prin intermediul unei matrici, numită *matricea corectitudinii clasificării* sau, mai simplu, *matricea clasificării*, care conține informațiile necesare pentru a aprecia corectitudinea clasificării obiectelor.

Dacă vom considera un eșantion format din T obiecte, care aparțin claselor $\omega_1, \omega_2, \dots, \omega_K$, atunci matricea de clasificare are forma din tabelul următor.

Matricea clasificării

Tabelul 8.2

Clase reale	Clase de predicție				Obiecte de clasificat
	ω_1	ω_2	...	ω_K	
ω_1	T_{11}	T_{12}	...	T_{1K}	$T_{1\cdot}$
ω_2	T_{21}	T_{22}	...	T_{2K}	$T_{2\cdot}$
...
ω_K	T_{K1}	T_{K2}	...	T_{KK}	$T_{K\cdot}$
Obiecte clasificate	$T_{\cdot 1}$	$T_{\cdot 2}$...	$T_{\cdot K}$	T

Un element T_{ij} al matricii de clasificare arată numărul de obiecte aparținând în mod real clasei ω_i și care, prin utilizarea tehnicii de recunoaștere a formelor, sunt clasificate în clasa ω_j . Definind în acest fel elementele matricii de clasificare, rezultă că numărul de obiecte clasificate corect este reprezentat de suma elementelor de pe diagonala principală a matricii clasificării, respectiv:

$$\left(\begin{array}{c} \text{Număr de obiecte} \\ \text{clasificate corect} \end{array} \right) = T_{11} + T_{22} + \dots + T_{KK} = \sum_{i=1}^K T_{ii}.$$

Similar, numărul de obiecte clasificate incorect este reprezentat de suma elementelor aflate în afara diagonalei principale a matricii clasificării, respectiv:

$$\left(\begin{array}{c} \text{Număr de obiecte} \\ \text{clasificate incorect} \end{array} \right) = T - (T_{11} + T_{22} + \dots + T_{KK}) = \sum_{i,j; i \neq j} T_{ij}.$$

Suma valorilor dintr-o linie a matricii de clasificare reprezintă numărul de obiecte din clasa de proveniență ce corespunde liniei respective, indiferent de clasele în care au fost clasificate acestea. Astfel, $T_{\cdot k}$ reprezintă numărul de obiecte din clasa de proveniență ω_k , indiferent de clasa în care acestea au fost clasificate. În mod similar, suma valorilor dintr-o coloană a matricii de clasificare reprezintă numărul de obiecte clasificate în clasa corespunzătoare coloanei, indiferent de clasa de proveniență a obiectelor. Rezultă că $T_{\cdot k}$ reprezintă numărul de obiecte clasificate în clasa ω_k , indiferent de clasa de proveniență a acestora.

Pe baza informațiilor din matricea de clasificare pot fi definiți o serie de indicatori care caracterizează corectitudinea clasificării. Printre aceștia menționăm:

- gradul de clasificare *corectă*:

$$p_c = \frac{\text{Număr de obiecte clasificate corect}}{\text{Număr total de obiecte clasificate}} = \frac{T_{11} + T_{22} + \dots + T_{KK}}{T},$$

- gradul de clasificare *incorRECTă*:

$$p_{inc} = \frac{\text{Număr de obiecte clasificate incorrect}}{\text{Număr total de obiecte clasificate}} = \frac{T - (T_{11} + T_{22} + \dots + T_{KK})}{T}.$$

Împreună cu alți indicatori specifici, cei doi indicatori definiți anterior sunt folosiți pentru a aprecia calitatea unui clasificator, adică măsura în care acesta reușește să detecteze în mod corect apartenența obiectelor la clasele populației analizate. O clasificare este cu atât mai corectă, cu cât valoarea indicatorului p_c este mai mare.

Totalitatea activităților desfășurate în contextul unui proces de recunoaștere a formelor, împreună cu mulțimea de metode și tehnici utilizate în scopul stabilirii apartenenței formelor la anumite clase sau grupe, determină conceptul cunoscut sub numele de *sistem de recunoaștere a formelor*.

8.5 Sisteme de recunoaștere a formelor

Complexitatea activităților care apar în cadrul soluționării oricărei probleme de recunoaștere a formelor, succesiunea și condiționarea fazelor care compun demersul logic întreprins în cadrul acestor probleme, precum și funcționalitatea specifică ce caracterizează acest demers, conferă procesului de recunoaștere a formelor un pronunțat caracter de sistem. Din acest motiv, totalitatea activităților implicate într-un proces de recunoaștere a formelor, ansamblul informațiilor manipulate în acest context și mulțimea procedurilor, algoritmilor, metodelor și tehnicii utilizate în acest scop, sunt privite ca reprezentând un sistem, numit *sistem de recunoaștere a formelor*.

Ca sistem de prelucrare informațională, un sistem de recunoaștere a formelor este format dintr-o mulțime de activități, reguli, proceduri, metode și tehnici, care au ca scop general identificarea apartenenței unui obiect sau unei forme la o anumită clasă bine determinată din populația analizată.

Funcționarea unui sistem de recunoaștere a formelor presupune existența apriorică a unor informații, care vor fi folosite în procesul de clasificare. Aceste informații pot fi reprezentate, după caz, fie de o întreagă populație de forme, fie numai de un eșantion de forme, extrase dintr-o populație de interes.

Intrările unui sistem de recunoaștere a formelor sunt reprezentate de vectorii de proprietăți ale obiectelor, adică de formele propriu-zise, iar ieșirile sistemului de recunoaștere a formelor sunt reprezentate de clasele de apartenență ale formelor de intrare, clase identificate cu ajutorul unor reguli specifice de clasificare. Vectorii de proprietăți ale obiectelor sunt rezultatul unor procese de observare, măsurare și înregistrare a nivelurilor caracteristicilor mulțimilor sau submulțimilor de obiecte, iar informațiile privind apartenența obiectelor la anumite clase sau categorii sunt rezultatul unor procese de evaluare complexă, bazate pe utilizarea unor proceduri și instrumente specifice, de natură statistică-matematică.

Există două tipuri fundamentale de sisteme de recunoaștere a formelor: *sisteme de recunoaștere necontrolată* și *sisteme de recunoaștere controlată*. Aceste două tipuri de sisteme de recunoaștere a formelor sunt determinate de scopurile urmărite, de natura informațiilor pe care le prelucrează, de specificitatea metodelor și instrumentelor utilizate, precum și de natura rezultatelor obținute cu ajutorul acestora.

8.5.1 Sisteme de recunoaștere necontrolată

Sistemele de recunoaștere necontrolată a formelor sunt sistemele în cadrul cărora nu se dispune de informații inițiale referitoare la numărul de clase și la apartenența formelor la anumite clase, construirea claselor făcându-se progresiv, pe măsura creșterii numărului de forme analizate, iar numărul de clase posibile fiind stabilit doar în faza finală a procesului de recunoaștere.

Caracteristica principală a sistemelor de recunoaștere necontrolată a formelor constă în faptul că *nu se cunoaște apartenența obiectelor analizate la o clasă sau alta*. Aceasta înseamnă că, în mod implicit, nu se cunoaște *cu precizie* nici numărul de clase. În legătură cu această ultimă afirmație, considerăm că este necesar să facem următoarea precizare importantă: o serie de algoritmi de clasificare necontrolată, cum ar fi de exemplu *algoritmi de partionare*, presupun fixarea apriorică a numărului de clase în care vor fi împărțite obiectele analizate. Aceasta nu înseamnă însă că este cunoscut, în mod real, și numărul de clase, ci doar că se face o presupunere cu privire la acest număr.

Principiile, procedurile, metodele și tehniciile aparținând sistemelor de recunoaștere necontrolată a formelor sunt cunoscute sub denumirea generală de *tehnici de clasificare, clasificare nesupervizată* sau *analiză cluster*.

Analiza cluster este o tehnică de clasificare caracterizată prin faptul că afectarea formelor sau obiectelor în clustere sau grupe se face progresiv și fără a cunoaște aprioric numărul de clase, în funcție de verificarea a două criterii fundamentale:

- a. obiectele sau formele clasificate în fiecare clasă să fie *cât mai similare* din punct de vedere al anumitor caracteristici;
- b. obiectele clasificate într-o clasă să se diferențieze *cât mai mult* de obiectele clasificate în oricare din celelalte clase.

Primul criteriu de afectare a formelor pe clase cere ca fiecare clasă să fie *cât mai omogenă* în raport cu caracteristicile luate în considerare pentru clasificarea obiectelor. Cel de-al doilea criteriu cere ca fiecare clasă să difere *cât mai mult* din punct de vedere al caracteristicilor de clasificare.

În funcție de caracteristicile procedurilor pe care le utilizează, de ipotezele inițiale pe care se bazează și de natura rezultatelor obținute cu ajutorul lor, metodele de analiză cluster se împart în două mari categorii: *metode de clusterizare*

ierarhică și metode de clasificare prin partităionare sau metode iterative.

Prima categorie include metodele de clusterizare prin *aggregare* și metodele de clusterizare prin *divizare*. Pentru fiecare dintre cele două tipuri de clusterizare există mai multe proceduri specifice, între care menționăm: metoda *agregării simple*, metoda *agregării complete*, metoda *agregării medii*, metoda lui *Ward* etc.

Cea de-a doua categorie include o serie de algoritmi, între care menționăm: algoritmul celor *K-medii*, algoritmul celor *K-medooizi*, algoritmul *CLARA*, algoritmul *fuzzy* etc.

În ceea ce privește rezultatele furnizate de sistemele de recunoaștere necontrolată a formelor, precizăm că ieșirile acestor sisteme nu se reduc, de regulă, la o unică și simplă configurare a obiectelor analizate pe clase, ci includ *mai multe variante* de configurare a obiectelor pe clase, variante conținute într-o entitate informațională numită *structură cluster* sau *ierarhie cluster*. Ierarhia cluster oferă posibilitatea cercetătorului de a alege o anumită configurare a obiectelor pe clase, ceea ce înseamnă, implicit, și alegerea unui anumit număr de clase.

Sistemele de recunoaștere necontrolată sunt utilizate mai mult pentru scopuri de sistematizare, grupare și sintetizare informațională, în situațiile în care sunt analizate cantități foarte mari de date și aceste date se caracterizează printr-un grad ridicat de eterogenitate. În acest sens, tehniciile de recunoaștere necontrolată a formelor sunt foarte utile și eficiente în activitățile de analiză preliminară a datelor. Utilizarea analizei cluster în această fază a analizei datelor este importantă deoarece ea permite organizarea mai eficientă a datelor eterogene. Regăsirea informațiilor în cadrul masivelor de date structurate cu ajutorul tehniciilor de analiză cluster devine mult mai ușoară, iar datele pot fi interpretate mult mai consistent.

8.5.2 Sisteme de recunoaștere controlată

Sistemele de recunoaștere controlată a formelor sunt acele sisteme în cadrul cărora se presupune existența apriorică a unui număr dat de clase și a unui set de forme, numite *prototipuri* sau *referințe*, a căror apartenență la aceste clase este cunoscută. Acest set de forme este reprezentat de eșantionul de obiecte extrase din populația supusă studiului, eșantion cunoscut și sub numele de *set de formare* sau *set de învățare*.

Definiție: *Setul de formare* sau *setul de învățare* este un eșantion de forme extrase din populația studiată, forme a căror apartenență la clasele populației este cunoscută și pe baza cărora sunt deduse criteriile formale de clasificare.

În cadrul sistemelor de recunoaștere controlată a formelor, datele reprezentate de setul de formare includ atât informații referitoare la proprietățile esențiale ale obiectelor supuse analizei, cât și informații referitoare la apartenența acestor obiecte la clasele existente. Pe baza acestor informații inițiale, se deduc regulile și criteriile de decizie pentru partităionarea sub formă de *regiuni* sau *clase* a mulțimii de obiecte supusă studiului sau a spațiului în care iau valori caracteristicile obiectelor.

De fapt, în cazul tehniciilor de acest fel informațiile conținute în setul de formare sunt folosite pentru a face *inferențe* cu privire la împărțirea populației totale pe clase. Mai mult decât atât, din aplicarea tehniciilor de clasificare controlată rezultă și un set de reguli și criterii formale de clasificare, adică un *clasificator*. Aceste reguli și criterii sunt folosite, în continuare, pentru clasificarea unor noi forme neclasificate încă, forme a căror apartenență este necunoscută, adică pentru a face *predicții* cu privire la apartenența noilor forme.

În mod uzual, setul inițial de forme este împărțit în două subseturi folosite în scopuri diferite: primul subset este numit *set de formare* și conține acele forme utilizate pentru deducerea regulilor și criteriilor de clasificare, adică pentru construirea *clasificatorului* propriu-zis; al doilea subset este numit *set de predicție* și conține acele forme utilizate pentru testarea clasificatorului construit pe baza setului de formare.

Definiție: *Sistemul de recunoaștere controlată a formelor* reprezintă totalitatea activităților și procedurilor care au ca scop deducerea unor criterii de *partajare* a unei populații de entități informaționale (obiecte sau variabile), sub forma unui număr cunoscut de clase, pe baza cunoașterii caracteristicilor și a apartenenței elementelor unui eșantion provenit din respectiva populație.

Spre deosebire de tehniciile de clasificare necontrolată, care se bazează, în principal, pe utilizarea conceptului de distanță, elementul fundamental al tehniciilor de clasificare controlată este un model formal, numit *clasificator*. În cazul analizei discriminante, clasificatorul este reprezentat de *funcțiile discriminant* sau de *funcțiile de clasificare*.

8.6 Analiza cluster

Preocupările legate de metodele și tehniciile de analiză cluster datează de peste o jumătate de secol. Primele și cele mai sistematice studii dedicate acestui domeniu sunt reprezentate de lucrările elaborate de Sokal și Sneath în anul 1963 și de Lance și Williams în anul 1967. Ulterior, preocupările științifice din domeniul analizei cluster s-au înmulțit aproape exponențial și s-au diversificat extrem de mult.

În multitudinea preocupărilor și lucrărilor dedicate domeniului analizei cluster pot fi identificate două importante curente științifice, reprezentate de *școala americană* și de *școala franceză*. Printre cei mai de seamă reprezentanți ai școlii franceze se numără: J. P. Benzecri, M. Jambu, L. Lebart, A. Morineau, B. Escofier, G. Saporta și M. Bardos.

Analiza cluster are ca scop căutarea și identificarea de *clase*, *grupe* sau *clustere* în cadrul unor mulțimi de obiecte sau forme, astfel încât elementele care aparțin aceleiași clase să fie cât mai *asemănătoare*, iar elementele care aparțin la clase diferite să fie cât mai *deosebite* între ele. Altfel spus, analiza cluster este o modalitate de examinare a similarităților și disimilarităților dintre obiectele aparținând unei anumite mulțimi, în scopul grupării acestor obiecte sub forma unor clase *distințe* între ele și *omogene* în interior.

Aceasta înseamnă că în toate situațiile, criteriul general de clasificare este, de fapt, un criteriu combinat, care poate fi formulat sub următoarea formă:

Criteriu general de clasificare: Clasificarea obiectelor în clase se face în aşa fel încât să se asigure o *variabilitate minimă în interiorul claselor și o variabilitate maximă între clase*.

Termenul de *analiză cluster* a fost utilizat pentru prima oară în anul 1939, de către R. C. Tyron, în lucrarea *“Cluster Analysis”*. Acest termen este folosit în prezent ca nume generic pentru o mulțime variată de proceduri și algoritmi de clasificare de tip necontrolat.

Prin intermediul analizei cluster fiecare obiect din mulțimea analizată este atribuit *unei singure* clase, iar mulțimea claselor este o mulțime *discretă și neordonabilă*. Clasele rezultate în urma utilizării analizei cluster au o semnificație concretă și generalizatoare, pe baza căreia pot fi efectuate o serie de interpretări și pot fi formulate o serie de concluzii importante pentru procesul de cunoaștere.

Clasele sau grupele sub forma cărora se structurează mulțimile de obiecte se mai numesc și *clustere*. Un cluster este o submulțime formată din obiecte similare, adică din obiecte care sunt suficient de asemănătoare între ele din punct de vedere al caracteristicilor care le definesc.

Definiție: *Clusterul* este o submulțime formată din obiecte care au proprietatea că gradul de disimilaritate dintre oricare două obiecte aparținând clusterului este *mai mic* decât gradul de disimilaritate dintre orice obiect care aparține clusterului și orice obiect care nu aparține clusterului respectiv.

Clusterul poate fi privit și ca reprezentând o *regiune* a unui spațiu multidimensional, caracterizată printr-o *densitate relativ mare* de puncte sau de obiecte. De exemplu, în cazul aplicațiilor informatici, clusterul poate să fie reprezentat de o submulțime de *documente* de același tip sau cu conținut asemănător. Aceste documente pot fi programe sursă, pagini WEB, fișiere de tip text, fișiere HTML etc. Un astfel de document poate fi privit ca un punct dintr-un spațiu multidimensional, în care fiecare dimensiune a spațiului este asociată cu un anumit cuvânt. Coordonatele care definesc poziția unui document în acest spațiu sunt reprezentate de frecvențele cu care apar diferențele cuvintelor în cadrul documentului.

Din punct de vedere geometric, ca mulțimi de puncte dintr-un anumit spațiu, clusterele pot avea forme foarte diferite, mai mult sau mai puțin regulate. Astfel, forma clusterelor poate să fie de tip *convex* sau *concav*, de tip *compact* sau de tip *alungit* etc. În figura următoare sunt ilustrate câteva dintre formele posibile ale clusterelor, pentru cazul particular al obiectelor de tip bidimensional.

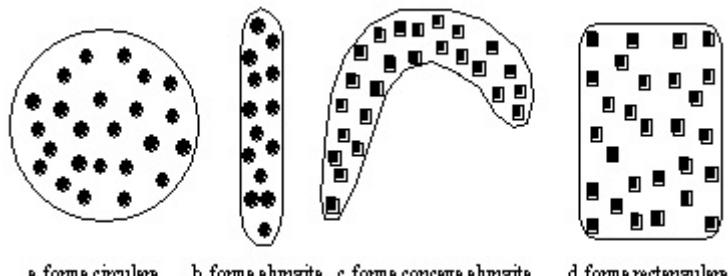


Figura 8.3: Forme posibile ale clusterelor de obiecte bidimensionale

Tipurile de forme pe care le pot avea clusterele în realitate sunt foarte importante în analiza cluster, deoarece atât eficiența procesului de clasificare, cât și calitatea soluțiilor, depind foarte mult de formele clusterelor, mai ales în cazul unor algoritmi de clasificare ierarhică prin agregare.

De regulă, analizele de tip cluster reprezintă proceduri de clasificare de tip *necontrolat*, în care nu este cunoscută aprioric nici apartenența anumitor obiecte la anumite clase, nici numărul de clase posibile. Numărul de clase sau clustere este variabil și este stabilit *concomitant* cu activitatea de clasificare propriu-zisă.

Definiție: *Analiza cluster* poate fi definită ca reprezentând o mulțime de principii, metode și algoritmi de clasificare, având ca scop organizarea datelor sub forma unor structuri informative semnificative, relevante.

Analiza cluster este o *analiză explorativă*, de tip multidimensional, care are ca scop gruparea unor entități informative, cu natură fizică sau abstractă, în clase sau clustere alcătuite din entități informative cu grad ridicat de similaritate.

Din punct de vedere concret, efectuarea unei clasificări cu ajutorul metodelor și tehniciilor de analiză cluster constă în obținerea unor *soluții cluster* sau a unor *partiții*, reprezentate de o mulțime de clase sau clustere notate cu $\omega_1, \omega_2, \dots, \omega_K$, care verifică proprietățile menționate anterior. În cazul anumitor metode de clasificare, rezultatele clasificării sunt reprezentate de *soluții cluster unice*, în timp ce în cazul altor metode de clasificare, cum ar fi metodele de clasificare ierarhică de tip aglomerativ, sunt reprezentate de *mulțimi de soluții cluster*, numite *ierarhii de soluții cluster* sau *ierarhii de partiții*. În aceste situații, este necesar să se aleagă din mulțimea de soluții cluster, adică din ierarhia de partiții, o singură soluție cluster sau o singură partiție.

Deși alegerea unei anumite partiții se face, în principal, în funcție de scopurile urmărite în analiză, pentru a se obține o clasificare consistentă și semnificativă, este necesar alegerea partiției să se bazeze pe o evaluare cât mai riguroasă a calității tuturor partițiilor care alcătuiesc ierarhia cluster.

Din punct de vedere strict teoretic, analiza cluster poate fi privită ca reprezentând o modalitate specifică de construire a uneia sau a mai multor partiții pe mulțimea obiectelor analizate. Orice partiție de acest fel definește o *soluție cluster*, adică un anumit mod de grupare pe clase a obiectelor mulțimii supuse studiului.

Din punct de vedere strict matematic, analiza cluster poate fi privită ca o modalitate de alegere a celei mai adecvate partiții sau *submulțimi* din cadrul *familiei de părți* a mulțimii de obiecte analizate.

În analiza cluster, *ierarhiile cluster* sunt formate dintr-un număr de T soluții cluster, fiecare soluție conținând clustere din ce în ce mai mari, respectiv clustere cu *niveluri de agregare din ce în ce mai ridicate*. O ierarhie cluster are o structură de forma următoare:

$$\begin{aligned} \text{nivel } 0: & \omega_1^{(0)}, \omega_2^{(0)}, \dots, \omega_{K_0}^{(0)} \\ \text{nivel } 1: & \omega_1^{(1)}, \omega_2^{(1)}, \dots, \omega_{K_1}^{(1)} \\ \text{nivel } 2: & \omega_1^{(2)}, \omega_2^{(2)}, \dots, \omega_{K_2}^{(2)} \quad , \\ & \vdots \\ \text{nivel } T-1: & \omega_1^{(T-1)}, \omega_2^{(T-1)}, \dots, \omega_{K_{T-1}}^{(T-1)} \end{aligned}$$

unde T este numărul de obiecte, iar K_i este numărul de clustere din soluția cluster de la nivelul i.

În cazul metodelor ierarhice aglomerative, numărul de clustere din prima partiție este egal cu numărul de obiecte, adică $K_0 = T$. De asemenea, numărul de clustere dintr-o partiție de la un anumit nivel este mai mic cu 1 decât numărul de clustere din partiția de la nivelul inferior și mai mare cu 1 decât numărul de clustere din partiția de la nivelul superior, respectiv:

$$K_{T-1} = K_{T-2} - 1 = K_{T-3} - 1 = \dots = K_1 - 1 = K_0 - 1.$$

Având în vedere că prima partiție obținută dintr-o clasificare ierarhică aglomerativă este soluție cluster de tip banal, reprezentată chiar de lista obiectelor supuse clasificării, rezultă că numărul de partiții propriu-zise, obținute ca soluții ale clasificărilor de acest tip, este egal cu T-1.

Analiza cluster se deosebește în mod fundamental de procedurile de natură statistică, cum ar fi cele care au ca scop verificarea semnificației, prin faptul că ea nu se bazează și nu presupune îndeplinirea apriorică a nici unei ipoteze specifice. În consecință, prin esența sa, analiza cluster constituie un important și eficient instrument de *analiză exploratorie*.

Se poate spune că scopul general al analizelor de tip cluster este acela de creare a aşa-numitelor *taxonomii* sau *tipologii*. Construcția tipologii este bazată pe analiza *asemănărilor* și *deosebirilor* existente între obiectele unei mulțimi date.

Necesitatea de a construi tipologia apare în cele mai diverse domenii de activitate, existența tipologii oferind largi posibilități pentru analiza și interpretarea fenomenelor aparținând acestor domenii.

Deși folosirea tehniciilor de analiză cluster nu este specifică doar pentru anumite domenii de activitate, totuși, utilizarea cea mai frecventă a acestora este întâlnită în domeniul marketingului, în investigațiile de natură psihosocială sau în evaluările econo-micosociale la nivel teritorial.

În domeniul marketingului, se dețează aplicațiile tehniciilor de analiză cluster în studierea comportamentului consumatorilor. Aceste aplicații vizează evaluarea șanselor pe care poate să le aibă lansarea unui produs nou, identificarea unor noi piețe, modalitățile de segmentare a pieței sau identificarea poziționării pe piață a produselor diferitelor producători. Posibilitatea de a deduce tipologia specifică pe mulțimea clienților unei firme este deosebit de importantă pentru fundamentarea și stabilirea politicilor comerciale ale firmei.

În cazul determinării poziționării pe piață a diferitelor mărci ale unui produs, analiza cluster este folosită pentru a clasifica mărcile de fabricație, în funcție de similitudinea sau disimilitudinea percepțiilor pe care le manifestă consumatorii față de aceste mărci. Pe baza modului în care se clasifică mărcile și a caracteristicilor consumatorilor care își manifestă preferințele, un producător poate identifica mărcile concurente și trăsăturile specifice ale categoriilor de consumatori care preferă produsul acestui producător. De exemplu, mărcile aflate în aceeași clasă cu marca unui producător sunt mărci concurente, deoarece ele se adresează aceluiași segment de consumatori.

Tehnicile specifice analizei cluster sunt deosebit de necesare și utile în orice proces de analiză a datelor, nu numai în cele care vizează în mod direct necesități legate de clasificare. De exemplu, utilizarea acestor tehnici este extrem de importantă pentru acele procese de analiză în care cantitatea de informație ce trebuie prelucrată este atât de mare și variată încât extragerea a ceea ce este logic, esențial și semnificativ în această cantitate informațională, devine imposibilă dacă nu sunt folosite instrumente corespunzătoare de sintetizare și structurare a informației brute. În acest context, tehniciile de analiză cluster sunt utilizate, cu precădere, pentru *sistematizarea* informațiilor supuse analizei, activitate care este strict necesară în faza de analiză preliminară a datelor.

Identificarea pe o mare cantitate de informații brute a unor categorii, clase sau grupe informaționale reprezentă unul dintre scopurile generale și, în același timp, principale ale oricărei analize cluster.

În mod sintetic, efectuarea unei analize cluster, având ca scop clasificarea unei mulțimi de obiecte, cuprinde următoarele etape:

- *alegerea caracteristicilor* în funcție de care se va face clasificarea;
- *alegerea tipului de măsură* pentru evaluarea proximității dintre obiecte;
- *stabilirea regulilor de formare a claselor sau clusterelor*;
- *construirea claselor*, adică încadrarea obiectelor în clase;
- *verificarea consistenței și semnificației* clasificării;

- alegerea unui număr optimal de clustere, în funcție de natura problemei de clasificare și de scopurile care se urmăresc;
- interpretarea semnificației clusterelor;

Rezultatele unei analize cluster sunt reprezentate fie de o singură soluție cluster, fie de ierarhii cluster, care conțin diferite modalități de configurare a obiectelor pe clase, adică mai multe soluții cluster. În cel de-al doilea caz, pe baza efectuării unei "tăieturi" în ierarhia cluster, utilizatorul are posibilitatea alegerii unei configurații a obiectelor pe un anumit număr dorit de clase.

Pe baza rezultatelor obținute în urma efectuării unei analize cluster, pot fi deduse anumite legități care guvernează evoluția unor populații de fenomene, pot fi identificate anumite principii utile pentru procesul de cunoaștere sau pot fi formulate o serie de concluzii științifice cu caracter de generalitate. În acest sens, analiza cluster și rezultatele obținute pe baza acesteia pot contribui la:

- definirea unor scheme de clasificare formală și a unor tipologii, pe baza cărora realitățile complexe pot fi mai bine cunoscute și înțelese;
- identificarea unor modele statistică-matematice cu ajutorul cărora mulțimi complexe și eterogene de fenomene și procese pot fi sintetizate și reprezentate sub o formă simplificată și inteligibilă;
- definirea mai corectă și mai completă a caracteristicilor fundamentale ale unor populații de fenomene și procese;
- deducerea unor măsuri numerice adecvate pentru caracterizarea dimensiunilor populațiilor de fenomene și pentru evidențierea modificărilor care au loc în nivelul și structura acestora;
- identificarea unor entități individuale care sunt reprezentative pentru clase și categorii complexe de fenomene și procese.

Din cele de mai sus, rezultă că analiza cluster poate fi privită, în general, ca un instrument care are ca scop reducerea unor mulțimi de obiecte, sau chiar de variabile, la un număr mai restrâns de entități informaționale, care sunt clasele sau clusterele. Din acest punct de vedere, se poate face o analogie între analiza cluster și analiza componentelor principale, cu menținerea că în analiza componentelor principale reducerea vizează, de regulă, variabilele.

În sensul său obișnuit, ca ansamblu de metode și tehnici de *clasificare a obiectelor*, analiza cluster este o analiză efectuată în *spațiul variabilelor*. Într-adevăr, cele mai multe utilizări ale tehnicii de analiză cluster sunt cele care au ca scop clasificarea obiectelor, și nu clasificarea variabilelor.

Există însă și situații în care analiza cluster este folosită pentru *clasificarea variabilelor* care caracterizează obiectele, adică situații în care analiza este efectuată în *spațiul obiectelor*. În aceste situații, analiza cluster poate servi ca instrument de agregare a caracteristicilor obiectelor, sub forma unor caracteristici generale și cu relevanță ridicată din punct de vedere al posibilităților de interpretare.

Remarcă: Analiza cluster poate fi utilizată atât pentru *clasificarea obiectelor*, cât și pentru *clasificarea variabilelor* care definesc obiectele.

Spre deosebire de utilizarea analizei cluster pentru clasificarea obiectelor, situație în care specificitatea este reprezentată de faptul că distanțele sunt evaluate pentru *perechi de obiecte*, în cazul utilizării analizei cluster pentru clasificarea variabilelor, evaluarea distanțelor se face pentru *perechi de variabile*.

8.6.1 Tipul informațiilor primare utilizate în analiza cluster

Problema cea mai importantă a oricărui tip de analiză cluster este aceea a modului în care poate fi măsurată *proximitatea*, respectiv *gradul de apropiere* sau *gradul de depărtare*, dintre obiecte și dintre clustere.

Orice proces de clasificare a obiectelor este definit în raport cu o anumită măsură a gradului de apropiere sau de depărtare dintre obiectele analizate, indiferent de metoda sau algoritmul pe care se bazează acest proces. Această măsură poate fi reprezentată fie de un indicator de similaritate, fie de un indicator de disimilaritate. Fiecare dintre cele două categorii de indicatori va fi definită și analizată în continuare.

În general, măsurarea gradului de proximitate dintre obiecte se face cu ajutorul a două grupe de indicatori, cunoscute sub numele de *indicatori de similaritate* și *indicatori de disimilaritate*. Indicatorii de similaritate și indicatorii de disimilaritate pot fi utilizati atât în analizele cluster efectuate pe obiecte, cât și în analizele cluster efectuate pe variabile.

Indicatorii de similaritate și de disimilaritate pot fi utilizati ca bază informațională în orice proces de clasificare datorită faptului că ei pot induce o *relație de ordine* pe mulțimea perechilor de obiecte sau de variabile și, în consecință, 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 evaluatează pot fi considerate a fi mai asemănătoare, respectiv mai apropiate. De asemenea, o valoare foarte mică a indicatorului de similaritate evidențiază faptul că cele două obiecte sau cele două variabile sunt mai depărtate între ele.

Indicatorii de disimilaritate sunt mărimi numerice care exprimă cât de *deosebite* sau cât de *depărtate* sunt două obiecte sau două variabile. Indicatorii de disimilaritate se mai numesc și indicatori sau coeficienți de *deosebire* sau de *distanțare* a obiectelor sau variabilelor. Cu cât valoarea unui indicator de disimilaritate este mai mare, cu atât cele două obiecte sau cele două variabile pentru care se calculează sunt mai diferite, adică mai distanțate între ele.

Cea mai importantă și cea mai utilizată categorie de indicatori de disimilaritate este reprezentată de indicatorii de tip *distanță*. De multe ori însă, conceptul de distanță este utilizat și pentru a desemna indicatori de similaritate, cu toate că aceștia exprimă gradul de apropiere dintre două entități informaționale.

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 mărimi mai potrivite pentru măsurarea proximității în cazul obiectelor cu caracteristici de tip *cantitativ*.

În legătură cu această deosebire, facem precizarea că există situații în care indicatorii de similaritate pot fi utilizati nu numai în cazul variabilelor de tip calitativ, ci și în cazul variabilelor de tip cantitativ. Acest lucru este posibil în situațiile în care variabilele de tip cantitativ sunt supuse unor transformări adecvate.

Cu toate că indicatorii de similaritate și indicatorii de disimilaritate sunt priviți, de regulă, ca fiind două categorii distincte, putem face afirmația că ambele categorii exprimă, într-un anumit fel, două fațete ale același lucru. Mai mult decât atât, în anumite condiții, indicatorii de similaritate pot fi transformați în indicatori de disimilaritate. Diferențele dintre aceste categorii de indicatori țin de natura variabilelor în raport cu care sunt evaluati și de modalitățile de calcul specifice fiecărui tip de indicator.

Informațiile utilizate, în ultimă instanță, în analiza cluster sunt reprezentate sub forma unor matrici simetrice de tip *obiecte × obiecte*, numite, după caz, *matrici de proximitate*, *matrici de similaritate*, *matrici de asociere*, *matrici de incidentă*, *matrici de disimilaritate* sau *matrici de distanțe*. Atât liniile, cât și coloanele matricilor de acest fel se referă la obiectele analizate, astfel încât numărul lor este egal cu numărul de obiecte supuse analizei. Elementele acestor matrici sunt mărimi numerice care exprimă proximitatea dintre perechile de *obiecte* care etichetează rândurile și coloanele matricilor.

În cazul particular al clasificării variabilelor, informațiile utilizate efectiv în analiză sunt reprezentate sub forma unor matrici de tipul *variabilă × variabilă*. Elementele acestor matrici sunt mărimi numerice care exprimă gradul de proximitate dintre perechile de *variabilă* aflate în liniile și coloanele acestor matrici.

Rezultă că matricile de proximitate conțin indicatori de disimilaritate (distanțe) sau indicatori de similaritate pentru toate perechile posibile de obiecte sau de variabile. În construirea matricilor de proximitate pot fi utilizate, în funcție de proprietățile obiectelor la care se referă, atât variabile de tip cantitativ, cât și variabile de tip calitativ.

Tipurile indicatorilor de similaritate sau de disimilaritate utilizati în evaluarea proximităților trebuie să fie adecvate și compatibile cu natura datelor existente. De asemenea, în evaluarea proximităților trebuie să se ia în considerare *toate variabilele care au o relevanță ridicată* din punct de vedere al clasificării. Omiterea unor variabile din calculul proximităților poate conduce la obținerea unor soluții inconsistente.

Datele din matricile de proximitate pot fi reprezentate sub forma unui graf specific, care evidențiază poziționarea spațială relativă a obiectelor sau a variabilelor și care oferă o imagine sugestivă, de ansamblu, asupra *distanțării* respectivelor entități informaționale.

Baza informațională pentru determinarea matricilor de proximitate o reprezintă așa-numitele *matrici de observații*, care sunt matrici de tipul *obiecte × variabilă* sau matrici de tipul *variabilă × obiecte*, în funcție de tipul analizei efectuate. În primul caz, rândurile matricilor de observații reprezintă obiectele analizate, iar coloanele acestor matrici reprezintă caracteristicile reținute în analiză, adică variabilele descriptor. În cel de-al doilea caz, interpretările rândurilor și coloanelor sunt inversate.

Entitățile informaționale supuse procesului de clasificare cu ajutorul metodelor și tehniciilor de analiză cluster sunt reprezentate de obiecte sau variabile. Obiectele implicate într-o analiză cluster se mai numesc indivizi, observații, articole sau înregistrări. Din punct de vedere al modului de reprezentare externă, mulțimile de informații referitoare la aceste entități sunt organizate sub forma unor *fișiere* sau *baze de date*. Fiecare înregistrare din cadrul unui fișier sau unei baze de date definește un anumit obiect. De obicei, în analiza cluster se presupune că toate obiectele sunt caracterizate prin intermediul *aceleiași mulțimi* de variabile descriptor. Variabilele descriptor utilizate în analiza cluster pot să fie de *același tip*, cantitativ sau calitativ, sau pot să fie de *tipuri diferite*. În fiecare dintre cele două cazuri, evaluarea gradului de proximitate dintre obiecte se face în mod diferit.

Cele mai mari probleme apar în cazul în care variabilele descriptor sunt de tipuri diferite, deoarece în acest caz proximitățile parțiale, evaluate în raport cu variabile diferite, au natură incompatibilă și nu pot fi agregate în mod direct în scopul obținerii unui indicator de proximitate la nivelul ansamblului de variabile. O astfel de situație apare, de exemplu, când unele variabile sunt de tip interval sau raport, iar altele sunt de tip nominal. Așa cum o să vedem în cadrul paragrafului 10.3.3.4, situațiile de acest fel impun utilizarea unor proceduri specifice de construire a indicatorilor de proximitate.

Matricile de observații pot conține fie rezultatele măsurătorilor directe, efectuate asupra variabilelor originale, fie rezultatele obținute în urma unor transformări specifice, efectuate asupra variabilelor originale. Mărimile din cea de-a două categorie sunt reprezentate de *scorurile componentelor principale* sau de *scorurile factorilor* și se obțin prin efectuarea, pe observațiile originale existente, a unei analize a componentelor principale sau a unei analize factoriale.

În analiza cluster, matricile de observații conțin informații cu caracter *complet*, adică informații referitoare la *întreaga mulțime de obiecte* supuse clasificării. Spre deosebire de aceasta, în cazul analizei discriminante informațiile conținute în matricea de observații sunt informații cu caracter *parțial*, referitoare la un eșantion de obiecte extrase din populația de obiecte supusă analizei.

8.6.2 Evaluarea distanțelor dintre obiecte și tipuri de distanțe

Prin natura lor *numerică*, variabilele de tip cantitativ, adică variabilele măsurate pe scalele de tip raport, interval și, eventual, ordinal, permit o definire mai naturală a conceptului de distanță. Pentru variabilele de tip nominal, inclusiv variabilele de tip binar, distanțele se calculează într-un mod specific, compatibil cu natura acestor variabile.

Pentru evaluarea disimilarităților dintre obiectele ale căror caracteristici sunt de tip cantitativ sau dintre variabile de tip cantitativ, pot fi folosite mai multe tipuri de distanțe, cum ar fi: distanța *Euclidiană* (simplă, ponderată sau pătrată), distanța *Manhattan*, distanța *Cebișev*, distanța *Minkovski*, distanța *Camberra*, distanța *Mahalanobis*, distanța *Pearson*, distanța *Jambu* etc.

• Distanța Euclidiană

Distanța Euclidiană, care mai este cunoscută și sub numele de *normă de tip L₂*, este distanța cea mai frecvent utilizată în problemele de analiză cluster. Ea se calculează ca rădăcină pătrată a sumei pătratelor diferențelor coordonatelor celor două obiecte sau variabile pentru care se evaluează distanța.

Distanța Euclidiană măsoară depărtarea dintre două obiecte sau dintre două variabile “în linie dreaptă” și este definită sub forma următoare:

$$d(o_i, o_j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}; \quad d(x_p, x_q) = \sqrt{\sum_{t=1}^T (y_{pt} - y_{qt})^2}.$$

Distanța Euclidiană exprimă proximitatea dintre obiecte ca distanță între două puncte din spațiul Euclidian, respectiv ca distanță măsurată în linie dreaptă. În acest sens, de exemplu, distanța dintre orașul București și orașul New-York nu este o distanță de tip Euclidian deoarece ea este exprimată de-a lungul curburii sau rotunjimii globului pământesc, și nu în linie dreaptă.

• Distanța Manhattan

Distanța Manhattan, numită și distanță rectangulară, distanță “City-Block” sau *normă de tip L₁*, se calculează ca sumă a valorilor absolute ale diferențelor coordonatelor celor două obiecte sau celor două variabile analizate și este definită de relațiile:

$$d(o_i, o_j) = \sum_{k=1}^n |x_{ik} - x_{jk}|; \quad d(x_p, x_q) = \sum_{t=1}^T |y_{pt} - y_{qt}|.$$

Deoarece diferențele de coordonate utilizate în calculul său nu sunt amplificate printr-o ridicare la o putere, distanța Manhattan este mai robustă în raport cu prezența în date a valorilor aberante.

Distanța Manhattan poate fi calculată și în varianta *ponderată*, calculul făcându-se în mod similar cu cel al distanței Euclidiene ponderate. De asemenea, distanța Manhattan poate fi utilizată în cazul în care obiectele au caracteristici care sunt măsurate pe scala de tip interval și pe scala de tip raport.

• Distanța Cebîșev

Distanța Cebîșev, cunoscută și sub numele de “maxim al dimensiunilor” sau *normă de tip L_∞*, este o distanță de tip *valoare absolută* și se determină ca fiind valoarea maximă a valorilor absolute ale diferențelor dintre coordonatele obiectelor sau variabilelor, respectiv:

$$d(o_i, o_j) = \max_k |x_{ik} - x_{jk}|; \quad d(x_p, x_q) = \max_t |y_{pt} - y_{qt}|.$$

Distanța Cebîșev poate fi utilizată atunci când se dorește ca două obiecte sau variabile să apară ca fiind diferite, dacă ele diferă chiar și doar din punct de vedere al unei caracteristici, respectiv al unui obiect. În alte situații, nu este recomandabil să se folosească acest tip de distanță.

• Distanța Mahalanobis

Distanța Mahalanobis este una dintre cele mai cunoscute, mai importante și mai frecvent utilizate distanțe. Ea este o formă generalizată a conceptului de distanță și se calculează sub formele următoare:

$$d(o_i, o_j) = (x^{(i)} - x^{(j)})^t \cdot \Sigma_{n \times n}^{-1} \cdot (x^{(i)} - x^{(j)}); \quad d(x_p, x_q) = (y^{(p)} - y^{(q)})^t \cdot \Sigma_{T \times T}^{-1} \cdot (y^{(p)} - y^{(q)}),$$

unde $x^{(i)}$ și $x^{(j)}$ sunt vectori coloană reprezentând liniile i și j din matricea de observații X, $y^{(p)}$ și $y^{(q)}$ sunt vectori coloană reprezentând liniile p și q din matricea de observații Y, iar Σ^{-1} este notația pentru inversa matricii de covarianță, matrice calculată în spațiul variabilelor - în primul caz, respectiv în spațiul observațiilor - în al doilea caz. Se poate observa că, în cazul în care matricea de covarianță Σ este egală cu matricea unitate, distanța Mahalanobis se reduce la distanța Euclidiană pătrată.

Distanța Mahalanobis reprezintă singurul tip de distanță care ia în considerare, într-o manieră completă, *gradul de dispersare* al mulțimii de obiecte sau al mulțimii de variabile analizate, precum și *gradul de corelare* al respectivelor entități informaționale. Utilizarea distanței Mahalanobis este recomandată, mai ales în situațiile în care variabilele care descriu obiectele sunt corelate între ele. Distanța Mahalanobis este utilizată și în cazul tehniciilor de clasificare controlată, pe baza acestei distanțe fiind dezvoltat chiar un criteriu operațional de discriminare.

8.6.3 Evaluarea distanțelor dintre clustere

O problemă dificilă care apare în analiza cluster, este legată de *necesitatea evaluării distanțelor dintre clase sau clustere*. Dificultatea acestei probleme este dată de faptul că distanțele dintre clase sau clustere sunt, de fapt, distanțe între *mulțimi de obiecte* sau distanțe între *mulțimi de variabile*.

Problema evaluării distanțelor dintre clustere apare în special în cazul *analizei cluster de tip ierarhic*, în care construirea

arborelui de clustere poate fi făcută pe baza *omasării successive* sau *divizării successive* a clusterelor. Comasarea clusterelor este numită *amalgamare* sau *aggregare*, iar divizarea clusterelor este numită *dezaggregare*.

Theoretic, procesul de aggregare sau dezaggregare succesivă a clusterelor se bazează pe definirea unei *distanțe limită* între clustere, distanță numită și *prag de aggregare*, respectiv *prag de dezaggregare*. În principiu, decizia de comasare a două clustere sau de divizare a unui cluster este luată numai dacă distanța dintre aceste clustere este mai mică, respectiv mai mare decât distanța limită fixată.

Dacă în cazul evaluării gradului de apropiere sau depărtare dintre două obiecte sunt relativ simple, fiind suficient să se calculeze una din distanțele menționate mai sus, în cazul în care este necesar să se calculeze distanța dintre două clusteră lucrurile devin ceva mai complicate și presupun existența unei metode specifice de evaluare.

Distanța dintre două clustere este, de fapt, o distanță dintre două *mulțimi de puncte*, adică o distanță mai dificil de evaluat. Ca distanță între două *mulțimi de puncte*, distanța dintre două clustere poate fi măsurată cu ajutorul uneia dintre multe metode posibile.

Dintre metodele propuse pentru evaluarea distanțelor dintre clustere menționăm: *metoda celor mai apropiati vecini*, *metoda celor mai depărtăti vecini*, *metoda distanței medii între perechi*, *metoda centroidului* și *metoda lui Ward* etc.

8.6.3.1 Metoda celor mai apropiati vecini

Metoda celor mai apropiati vecini evaluatează distanța dintre două clustere ca fiind *distanța minimă* dintre toate perechile posibile de forme din cele două clustere. Aceasta înseamnă că distanța dintre două clustere este măsurată prin *distanța dintre cele mai apropiate obiecte aparținând celor două clase*.

Definiție: *Metoda celor mai apropiati vecini* evaluatează distanța dintre două clustere ca distanță între două obiecte, unul din primul cluster, iar celălalt din cel de-al doilea cluster, care sunt *cele mai apropiate* între ele în sensul distanței utilizate.

În figura următoare este vizualizată distanța dintre două clustere, evaluată după metoda celor mai apropiati vecini.

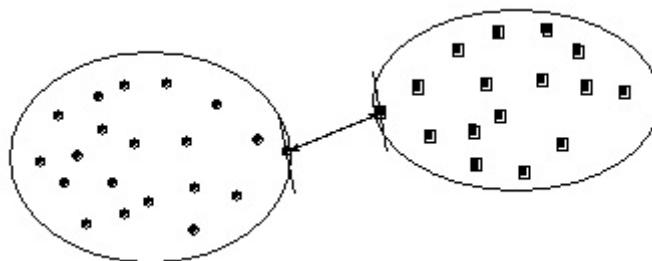


Figura 8.4: Distanța dintre două clustere în cazul metodei *celor mai apropiati vecini*

8.6.3.2 Metoda celor mai depărtăti vecini

Metoda celor mai depărtăti vecini este metoda după care distanța dintre două clase este măsurată prin *distanța dintre cele mai depărtate obiecte aparținând celor două clustere*. Pe baza acestei metode, două clustere sunt considerate a fi mai apropiate sau mai depărtate, în funcție de proximitatea dintre cele mai depărtate obiecte din cele două clustere.

Definiție: *Metoda celor mai depărtăti vecini* evaluatează distanța dintre două clustere ca distanță între două obiecte, unul din primul cluster, iar celălalt din cel de-al doilea cluster, care sunt *cel mai depărtat* între ele în sensul distanței utilizate.

Calculul distanței dintre două clustere cu ajutorul metodei celor mai depărtăti vecini se face pe baza datelor din matricea distanțelor dintre obiectele din cele două clustere, prin identificarea în această matrice a elementului cu valoarea *cea mai mare*.

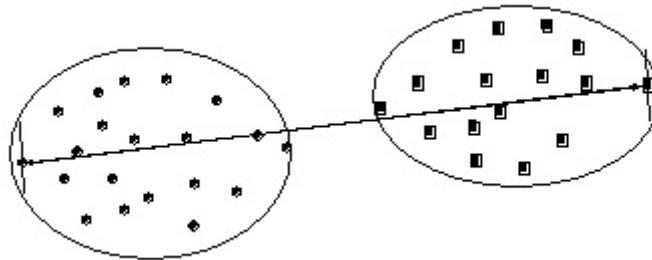


Figura 8.5: Distanța dintre două clustere în cazul metodei *celor mai depărtăti vecini*

Pentru evaluarea distanțelor dintre obiectele cele mai depărtate din cele două clustere poate fi utilizată oricare dintre metodele cunoscute de calcul a distanțelor dintre obiecte, în funcție de natura variabilelor care definesc obiectele supuse clasificării.

8.6.3.3 Metoda distanței medii dintre perechi

Metoda *distanței medii dintre perechile de obiecte* evaluatează distanța dintre două clustere prin intermediul distanței medii dintre toate perechile posibile de obiecte care aparțin celor două clustere.

Definiție: Metoda distanței medii dintre perechi evaluatează distanța dintre două clustere ca *medie* a distanțelor dintre oricare două obiecte care aparțin celor două clustere, unul primului cluster, iar celălalt din celui de-al doilea cluster.

Evaluarea distanței dintre două clustere cu ajutorul metodei distanței medii între perechile de obiecte se face pe baza datelor din matricea distanțelor dintre obiectele din cele două clustere, calculând media acestor distanțe.

În figura următoare este sugerată o interpretare geometrică a modului de calcul a distanței dintre clustere cu ajutorul metodei distanței medii dintre perechi.

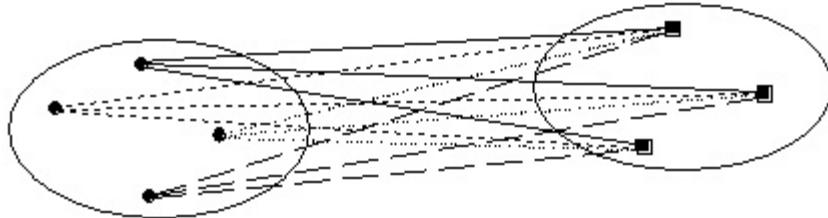


Figura 8.6: Ilustrarea grafică a metodei distanței medii dintre perechi

Ca și în cazul celorlalte două metode, pentru evaluarea distanțelor dintre obiectele celor două clustere, poate fi utilizată oricare dintre metodele cunoscute de calcul al distanțelor dintre obiecte.

8.6.3.4 Metoda centroidului

Metoda centroidului este metoda după care distanța dintre două clustere este măsurată ca distanță între centroizii celor două clustere. În acest fel, două clustere sunt considerate mai apropiate sau mai depărtate, în funcție de gradul de apropiere sau de depărtare dintre centroizii lor.

Centroidul sau *centrul de greutate* al unui cluster reprezintă obiectul, real sau abstract, ale căruia caracteristici au ca valori chiar mediile caracteristicilor obiectelor care compun clusterul respectiv.

Definiție: Metoda centroidului evaluatează distanța dintre două clustere ca distanță între centroizii celor două clustere.

Evaluarea distanței dintre două clustere cu ajutorul metodei centroidului se face calculând mai întâi centroizii celor două clustere, după care se evaluatează distanța dintre clustere ca distanță între acești centroizi.

Figura următoare ilustrează interpretarea geometrică a calculului distanțelor dintre clustere cu ajutorul metodei centroidului. În această figură, centroizii celor două clustere sunt marcați prin cele două puncte de dimensiune mai mare.

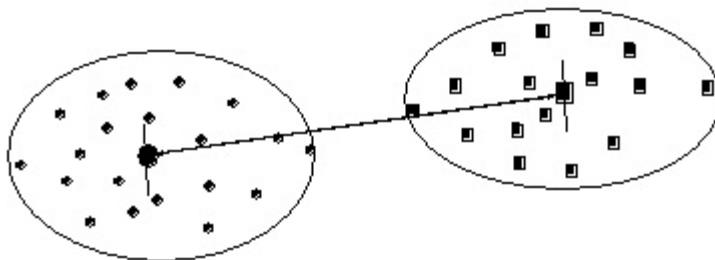


Figura 8.7: Distanța dintre clustere în cazul metodei centroidului

Deoarece centroidul este vectorul mediilor corespunzătoare tuturor obiectelor dintr-un cluster, în calculul distanței dintre două clustere cu ajutorul metodei centroidului sunt luate în considerare, în mod implicit, toate obiectele din fiecare cluster.

8.6.3.5 Metoda lui Ward

Metoda lui Ward este o metodă de evaluare a distanței dintre două clustere, care se bazează pe *maximizarea gradului de omogenitate a clusterelor* sau, ceea ce este același lucru, pe *minimizarea variabilității intracluster*. De regulă, gradul de omogenitate a unui cluster se consideră a fi *cu atât mai mare, cu cât suma totală a pătratelor abaterilor intracluster este mai mică*.

Elementul caracteristic al metodei lui Ward este reprezentat de faptul că prin comasarea a două clustere se urmărește obținerea unei omogenități maxime la nivelul tuturor clusterelor care aparțin unei configurații date a obiectelor pe clustere. În acest sens, se poate spune că distanța Ward dintre două clustere măsoară *variabilitatea intracluster cumulată*, pe care o induce *comasarea celor două clustere la nivelul configurației cluster rezultante*. În acest sens, distanța Ward poate fi definită sub forma următoare:

Definiție: Metoda lui Ward evaluatează distanța dintre două clustere *sumă totală a pătratelor abaterilor la nivelul configurației cluster rezultante din comasarea celor două clustere pentru care se evaluatează distanța*.

Spre deosebire de alte metode de calcul a distanțelor între clustere, distanța Ward oferă o serie de avantaje. Aceste avantaje decurg din faptul că ea este singura dintre metodele de evaluare a distanțelor dintre clustere, care exprimă distanțele din punct de vedere al minimizării variabilității intracluster sau, ceea ce înseamnă același lucru, din punct de vedere al maximizării variabilității intercluster.

8.6.4 Metode și tehnici de analiză cluster

Analiza cluster are ca scop căutarea și identificarea în datele supuse analizei a unor grupuri sau clustere, în funcție de similaritățile și disimilaritățile dintre obiectele la care se referă respectivele date.

Cea mai importantă etapă din cadrul unei analize cluster este cea a formării clusterelor sau claselor. Algoritmii care pot fi utilizati pentru realizarea activității de construire a clusterelor cunosc o mare varietate, care include algoritmi euristici, algoritmi de optimizare și algoritmi fuzzy. Diferențele dintre modul de construire a clusterelor după un algoritm sau altul, sunt determinate, în principal, de natura metodei utilizate pentru evaluarea distanțelor între clustere. Mai mult, chiar tipul analizei cluster rezultă din natura algoritmului utilizat pentru construirea clusterelor.

Din punct de vedere al naturii lor, al modului de operare și al tipului de soluții pe care le furnizează, metodele de analiză cluster pot fi împărțite în două mari categorii: *metode de tip ierarhic și metode de tip iterativ sau de partitioanare*.

Algoritmii sau metodele de tip ierarhic au ca scop producerea *mai multor* soluții cluster, soluții numite *ierarhii cluster*. Caracteristica principală a acestor algoritmi constă în faptul că numărul de clustere nu este cunoscut aprioric.

Remarcă: În cazul metodelor de clasificare ierarhică, *numărul de clustere nu este cunoscut aprioric*.

Există două categorii de algoritmi de clasificare ierarhică: *algoritmi de agregare și algoritmi de dezagregare*.

Rezultatele furnizate de algoritmii de clasificare ierarhică includ mai multe variante de clasificare a obiectelor, fiecare variantă de clasificare conținând structuri cluster cu un *număr variabil* de clustere. Structurile cluster obținute cu ajutorul algoritmilor de acest fel se numesc *structuri cluster multinivel*.

Remarcă: Algoritmii de clasificare ierarhică furnizează *mai multe soluții*, de tip *multinivel*, care se numesc *ierarhii cluster* și care diferă între ele prin numărul de clustere pe care le includ și prin gradul de agregare al clusterelor.

Cea mai sintetică soluție a unei structuri cluster obținute cu ajutorul metodelor de clasificare ierarhică este formată *dintr-un singur cluster*, care include toate obiectele analizate. Cea mai detaliată soluție a unei structuri cluster de acest fel include un număr maxim de clustere, *egal cu numărul de obiecte analizate*, fiecare cluster conținând un singur obiect. Aceasta înseamnă că numărul posibil de soluții dintr-o structură cluster obținută cu ajutorul algoritmilor ierarhici este mai mic cu unu decât numărul de obiecte supuse clasificării. Acest număr este determinat de numărul de *nivele ierarhice* ale soluției și este dat de relația următoare:

$$N_s = T - 1.$$

Alegerea dintre cele $T - 1$ soluții ale unei structuri cluster a celei mai potrivite soluții cluster rămâne la latitudinea cercetătorului și se face, în principal, în funcție de obiectivele urmărite în analiză.

Algoritmii sau metodele de tip iterativ au ca scop producerea unei structuri cluster formată dintr-o *singură* soluție cluster. O astfel de structură cluster se numește *structură cluster uninivel* și conține o singură cluster, care include un număr *fixat* de clustere.

Remarcă: Algoritmii de clasificare prin partitioanare furnizează *soluții unice*, adică soluții de tip *uninivel*.

Caracteristica principală a algoritmilor de partitioanare este data de faptul că *numărul de clustere este fixat aprioric* de către analistul de informații. Algoritmii din această categorie mai sunt cunoscuți și sub numele de *algoritmi de partitioanare*.

Remarcă: În cazul metodelor de clasificare prin partitioanare, *numărul de clustere este cunoscut aprioric*.

În funcție de natura criteriului utilizat în procesul propriu-zis de clasificare, metodele de analiză cluster pot fi împărțite în două categorii: *metode euristică și metode algoritmice*.

Metodele euristice includ procedurile de clasificare dezvoltate pe baza unei anumite *euristici*. O euristică este o modalitate intuitivă de soluționare a unei anumite probleme particulare. Euristicile reprezintă seturi de reguli sau de recomandări cu caracter general, deduse pe baza unor raționamente teoretice sau pe baza unor observații statistice. În general, conceptul de euristică este opus conceptului de algoritm și este utilizat pentru a defini metode și tehnici non-algoritmice.

Prin natura lor, metodele de clasificare ierarhică sunt metode euristice. Astfel, metoda agregării simple, metoda agregării complete, metoda agregării medii, metoda centroidului sau metoda lui Ward, sunt metode de tip euristic.

Metodele algoritmice includ procedurile de clasificare de tip formal, bazate pe existența unui anumit algoritm de soluționare a problemei. Un algoritm este o mulțime de finită și complet definită de operații, pași sau proceduri, a căror execuție determină obținerea unui anumit rezultat sau a unei anumite soluții. Orice algoritm se compune din trei părți esențiale: *initializarea*, *procedura* sau *schema iterativă* și *criteriul de oprire*.

Deoarece includ toate componentele caracteristice unui algoritm, metodele de clasificare prin partitioanare sunt metode cu natură preponderent algoritmică. Spre deosebire de aceste metode, cele trei componente ale unui algoritm nu se regăsesc în mod explicit și în cazul metodelor de clasificare ierarhică.

8.6.4.1 Analiza cluster de tip ierarhic

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

Ipozitia fundamentală a analizei cluster de tip ierarhic este aceea la nivelul mulțimilor supuse studiului există mai multe niveluri de structurare naturală a obiectelor pe grupe sau clase, evidențiuindu-se o imbricare sau o includere, de tip arborescent, a structurilor conținute la nivel latent în cadrul acestor mulțimi.

Ipoteza de bază a clasificării ierarhice: În cadrul mulțimilor de obiecte analizate se diferențiază o multitudine de structuri de tip latent, care sunt caracterizate printr-o *imbricare de natură arborescentă*.

În cea mai mare parte a lor, algoritmii de clasificare ierarhică sunt algoritmi de tip *heuristic*. Există însă și o categorie aparte de algoritmi de clasificare ierarhică, reprezentată de algoritmii de tip *model formal*, care generează structurile cluster pe baza maximizării verosimilității.

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

Structurile cluster care alcătuiesc arborerile de clasificare includ un număr de clustere diferit. O soluție cluster ce corespunde unui nivel mai ridicat de agregare conține un număr de clustere mai mic cu 1 decât o soluție cluster corespunzătoare proximului nivel ierarhic inferior. Aceasta înseamnă că structurile cluster de tip ierarhic sunt caracterizate prin nivele diferite de agregare, cuprinse între un nivel minim și un nivel maxim.

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.

Numărul de clustere din două structuri cluster succesive diferă printr-o unitate, structura cluster cu nivel mai înalt de agregare conținând cu un cluster mai puțin decât structura cluster precedentă.

Cu cât nivelul de agregare al structurilor cluster este mai ridicat, cu atât similaritățile dintre obiectele unui cluster sunt mai reduse, adică clusterele sunt mai eterogene. Acest lucru se explică prin faptul că un cluster de la un nivel de agregare mai înalt conține un număr mai mare de obiecte decât un cluster de la un nivel de agregare mai redus.

În funcție de condițiile inițiale de la care se pornește în construirea structurilor cluster și de sensul în care se desfășoară construirea acestora, algoritmii de clasificare de tip ierarhic pot fi împărțiti în două mari categorii:

- algoritmi de clasificare prin *agregare, amalgamare sau combinare*;
- algoritmi de clasificare prin *dezagregare sau divizare*.

Algoritmii de dezagregare construiesc clusterele într-o manieră *descendentă*, pornind cu toate obiectele într-un singur cluster și continuând, prin divizarea succesivă a acestuia, până la obținerea unor clustere care conțin câte un singur obiect.

Algoritmii de agregare sau de amalgamare construiesc clustere într-o manieră *ascendentă*, pornind de la clustere care conțin câte un singur obiect și continuând, prin comasarea succesivă a clusterelor, până la obținerea 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 și 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 și unui obiect, fie sub forma a două obiecte.

Numărul de pași necesari pentru obținerea unei soluții cluster de tip ierarhic depinde de numărul de obiecte supuse clasificării și este diferit pentru cele două categorii de metode de clasificare ierarhică.

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

8.6.4.1.1 Metode de clasificare ierarhică prin agregare

Algoritmii de agregare sau amalgamare se bazează în mod exclusiv pe evaluarea disimilarităților dintre clustere, adică pe evaluarea de distanțe intercluster. Datorită simplității lor, rezultatul din naturalețea ideii de comasare, algoritmii de agregare sunt mai frecvent utilizați în activitățile de clasificare, în comparație cu algoritmii bazați pe dezagregare. Indiferent de tipul algoritmului concret care este utilizat, în orice procedură bazată pe agregare construcția *arborelui ierarhic* de clustere presupune parcurgerea următoarelor etape generale:

- inițial se pornește cu un număr de clustere egal cu numărul de obiecte, fiecare cluster fiind alcătuit dintr-un singur obiect, respectiv:

$$\omega_1^{(1)} = \{o_1\}, \quad \omega_2^{(1)} = \{o_2\}, \dots, \omega_T^{(1)} = \{o_T\};$$

- ulterior, de-a lungul a mai multor etape, clusterele inițiale sunt succesiv aggregate în vederea obținerii unor clase din ce în ce mai complexe. Numărul total de etape ale procedurii este egal cu $T - 1$. Agregarea este făcută pe baza unei măsuri de disimilaritate între clusterele existente la un moment dat, respectiv pe baza uneia dintre distanțele specifice. În fiecare etapă, pe care o vom nota cu t , sunt aggregate doar două clustere, respectiv acele clustere pentru care distanța dintre ele este *minimă*, în comparație cu distanțele dintre orice două clustere existente în acea etapă. Această distanță se numește *distanță de agregare* și poate fi definită astfel:

$$d_{\text{aggregare}}^{(t)} = \min_{i,j; i \neq j} \{d_{\omega_i^{(t)}, \omega_j^{(t)}}\} \quad i,j = 1,2,\dots,k_t,$$

unde k_t reprezintă numărul de clustere existente în etapa t . Cele două clustere care se comasează într-unul singur sunt clusterele pentru care se obține distanța de agregare. Distanța de agregare se numește *prag de agregare* și este specifică fiecărei etape între care există o distanță egală cu distanța de agregare. Structura cluster obținută în etapa t este de forma:

$$\omega_1^{(t)} = \{o_1, o_2, \dots, o_{T_1}\} \quad \omega_2^{(t)} = \{o_{T_1+1}, o_{T_1+2}, \dots, o_{T_2}\} \dots, \omega_{k_t}^{(t)} = \{o_{T_{k_t-1}+1}, o_{T_{k_t-1}+2}, \dots, o_T\}.$$

Pe măsura construirii ierarhiei cluster, pragul de agregare crește continuu, iar numărul de clustere se reduce cu 1 în fiecare etapă. Ca urmare a relaxării succesive a pragului de agregare, gradul de agregare a obiectelor în clustere crește continuu.

- în ultima etapă a agregării toate obiectele sunt incluse într-un singur cluster, respectiv:

$$\omega^{(T-1)} = \{o_1, o_2, \dots, o_T\}.$$

Această procedură de clasificare pe bază de *agregare* este comună tuturor algoritmilor din această categorie. Diferențele

dintre algoritmii de clasificare ierarhică prin agregare sunt date doar de modul specific în care sunt evaluate distanțele dintre clustere.

În cadrul figurii următoare sunt vizualizate etapele necesare pentru o clasificare de tip ierarhic prin metode de agregare.

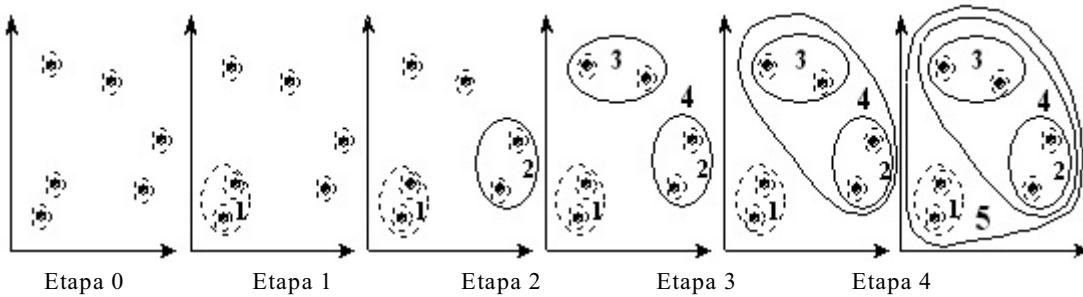


Figura 8.8: Ilustrarea grafică a etapelor clasificării ierarhice prin agregare

Evaluarea distanțelor dintre clusterele obținute la un moment dat din desfășurarea analizei cluster de tip agregare ierarhică, exceptând prima etapă în care clasele sunt alcătuite din câte un singur obiect, poate fi făcută folosind oricare dintre metodele de măsurare a distanțelor dintre clustere, metode prezentate anterior.

Spre deosebire de cazul clasificării ierarhice prin agregare, în cazul procedurii bazate pe dezagregare se procedează oarecum invers. Se pornește cu un cluster care include toate obiectele și din acesta sunt diferențiate clustere din ce în ce mai mici, până când se obțin clustere formate din câte un singur obiect.

Datorită faptului că tehniciile de clusterizare bazate pe agregare sunt cele mai frecvent utilizate, vom prezenta în continuare principalele tipuri ale acestora. Construirea arborilor de clustere prin dezagregare este similară celei obținute prin agregare.

În funcție de tipul distanțelor utilizate pentru agregarea clusterelor, există patru metode euristiche de clasificare ierarhică: *metoda agregării simple*, *metoda agregării complete*, *metoda agregării medii*, *metoda centroidului* și *metoda lui Ward* sau *metoda varianței*.

8.8.4.1.1.1 Metoda agregării simple

În analiza cluster bazată pe *agregare simplă* afectarea unui obiect la un cluster se face numai dacă acel obiect are un anumit grad de disimilaritate cu unul dintre obiectele care aparțin deja clusterului. Clusterizarea de acest tip se mai numește și *analiză cluster de distanță minimă* sau *analiză cluster de tip MIN*.

Metoda agregării simple se bazează pe exprimarea proximității dintre două clustere prin intermediul distanței dintre cele mai apropiate obiecte din cele două clustere. Evaluarea acestei distanțe se face cu ajutorul *metodei celor mai apropiatați vecini*.

Definiție: *Metoda agregării simple* este o metodă de clasificare ierarhică de tip ascendent, care comasează în fiecare etapă a clasificării acele două clustere pentru care *distanța dintre cei mai apropiatați vecini este cea mai mică*, în comparație cu alte perechi de clustere.

În figura următoare, este ilustrat felul în care sunt comasate două clustere în cazul utilizării metodei agregării simple.

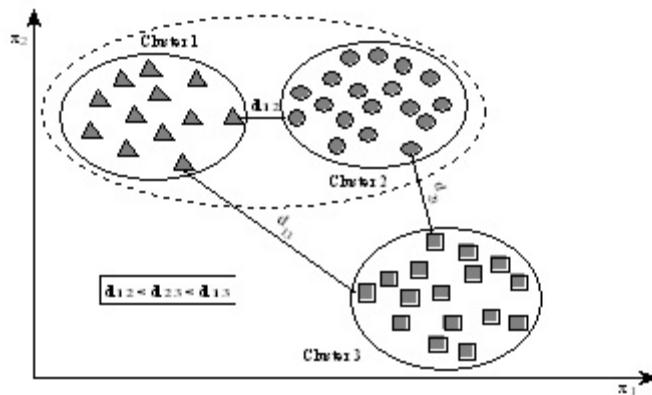


Figura 8.9: Comasarea clusterelor după *metoda agregării simple*

Cea mai mică distanță între cei mai apropiatați vecini din cele trei posibile de clustere este distanța d_{12} , care corespunde perechii de clustere (**cluster1, cluster2**). Ca urmare, clusterul 1 va fi comasat cu clusterul 2, rezultând un nou cluster, care va conține obiectele din cele două clustere.

8.8.4.1.1.2 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 distanțe de agregare care este distanța dintre *cele mai depărtate obiecte* din acele clustere. Clusterizarea de acest tip se mai numește și *analiză cluster de distanță maximă* sau *analiză cluster de tip MAX*.

În cazul metodei agregării complete evaluarea distanțelor dintre clustere se face cu ajutorul metodei celor mai depărtați vecini. Aceasta înseamnă că distanța dintre două clustere este considerată a fi în acest caz distanța cea mai mare dintre oricare două puncte aparținând celor două clustere.

Definiție: Metoda agregării complete este o metodă de clasificare ierarhică de tip ascendent, care comasează în fiecare etapă a clasificării acele două clustere pentru care distanța dintre cei mai depărtați vecini este cea mai mică, în comparație cu alte perechi de clustere.

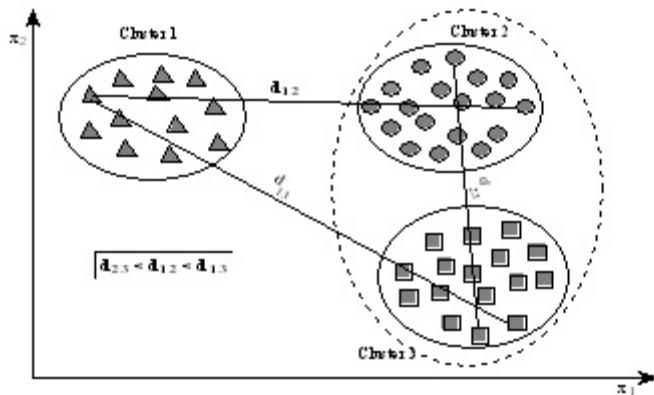


Figura 8.10: Comasarea clusterelor după metoda agregării complete

8.8.4.1.3 Metoda agregării medii

Metoda agregării medii este o metodă de clusterizare similară celor două metode menționate anterior, cu deosebirea că evaluarea distanței dintre două clustere este considerată a fi media distanțelor care separă obiectele aparținând celor două clustere.

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

Definiție: Metoda agregării medii este o metodă de clasificare ierarhică de tip ascendent, care comasează în fiecare etapă a clasificării acele două clustere pentru care distanța medie dintre toate perechile formate cu obiecte din cele două clustere este cea mai mică, în comparație cu alte perechi de clustere.

8.8.4.1.4 Metoda centroidului

Metoda centroidului este o metodă de clasificare ierarhică ascendentă, în care distanțele dintre clustere sunt evaluate cu ajutorul metodei centroidului. Ideea de bază a metodei centroidului este aceea de obținere a unui nou cluster prin comasarea a două clustere existente, în funcție de distanța cea mai mică dintre centroizii clusterelor care sunt verificate în scopul comasării.

Definiție: Metoda centroidului este o metodă de clasificare ierarhică de tip ascendent, care comasează în fiecare etapă a clasificării acele două clustere pentru care distanța dintre centroizii celor două clustere este cea mai mică, în comparație cu alte perechi de clustere.

Două clustere sunt comasate într-un nou cluster dacă și numai dacă distanța dintre centroizii lor este cea mai mică dintre toate distanțele dintre centroizii oricărora două clustere care aparțin configurației cluster disponibile. În figura următoare este vizualizat modul de comasare a două clustere folosind metoda centroidului.

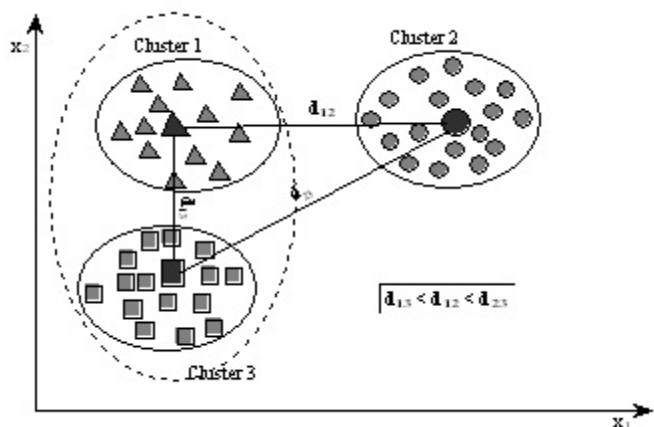


Figura 8.11: Ilustrarea metodei centroidului

8.8.4.1.1.5 Metoda lui Ward

Metoda lui Ward, cunoscută și sub numele de metoda minimei varianțe intracluster, este una dintre cele mai cunoscute și mai eficiente metode de clasificare ierarhică prin agragare.

În virtutea acestei metode atribuirea unui obiect la un cluster se face numai dacă această atribuire minimizează varianța din cadrul clusterului. Pe post de criteriu general de afectare a obiectelor la clustere este considerată minimizarea sumelor elementelor de pe diagonala matricii comune de covarianță a clusterelor, adică *minimizarea trasei matricii de covarianță intraclace*.

Metoda lui Ward este o metodă de evaluare a distanței dintre două clustere care se bazează pe maximizarea gradului de omogenitate a clusterelor.

Definiție: *Metoda lui Ward* este o metodă de clasificare ierarhică de tip ascendent, care comasează în fiecare etapă a clasificării cele două clustere pentru care suma pătratelor abaterilor la nivelul clusterului rezultat din comasare este cea mai mică, în comparație cu alte perechi de clustere.

Metoda lui Ward nu este o metodă propriu-zisă de calcul a distanțelor dintre clustere, ci o metodă de formare a clusterelor pe baza *maximizării gradului de omogenitate a clusterelor*.

Ca măsură a gradului de omogenitate a clusterelor este utilizată suma pătratelor abaterilor, numită *suma pătratelor abaterilor intracluster*. Gradul de omogenitate a unui cluster se consideră a fi cu atât mai mare cu cât suma abaterilor intracluster este mai mică.

Distanța Ward se evaluează pentru toate combinațiile posibile de comasare într-un singur cluster a oricărora două clustere din configurația inițială.

8.6.4.1.2 Metode de divizare

Metodele de clasificare prin divizare, numite și metode de tip *descendent*, sunt analoage cu metodele aglomerative, cu deosebirea că derularea acestora se desfășoară într-o manieră *inversă*. Ca și în cazul metodelor de agregare, soluțiile obținute cu ajutorul metodelor divizative sunt *ierarhii de clustere*, care pot fi reprezentate prin intermediul arborilor cluster sau dendrogramelor.

Algoritmii de clasificare ierarhică pe bază de divizare se caracterizează prin faptul că, inițial, se pornește cu un singur cluster, care conține toate obiectele care trebuie supuse clasificării. Ulterior, clusterul inițial este divizat succesiv, până când se obțin clustere formate dintr-un singur obiect.

Spre deosebire de metodele de clasificare ierarhică prin agregare, care sunt, într-un anumit fel mai naturale, metodele de clasificare ierarhică prin divizare sunt caracterizate de o complexitate mai ridicată. O metodă de clasificare ierarhică prin divizare produce o structură cluster în $2 \cdot (T - 1)$ etape, unde T este numărul de obiecte clasificate.

Datorită existenței unui număr suficient de mare de algoritmi de clasificare ierarhică prin agregare, numărul de algoritmi dezvoltăți pentru clasificarea ierarhică prin divizare este extrem de mic. Unul dintre cei mai cunoscuți algoritmi din această categorie este algoritmul DIANA. Un alt algoritm de acest tip este cel bazat pe metoda clasificării monotetice, însă acesta poate fi utilizat numai în cazul în care variabilele care descriu obiectele sunt de tip binar. Vom prezenta în cele ce urmează fiecare dintre cei doi algoritmi de divizare menționați.

8.6.4.2 Algoritmi de partitioanare

Algoritmii de partitioanare includ o serie de metode de analiză cluster, cu mult mai performante decât metodele de clasificare ierarhică. Dintre cei mai importanți algoritmi de partitioanare, menținăm: algoritmul celor *K*-medii și algoritmul celor *K*-medoizi.

8.7. Recunoașterea formelor cu ajutorul tehniciilor de analiză discriminantă

În mod frecvent, în analiza datelor apare necesitatea studierii unor populații care sunt *eterogene* din punct de vedere al caracteristicilor analizate, fapt care complică procesul de cunoaștere a acestor populații și impune efectuarea unui demers științific specific. Expresia cea mai semnificativă a populațiilor de tip eterogen este întâlnită în special în domeniul statisticii, econometriei și analizei datelor, fiind reprezentată chiar de cantitățile foarte mari de informație care trebuie prelucrată, sintetizată și interpretată.

În cazul cercetării unor populații de acest tip, pentru ca rezultatele investigării să capete consistență și relevanță, este necesară o împărțire, o divizare a acestor populații în subpopulații cu un anumit grad de omogenitate, urmând ca analizele și procesul de modelare implicate în studierea respectivei populații să se facă în mod diferențiat, pentru fiecare subpopulație în parte.

Formularea unor concluzii corecte și robuste cu privire la manifestarea populațiilor caracterizate de un grad mai mare sau mai mic de eterogenitate nu este posibilă decât dacă analiza ia în considerare structurarea acestor populații pe categorii.

În alte situații, cum sunt cele în care sunt analizate diverse entități economico-sociale, considerate a proveni din populații cu caracteristici foarte diferite, există interesul de a identifica, de a recunoaște, originea acestor entități, și de a obține o încadrare corectă a acestora în anumite clase reprezentative pentru populația de origine. Situațiile acest fel depășesc sfera economico-financiară, ele întâlnindu-se în mod frecvent într-o mare varietate de alte domenii importante ale științei, cum ar fi: informatică, biologie, antropologie, medicina, sociologia, geologia, meteorologia etc.

În domeniul economico-financiar, entitățile care fac obiectul problemelor legate de stabilirea apartenenței la o anumită grupă sau clasă pot fi firme, clienți ai unei bănci, cumpărători ai unui produs, unități administrativ-teritoriale, piețe de bunuri sau servicii etc.

Procedura generală de stabilire, pe baza unor caracteristici definitorii și utilizând metode și tehnici specifice, a apartenenței unor obiecte la anumite grupe sau clase dinainte cunoscute poartă numele de **analiza discriminantă**.

Analiza discriminantă reprezintă procesul de utilizare a unei game variate de metode, tehnici și algoritmi în scopul de a determina care dintre caracteristicile unor anumite obiecte au cea mai mare relevanță din punct de vedere al recunoașterii apartenenței acestor obiecte la anumite clase aprioric definite și de a stabili apartenența cea mai probabilă a obiectelor la diferite clase.

Stabilirea apartenenței obiectelor unei populații la anumite clase are la bază proprietățile sau caracteristicile obiectelor respective, care sunt reprezentate la nivel formal prin intermediul unor variabile, notate cu x_1, x_2, \dots, x_N .

În general, se poate spune că analiza discriminantă se ocupă cu rezolvarea următoarelor trei categorii de probleme:

- determinarea aceluia set optimal de caracteristici ale unor obiecte, care să permită *cea mai bună discriminare* între două sau mai multe tipuri de obiecte;
- utilizarea variabilelor din setul optimal de caracteristici pentru deducerea unor *criterii* sau *reguli* pe baza cărora se poate face *separarea populației* studiate pe clase sau grupe distincte;
- utilizarea setului de caracteristici cu cea mai mare putere discriminatoare și a criteriilor de separare identificate pentru *clasificarea* unor obiecte, a căror apartenență nu este cunoscută, în clasele grupele sau clasele populației studiate; clasificarea de noi obiecte, pe baza variabilelor discriminante și a criteriilor de separare, este cunoscută sub numele de *predicție*.

Variabilele din setul optimal de caracteristici se numesc *variabile descriptor* și pot fi reprezentate fie de întreaga mulțime de variabile care descriu obiectele, fie doar de o submulțime a acesteia. Aceasta înseamnă că mulțimea variabilelor descriptor este o mulțime de forma:

$$\{x_1, x_2, \dots, x_n\} \subseteq \{x_1, x_2, \dots, x_N\}.$$

Variabilele descriptor nu sunt folosite în procesul de clasificare în mod direct, ca atare, ci sub o formă transformată, reprezentată de variabilele discriminante.

Criteriile care trebuie deduse în vederea separării claselor din populația analizată sunt utilizate pentru construirea unor ecuații sau funcții, care definesc *puncte*, *curbe* sau *suprafețe* de separare a acestor clase. Ecuațiile sau funcțiile utilizate pentru separarea claselor sunt cunoscute și sub numele de *clasificatori*.

Funcțiile pe baza cărora se face separarea claselor se numesc *funcții discriminante*, *funcții de clasificare* sau *funcții scor*, sunt definite în raport cu *variabilele descriptor* ale obiectelor și servesc la determinarea unor noi variabile, numite *variabile discriminante* sau *variabile scor*.

Legătura dintre cele trei categorii de elemente informaționale ale analizei discriminante, respectiv variabilele descriptor, variabilele discriminante și funcția discriminantă este dată de relația:

$$d_i = D_i(x_1, x_2, \dots, x_n), \quad i=1,2,\dots,p,$$

unde x_1, x_2, \dots, x_n reprezintă *variabilele descriptor*, d_i este *variabila discriminantă*, iar D_i este *funcția discriminantă*.

După cum o să vedem în continuare, în majoritatea covârșitoare a cazurilor de folosire a analizei discriminante, funcțiile discriminante sunt *funcții liniare* de forma:

$$d_i = \beta_1^{(i)} \cdot x_1 + \beta_2^{(i)} \cdot x_2 + \dots + \beta_n^{(i)} \cdot x_n = \beta^i \cdot x, \quad i=1,2,\dots,p.$$

Numărul de funcții discriminante, adică p , este determinat de numărul variabilelor descriptor și de numărul claselor existente la nivelul populației studiate.

Variabilele discriminante d_1, d_2, \dots, d_p determină un nou spațiu p -dimensional, numit *spațiu discriminant*, ale căruia axe sunt reprezentate de vectorii $\beta^{(i)}$ și în contextul căruia se va face clasificarea efectivă obiectelor. Valorile variabilelor discriminante d_i sunt rezultatul evaluării funcțiilor discriminante D_i pentru un anumit obiect fixat și se numesc *scoruri discriminante*. Scorurile discriminante sunt utilizate ca *indicatori* în clasificarea propriu-zisă a obiectelor.

Funcțiile discriminante se identifică pe baza unor criterii specifice de discriminare, iar parametrii acestor funcții se estimează pe baza informațiilor conținute de un eșantion particular de observații (obiecte, forme etc.), extras din populația analizată.

După ce au fost selectate variabilele discriminante și au fost construite funcțiile discriminante, acestea pot să fie utilizate în efectuarea de *predicții* cu privire la apartenența la o clasă sau alta a unor noi obiecte.

Din formularea acestor trei categorii de probleme, rezultă că analiza discriminantă poate fi folosită atât în scopul *descrierii și studierii unor populații eterogene*, prin intermediul unor variabile relevante, cât și în scopul *realizării de predicții* cu privire la apartenența unor obiecte la clasele acestor populații.

Prin conținutul său și prin natura procedurilor și tehniciilor pe care le utilizează, analiza discriminantă este echivalentă cu rezolvarea unei probleme de *predicție*, rezultatul predicției constând din identificarea apartenenței unui obiect la o anumită clasă dintr-o mulțime cunoscută de clase posibile.

Facilitățile deosebite pe care le oferă analiza discriminantă, ca instrument de investigare științifică, au o importanță specială pentru problematica domeniului economico-financiar, domeniu în care utilizarea acestui instrument prevalează în raport cu alte instrumente similare.

De la domeniul finanțier-bancar, în care analiza discriminantă este utilizată, cu precădere, pentru clasificarea firmelor solicitatoare de credite, și până la domeniul marketingului, în care analiza discriminantă este utilizată, printre altele, în probleme de segmentare a pieții, analiza discriminantă oferă multiple și interesante posibilități de analiză și cunoaștere.

În domeniul economic, cele mai multe, mai utile și mai interesante aplicații ale analizei discriminante sunt cele legate de evaluarea șanselor de viabilitate pe care le au diferite activități sau firme în care se pot face investiții sau cărora băncile le pot acorda credite. În acest sens, analiza discriminantă poate fi folosită pentru fundamentarea unor decizii cum ar fi: vânzarea sau cumpărarea de acțiuni, acordare de credite, cumpărarea sau vânzarea de firme etc.

8.7.1. Definirea problemei analizei discriminante

Privită într-un mod foarte general, rezolvarea unei probleme de clasificare cu ajutorul analizei discriminante presupune deducerea unor reguli sau criterii astfel încât, după cunoașterea vectorului x de proprietăți ale unui obiect care aparține unei populații Ω , să se poată lua o decizie cu privire la clasificarea respectivului obiect într-o din cele K clase posibile sub care poate fi structurată populația Ω .

Ipoteza fundamentală a analizei discriminante este aceea că mulțimea Ω este formată din elemente eterogene și că, în mod implicit, în cadrul mulțimii Ω există un număr de K clase, notate cu $\omega_1, \omega_2, \dots, \omega_K$ și numite *clase reale* sau *clase inițiale*, a căror alcătuire nu este cunoscută complet și care au următoarele proprietăți:

- $\omega_k \subset \Omega, \quad k=1,2,\dots,K$
- $\omega_1 \cup \omega_2 \cup \dots \cup \omega_K = \Omega$

Reamintim faptul că, în general, clasele inițiale ale mulțimii Ω sunt considerate a fi *nedisjuncte*, adică există posibilitatea ca:

$$\omega_i \cap \omega_j \neq \emptyset.$$

Distribuția statistică a obiectelor în cadrul fiecărei clase reale ω_k este descrisă cu ajutorul densităților multidimensionale de probabilitate condiționată ale claselor, adică cu ajutorul funcțiilor $f_{\omega_k}(x)$, a căror formă se presupune a fi cunoscută.

În cadrul analizei discriminante, cea mai mare importanță, atât din punct de vedere teoretic, cât și din punct de vedere practic, o au nu clasele reale, ci clasele de predicție, pe care le vom defini în continuare.

Vom considera mulțimea Ω și clasele reale $\omega_1, \omega_2, \dots, \omega_K$ din care aceasta este formată inițial. Scopul principal al analizei discriminante este acela de a identifica o modalitate eficientă de structurare a mulțimii Ω sub forma unui număr de K regiuni sau clase. Regiunile sub forma cărora trebuie partajată mulțimea Ω în cazul problemei analizei discriminante, pe care le vom nota cu $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_K$ se numesc *clase de predicție* sau *clase de clasificare* și au următoarele proprietăți:

- $\tilde{\omega}_k \subset \Omega, \quad k=1,2,\dots,K$
- $\tilde{\omega}_1 \cup \tilde{\omega}_2 \cup \dots \cup \tilde{\omega}_K = \Omega$
- $\tilde{\omega}_i \cap \tilde{\omega}_j = \emptyset, \quad \forall i,j=1,2,\dots,K; i \neq j$

Cea de-a doua proprietate se referă la faptul că orice obiect din mulțimea Ω trebuie să fie clasificat. Proprietatea a treia impune necesitatea ca oricare dintre obiectele mulțimii Ω să fie clasificat *numai într-o singură clasă*.

O clasificare poate fi considerată ca fiind perfectă, adică neafectată de erori, dacă și numai dacă există o coincidență perfectă între orice clasă de predicție $\tilde{\omega}_k$ și clasa reală omoloagă ω_k . Acest lucru nu este totdeauna posibil din cauza consecințelor pe care le implică proprietatea de disjuncție a claselor de predicție. După cum se poate observa, spre deosebire de clasele reale $\omega_1, \omega_2, \dots, \omega_K$, care pot avea anumite suprapunerii, clasele de predicție $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_K$ trebuie să fie disjuncte două câte două, adică să nu aibă obiecte comune.

Deoarece clasele de predicție $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_K$ sunt disjuncte două câte două, ele apar ca fiind niște *trunchieri* ale claselor reale $\omega_1, \omega_2, \dots, \omega_K$, ceea ce înseamnă că ele pot fi privite ca fiind definite sub forma unor restricții impuse asupra claselor reale. Ca rezultat al trunchierii claselor reale omoloage, clasele de predicție vor dифeri, mai mult sau mai puțin, de clasele reale, astfel încât între o clasă de predicție și o clasă reală omoloagă vom avea relația:

$$\tilde{\omega}_k \neq \omega_k, \quad k=1,2,\dots,K.$$

Diferențele care există între clasele de predicție și clasele reale, rezultate din faptul că o clasă de predicție este obținută prin trunchierea unei clase reale, reprezentă expresia posibilității ca anumite obiecte să fie clasificate incorrect.

O consecință imediată a modului în care sunt definite clasele de predicție este aceea că fiecare clasă de predicție reprezintă, de fapt, o submulțime a clasei reale omoloage, adică:

$$\tilde{\omega}_k \subseteq \omega_k, \quad k=1,2,\dots,K.$$

Pe de altă parte, deoarece clasele de predicție trebuie să includă toate obiectele mulțimii Ω , se verifică relația:

$$\tilde{\omega}_1 \cup \tilde{\omega}_2 \cup \dots \cup \tilde{\omega}_K = \omega_1 \cup \omega_2 \cup \dots \cup \omega_K = \Omega.$$

În aceste condiții, este evident că, atâtă timp cât fiecare clasă reală reprezintă un câmp complet de evenimente, orice clasă de predicție, care este o submulțime a clasei reale omoloage, apare ca fiind un câmp *incomplet* de evenimente.

8.7.2 Separarea claselor în spațiul formelor

Prima și cea mai dificilă problemă care trebuie rezolvată în analiza discriminantă este aceea a separării claselor de predicție în cadrul mulțimii Ω . Modalitatea cea mai directă de separare a claselor mulțimii Ω este reprezentată de definirea în spațiu a unor *suprafețe de separare* sau *suprafețe de decizie*. Aceste suprafețe de separare sunt cele care determină trunchierea claselor de predicție $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_K$ și ele trec, cu necesitate, prin mulțimea de obiecte care aparțin intersecției claselor pe care le separă.

Din considerente legate de simplificarea proceselor de clasificare, sunt utilizate, de regulă, suprafețele de separare de tip *liniar*, cum ar fi dreptele, planele sau hiperplanele. Suprafețele de separare sunt definite prin intermediul unor funcții cunoscute sub numele de *funcții discriminant*.

Informațiile necesare construirii suprafețelor de separare a claselor de predicție sunt reprezentate de un eșantion de volum T de obiecte extrase din populația Ω , obiecte a căror apartenență la clasele $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_K$ este cunoscută aprioric și cu exactitate.

Găsirea unei modalități eficiente de separare pe clase disjuncte a elementelor mulțimii Ω constituie o problemă dificilă, mai ales din cauza existenței în mulțimea Ω a unor obiecte care aparțin simultan la două clase reale diferite. Afectarea obiectelor de acest tip la o clasă sau alta se poate face numai sub rezerva calculului probabilistic.

După identificarea unei modalități corespunzătoare de separare a elementelor mulțimii Ω pe clasele de predicție $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_K$, sarcina principală a analizei discriminante este aceea de a decide cu privire la apartenența la cele K clase a unor noi obiecte din mulțimea Ω sau de a face predicii privind apartenența acestor obiecte. Aceasta înseamnă că problema de clasificare cu ajutorul analizei discriminante poate fi formulată astfel: dându-se un obiect pentru care se cunoaște vectorul x de valori ale caracteristicilor acestuia, se cere să se determine apartenența acestui obiect la una dintre cele K clase posibile, $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_K$, ale mulțimii Ω .

Principala problemă care trebuie rezolvată în cadrul analizei discriminante este aceea a construirii criteriilor sau regulilor de clasificare, pe baza cărora se pot face predicii privind apartenența unor forme noi, cu apartenență inițială necunoscută. Criteriile de clasificare mai sunt cunoscute și sub numele de *clasificator*, iar deducerea acestor criterii se numește *formare* a clasificatorului.

Clasificatorul este, de fapt, un algoritm cu ajutorul căruia se stabilește apartenența cea mai probabilă a unei forme la o anumită clasă de predicție. Formarea clasificatorului se face pe baza informațiilor conținute într-un eșantion de forme a căror apartenență este cunoscută aprioric și care se mai numește *set de formare*.

Eșantionul care reprezintă setul de formare este extras din populația analizată și conține datele primare utilizate în orice analiză discriminantă. În anumite situații, pentru formarea clasificatorului poate fi folosită, în mod efectiv, numai o parte a eșantionului disponibil, cealaltă parte urmând să fie utilizată pentru testarea și validarea abilității pe care o are clasificatorul obținut pe baza setului de formare de a clasifica în mod corect forme a căror apartenență este cunoscută. În acest fel, setul de formare poate să reprezinte doar o parte a eșantionului disponibil. Partea eșantionului utilizată pentru testarea și validarea puterii de discriminare a clasificatorului se numește *set de predicție*. De multe ori, întregul eșantion disponibil poate fi folosit atât ca set de formare, cât și ca set de predicție, ceea ce înseamnă că cele două seturi pot să coincidă.

Există mai multe modalități de abordare care pot fi utilizate pentru formarea clasificatorului. Printre acestea menționăm: criteriul minimizării costului clasificării, criteriul lui Bayes sau criteriul probabilităților a posteriorice, criteriul funcțiilor discriminante liniare ale lui Fisher, criteriul metric sau criteriul distanței Mahalanobis, criteriul raportului de verosimilitate etc. Utilizarea fiecărui dintre criteriile menționate conduce la obținerea unui clasificator, a căruia esență este în principiu aceeași pentru majoritatea criteriilor menționate.

8.7.3 Clasificatorii de tip liniar

Prima modalitate de abordare a problemelor de clasificare cu ajutorul tehniciilor de analiză discriminantă datează din anul 1933 și a fost propusă de Fisher. Ulterior abordările de acest tip s-au dezvoltat în mod constant, iar aplicațiile bazate pe analiza discriminantă s-au extins la din ce în ce mai multe domenii de activitate și s-au diversificat din ce în ce mai mult.

Cele mai multe și cele mai utile aplicații ale analizei discriminante bazată pe criteriul lui Fisher sunt întâlnite în domeniul financiar-bancar, domeniu în care tehniciile de tip se numesc *tehnici de credit-scoring* și constituie cele mai importante instrumente pentru fundamentarea deciziilor privind acordarea de credite.

Metoda de analiză discriminantă propusă de Fisher este o metodă parametrică, caracterizată prin simplitate și robustețe, și care oferă posibilități de interpretare foarte utile pentru analiză. Simplitatea acestei metode decurge din faptul că utilizarea sa nu necesită decât evaluarea unor estimații pentru parametrii populației și claselor acesteia, parametri reprezentanți de medii, varianțe sau covarianțe. Aceasta reprezintă un avantaj foarte important al analizei discriminante de tip Fisher, în comparație, de exemplu, cu tehniciile de analiză discriminantă bazate pe criteriul Bayes-ian, tehnici a căror utilizare presupune cunoașterea probabilităților apriorice.

Fundamentalul teoretic al analizei discriminante de tip Fisher este reprezentat de analiza varianței. Criteriul lui Fisher definește o modalitate de deducere a funcțiilor discriminante pe baza analizei comparative dintre *variabilitatea intragrupală* și *variabilitatea intergrupală*, la nivelul claselor sau grupelor populației analizate. Funcțiile discriminante deduse pe baza criteriului lui Fisher se mai numesc și *funcții scor* și sunt *funcții liniare*.

După cum am mai menționat, criteriul fundamental care stă la baza împărțirii mulțimii de obiecte Ω în submulțimile $\tilde{\omega}_1, \tilde{\omega}_2, \dots, \tilde{\omega}_K$ este un criteriu mixt, care urmărește *minimizarea variabilității intragrupale* și *maximizarea variabilității*

intergrupale. Utilizarea acestui criteriu combinat asigură cea mai bună diferențiere a claselor sau grupelor populației Ω .

Ideea care stă la baza criteriului lui Fisher este aceea a determinării unor *direcții* sau *axe*, astfel încât, de-a lungul acestora, clasele multimii Ω să se diferențieze *cât mai mult* între ele și, în același timp, fiecare clasă să aibă un *grad de omogenitate* *cât mai mare*. Cu alte cuvinte, criteriul lui Fisher are ca scop determinarea unor *direcții* de-a lungul cărora variabilitatea intergrupală să fie *cât mai mare*, iar variabilitatea intragrupală să fie *cât mai mică*. Proiecțiile obiectelor pe axele definite de aceste direcții reprezintă sunt noi coordonate ale obiectelor și se numesc *scoruri discriminant*.

Dintron un anumit punct de vedere, analiza discriminantă poate fi considerată ca fiind asemănătoare cu analiza componentelor principale, care are ca scop general identificarea unor axe în raport cu care variabilitatea obiectelor să fie maximă. Deosebirea principală dintre analiza discriminantă și analiza componentelor principale este legată de faptul că în cadrul analizei componentelor principale spațiul cauzal este considerat în integralitatea sa, fără a se face nici o diferențiere între elementele acestuia din punct de vedere al unui anumit criteriu.

În cazul analizei componentelor principale variabilitatea este privită ca o caracteristică generală a populației analizate, fără a se ține seama de existența unei eventuale structurări a acestei populații pe grupe sau clase. În consecință, variabilitatea care face obiectul analizei componentelor principale este considerată ca un tot unitar, fără a exista posibilitatea descompunerii acesteia în raport cu o anumită structură a spațiului cauzal analizat.

Spre deosebire de aceasta, în cazul analizei discriminante se consideră că populația analizată este structurată pe grupe sau clase, iar variabilitatea acestei populații poate fi descompusă sub forma a două componente importante: *variabilitatea intergrupală* și *variabilitatea intragrupală*.

În plus, față de diferența menționată, în analiza discriminantă noile direcții care trebuie identificate nu trebuie să fie în mod obligatoriu ortogonale, spre deosebire de analiza componentelor principale în care direcțiile de variabilitate maximă trebuie să verifice proprietatea de ortogonalitate.

Cea mai importantă problemă a criteriului lui Fisher de discriminare între clasele unei populații Ω este legată de descompunerea variabilității acestei populații. Vom detalia modul în care poate fi descompusă variabilitatea populației în raport cu cele două sensuri ale acesteia: *variabilitatea simplă* - exprimată prin intermediul sumei totale a pătratelor abaterilor și *variabilitatea mixtă* sau *compusă* - măsurată prin intermediul matricii produselor mixte ale abaterilor. Este evident că variabilitatea mixtă poate fi definită numai pentru cazul obiectelor multidimensionale.

Așa cum am precizat mai înainte, determinarea funcțiilor discriminant este echivalentă cu găsirea unor direcții, sau vectori, în raport cu care variabilitatea intragrupală să fie minimă, iar variabilitatea intergrupală să fie maximă. Aceste direcții vor defini axele spațiului discriminant și pot fi identificate sub forma unor combinații liniare de variabilele descriptor selectate în analiză.

Pentru determinarea acestor direcții, vom considera că variabilele descriptor implicate în analiză sunt variabilele $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, ale căror medii sunt $\mu_1, \mu_2, \dots, \mu_n$. Vom nota cu $\mathbf{x}_1^c, \mathbf{x}_2^c, \dots, \mathbf{x}_n^c$ variabilele centrate corespunzătoare variabilelor descriptor, adică:

$$\mathbf{x}_i^c = \mathbf{x}_i - \mu_i, \quad i=1,2,\dots,n,$$

ceea ce înseamnă că:

$$E(\mathbf{x}_i^c) = 0, \quad i=1,2,\dots,n.$$

Vom nota matricea de covarianță a vectorului aleator centrat $\mathbf{x}^c = (\mathbf{x}_1^c, \mathbf{x}_2^c, \dots, \mathbf{x}_n^c)^t$ cu $\Sigma_{n \times n}$, iar matricea produselor mixte ale abaterilor, corespunzătoare realizărilor vectorului aleator \mathbf{x}^c , cu \mathbf{C}_T .

Problema care se pune în cadrul analizei discriminate este aceea de a determina o nouă variabilă \mathbf{d}^c , ca o combinație liniară de forma:

$$\mathbf{d}^c = \beta_1 \cdot \mathbf{x}_1^c + \beta_2 \cdot \mathbf{x}_2^c + \dots + \beta_n \cdot \mathbf{x}_n^c,$$

unde ponderile $\beta_1, \beta_2, \dots, \beta_n$ se determină în aşa fel încât pentru noua variabilă suma pătratelor abaterilor intraclasă să fie minimă, iar suma pătratelor abaterilor interclasă să fie maximă. Această variabilă se numește *variabilă discriminant (centrată)*.

Dacă vom nota cu β vectorul ponderilor $\beta_1, \beta_2, \dots, \beta_n$, atunci variabila \mathbf{d}^c poate fi scrisă sub forma:

$$\mathbf{d}^c = \beta^t \cdot \mathbf{x}^c.$$

Privită ca o funcție de vectorul \mathbf{x}^c , această combinație liniară definește, de fapt, o *funcție discriminant* sau o *funcție scor*, de forma:

$$D(\mathbf{x}^c) = \beta^t \cdot \mathbf{x}^c.$$

Prin urmare, procedura de construire a unei funcții discriminant se reduce la determinarea vectorului β , adică a ponderilor $\beta_1, \beta_2, \dots, \beta_n$.

Trebuie să facem precizarea că natura liniară a funcției discriminant este impusă ca ipoteză inițială și ea nu trebuie considerată ca rezultând din impunerea unui anumit criteriu de performanță privind separabilitatea claselor.

Deoarece variabila \mathbf{d}^c este definită ca o combinație liniară de variabile care au media nulă, rezultă că și media acesteia este nulă, adică:

$$E(\mathbf{d}^c) = \beta_1 \cdot E(\mathbf{x}_1^c) + \beta_2 \cdot E(\mathbf{x}_2^c) + \dots + \beta_n \cdot E(\mathbf{x}_n^c) = 0.$$

În aceste condiții, suma totală a pătratelor abaterilor pentru noua variabilă discriminant \mathbf{d}^c este definită de relația:

$$SP_{T_{d^c}} = \sum_{t=1}^N (d_t^c)^2 = N \cdot \sigma_{d^c}^2,$$

unde $\sigma_{d^c}^2$ reprezintă varianța variabilei \mathbf{d}^c la nivel de populație.

Considerând populația Ω structurată pe clase, suma totală a pătratelor abaterilor variabilei \mathbf{d}^c poate fi descompusă sub forma sumei dintre suma pătratelor abaterilor intragrupale și suma pătratelor abaterilor intergrupale, respectiv:

$$SP_{T_{d^c}} = SP_{w_{d^c}} + SP_{b_{d^c}}.$$

În mod similar, varianța variabilei discriminant \mathbf{d}^c poate fi descompusă ca sumă dintre varianța intragrupală și varianța intergrupală, adică:

$$\sigma_{d^c}^2 = \sigma_{w_{d^c}}^2 + \sigma_{b_{d^c}}^2.$$

Pe de altă parte, având în vedere relația care definește variabila discriminant \mathbf{d}^c , rezultă că varianța acesteia poate fi scrisă sub formă:

$$Var(d^c) = E[d^c \cdot (d^c)^t] = \beta^t \cdot E[x^c \cdot (x^c)^t] \cdot \beta = \beta^t \cdot \Sigma \cdot \beta.$$

Ținând seama de relația de descompunere a matricii de covarianță Σ , respectiv de relația:

$$\Sigma = \Sigma_w + \Sigma_b,$$

varianța variabilei scor \mathbf{d}^c devine:

$$Var(d^c) = \beta^t \cdot \Sigma \cdot \beta = \beta^t \cdot (\Sigma_w + \Sigma_b) \cdot \beta = \beta^t \cdot \Sigma_w \cdot \beta + \beta^t \cdot \Sigma_b \cdot \beta.$$

Luând în considerare descompunerea de mai sus a varianței variabilei scor \mathbf{d}^c , rezultă că mărimele scalare $\beta^t \cdot \Sigma_w \cdot \beta$ și $\beta^t \cdot \Sigma_b \cdot \beta$ reprezintă varianța intragrupală, respectiv varianța intergrupală, corespunzătoare variabilei \mathbf{d}^c .

În aceste condiții, coeficienții combinației liniare care definește variabila discriminant \mathbf{d}^c se determină astfel încât variabila \mathbf{d}^c să aibă o varianță intragrupală minimă și o varianță intergrupală maximă, adică astfel încât raportul:

$$\lambda = \frac{\beta^t \cdot \Sigma_b \cdot \beta}{\beta^t \cdot \Sigma_w \cdot \beta},$$

să fie maxim.

Rezultă că determinarea coeficienților funcției discriminant \mathbf{d}^c poate fi formulată sub formă următoarei probleme de extrem:

$$\max_{\beta} \left\{ \lambda = \frac{\beta^t \cdot \Sigma_b \cdot \beta}{\beta^t \cdot \Sigma_w \cdot \beta} \right\}.$$

Condițiile necesare de extrem pentru această problemă sunt date de anularea derivatelor parțiale ale funcției λ în raport cu componentele vectorului β , ceea ce în exprimare vectorială înseamnă:

$$\frac{\partial \lambda}{\partial \beta} = \frac{2(\Sigma_b \cdot \beta)(\beta^t \cdot \Sigma_w \cdot \beta) - 2(\beta^t \cdot \Sigma_b \cdot \beta)(\Sigma_w \cdot \beta)}{(\beta^t \cdot \Sigma_w \cdot \beta)^2} = 0.$$

Înmulțind condiția obținută cu $\beta^t \cdot \Sigma_w \cdot \beta$ și ținând seama de definirea lui λ , rezultă următoarea formă a condiției de extrem:

$$\frac{(\Sigma_b - \lambda \cdot \Sigma_w) \cdot \beta}{\beta^t \cdot \Sigma_w \cdot \beta} = 0.$$

În concluzie, se poate spune că vectorul β asigură maximizarea raportului λ dacă el este soluție a ecuației:

$$(\Sigma_b - \lambda \cdot \Sigma_w) \cdot \beta = 0,$$

sau a ecuației:

$$(\Sigma_w^{-1} \cdot \Sigma_b - \lambda \cdot I) \cdot \beta = 0.$$

Această ultimă formă a condiției necesare de extrem arată că β este un vector propriu al matricii $\Sigma_w^{-1} \cdot \Sigma_b$, asociat valorii proprii λ a aceleiași matrici. Pentru ca vectorul β , ca soluție a unui sistem omogen, să fie diferit de vectorul nul, este necesar ca valoarea proprie λ să verifice condiția:

$$|\Sigma_w^{-1} \cdot \Sigma_b - \lambda \cdot I| = 0.$$

Deoarece, ca matrice de covarianță, matricile Σ_w și Σ_b sunt simetrice, rezultă că și matricea $\Sigma_w^{-1} \cdot \Sigma_b$ este simetrică, ceea ce înseamnă că valorile proprii ale acesteia sunt reale.

În aceste condiții, a maximiza raportul de mai sus echivalează cu a alege cea mai mare valoare proprie a matricii $\Sigma_w^{-1} \cdot \Sigma_b$ și vectorul propriu corespunzător.

Dacă $\tilde{\lambda}$ este cea mai mare valoare proprie a matricii $\Sigma_w^{-1} \cdot \Sigma_b$ și $\tilde{\beta}$ este vectorul propriu asociat acesteia, adică:

$$\begin{cases} \tilde{\lambda} = \max_{\lambda} \{\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(n)}\} \\ (\Sigma_w^{-1} \cdot \Sigma_b) \cdot \tilde{\beta} = \tilde{\lambda} \cdot \tilde{\beta} \end{cases},$$

atunci raportul este maxim, iar valoarea de maximă a acestuia este $\tilde{\lambda}$, adică:

$$\frac{\tilde{\beta}^t \cdot \Sigma_b \cdot \tilde{\beta}}{\tilde{\beta}^t \cdot \Sigma_w \cdot \tilde{\beta}} = \max_{\beta} \frac{\beta^t \cdot \Sigma_b \cdot \beta}{\beta^t \cdot \Sigma_w \cdot \beta} = \tilde{\lambda}.$$

Componentele vectorului $\tilde{\beta}$ reprezintă coeficienții funcției discriminant liniare $D(\mathbf{x}^c)$, ceea ce înseamnă că funcția discriminant are forma:

$$D(\mathbf{x}^c) = \tilde{\beta}_1 \cdot \mathbf{x}_1^c + \tilde{\beta}_2 \cdot \mathbf{x}_2^c + \dots + \tilde{\beta}_n \cdot \mathbf{x}_n^c.$$

Înlocuind variabilele centrate \mathbf{x}_i^c cu $\mathbf{x}_i - \boldsymbol{\mu}_i$, vom obține exprimarea funcției discriminat în funcție de variabilele discriminant originale, respectiv:

$$D(\mathbf{x}) = \tilde{\beta}_1 \cdot (\mathbf{x}_1 - \boldsymbol{\mu}_1) + \tilde{\beta}_2 \cdot (\mathbf{x}_2 - \boldsymbol{\mu}_2) + \dots + \tilde{\beta}_n \cdot (\mathbf{x}_n - \boldsymbol{\mu}_n).$$

Izolând termenii care conțin mediile variabilelor descriptor, funcția discriminant poate fi scrisă sub forma:

$$D(\mathbf{x}) = -(\tilde{\beta}_1 \cdot \boldsymbol{\mu}_1 + \tilde{\beta}_2 \cdot \boldsymbol{\mu}_2 + \dots + \tilde{\beta}_n \cdot \boldsymbol{\mu}_n) + \tilde{\beta}_1 \cdot \mathbf{x}_1 + \tilde{\beta}_2 \cdot \mathbf{x}_2 + \dots + \tilde{\beta}_n \cdot \mathbf{x}_n.$$

În concluzie, putem spune că funcțiile discriminant ale lui Fisher sunt funcții liniare de forma următoare:

$$D(\mathbf{x}) = \beta_0 + \beta_1 \cdot \mathbf{x}_1 + \beta_2 \cdot \mathbf{x}_2 + \dots + \beta_n \cdot \mathbf{x}_n,$$

unde $\beta_0 = -(\tilde{\beta}_1 \cdot \boldsymbol{\mu}_1 + \tilde{\beta}_2 \cdot \boldsymbol{\mu}_2 + \dots + \tilde{\beta}_n \cdot \boldsymbol{\mu}_n)$ reprezintă termenul liber, iar coeficienții $\beta_1, \beta_2, \dots, \beta_n$ sunt componente ale unui vector propriu al matricii $\Sigma_w^{-1} \cdot \Sigma_b$.

În consecință, variabila discriminant corespunzătoare funcției discriminant $D(\mathbf{x})$ este definită astfel:

$$\mathbf{d} = \beta_0 + \beta_1 \cdot \mathbf{x}_1 + \beta_2 \cdot \mathbf{x}_2 + \dots + \beta_n \cdot \mathbf{x}_n,$$

iar valoarea acesteia pentru o anumită formă \mathbf{x} , adică scorul discriminant, reprezintă evaluarea funcției discriminat D în punctul respectiv. Media și varianța variabilei discriminant \mathbf{d} (necentrată) sunt definite de următoarele relații:

$$E(\mathbf{d}) = \beta_0 + \beta_1 \cdot \boldsymbol{\mu}_1 + \beta_2 \cdot \boldsymbol{\mu}_2 + \dots + \beta_n \cdot \boldsymbol{\mu}_n = \beta_0 + \beta^t \cdot \boldsymbol{\mu}$$

$$\text{Var}(\mathbf{d}) = \beta^t \cdot E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] \cdot \beta = \beta^t \cdot \Sigma \cdot \beta = \beta^t \cdot \Sigma_w^{-1} \cdot \Sigma_b \cdot \beta + \beta^t \cdot \Sigma_b \cdot \beta.$$

Din modul în care este definită funcția discriminant de tip Fisher, rezultă că numărul posibil de funcții discriminant este egal, în principiu, cu numărul de vectori proprii ai matricii $\Sigma_w^{-1} \cdot \Sigma_b$. Teoretic, se poate defini câte o variabilă și o funcție discriminant pentru fiecare vector propriu relevant al acestei matrici. Vectorii proprii ai matricii $\Sigma_w^{-1} \cdot \Sigma_b$ vor defini axele spațiului discriminant.

O problemă importantă legată de definirea variabilelor discriminant și a funcțiilor discriminat este aceea că matricea $\Sigma_w^{-1} \cdot \Sigma_b$ nu este o matrice simetrică, ceea ce înseamnă că este posibil ca nu toate valorile ei proprii să fie reale. Mai mult decât atât, vectorii proprii ai acestei matrici nu mai au proprietatea de ortogonalitate, ceea ce înseamnă că axele spațiului discriminant nu sunt ortogonale.

8.7.4 Definirea funcțiilor discriminant ale lui Fisher

Am prezentat mai înainte modul în care poate fi dedusă o funcție discriminant de tip Fisher. Criteriul pe baza căruia a fost dedusă o funcție discriminant de acest tip este un criteriu mixt, care vizează în mod simultan două aspecte: minimizarea variabilității intragrupale și maximizarea variabilității intergrupale.

O funcție discriminant de tip Fisher se determină ca o combinație liniară de variabilele discriminant, combinație ai cărei coeficienți sunt componente ale unui vector propriu al matricii $\Sigma_w^{-1} \cdot \Sigma_b$. Din această modalitate de definire rezultă, în mod implicit, că pot fi identificate mai multe funcții discriminant.

Numărul maxim posibil de funcții discriminant care pot fi identificate pe baza criteriului lui Fisher este egal cu numărul de valori proprii *distincte* și *strict pozitive* ale matricii $\Sigma_w^{-1} \cdot \Sigma_b$. Deoarece această matrice este de dimensiune $n \times n$, în situația în care ea este strict pozitiv definită și are rangul maxim, rezultă că numărul total de funcții discriminant care pot fi determinate este egal cu n .

Vom prezenta în continuare modul în care pot fi determinate toate funcțiile discriminant posibile. Pentru aceasta vom nota cele n valori proprii ale matricii $\Sigma_w^{-1} \cdot \Sigma_b$ cu $\lambda_1, \lambda_2, \dots, \lambda_n$ și vom presupune că ele sunt ordonate din punct de vedere al valorilor pe care le au astfel:

$$\lambda_1 > \lambda_2 > \dots > \lambda_n > 0.$$

Vom nota cu $\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(n)}$ cei n vectori proprii ai matricii $\Sigma_w^{-1} \cdot \Sigma_b$, asociați, în ordine, cu valorile proprii $\lambda_1, \lambda_2, \dots, \lambda_n$.

Prima funcție discriminant se definește cu ajutorul vectorului propriu $\beta^{(1)}$, care corespunde celei mai mari valori proprii, și are forma următoare:

$$D_1(x) = \beta_0^{(1)} + \beta_1^{(1)} \cdot x_1 + \beta_2^{(1)} \cdot x_2 + \dots + \beta_n^{(1)} \cdot x_n.$$

Deoarece această funcție corespunde celei mai mari valori posibile a raportului dintre varianța intergrupală și varianța intragrupală, ea asigură *cea mai bună separabilitate* a claselor, din punct de vedere al criteriului mixt menționat mai sus. Aceasta înseamnă că proiecțiile obiectelor pe noua axă determinată de vectorul de coeficienți $\beta^{(1)}$ pot fi separate pe clase care se diferențiază în cel mai mare grad posibil și care au cel mai mare grad posibil de omogenitate.

În mod similar, cea de-a doua funcție discriminant se definește cu ajutorul vectorului propriu care corespunde celei de-a doua valori proprii, respectiv:

$$D_2(x) = \beta_0^{(2)} + \beta_1^{(2)} \cdot x_1 + \beta_2^{(2)} \cdot x_2 + \dots + \beta_n^{(2)} \cdot x_n.$$

Fiind determinată pe baza celei de-a doua valori proprii a matricii $\Sigma_w^{-1} \cdot \Sigma_b$, această funcție discriminant corespunde unei valori mai reduse a raportului dintre varianța intergrupală și varianța intragrupală. În consecință, ea asigură o rezoluție mai mică din punct de vedere al separabilității claselor mulțimii Ω . Din acest punct de vedere, este posibil ca proiecțiilor obiectelor pe noua axă care are ca suport vectorul de $\beta^{(2)}$ să le corespundă clase care sunt și mai puțin omogene și se diferențiază și mai puțin între ele.

În sfârșit, cu ajutorul vectorului propriu asociat cu cea mai mică valoare proprie, adică vectorul $\beta^{(n)}$, se determină ultima funcție discriminant, respectiv:

$$D_n(x) = \beta_0^{(n)} + \beta_1^{(n)} \cdot x_1 + \beta_2^{(n)} \cdot x_2 + \dots + \beta_n^{(n)} \cdot x_n.$$

Prin comparație cu celelalte funcții discriminant, această ultimă funcție discriminant asigură cea mai proastă separabilitate între clasele mulțimii Ω .

Puterea de separabilitate din ce în ce mai mică pe care o au funcțiile discriminant d_1, d_2, \dots, d_n , conduce la ideea necesității de a selecta în analiză numai *un anumit număr de funcții discriminant*, în ordinea puterii lor de discriminare.

Numărul efectiv al funcțiilor discriminant care trebuie reținute în analiză, depinde în mod direct de numărul de clase și de numărul de variabile discriminant.

În concluzie la cele arătate mai înainte, putem defini funcțiile discriminant liniare și variabilele discriminant sub forma următoare:

Funcțiile discriminant (Fisher) sunt combinații liniare de variabilele descriptor, de forma:

$$D(x) = \beta_0 + \beta^t \cdot x,$$

unde x este vectorul variabilelor descriptor, iar β este vector propriu al matricii $\Sigma_w^{-1} \cdot \Sigma_b$. Valorile funcțiilor discriminant se numesc *scoruri discriminant*.

Variabilele discriminant sunt combinații liniare de variabilele descriptor, de forma:

$$d = \beta_0 + \beta^t \cdot \mu,$$

unde x și β au semnificația din definiția precedentă. Media și varianța variabilelor discriminant sunt:

$$E(d) = \beta_0 + \beta^t \cdot \mu$$

$$\text{Var}(d) = \beta^t \cdot \Sigma \cdot \beta = \beta^t \cdot \Sigma_w \cdot \beta + \beta^t \cdot \Sigma_b \cdot \beta$$

Odată ce funcțiile discriminant au fost estimate, ele pot fi utilizate pentru efectuarea de predicții cu privire la apartenența unor noi obiecte la clasele de predicție.

Exemplul 8.1

În scopul evidențierii modului clasificare cu ajutorul clasificatorilor liniari, vom considera cazul unei populații de firme comerciale care se grupează în două clase: *firme performante și firme neperformante*. Vom presupune că performanțele firmelor sunt apreciate pe baza a doi indicatori economico-financiari: *rata profitului și rata profitului investit*. De asemenea, vom mai presupune că dispunem de un eșantion de 10 firme din populația de firme analizate, dintre care 6 sunt firme performante, iar 4 sunt firme neperformante. Valorile convenționale ale celor doi indicatori de performanță pentru firmele din cele două categorii se găsesc în tabelul următor.

Firme performante			Firme neperformante		
Firma	Rata profitului	Rata profitului investit	Firma	Rata profitului	Rata profitului investit
f_1	13,0	43,0	f_7	4,0	31,0
f_2	19,0	28,0	f_8	7,0	36,0
f_3	12,0	35,0	f_9	2,0	11,0

f_4	9,0	38,0	f_{10}	11,0	17,0
f_5	9,0	56,0			
f_6	17,0	39,0			
Media	13,17	39,83	Media	6,00	23,75
Varianță	16,9667	87,7667	Varianță	15,33	136,917

Vom mai presupune că densitățile de probabilitate ale claselor sunt de tip normal, adică sunt de forma următoare:

$$f_{\omega_i}(x) = \frac{1}{(\sqrt{2\pi})^n \cdot |\Sigma_i|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(\mathbf{x} - \mu^0)^t \cdot \Sigma_i^{-1} \cdot (\mathbf{x} - \mu^0)} \quad i=1,2$$

În graficul din figura următoare sunt reprezentate densitățile de probabilitate bidimensionale ale celor două clase de predicție, în ipoteza de normalitate, și planul de decizie care asigură separarea acestor clase.

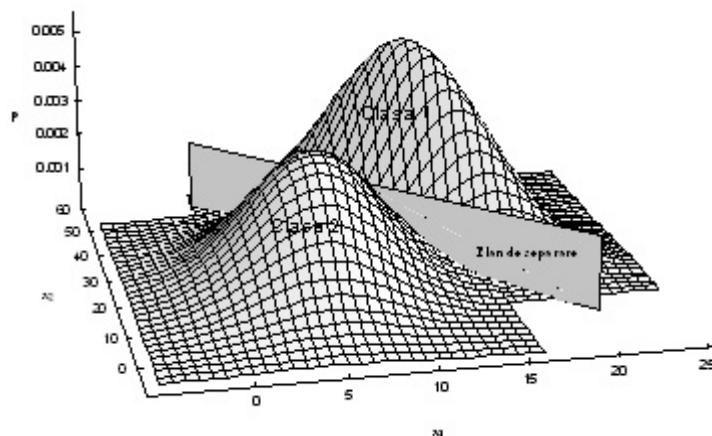


Figura 8.12: Separarea claselor de predicție cu ajutorul planului de decizie

Planul de decizie care separă cele două clase a fost determinat cu ajutorul funcțiilor discriminant pe care le vom calcula în continuare. Pentru a calcula valorile proprii ale matricii $\hat{\Sigma}_w^{-1} \cdot \hat{\Sigma}_b$, va trebui să calculăm, mai întâi, inversa matricii de covarianță intragrupală $\hat{\Sigma}_w$. Vom avea:

$$\hat{\Sigma}_w^{-1} = \begin{pmatrix} 14,537 & -12,426 \\ -12,426 & 94,398 \end{pmatrix}^{-1} = \begin{pmatrix} 0,0775 & 0,0102 \\ 0,0102 & 0,0119 \end{pmatrix},$$

și deci:

$$\hat{\Sigma}_w^{-1} \cdot \hat{\Sigma}_b = \begin{pmatrix} 0,0775 & 0,0102 \\ 0,0102 & 0,0119 \end{pmatrix} \cdot \begin{pmatrix} 13,709 & 30,745 \\ 30,745 & 68,951 \end{pmatrix} = \begin{pmatrix} 1,3763 & 3,0866 \\ 0,5068 & 1,1367 \end{pmatrix}.$$

Vectorii proprii și valorile proprii pentru matricea $\hat{\Sigma}_w^{-1} \cdot \hat{\Sigma}_b$ sunt:

$$\lambda_1 = 2,51303; \quad \beta^{(1)} = (0,938386 \quad 0,345588)^t$$

$$\lambda_2 = 0,000000209; \quad \beta^{(2)} = (-0,913319 \quad 0,407245)^t,$$

iar termenii liberi ai funcțiilor discriminant sunt:

$$\beta_0^{(1)} = -(0,938 \cdot 13,167 + 0,346 \cdot 39,833) = -21,208$$

$$\beta_0^{(2)} = -(-0,913 \cdot 13,167 + 0,407 \cdot 39,833) = -4,195$$

Pe baza termenilor liberi și a celor doi vectori proprii $\beta^{(1)}$ și $\beta^{(2)}$ vom putea defini cele două funcții discriminant liniare astfel:

$$D_1(x) = \beta_0^{(1)} + (\beta^{(1)})^t \cdot x = -21,208 + 0,938 \cdot x_1 + 0,346 \cdot x_2$$

$$D_2(x) = \beta_0^{(2)} + (\beta^{(2)})^t \cdot x = -4,195 - 0,913 \cdot x_1 + 0,407 \cdot x_2$$

Deoarece cea de-a două funcție discriminant corespunde unei valori aproape neglijabile, relevanța sa este minimă și se poate renunța la ea. Ecuatiile discriminat care definesc planele de separare a celor două clase sunt:

$$-21,208 + 0,938 \cdot x_1 + 0,346 \cdot x_2 = 0$$

$$-4,195 - 0,913 \cdot x_1 + 0,407 \cdot x_2 = 0$$

După cum se poate observa, centroidul populației de obiecte aparține primului plan de separare, deoarece el verifică prima ecuație discriminat:

$$-21,208 + 0,938 \cdot \bar{x}_1 + 0,346 \cdot \bar{x}_2 = -21,208 + 0,938 \cdot 10,3 + 0,346 \cdot 33,4 = 0.$$

În figura următoare sunt reprezentate grafic elementele esențiale legate de separarea claselor cu ajutorul primei funcții discriminant.

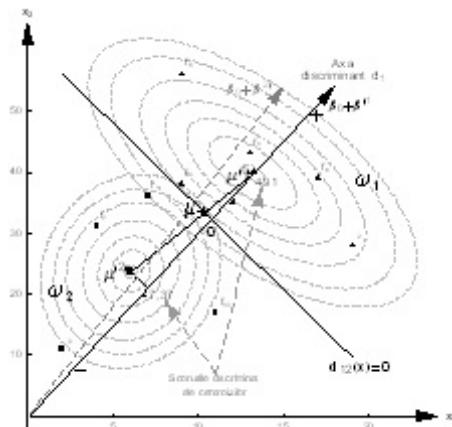


Figura 8.13: Graficul separării claselor de predicție $\tilde{\omega}_1$ și $\tilde{\omega}_2$

Prima axă a spațiului discriminat, respectiv d_1 , are ca suport vectorul propriu $\tilde{\beta}^{(1)}$ și este ortogonală cu dreapta de decizie $d_{12}(x) = 0$. În raport cu această axă se determină scorurile obiectelor, ca proiecții ale obiectelor pe această axă. Aceste scoruri sunt valori ale variabilei discriminant d_1 . Scorurile celor 10 obiecte în spațiul discriminat, calculate cu ajutorul funcțiilor discriminant $D_1(x)$ și $D_2(x)$, sunt prezentate în tabelul următor.

Firme performante			Firme neperformante		
Firma	Scoruri discriminant		Firma	Scor discriminant	
	d_1	d_2		d_1	d_2
f_1	5,851	1,443	f_7	-6,741	4,777
f_2	6,298	-10,145	f_8	-2,198	4,073
f_3	2,148	-0,901	f_9	-15,530	-1,542
f_4	0,370	3,060	f_{10}	-5,011	-7,318
f_5	6,590	10,391			
f_6	8,222	-3,838			
Media	4,913	0,002	Media	-7,370	-0,003
Varianța	8,970	47,579	Varianța	33,097	31,778

Mediile scorurilor pentru toate cele 10 obiecte sunt nule, iar varianțele sunt egale cu 56,250, în cazul variabilei discriminant d_1 , respectiv cu 37,025, în cazul variabilei discriminant d_2 . Estimăriile pentru matricile de covarianta ale variabilelor discriminant d_1 și d_2 , calculate pe baza scorurilor, la nivel de ansamblu și pe clase, sunt următoarele:

$$\hat{\Sigma}^d = \begin{pmatrix} 56,250 & 0,014 \\ 0,014 & 37,025 \end{pmatrix}; \quad \hat{\Sigma}_1^d = \begin{pmatrix} 8,970 & -3,876 \\ -3,876 & 47,579 \end{pmatrix}; \quad \hat{\Sigma}_2^d = \begin{pmatrix} 33,097 & 6,461 \\ 6,461 & 31,778 \end{pmatrix}.$$

Estimăriile pentru matricile de covarianta intraclasă și interclasă ale variabilelor discriminant sunt următoarele:

$$\hat{\Sigma}_w^d = \begin{pmatrix} 16,016 & 0,000 \\ 0,000 & 37,025 \end{pmatrix}; \quad \hat{\Sigma}_b^d = \begin{pmatrix} 40,235 & 0,014 \\ 0,014 & 0,000 \end{pmatrix}.$$

Figura următoare conține reprezentarea grafică a obiectelor în spațiul discriminant, ale cărui axe sunt d_1 și d_2 . În cadrul figurii sunt reprezentate și curbele de nivel ale densităților de probabilitate ale celor două clase.

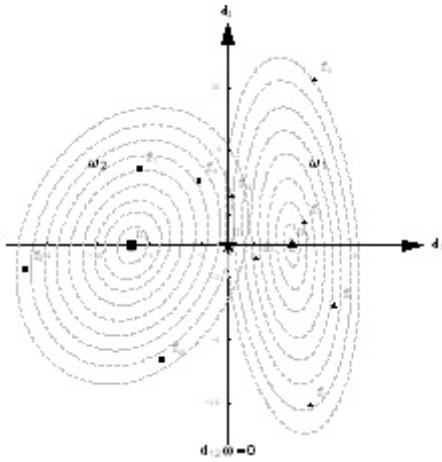


Figura 8.14: Reprezentarea obiectelor în spațiul discriminant

Pentru variabila discriminat d_1 , care este cea mai semnificativă, varianțele intragrupală și intergrupală sunt:

$$\sigma_{w_d}^2 = \frac{5}{9} \cdot 8,970 + \frac{3}{9} \cdot 33,097 = 16,016$$

$$\sigma_{b_d}^2 = \frac{1}{9} \cdot (6 \cdot 4,913^2 + 4 \cdot (-7,370)^2) = 40,233$$

Ca mărime care aproximează calitatea separării claselor, raportul dintre varianța intergrupală și varianța intragrupală este:

$$\lambda_d = \frac{\sigma_{b_d}^2}{\sigma_{w_d}^2} = \frac{40,233}{16,016} = 2,512.$$

Pentru a evidenția avantajele pe care le are utilizarea scorurilor discriminant în raport cu observațiile originale, vom compara raportul în care se află varianța intergrupală și varianța intragrupală la nivelul *observațiilor originale*, pe de o parte, cu raportul în care se află varianța intergrupală și varianța intragrupală la nivelul *scorurilor discriminant*, pe de altă parte.

Deoarece în varianta spațiului inițial obiectele sunt reprezentate exprimate prin intermediul a două variabile descriptor, iar în spațiul discriminant prin intermediul unei singure variabile, pentru a putea face comparația va trebui să exprimăm variabilitatea din spațiul inițial în mod unic, prin intermediul varianței totale. În cazul spațiului cauzal inițial, varianța totală intergrupală este reprezentată de suma elementelor de pe diagonala principală a matricii de covarianță intergrupală $\hat{\Sigma}_b$, respectiv:

$$\hat{V}_T^b = \delta_{11}^b + \delta_{22}^b = 13,709 + 68,951 = 82,660,$$

iar varianța totală intragrupală este reprezentată de suma elementelor de pe diagonala principală a matricii de covarianță intragrupală $\hat{\Sigma}_w$, respectiv:

$$\hat{V}_T^w = \delta_{11}^w + \delta_{22}^w = 14,357 + 94,3982 = 108,935.$$

Este evident că suma celor două varianțe este egală cu varianța totală din spațiul inițial, adică suma elementelor de pe diagonala principală a matricii de covarianță $\hat{\Sigma}$, respectiv:

$$\hat{V}_T^b + \hat{V}_T^w = 82,660 + 108,935 = 191,595 \sim 191,611 = 28,233 + 163,378 = \hat{V}_T.$$

Raportul dintre varianța totală intergrupală și varianța totală intragrupală, corespunzătoare spațiului inițial, este:

$$\lambda_{v_T} = \frac{\hat{V}_T^b}{\hat{V}_T^w} = \frac{82,660}{108,935} = 0,7588.$$

Deoarece $\lambda_d = 2,51 > 0,7588 = \lambda_{v_T}$, putem spune că în spațiul discriminant se obține o mai mare omogenitate intraclasă și o mai mare eterogenitate interclasă, ceea ce corespunde cu criteriul de optimalitate a clasificării.

În concluzie, se poate spune că în spațiul discriminant, atât omogenitatea intraclasă, cât și eterogenitatea interclasă, sunt sensibil mai mari, în comparație cu spațiul inițial.

Evaluarea scorurilor pentru noi obiecte, pe baza funcțiilor discriminat, poate permite stabilirea clasei de apartenență a acestor obiecte. Dacă scorurile obiectelor vor fi negative, obiectele vor aparține clasei situate la stânga în raport cu centroidul populației, iar dacă vor fi pozitive, obiectele vor fi atribuite clasei situate la dreapta față de centroidul populației.

Tema 9: Modelarea cu ajutorul rețelelor neuronale

Rețelele neuronale constituie o modalitate, relativ nouă, de abordare cantitativă a problemelor de analiză și, mai ales, de predicție din domeniul economico-financiar, abordare care se caracterizează, în principal, prin simplitate, robustețe și eficiență.

Odată cu apariția primelor calculatoare electronice, s-au conturat două modalități diferite de abordare în soluționarea problemelor de natură științifică. Prima dintre acestea, a avut la bază ideea achiziționării de cunoștiințe dintr-un anumit domeniu, a constituirii acestora sub forma unor obiecte semantice sau simboluri și a utilizării acestora pentru rezolvarea unor probleme complexe, pe baza definirii unui set de reguli formale. Această modalitate de abordare este cunoscută sub numele generic de *paradigmă simbolică* și este circumscrisă domeniului cunoscut sub numele de *Inteligenta Artificială*.

Cea de-a doua modalitate de abordare, se bazează pe dezvoltarea unor proceduri de soluționare a problemelor științifice, într-o manieră similară cu procesele care au loc la nivelul creierului uman. Această abordare este cunoscută sub numele generic *paradigma conexionalistă*, iar domeniul în care se înscrie poartă numele de *Calcul Neuronal* sau *Rețele Neuronale*.

9.1 Apariția și dezvoltarea calculului neuronal

Începuturile abordării bazate pe modelul neuronal datează din anul 1943, odată apariția primului model al neuronului, model propus de neurofiziologul W. S. McCulloch și de matematicianul W. Pitts.

Un interes deosebit pentru modelul neuronal s-a manifestat după apariția primelor lucrări privind modelarea matematică a proceselor de învățare. O primă apariție de acest fel a avut loc în anul 1947, fiind reprezentată de modelul de învățare al lui D. O. Hebb, care a deschis direcții de nebănuite în domeniul calculului neuronal.

Un alt pas important pe drumul dezvoltării abordărilor de tip neuronal a fost făcut în 1957, o dată cu apariția lucrării lui Frank Rosenblatt, dedicată unui model neuronal simplificat, de natură probabilistică, cunoscut sub numele de *perceptron*. În aceeași direcție de dezvoltare, se încadrează și apariția modelului neuronal *ADALINE* (ADaptive Linear Neuron), model dezvoltat la începutul anilor 60' de către B. Widrow.

Rețele neuronale artificiale au fost create pentru a încerca reproducerea mecanismelor neurofiziologice care au loc la nivelul creierului uman. Prin combinarea unui număr foarte mare de unități de calcul, numite neuroni, sub forma unor sisteme cu grad foarte ridicat de interconectare, se încercă reproducerea unor fenomene complexe, specifice creierului uman, cum ar fi memorarea și inteligența.

Indiferent de faptul că rețelele neuronale pot sau nu pot fi caracterizate în mod efectiv prin inteligență, ele sunt foarte utile pentru activitățile de analiză și predicție, ca modele de natură statistică.

Rețelele neuronale pot fi privite ca reprezentând clase de modele de regresie, de analiză discriminantă sau de reducere a dimensionalității. Deoarece cu ajutorul rețelelor neuronale pot fi detectate relații și structuri complexe conținute în date, ele pot fi utilizate și ca instrumente de predicție pentru numeroase categorii de fenomene reale.

Utilitatea rețelelor neuronale pentru problemele de predicție rezultă, în primul rând, din faptul că utilizarea acestora nu presupune necesitatea formulării unui model matematic explicit, care să descrie legăturile dintre o mulțime de input-uri și o mulțime de output-uri. Pe de altă parte, abordarea bazată pe rețele neuronale este indicată în numeroase situații, ca urmare a faptului că, de cele mai multe ori, activitatea de predicție este o activitate mai importantă din punct de vedere științific, decât explicarea legăturilor dintre fenomene. Deși în principiu, activitatea de predicție presupune cunoașterea și formalizarea legăturilor de cauzalitate, în contextul modelării neuronale efectuarea de predicții poate fi realizată fără o cunoaștere prealabilă a legăturilor de cauzalitate.

Cea mai importantă categorie de rețele neuronale este cea reprezentată de rețelele de tip "reacție directă", sau "feedforward". În rândul acestora, categoria cea mai importantă este aceea a rețelelor de rețele de tip "propagare inversă" sau "backpropagation". Această propagare inversă se referă la modul în care se calculează gradientul funcției eroare, pe baza căruia are loc ajustarea ponderilor sinaptice ale rețelei neuronale. "Back-propagarea" se referă și la diferențele metode de învățare, de instruire sau de antrenare a rețelelor neuronale.

Back-propagarea standard este de fapt o denumire alternativă pentru regula cunoscută în teoria rețelelor neuronale sub numele de "regula Delta Generalizată", care este o tehnică de învățare propusă de Rumelhart, Hinton și Williams în anul 1986, și care este cea mai răspândită tehnică de învățare supervizată a rețelelor neuronale.

Există două variante ale back-propagării standard: propagarea inversă de tip *batch* și propagarea inversă de tip *punctual* sau *incremental*.

Propagarea inversă de tip *batch* este o tehnică de optimizare în care actualizarea ponderilor este bazată pe utilizarea ciclică a întregului set de învățare. Spre deosebire de propagarea inversă de tip *batch*, propagarea inversă de tip *punctual* sau *incremental* se bazează pe actualizarea ponderilor după utilizarea fiecărei forme din setul de învățare.

9.2 Analogia dintre modelele neuronale și creierul uman

Majoritatea preocupărilor și abordările științifice din domeniul inteligenței artificiale au avut ca premiză încercarea de a reproduce procesele informaționale și decizionale care au loc în creierul uman.

Capacitatea creierului uman de a abstractiza, de a face generalizări și de a deduce principii și legități, de a cunoaște, ierarhiza și clasifica pe bază asociere, de a face identificări informaționale exacte și corecte pe baza unor criterii aproximative și vagi, în condițiile existenței unor informații parțiale și cu natură incertă, de a învăța atât în plan teoretic, cât și prin experiență practică, de a stoca o cantitate uriașă de informație extrem de variată a inspirat și a influențat totdeauna atât modalitățile generale de abordare, dezvoltate de umanitate în scopul cunoașterii materiale și spirituale, cât și crearea, dezvoltarea și perfecționarea procedurilor, instrumentelor, metodelor și tehnicielor din domeniul cunoașterii științifice.

Una dintre caracteristicile cele mai interesante ale raționamentului uman este aceea a stabilității și robusteții pe care o au activitățile cerebrale, chiar în condiții de distorsionare puternică a informațiilor sau de accentuată penuria informațională. Mecanismele neuronale au posibilitatea extraordinară de a reconstituire integralitatea sa un anumit context informațional, doar pe baza unor informații parțiale, distorsionate, colaterale sau indirecte.

Ca rezultat al încercărilor de a dezvolta metode și tehnici de cunoaștere care să reproducă mecanismele informaționale și decizionale care au loc la nivelul creierului uman, s-a dezvoltat, începând cu anii 60, un nou domeniu important și vast al cunoașterii umane, inteligența artificială.

Un prim rezultat al preocupărilor inițiate în domeniul inteligenței artificiale a fost concretizat sub forma *sistemelor expert*.

Sistemele expert sunt modele logico-simbolistice care încearcă să simuleze succesiunea de raționamente pe care creierul uman o desfășoară atunci când ia o anumită decizie sau când soluționează o anumită problemă. Încorporând o mulțime de elemente, cunoștințe, proceduri, reguli și principii analoage raționamentului uman, sistemele expert au ca scop luarea unor decizii sau deducerea de răspunsuri.

Luarea deciziilor sau formularea răspunsurilor de către un sistem expert este bazată pe o succesiune de analogii, asocieri și deducții și presupune derularea unor complexe proceduri inferențiale și efectuarea unor lăbiorioase calcule de tip simbolic.

Deși dezvoltarea de sisteme expert a înregistrat o serie de succese în anumite domenii cum ar fi, de exemplu, medicina, în timp s-a văzut că sistemele expert reprezintă o aproximare prea vagă și prea săracă a complexelor și subtilelor procese și mecanisme care caracterizează raționamentul uman.

Cu toate avantajele care decurg din faptul că un sistem expert se referă doar la un segment strict limitat de realitate, sistemul expert rămâne doar o replică mai puțin decât mult decât parțială a mecanismelor gândirii umane.

O altă direcție de abordare în domeniul inteligenței artificiale este aceea bazată pe rețelele neuronale. Cel puțin până în prezent, această modalitate de abordare s-a dovedit a fi mai eficientă și mai apropiată de modul în care lucrează creierul uman, în comparație cu direcția reprezentată de sistemele expert.

Analogia mai mare dintre rețelele neuronale și mecanismul gândirii umane, precum măsura mai mare în care rețelele neuronale reușesc să reproducă inteligența umană rezultă din faptul că rețeaua neuronală are o structură mult mai apropiată de arhitectura biologică a creierului uman.

9.2.1 Structura și funcționalitatea neuronului biologic

Creierul uman este alcătuit din peste 10 miliarde de celule sau unități elementare, cunoscute sub numele de *neuroni*. Aceste unități elementare sunt puternic interconectate între ele, numărul de conexiuni ale fiecărui neuron cu alți neuroni fiind de ordinul zecilor de mii.

Un neuron poate fi privit ca fiind o unitate de prelucrare informațională, care generează o serie de informații de ieșire, pe baza unor informații specifice de intrare. Din punct de vedere fizic, atât informațiile de intrare, primite de un neuron, cât și informațiile de ieșire, transmise de un neuron, sunt reprezentate de semnale electro-chimice.

Structura de intrare a unui neuron este formată dintr-o mulțime de ramuri de conectare numite *dendrite*, iar structura sa de ieșire este formată dintr-o mulțime de ramuri de conectare numite *axoni*. Dendritele unui anumit neuron sunt conectate cu axoni neuronilor de la care acesta primește semnale, după cum axonii neuronului sunt conectați cu dendritele neuronilor către care el transmite semnale.

Punctele în care axonii unor neuroni se conectează cu dendritele altor neuroni se numesc *sinapse neuronale*. Sinapsile sunt localizate pe dendritele unui neuron și sunt de fapt *terminale* la care ajung impulsurile pe care neuronul le primește de la alți neuroni.

Fiecare impuls care ajunge la o sinapsă a unui neuron determină degajarea unei mici cantități de substanță electro-chimică, numită *neurotransmițor*. Această substanță electro-chimică este transmisă în continuare, prin canalele sinaptice, către un *receptor post-sinaptic*, care, în funcție de cantitatea de substanță primită, determină o schimbare a potențialului suprafetei sau membranei dendritice, numit *potențial post-sinaptic*.

Această schimbare de potențial post-sinaptic poate să determine, la rândul său, o creștere sau o scădere a *polarizării*

membranei dendritice. Creșterea polarizării membranei dendritice va contribui, împreună cu modificările de polarizare ale celorlalte dendrite ale neuronului, la o inhibare a generării de semnale de către neuron. Similar, scăderea polarizării membranei dendritice va determina stimularea neuronului de a genera semnale. Mărimea și tipul potențialului post sinaptic depinde atât de cantitatea și tipul substanței neurotransmițătoare, cât și de natura și structura sinapsei.

Funcționalitatea unui neuron poate fi privită ca reprezentând o transformare a semnalelor de intrare în semnale de ieșire. De fapt, activitatea internă a neuronului constă dintr-o sumare sau integrare specifică a semnalelor de intrare, reprezentate de miile de potențiale post-sinaptice, și crearea, pe această bază, a unui potențial electro-chimic, numit *stare internă a neuronului*.

În funcție de mărimea acestui potențial, neuronul *generează* sau *nu generează* un anumit semnal electrochimic, pe care îl transmite, în continuare, neuronilor ale căror dendrite sunt conectate la axonii săi. Generarea de către neuron a semnalului de ieșire, are loc dacă și numai dacă potențialul său, rezultat din sumarea intrărilor, depășește o anumită valoare specifică, numită *prag de excitație*.

Spre deosebire de *logica von Neumann*, pe care se bazează arhitectura și funcționalitatea calculatoarele convenționale actuale, în contextul căreia informația disponibilă este memorată la nivelul unor structuri de stocare bine determinate, în cadrul rețelelor neuronale informația disponibilă este reprezentată sub o formă funcțională, prin intermediul așa-numitelor *ponderilor sinaptice*, asociate conexiunilor rețelei neuronale, fiind distribuită peste întreaga mulțime de conexiuni ale rețelei.

O altă caracteristică importantă a rețelelor neuronale o reprezintă activitatea de *adaptabilitate* continuă a acestora, în raport cu exemple reale, adaptabilitate cunoscută sub numele de *învățare sau instruire*. Din acest punct de vedere, rețelele neuronale reprezintă mecanisme de modelare bazate pe “*învățarea prin exemple*”.

În abordarea tradițională, rezolvarea unei anumite probleme presupune parcurgerea unor etape specifice, cum ar fi: *identificarea modelului, estimarea parametrilor* acestuia, *validarea modelului și utilizarea modelului pentru analiză și predicție*. În cadrul unei astfel de abordări, rezolvarea problemelor specifice fiecărei etape și parcurgerea succesivă a acestor etape au o natură pur *algoritmică* și se bazează pe o cunoaștere cât mai bună a structurilor cauzale ce caracterizează fenomenele modelate. Din aceste motive, în cazul unor probleme complexe sau în cazul în care nu este posibilă cunoașterea sensului și structurilor relațiilor dintre fenomenele modelate, abordarea tradițională devine ineficientă sau imposibil de utilizat.

Din punct de vedere calculatoriu, în contextul abordării tradiționale este presupusă cunoașterea apriorică și exactă a pașilor algoritmului utilizat, cunoaștere concretizată sub forma *programului de calcul*.

Abordarea bazată pe rețele neuronale permite soluționarea unor probleme extrem de complexe, fără a necesita cunoașterea cu precizie a structurii cauzalității fenomenelor implicate în respectiva problemă. Pe baza unei mulțimi de observații empirice referitoare la fenomenele studiate, rețelele neuronale reușesc să reproducă comportamentul real al fenomenelor, fără a mai fi nevoie să se formuleze un model specific pentru aceste fenomene, algoritmul de rezolvare a problemei fiind construit în mod treptat, pe măsura avansării procesului de învățare.

În cadrul rețelelor neuronale, modelul problemei și algoritmul de soluționare a modelului sunt approximate și reproduse într-o manieră simplificată, pe baza definirii unui ansamblu de legături de tip artificial, între o mulțime de unități abstrakte, care sunt neuronii rețelei. Din acest punct de vedere, structura și funcționalitatea unei rețele neuronale sunt relativ independente în raport cu structura problemei ce trebuie să fie rezolvată și cu algoritmul specific acestei rezolvări.

Dacă în cazul abordării tradiționale modelele diferă de la o categorie de probleme la alta, în cazul rețelelor neuronale putem vorbi de o anumită *universalitate* a acestora. Din acest punct de vedere, și păstrând rezervele necesare, putem spune că modelul neuronal are natura unui adevarat panaceu.

Spre deosebire de abordarea de tip algoritmic, în contextul căreia pot să apară o serie de dificultăți de natură numerică, rețeaua neuronală se caracterizează printr-o robustețe mai ridicată, sensibilitatea rezultatelor produse cu ajutorul unei rețele neuronale, în raport cu acuratețea informațiilor de intrare, fiind mult mai redusă.

Exceptând complexitatea rețelelor neuronale, complexitatea dată de numărul de neuroni, de numărul de straturi și de numărul de conexiuni, modelul neuronal are aproximativ aceeași structură generală, indiferent de natura problemei studiate. Cu alte cuvinte, modelul neuronal este relativ independent în raport cu natura fenomenelor pe care le vizează. Compatibilitatea dintre un model neuronal și realitate este realizată prin definirea unui număr mai mic sau mai mare de neuroni și de straturi neuronale, fără a fi necesar să se schimbe natura modelului în funcție de schimbarea realității modelate.

Având în vedere flexibilitatea abordării neuronale, se poate spune că aplicațiile posibile ale rețelelor neuronale sunt practic nelimitate, întinzându-se de la domeniul recunoașterii formelor și până la efectuarea de predicții pentru cele mai diferite fenomene și procese.

Construirea unei rețele neuronale presupune definirea unei *configurații de neuroni*, organizați sub forma unor straturi sau plane de neuroni, și a unor *moduri de conectare* a acestora, în funcție de structura intrărilor și ieșirilor, de natura și de complexitatea problemei care trebuie rezolvate. Ponderile sinaptice asociate conexiunilor rețelei neuronale sunt actualizate în mod iterativ, pe baza unei analize secvențiale a informațiilor reprezentate de formele care alcătuiesc setul de învățare. Regula de actualizare a ponderilor sinaptice sau regula de învățare, este dedusă din compararea valorilor efective de ieșire ale rețelei neuronale cu o serie de valori de prestabilitate, numite *valori de referință*. Deducerea regulii de învățare are loc în aşa manieră, încât valorile efective de ieșire ale rețelei neuronale, să conveargă către valorile de referință.

Regula de învățare se utilizează în mod secvențial, repetat, pe măsura analizării fiecărui exemplu disponibil, iar procesul de învățare se termină atunci când valorile efective de ieșire ale rețelei neuronale aproximează suficient de bine valorile de referință. După aceasta, rețeaua neuronală poate să fie utilizată pentru efectuarea de calcule, analize sau predicții specifice clasei de probleme pentru care ea a fost construită și instruită.

9.3 Avantaje ale modelării cu ajutorul rețelelor neuronale

În ultimele decenii modelarea bazată pe conceptul de rețea neuronală a cunoscut o foarte mare dezvoltare. Numeroase probleme din domeniul clasificării, controlului sau predicției au început să fie soluționate cu ajutorul tehnicilor bazate pe calculul neuronal.

Aplicațiile rețelelor neuronale sunt întâlnite în domenii de o maximă varietate, cum ar fi economia, finanțele, informatică, medicina, ingineria, biologia etc., iar rezultatele obținute în cazul abordărilor bazate pe calculul neuronal sunt cu adevărat spectaculoase.

Atât dezvoltarea în sine a tehnicilor de calcul neuronal, cât și extinderea utilizării lor în domenii din ce în ce mai variate și mai numeroase se datorează, în cea mai mare parte, simplității acestor instrumente și eficienței care se obține în urma utilizării lor.

Rețelele neuronale sunt instrumente puternice și eficiente de modelare, cu ajutorul cărora pot fi soluționate probleme de o mare complexitate, probleme care, de cele mai multe ori, nu pot fi rezolvate satisfăcător, facil și eficient, cu ajutorul altor metode și tehnici.

Cu toată puterea și eficiența pe care le au, rețelele neuronale se caracterizează printr-o extraordinară simplitate algoritmică și prin ușurința cu care pot fi folosite. Rețelele neuronale sunt, probabil, singurele instrumente de modelare care, în același context logic, pot fi configurate de la cea mai simplă și până la cea mai complexă formă, în funcție de natura și de complexitatea problemei abordate.

Capacitatea pe care o au rețelele neuronale de fi configurate în funcție de natura problemei abordate oferă posibilități de soluționare pentru cele mai variate tipuri de probleme și face din acestea instrumente de modelare cu caracter aproape universal.

Puterea pe care o au tehnicilor neuronale este și rezultatul faptului că, prin structura și funcționalitatea lor, rețelele neuronale pot fi echivalate cu *modele de tip neliniar*.

Ca o consecință a acestui fapt, în modelarea neuronală nu mai apare necesitatea formulării ipoteze simplificatoare și a acceptării unor aproximări mai mult sau mai puțin realiste, cum se întâmplă în cazul modelării de tip liniar.

În modelarea de tip neliniar din abordările tradiționale modelele sunt reprezentate prin intermediul unui mare număr de ecuații de tip algebric sau diferențial, a căror construcție presupune un laborios proces analiză și formalizare și care ridică numeroase dificultăți privind soluționarea analitică sau numerice.

Spre deosebire de modelarea neliniară tradițională, modelarea neuronală este caracterizată prin suplete și naturalețe, construirea modelului neuronal este foarte simplă și presupune un efort minimal, iar soluționarea problemelor pe această cale se reduce la efectuarea unor calcule numerice de tip elementar.

Datorită simplității rețelelor neuronale și ușurinței cu care acestea pot fi utilizate, dificultățile mari și inerente pe care le ridică problema complexității în cadrul modelării neliniare tradiționale sunt puternic estompată sau chiar eliminate în cazul modelării cu ajutorul rețelelor neuronale.

Un alt avantaj important care rezultă din simplitatea modelelor de tip neuronal constă în faptul că modelarea cu ajutorul rețelelor neuronale nu presupune un nivel foarte ridicat de pregătire științifică a utilizatorului în domeniul modelării matematice. Din acest punct de vedere, se poate spune că accesul la utilizarea tehnicilor neuronale este cu mult mai facil în comparație cu oricare alte metode și tehnici de modelare statistică-matematică.

Poate că cea mai importantă caracteristică a rețelelor neuronale, care, de asemenea, le deosebește în mod fundamental de alte instrumente de modelare, este aceea a adaptabilității dinamice, a capacitatei acestora de a învăța, de a se autoinstrui. Rețelele neuronale au capacitatea de a învăța din informația conținută în datele primare care definesc intrările acestor rețele.

Procesele de învățare din cadrul rețelelor neuronale sunt definite prin intermediul unor algoritmi specifici, numiți algoritmi de învățare, și care permit rețelelor neuronale să învețe să recunoască, din ce în ce mai bine, structurile informative conținute într-o formă invizibilă în datele de intrare.

9.4 Definirea și structura unei rețele neuronale generale

În general, rețeaua neuronală reprezintă un model cu ajutorul căruia poate fi rezolvată o anumită problemă particulară.

Definiție: *Rețeaua neuronală* reprezintă un ansamblu de unități elementare de procesare, numite *neuroni* sau *noduri*, a cărui capacitate de procesare este memorată la nivelul conexiunilor neuronale, sub forma unor ponderi sinaptice, și rezultă în urma unui proces iterativ de adaptare pe baza informațiilor dintr-un set de învățare.

O rețea neuronală este alcătuită din două elemente fundamentale: o mulțime de *unități funcționale*, numite *neuroni*; o mulțime de *legături* între neuroni, numite *conexiuni*. În cadrul paragrafelor următoare, vom defini și vom analiza fiecare dintre cele două elemente fundamentale ale unei rețele neuronale.

Cea mai simplă rețea neuronală este rețeaua care include: un singur neuron de intrare, un singur neuron de ieșire și o singură mărime de referință. Reprezentarea grafică a unei rețele neuronale de tip elementar este cea din figura următoare.

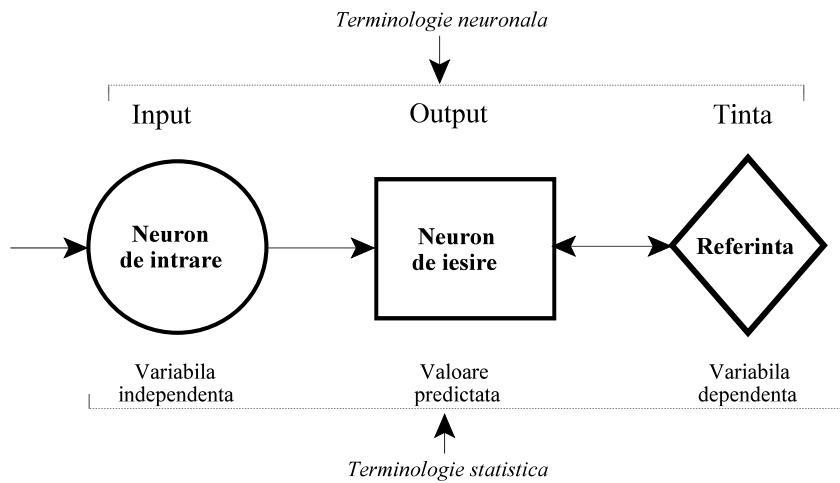


Figura 9.1: Structura unei rețele neuronale elementare

În termenii modelării econometrice, neuronul de intrare poate fi asociat cu o variabilă independentă, valoarea de referință poate fi asociată cu variabila dependentă, iar neuronul de ieșire poate fi asociat cu valoarea predictată.

9.4.1 Unitățile rețelei neuronale

Elementul fundamental al oricărei rețele neuronale, este *neuronul artificial*. Neuronii care intră în componența unei rețele neuronale, au funcții diferite, ei fiind specializați în efectuarea anumitor tipuri de activități. Din acest punct de vedere, o rețea neuronală conține trei tipuri fundamentale de neuroni:

- neuroni de *intrare*, care achiziționează valorile variabilelor de intrare sau valorile standardizate ale variabilelor de intrare; aceasta înseamnă că neuronii de intrare nu au o funcționalitate calculatorie propriu-zisă, ci doar un rol de interfațare, de receptare a valorilor din mediul extern; de regulă, numărul de neuroni de intrare ai unei rețele neuronale, este egal cu numărul de caracteristici ale formelor de intrare analizate; ansamblul neuronilor de intrare formează așa-numitul *stratul de intrare* sau *nivelul de intrare*;
- neuroni *intermediari* sau neuroni *ascunși*, sunt neuroni sunt dispuși între stratul de intrare și stratul de ieșire, care au o funcționalitate pur calculatorie; funcționalitatea neuronilor ascunși asigură caracterul de neliniaritate al unei rețele neuronale, ceea ce înseamnă că determină puterea modelatorie a unei rețele; neuroni ascunși pot fi dispuși pe straturi sau nivele, numite *straturi ascunse*; în funcție de complexitatea unei rețele neuronale, pot exista unul sau mai multe straturi ascunse;
- neuroni de *ieșire*, care calculează valorile predictate cu ajutorul rețelei neuronale și care compară aceste valori cu anumite valori țintă sau valori de referință; în funcție de rezultatul comparației, ponderile conexiunilor sunt sau nu actualizate.

Fiecare unitate elementară a unei rețele neuronale, respectiv fiecare neuron, are una sau mai multe, o stare internă și o ieșire. Funcționalitatea unui neuron constă în aceea că el produce o singură ieșire, reprezentată printr-o singură valoare numerică, în funcție de natura sau de starea respectivei unități, stare determinată pe baza informațiilor de intrare ale neuronului respectiv.

Comportamentul neuronilor unei rețele poate fi caracterizat cu ajutorul unui model de activare a output-ului, de forma generală:

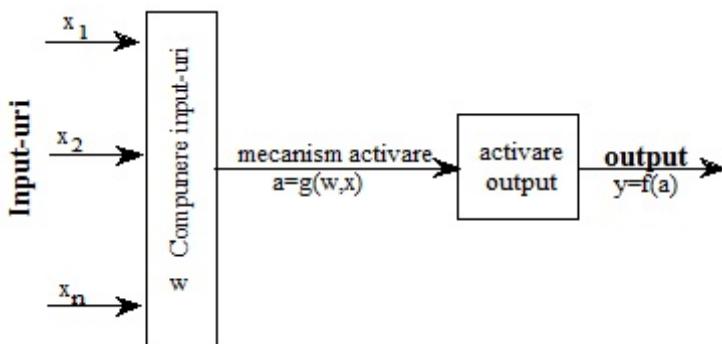


Figura 9.2: Structura și orientarea unei rețele neuronale generale

Funcția $g(\cdot)$ este numită *funcție de compunere* și definește modalitatea de sumare a input-urilor, sub forma unei mărimi a cărei valoare poate determina activarea output-ului, adică producerea de către neuron a unui output. În compunerea input-urilor intervin o serie de parametri, notați cu w_1, w_2, \dots, w_n , care se numesc *ponderi sinaptice*. Deoarece funcția de compunere a unui neuron are, în cele mai multe cazuri, o formă liniară, partea de activitate a unui neuron, legată de modalitatea de receptare a intrărilor, este cunoscută sub numele de componentă liniară.

Funcția $f(\cdot)$ este numită *funcție de activare* și definește modul de formare a output-ului neuronului, în funcție de nivelul obținut prin compunerea intrărilor.

9.4.2 Conexiunile rețelei neuronale

Circulația informației între neuronii unei rețele neuronale, are loc prin intermediul unor legături, cunoscute sub numele de *conexiuni*.

În general, conexiunile unei rețele neuronale sunt de tip *orientat* sau *direcțional*, ceea ce înseamnă că informația circulă de la un neuron la altul, numai într-un anumit sens, care indică direcția fluxului calculelor într-o rețea. În cadrul rețelelor de tip "înainte", conexiunile nu pot forma cicluri sau bucle.

Din punct de vedere topologic, sensul în care informația dintr-o rețea neuronală circulă și este prelucrată, este de tip "înainte" sau "de la stânga la dreapta". Aceste rețele se numesc rețele neuronale de tip "feedforward".

Conectarea elementelor unei rețele neuronale se face într-o manieră bine determinată, respectând anumite condiții, dintre care menționăm:

- neuronii de intrare sunt pot fi conectați cu neuronii ascunși sau cu neuronii de ieșire ai rețelei;
- neuronii ascunși pot fi conectați cu alți neuronii ascunși sau, direct, cu neuronii de ieșire;
- neuronii de ieșire nu pot fi conectați cu alți neuronii ai rețelei, cu excepția neuronilor din straturile ascunse anterioare;
- numărul de conexiuni dintr-o rețea este teoretic nelimitat, eventualele limitări ale numărului maxim de conexiuni fiind determinate doar de capacitatea de memorare și de viteza de prelucrare a sistemelor de calcul pe care rețelele neuronale sunt implementate.

În cazul unităților sau neuronilor de intrare, valoarea produsă este chiar valoarea achiziționată din mediul extern, fără nici o altă transformare, iar în cazul neuronilor ascunși și neuronilor de ieșire, valoarea produsă este rezultatul unui calcul, unei transformări a intrărilor acestora unități, în funcție de potențialul de activare al acestora. Valorile produse de către neuronii de ieșire se numesc *valori predictate*.

În cazul unităților de intrare, valoare de ieșire este reprezentată chiar de valoarea variabilei cu care unitatea este asociată, aceste unități având doar rolul de a *transfera* respectiva valoare către alte unități ale rețelei, unități care pot fi reprezentate de neuronii ascunși sau de neuronii de ieșire.

Valorile neuronilor de ieșire, adică valorile predictate, sunt comparate cu anumite valori de referință, comparații care sunt definite sub forma normelor. Valoarea diferenței totale dintre valorile predictate și valorile de referință pentru ieșiri, este descrisă prin intermediul unei funcții reale, cunoscută sub numele de *funcție eroare*, funcție care definește performanța unei rețele neuronale.

Cu cât valorile funcției eroare sunt mai mici, cu atât performanța rețelei neuronale este mai mare. Funcția eroare reprezintă criteriul în funcție de minimizarea căruia sunt dezvoltate activitățile de învățare sau de antrenare a rețelei neuronale, adică sunt actualizate ponderile sinaptice ale rețelei neuronale.

Conexiunile unei rețele neuronale sunt asociate cu anumite valori reale, cunoscute sub numele de *ponderi* sau *parametri*. Pentru ca rețeaua neuronală să poată fi utilizată pentru soluționarea unor probleme concrete, este necesar să se facă o estimare corespunzătoare a ponderilor sau parametrilor, astfel încât să se realizeze o minimizare a funcției eroare asociată cu ieșirile rețelei neuronale. Această estimare este cunoscută sub numele de *înstruire*, *învățare* sau *antrenare* a rețelei

neuronale.

Deși există mai multe modalități de învățare a rețelei neuronale, principial toate modelele au ca scop minimizarea a funcției eroare, printr-o modificare sau ajustare iterativă a ponderilor sinaptice.

În afara intrărilor pe care le primește de la alți neuroni, anumite unități ale rețelei sunt asociate și cu alte două intrări speciale, numite *deplasare* (prag, zgomot, bruijaj, perturbație) și *altitudine*, care sunt estimate tot prin intermediul tehniciilor de învățare, în mod similar cu estimarea ponderilor sinaptice. Din punct de vedere al funcțiilor analitice, deplasarea are o natură similară cu cea pe care o are termenul liber sau interceptul, în cazul modelelor econometrice.

Mecanismul de funcționare a neuronilor din straturile ascunse și din stratul de ieșire, sunt descrise prin intermediul a două funcții specifice, implicate în producerea ieșirilor acestor neuroni, cunoscute sub numele de *funcție de compunere*, respectiv *funcție de activare*.

9.4.3 Funcția de compunere

Prima funcție a unui neuron calculatoriu este funcția de agregare, numită și funcție de compunere sau funcție de combinare. Această funcție are ca scop combinarea tuturor valorilor de intrare ale unui neuron de acest tip într-o singură valoare și este definită sub forma următoare:

$$g : \mathbb{R}^m \rightarrow \mathbb{R}$$

Argumentele funcției de combinare sunt reprezentate de ieșirile altor neuroni, de ponderile sinaptice asociate cu aceste ieșiri, de deplasare (prag) și de altitudine.

Există două modalități de combinare a intrărilor unui neuron calculatoriu pentru a forma valoarea de activare a acestui neuron: *combinarea de tip liniar* și *combinarea de tip radial*.

Cele două modalități de combinare a input-urilor sunt definite prin intermediul a două clase de funcții de compunere: *funcții de compunere liniară* și *funcții de compunere radială*.

9.4.3.1 Funcții de compunere liniară

Funcțiile de compunere liniară calculează valoarea de activare sub forma unei combinații liniare de intrările neuronului (semnale de la alți neuroni, ponderi sinaptice etc.).

Dintre cele mai utilizate funcții de compunere în rețelele neuronale menționăm următoarele tipuri:

a. funcții *aditive*:

$$\sum_{i=1}^n x_i$$

b. funcții *liniare*:

$$b_j + \sum_{i=1}^n w_{ij} x_i$$

9.4.3.2 Funcții de compunere radială

Funcțiile de compunere radială au natura unor distanțe Euclidiene pătrate, calculate între vectorul ponderilor sinaptice și vectorul de valori de intrare în neuron, mărită prin pătratul deplasării sau pragului. În cadrul acestor funcții, deplasarea sau pragul acționează ca un factor de scalare sau de lățime inversă.

Dintre clasele de funcții cele mai frecvent utilizate ca funcții de compunere radială, menționăm:

a. funcții *X-radiale*:

$$f \cdot \log(a_j) - b_j^2 \sum_{i=1}^n (w_{ij} - x_i)^2$$

b. funcții *EH-radiale*:

$$- b_j^2 \sum_{i=1}^n (w_{ij} - x_i)^2$$

c. funcții *EV-radiale*:

$$f \cdot \log(b_j) - b_j^2 \sum_{i=1}^n (w_{ij} - x_i)^2$$

d. funcții *EW-radiale*:

$$f \cdot \log(a_j) - b_j^2 \sum_{i=1}^n (w_{ij} - x_i)^2$$

e. funcții *EQ-radiale*:

$$= b^2 \sum_{i=1}^n (w_{ij} - x_i)^2$$

Există de asemenea o funcție de combinare aditivă, care nu utilizează ponderi sau deplasări sau praguri.

9.4.4 Funcția de activare

Funcția de activare este o funcție definită pe mulțimea valorilor funcției de compunere și are ca scop transformarea valorii rezultate din compunere, adică a nivelului de activare, într-o valoare numită activare. După cum se poate înțelege, funcția de activare este definită pe o mulțime de scalari, care sunt nu depinde de ponderile sinaptice sau de alți parametri.

Funcția de activare descrie un mecanism de declanșare a neuronului, declanșare care are loc atunci când nivelul de activare depășește un anumit prag critic. Rezultatul acestei declanșări este reprezentat de producerea de către neuron a unui semnal de ieșire netrivial.

Există mai multe clase de funcții de activare:

a. Funcția identitate

Funcția identitate este o funcție liniară, care nu schimbă valoarea intrării și ale cărei valori sunt nemărginite atât inferior, cât și superior; forma generală a funcției identitate este:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = g$$

b. funcții sigmoide

Funcțiile "sigmoide" sunt funcții cu graficul de formă "S", ca de exemplu funcția logistică și funcția tangentă hiperbolică; funcțiile sigmoide au valorile în intervale de formă $[0, 1]$ sau de formă $\left[\frac{1}{1+e^{-g}}, \frac{1}{1+e^{-g}} \right]$; printre cele mai utilizate funcții de activare de tip sigmoid, se numără funcția logistică, a cărei formă este:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = \frac{1}{1 + e^{-g}}$$

c. funcții "softmax"

Funcțiile "softmax" sunt funcții logistice multidimensionale, adică generalizări ale funcției logistice, care afectează simultan mai multe unități sau neuroni, forțând ca suma valorilor de ieșire ale acestora să fie egală cu unitatea; forma generală a unei funcții "softmax" este:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = \frac{e^g}{\sum e^g}$$

d. funcții "valoare"

Funcțiile "valoare" sunt funcții care au graficul în formă de "clopot", cum ar fi, de exemplu, funcția lui Gauss; forma generală a funcției lui Gauss este următoarea:

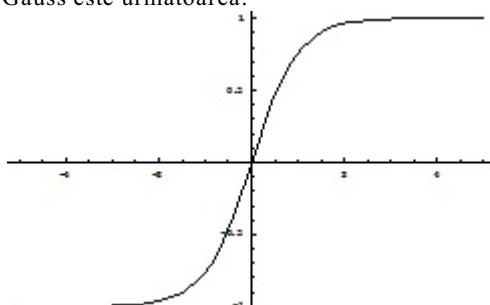


Figura 10.35: Graficul funcției "tanh"

$$f(g) = e^{-g^2}$$

e. funcții exponențiale:

Funcțiile exponențiale sunt funcții mărginite inferior de zero, dar nemărginite superior; funcțiile exponențiale au forma generală:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = e^g$$

f. funcții reciproce:

Funcțiile reciproce sunt funcții mărginite inferior de zero și nemărginite superior, de forma următoare:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = \frac{1}{g}$$

g. funcția *pătratică*:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = g^2$$

h. funcția *sinus*:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = \sin(g)$$

i. funcția *cosinus*:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = \cos(g)$$

j. funcția *Elliot*:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = \frac{1}{1 + |g|}$$

k. funcția *tangentă hiperbolică*:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = \tanh(g) = 1 - \frac{2}{1 + e^{2g}}$$

l. funcția *arctangentă*:

$$f : (-\infty, +\infty) \rightarrow \mathbb{R}; \quad f(g) = \frac{2}{\pi} \arctan(g)$$

9.4.5 Tipuri de straturi ale unei rețele neuronale

Pentru a simplifica procedurile de construcție a rețelilor neuronale și pentru a stăpâni mai ușor funcționalitatea acestora, neuronii unei rețele sunt priviți ca fiind organizați sub forma unor *straturi*, a căror natură este determinată de natura neuronilor pe care îi includ.

Din acest punct de vedere, există trei tipuri de straturi neuronale: straturi de intrare, straturi ascunse și straturi de ieșire. Din punct de vedere topologic, aceste straturi sunt dispuse de la stânga la dreapta, în ordinea: straturi de intrare, straturi ascunse, straturi de ieșire.

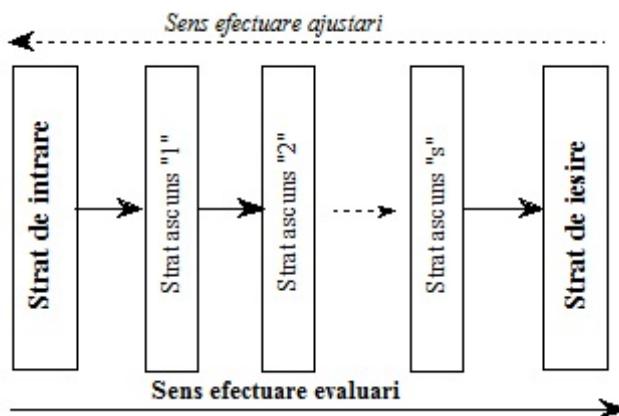


Figura 9.3: Structura și orientarea unei rețele neuronale generale

Din punct de vedere al conectării neuronilor, trebuie precizat că fiecare neuron dintr-un strat anterior este conectat cu fiecare neuron din stratul care îl succede.

Caracteristica fundamentală a straturilor unei rețele neuronale constă în faptul că toate unitățile sau neuronii unui anumit strat au aceleași caracteristici. În cazul stratului de intrare, aceasta înseamnă că toți neuronii acestui strat au același nivel de măsurare și aceeași metodă de standardizare. În cazul straturilor ascunse, comunălitatea caracteristicilor se referă la faptul că toți neuronii unui strat au aceeași funcție de combinare și aceeași funcție de activare. De asemenea, în cazul stratului de ieșire, toți neuronii care aparțin acestui strat au aceeași funcție de combinare, aceeași funcție de activare și aceeași funcție de eroare.

9.4.6 Funcția eroare

Evaluarea în cadrul unei rețele neuronale a gradului de depărtare dintre valorile predictate, pe de o parte, și valorile de referință, pe de altă parte, se face cu ajutorul unei funcții numite *funcție eroare*.

Funcția eroare totală definește o sumare peste toate formele din setul de învățare și peste toate variabilele de referință

ale unei funcții individuale eroare, care se aplică la fiecare valoare de referință și la output-ul corespunzător.

Funcția de penalizare este definită ca fiind produsul dintre constanta de atenuare (decădere, îmbătrânire) a ponderii și suma tuturor ponderilor rețelei, exceptând deplasarea.

Funcția de eroare individuală este evaluată pentru o singură valoare a variabilei de referință și poate avea două tipuri: verosimilitate și abatere.

Forme ale funcției eroare:

a. Normală:

$$E_L(y, \tilde{y}, \sigma) = \ln \sigma + \frac{1}{2} \left[\ln 2\pi + \left(\frac{\tilde{y} - y}{\sigma} \right)^2 \right]$$

$$E_D(y, \tilde{y}) = (\tilde{y} - y)^2$$

b. Gama:

$$E_L(y, \tilde{y}, \sigma) = \ln \Gamma(\sigma) + \frac{\tilde{y}}{y} - \sigma \ln \left(\frac{\tilde{y}}{y} \right)$$

$$E_D(y, \tilde{y}) = -2 \left[1 - \frac{\tilde{y}}{y} + \ln \left(\frac{\tilde{y}}{y} \right) \right]$$

c. Poisson:

$$E(y, \tilde{y}) = \ln \Gamma(\tilde{y} + 1) + y - \tilde{y} \ln y$$

$$E(y, \tilde{y}) = 2 \left[y - \tilde{y} + \ln \left(\frac{\tilde{y}}{y} \right) \right]$$

d. Bernoulli:

$$E(y, \tilde{y}) = \begin{cases} -\ln(1-y), & \text{dacă } \tilde{y}=0 \\ -\ln y, & \text{dacă } \tilde{y}=1 \end{cases}$$

$$E(y, \tilde{y}) = \begin{cases} -2\ln(1-y), & \text{dacă } \tilde{y}=0 \\ -2\ln y, & \text{dacă } \tilde{y}=1 \end{cases}$$

e. Bernoulli multiplă:

$$E(y, \tilde{y}) = \begin{cases} 0, & \text{dacă } \tilde{y}=0 \\ -\ln y, & \text{dacă } \tilde{y}=1 \end{cases}$$

$$E(y, \tilde{y}) = \begin{cases} 0, & \text{dacă } \tilde{y}=0 \\ -2\ln y, & \text{dacă } \tilde{y}=1 \end{cases}$$

f. binomială:

$$E(y, \tilde{y}, n) = -[\tilde{y} \ln y + (n-\tilde{y}) \ln(1-y)]$$

$$E(y, \tilde{y}, n) = 2 \left[\tilde{y} \ln \left(\frac{\tilde{y}}{ny} \right) + (n-\tilde{y}) \ln \left(\frac{n-\tilde{y}}{n(1-y)} \right) \right]$$

g. multinomială:

$$E(y, \tilde{y}) = \tilde{y} \ln y$$

$$E(y, \tilde{y}, n) = 2\tilde{y} \ln \left(\frac{\tilde{y}}{ny} \right)$$

h. entropie:

$$E(y, \tilde{y}) = \tilde{y} \ln y + (1-\tilde{y}) \ln(1-y)$$

$$E(y, \tilde{y}) = 2 \left[\tilde{y} \ln \left(\frac{\tilde{y}}{y} \right) + (1-\tilde{y}) \ln \left(\frac{1-\tilde{y}}{1-y} \right) \right]$$

i. entropie multiplă:

$$E(y, \tilde{y}) = -\tilde{y} \ln y$$

$$E(y, \tilde{y}) = 2\tilde{y} \ln \left(\frac{\tilde{y}}{y} \right)$$

j. Cauchy:

$$E(y, \tilde{y}, \sigma) = \ln \left[\pi \sigma + \pi \sigma \left(\frac{\tilde{y} - y}{\sigma} \right)^2 \right]$$

k. Logistică:

$$E(y, \tilde{y}, \sigma) = -\ln \left[\frac{1}{\sigma} \frac{e^{\frac{\tilde{y}-y}{\sigma}}}{\left[1 + e^{\frac{\tilde{y}-y}{\sigma}} \right]^2} \right]$$

1. M-estimator:

$$E(y, \tilde{y}, M_a, M_c, \sigma) = \sigma \left[\rho \left(\frac{\tilde{y}-y}{\sigma M_c} \right) M_c^2 + M_a \right]$$

unde funcția ρ poate să aibă una dintre următoarele forme:

- în cazul M-estimatorului *Hubert*:

$$\rho(z) = \begin{cases} \frac{1}{2} z^2, & \text{dacă } |z| < 1 \\ |z| - \frac{1}{2}, & \text{dacă } |z| \geq 1 \end{cases}$$

- în cazul M-estimatorului *dublu-ponderat*:

$$\rho(z) = \begin{cases} \frac{1}{6} [1 - (1 - z^2)^3], & \text{dacă } |z| < 1 \\ \frac{1}{6}, & \text{dacă } |z| \geq 1 \end{cases}$$

- în cazul M-estimatorului de tip “*undă*”:

$$\rho(z) = \begin{cases} \frac{1}{\pi^2} [1 - \cos(\pi z)], & \text{dacă } |z| < 1 \\ \frac{2}{\pi^2}, & \text{dacă } |z| \geq 1 \end{cases}$$

9.5 Modelul general al unei rețele neuronale

Rețeaua neuronală poate fi definită ca fiind un ansamblu de unități elementare de prelucrare, a căror funcționalitate este similară cu aceea a neuronilor care aparțin creierului uman, unități care se numesc și *neuroni artificiali*. Ea constituie o modalitate specifică de memorare a informației, modalitate bazată pe o memorare distribuită a informației, la nivel de conexiuni neuronale.

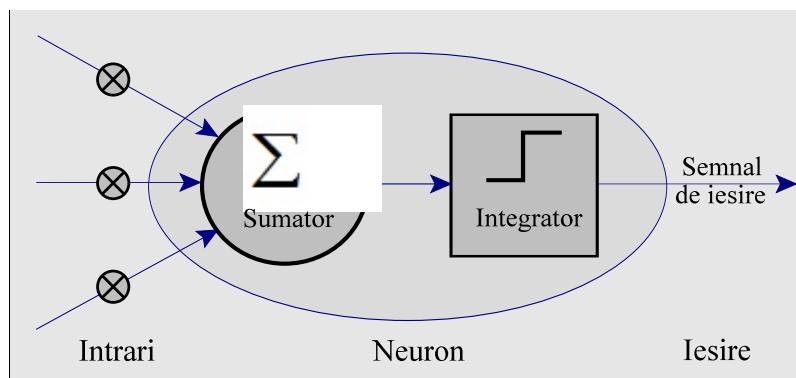


Figura 9.4: Structura funcțională a unui neuron

În cadrul figurii precedente, simbolul “ \otimes ” definește un operator de *multiplicare*, simbolul “ Σ ” definește un operator de *sumare*, iar simbolul “ \square ” definește un operator de tip *integrator* sau de tip *limitator hardware*.

Capacitatea de prelucrare a unei rețele neuronale este reprezentată de informația conținută la nivelul multiplelor conexiuni dintre unitățile care alcătuiesc respectiva rețea.

Caracteristica principală a unei conexiuni o reprezintă intensitatea acesteia și se exprimă prin intermediul unor mărimi numerice numite **ponderi**. Aceste ponderi sunt rezultatul unui proces de *adaptare*, de *instruire* sau de *învățare*, proces în care este utilizată informația dintr-un așa-numit *set de instruire* sau *set de învățare*, care este un eșantion de forme cu apartenență cunoscută..

În principiu, un neuron are un număr de n intrări, notate cu x_1, x_2, \dots, x_n , și o singură ieșire, notată cu . Intrările unu-

neuron sunt informații primite din mediul extern sau de la alți neuroni, numite și semnale de intrare, iar ieșirea unui neuron este o informație pe care o generează neuronul în urma prelucrării informațiilor de intrare, un semnal de ieșire.

Prin acumularea informațiilor de intrare, la nivelul unui neuron se formează un potențial numit *stare internă* a neuronului, *excitație* sau *activare totală* a neuronului.

Fiecare intrare a unui neuron corespunde unui anumit tip de informație și are o importanță mai mare sau mai mică în transmiterea de informație către neuron, în funcție de frecvența cu care informația de acest tip a contribuit anterior la excitarea neuronului, adică la formarea sării sale interne.

Intrările neuronului sunt de astăzi natură încât pot *amplifica* sau pot *atenua* informațiile pe care ele le transmit către neuron. Modelarea acestui proces de amplificare sau de atenuare a informațiilor care circulă de-a lungul sinapselor, se face prin intermediul definirii aşa-numitelelor *ponderi sinaptice*, care sunt mărimi numerice reale, fiecare intrare a neuronului fiind ponderată cu o astfel de mărime.

Starea internă a unui neuron este definită sub forma unei combinații liniare de intrările neuronului, coeficienții combinației liniare fiind chiar ponderile sinaptice. Aceasta înseamnă că starea internă a unui neuron este *o sumă ponderată* a informațiilor de intrare ale neuronului, adică o sumă ponderată a semnalelor de intrare.

Dacă x_1, x_2, \dots, x_n sunt cele n intrări ale unui neuron, iar $w_0, w_1, w_2, \dots, w_n$ sunt ponderile sinaptice asociate acestor intrări, atunci starea unui neuron poate fi definită prin intermediul unei mărimi scalare s , definită sub forma următoare:

$$s = w_0 + \sum_{i=1}^n w_i \cdot x_i$$

unde w_0 este un *prag de activare* sau un *prag de excitație* a neuronului.

Ieșirea unui neuron este o entitate informațională generată de neuron, în funcție de mărimea potențialului format la nivelul neuronului, pe baza informațiilor primite de către neuron prin intermediul conexiunilor sinaptice. În acest fel, ieșirea unui neuron poate fi descrisă cu ajutorul unei funcții speciale, a cărei valoare depinde de starea internă a neuronului și care se numește *funcție de transfer*, *funcție de activare* sau *funcție de răspuns*.

Funcția de activare, este definită pe mulțimea intrărilor posibile ale unui neuron și are valori reale, adică este definită sub forma:

$$f : X \rightarrow \mathbb{R}$$

unde X definește mulțimea intrărilor neuronale și este o submulțime a spațiului real n -dimensional, respectiv $X \subseteq \mathbb{R}^n$.

Dacă $f(\cdot)$ este funcția de activare a neuronului, atunci ieșirea unui neuron poate fi definită sub forma următoare:

$$y = f(s) = f\left(w_0 + \sum_{i=1}^n w_i \cdot x_i\right)$$

Având în vedere cele menționate anterior, funcționalitatea unui neuron poate fi ilustrată din punct de vedere sinoptic sub forma următoare:

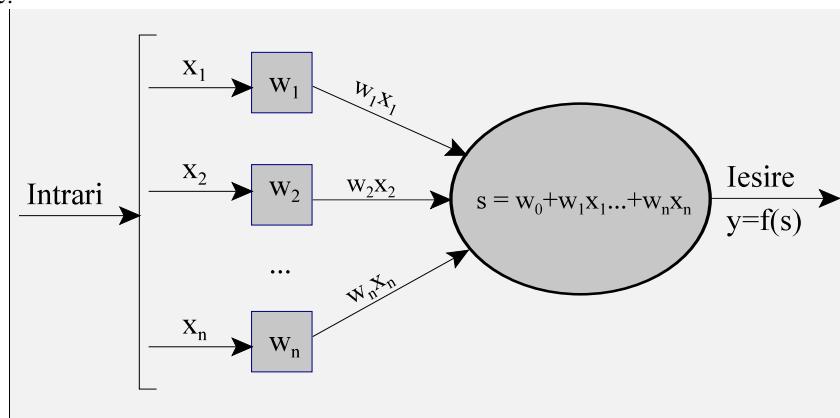


Figura 9.5: Arhitectura rețelei de tip perceptron, cu un singur strat

Pentru a asigura o scriere mai condensată a relațiilor care definesc funcționalitatea unui neuron, vom asimila pragul de excitație w_0 unei intrări virtuale, ale cărei valori sunt egale totdeauna cu 1 și care are ca pondere chiar valoarea w_0 a pragului. Aceasta înseamnă că starea internă a neuronului poate fi scrisă sub forma:

$$s = \sum_{i=0}^n w_i \cdot x_i,$$

iar output-ul neuronului devine:

$$y = f(s) = f\left(\sum_{i=0}^n w_i \cdot x_i\right).$$

Considerând pragul de excitație ca fiind încorporat în mulțimea intrărilor, schema funcționalității neuronului capătă următoarea formă:

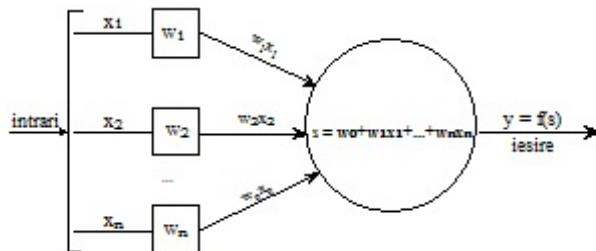


Figura 9.6: Schema funcțională a unui neuron cu prag de excitație încorporat în intrări

Având în vedere natura ieșirii unui neuron, funcția de răspuns, adică funcția care definește output-ul neuronului poate avea mai multe forme. De regulă, funcția de răspuns este *nelinieră* în raport cu variabila care descrie starea internă a neuronului.

9.6 Modelul general al unei rețele neuronale cu n neuroni

Vom considera o rețea neuronală alcătuită din n neuroni. Funcționalitatea unei astfel de rețele poate fi reprezentată ca în figura următoare.

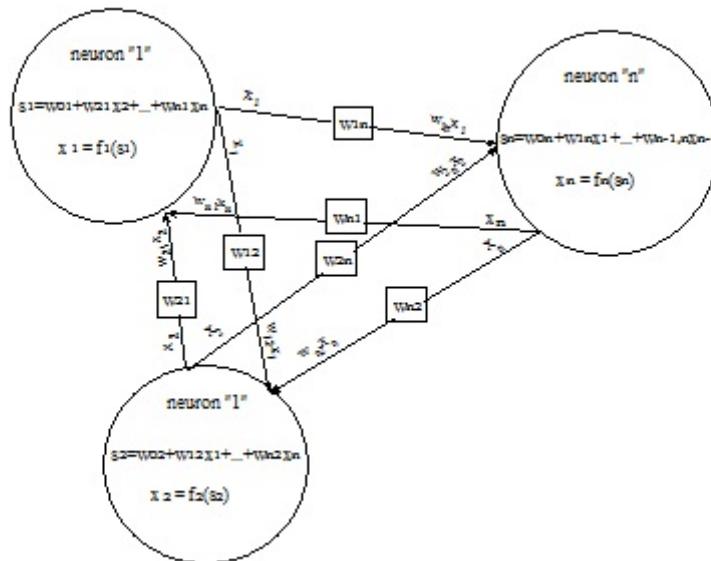


Figura 9.7: Schema funcțională a unei rețele cu n neuroni

Starea internă a unui neuron notat cu j va fi dată de sumarea intrărilor ponderate ale acestui neuron, respectiv:

$$s_j = w_{0j} + \sum_{i=1}^n w_{ij} \cdot x_i$$

În cazul în care pragul w_{0j} este încorporat în mărimile de intrare, starea internă sau potențialul unui neuron este:

$$s_j = \sum_{i=0}^n w_{ij} \cdot x_i$$

unde $x_0 = 1$ este o intrare fictivă.

Dacă vom nota cu x vectorul ieșirilor posibile ale celor n neuroni, inclusiv ieșirea fictivă x_0 , și cu $w^{(j)}$ vectorul extins

al celor $n+1$ ponderi ale intrărilor neuronului j , adică:

$$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}; \quad \mathbf{w}^{(j)} = \begin{pmatrix} w_{0j} \\ w_{1j} \\ \vdots \\ w_{nj} \end{pmatrix}, \quad j=1,2,\dots,n$$

atunci starea internă a neuronului j va putea fi scrisă sub formă produsului scalar:

$$s_j = (\mathbf{w}^{(j)})^t \cdot \mathbf{x}$$

Starea internă a tuturor celor n neuroni ai rețelei poate fi scrisă sub următoarea formă matricială:

$$\begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_j \\ \vdots \\ s_n \end{pmatrix} = \begin{pmatrix} w_{01} & w_{11} & \dots & w_{i1} & \dots & w_{n1} \\ w_{02} & w_{12} & \dots & w_{i2} & \dots & w_{n2} \\ \vdots & & & & & \\ w_{0j} & w_{1j} & \dots & w_{ij} & \dots & w_{nj} \\ \vdots & & & & & \\ w_{0n} & w_{1n} & \dots & w_{in} & \dots & w_{nn} \end{pmatrix} \cdot \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix}$$

Dacă vom considera că cei n vectori de ponderi, $\mathbf{w}^{(j)}$, $j=1,2,\dots,n$, sunt coloanele unei matrice \mathbf{W} , de forma:

$$\mathbf{W}_{(n+1) \times n} = \begin{pmatrix} w_{01} & w_{02} & \dots & w_{0j} & \dots & w_{0n} \\ w_{11} & w_{12} & \dots & w_{1j} & \dots & w_{1n} \\ \vdots & & & & & \\ w_{i1} & w_{i2} & \dots & w_{ij} & \dots & w_{in} \\ \vdots & & & & & \\ w_{n1} & w_{n2} & \dots & w_{nj} & \dots & w_{nn} \end{pmatrix}$$

atunci starea internă a tuturor celor n neuroni ai rețelei poate fi scrisă sub formă condensată:

$$\boxed{s = \mathbf{W}^t \cdot \mathbf{x}},$$

unde s este vectorul stăriilor interne ale celor n neuroni ai rețelei.

În continuare, vom nota cu $f_j(\cdot)$ funcția de transfer a celui de-al j -lea neuron, astfel încât ieșirea acestuia va fi:

$$\boxed{x_j = f_j(s_j)}.$$

Presupunând că procesul de prelucrare a intrărilor unui neuron și de generare a ieșirii acestui neuron implică trecerea unei anumite perioade de timp, vom obține o decalare în timp, o defazare a ieșirii unui neuron în raport cu intrările acestuia. Luând în considerare această decalare, output-ul celui de-al j -lea neuron poate fi scris sub formă:

$$\boxed{x_j(t+1) = f_j(s_j(t))}.$$

Această relație este cunoscută sub numele de *ecuație de evoluție* a celui de-al j -lea neuron al rețelei.

Pentru a deduce ecuația de evoluție a întregii rețele neuronale, vom nota cu $\mathbf{f}(s)$ vectorul de funcții de transfer definit astfel:

$$\mathbf{f}(s) = \begin{pmatrix} f_1(s_1(t)) \\ f_2(s_2(t)) \\ \vdots \\ f_n(s_n(t)) \end{pmatrix}$$

În aceste condiții, *ecuația de evoluție* a rețelei neuronale, adică ecuația care descrie modificarea stăriilor unei rețele neuronale poate fi scrisă sub formă:

$$\boxed{\mathbf{x}(t+1) = \mathbf{f}(\mathbf{W}^t \mathbf{x}(t))}.$$

În cadrul cursului urmator, vom analiza procesele specifice de învățare, de antrenare sau de instruire ale unei rețele neuronale.

9.7 Instruirea sau antrenarea unei rețele neuronale

Vom presupune că eșantionul care reprezintă setul de învățare sau de instruire a unei rețele neuronale de tip *perceptron simplu*, include un număr de T forme, fiind reprezentat de mulțimea de vectori n -dimensionali:

$$X = \{x^{(1)}, x^{(2)}, \dots, x^{(T)}\}.$$

Instruirea unei rețele neuronale presupune efectuarea unor succesiuni de operații, succesiune cunoscută sub numele de *algoritm de învățare*, care are ca scop configurarea rețelei în vederea efectuării unor anumite operații specifice.

În mod concret, învățarea unei rețele neuronale este echivalentă cu un proces iterativ de *ajustare a ponderilor sinaptice* ale acesteia, astfel încât rețeaua neuronală să poată realiza o anumită funcție, pornind de la o serie de valori cunoscute aprioric, numite exemple, valori care sunt cunoscute sub numele de *set de formare*, *set de instruire* sau *set de învățare*.

Un element al setului de instruire, notat în mod generic cu x , se numește *formă* sau *configurație* și se reprezintă prin intermediul unui vector n -dimensional, ale cărui elemente sunt *caracteristicile* sau *atributele* formei respective. În acest fel, orice formă x apare ca fiind un punct dintr-un spațiu vectorial real n -dimensional \mathbb{R}^n și numit *spațiul formelor*.

Din punct de vedere strict formal, o rețea neuronală poate fi privită ca fiind o transformare a unui spațiu de intrare \mathbb{R}^n , într-un spațiu de ieșire \mathbb{R}^m .

9.8 Instruirea rețelei neuronale de tip perceptron

Perceptronul este o rețea neuronală formată dintr-un singur strat și care are ca intrări variabile ale caror valori sunt de tip binar sau numere reale. Dacă vom considera că intrările perceptronului sunt elemente ale vectorului n -dimensional x , adică:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in \mathbb{R}^n,$$

iar cele n ponderi sinaptice sunt elemente ale vectorului n -dimensional w , respectiv:

$$w = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_n \end{pmatrix} \in \mathbb{R}^n,$$

atunci activarea neuronului poate să fie scrisă sub forma:

$$s = w_0 + \sum_{i=1}^n w_i x_i = w_0 + w^t x,$$

unde w_0 este mărimea scalară care reprezintă pragul de activare.

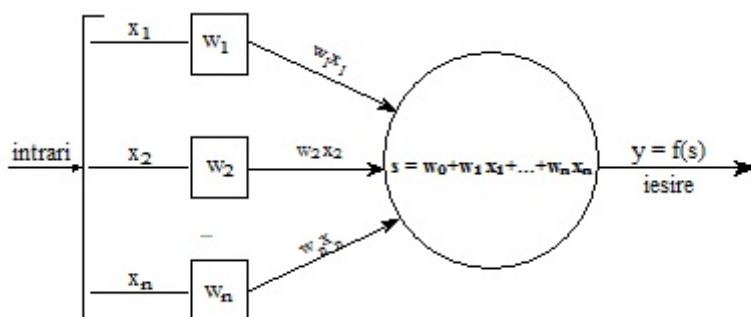


Figura 9.8: Schema perceptronului cu un singur strat

Utilizarea rețelei neuronale de tip *perceptron* pentru rezolvarea problemei privind predicția apartenenței unor forme la anumite clase, este echivalentă cu determinarea în spațiul formelor \mathbb{R}^n , a unei suprafețe liniare, definită de mulțimea:

$$H_w = \{z \in \mathbb{R}^n \mid w_0 + w^t z = 0\},$$

unde w este vectorul de ponderi, iar w_0 reprezintă pragul de activare a neuronului, suprafață care să permită o separare cât mai bună a formelor din spațiul formelor \mathbb{R}^n , sub forma a două regiuni distincte. Suprafața definită de ecuația:

$$w_0 + \mathbf{w}^t \mathbf{x} = 0,$$

se numește **hiperplan** de dimensiune $n-1$.

Din punct de vedere geometric, hiperplanul H_w reprezintă o *suprafață liniară* în spațiul n -dimensional al formelor, suprafață care, în general, nu trece însă prin originea acestui spațiu, deoarece vectorul nul $\mathbf{0}$ nu verifică ecuația hiperplanului.

Spre deosebire de un hiperplan dintr-un spațiu vectorial n -dimensional, un *subspațiu* de dimensiune $n-1$ al respectivului spațiu vectorial, este reprezentat de mulțimea:

$$L_w = \{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{w}^t \mathbf{x} = 0 \}.$$

Din punct de vedere geometric, subspațiul L_w reprezintă tot o *suprafață liniară* în spațiul n -dimensional al formelor. Deoarece vectorul nul $\mathbf{0}$, verifică ecuația care definește subspațiul L_w , această suprafață trece însă prin originea spațiului.

Având în vedere definirea hiperplanului, putem spune că subspațiul este un hiperplan care trece prin originea spațiului vectorial.

Mai mult decât atât, orice hiperplan poate fi privit ca reprezentând o translație a unui subspațiu, translație efectuată după un anumit vector. În acest fel, hiperplanul H_w apare ca fiind definit de o mulțime de forma:

$$H_w = \{ \mathbf{z} = \mathbf{x} + \mathbf{z}^{(0)} \mid \mathbf{x} \in L_w, \mathbf{z}^{(0)} \notin L_w \},$$

unde vectorul $\mathbf{z}^{(0)} \in \mathbb{R}^n$, diferit de vectorul nul $\mathbf{0}$, se numește *vector de translație* a subspațiului L_w .

Se poate arăta că, fiind dat un hiperplan H_w , există un unic subspațiu L_w , prin căruia translație se obține hiperplanul H_w . Cu toate acestea, pentru orice hiperplan H_w , care este o translație a unui subspațiu unic L_w , vectorul de translație $\mathbf{z}^{(0)} \in \mathbb{R}^n$ nu este unic, existând o infinitate de vectori de translație.

Așa cum o să arătăm ulterior, unul dintre vectorii de translație a subspațiului L_w , definit de ecuația:

$$\mathbf{w}^t \mathbf{x} = 0,$$

în vederea obținerii hiperplanului H_w , definit de ecuația:

$$w_0 + \mathbf{w}^t \mathbf{x} = 0,$$

este de forma:

$$\mathbf{z}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -\frac{w_0}{w_n} \end{pmatrix}.$$

Utilizarea unui vector de această formă pentru a exprima legătura dintre subspațiul L_w și hiperplanul H_w , este mult mai convenabilă pentru descrierea rețelelor neuronale, deoarece acest vector depinde doar de un element al vectorului de ponderi \mathbf{w} și de pragul de excitare w_0 .

Din modul în care este definit hiperplanul H_w , este evident că extremitatea vectorului de translație $\mathbf{z}^{(0)}$, aparține hiperplanului H_w . Într-adevăr, deoarece vectorul nul aparține subspațiului L_w , adică $\mathbf{0} \in L_w$, vom avea:

$$\mathbf{0} + \mathbf{z}^{(0)} \in H_w \Rightarrow \mathbf{z}^{(0)} \in H_w.$$

În mod similar, dacă H_w este un hiperplan care reprezintă o translație a subspațiului L_w , cu vectorul $\mathbf{z}^{(0)}$, atunci subspațiul L_w poate fi definit sub forma:

$$L_w = \{ \mathbf{x} = \mathbf{z} - \mathbf{z}^{(0)} \mid \mathbf{z} \in H_w, \mathbf{z}^{(0)} \notin L_w \}.$$

În cadrul figurii următoare, este reprezentat grafic subspațiul L_w , asociat cu vectorul \mathbf{w} , precum și hiperplanul H_w , rezultat din translația subspațiului L_w cu vectorul $\mathbf{z}^{(0)}$, pentru cazul particular bidimensional.

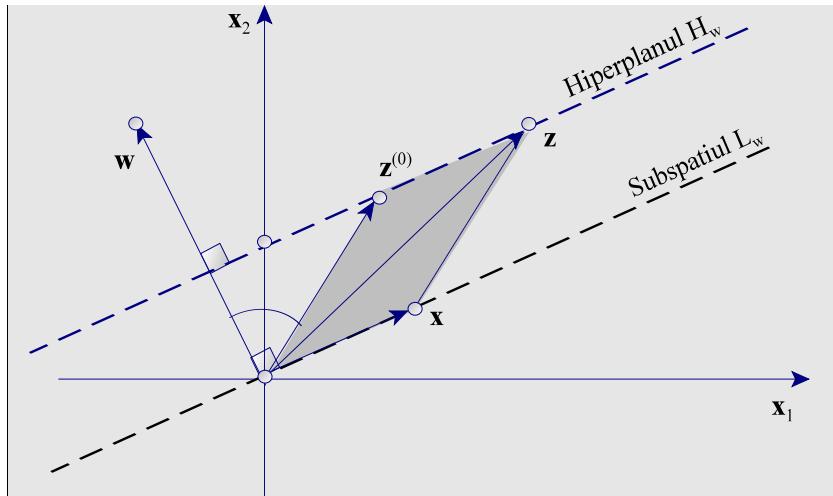


Figura 9.9: Obținerea unui hiperplan printr-o translație a unui subspațiu

Din modul în care este definit subspațiuul L_w , rezultă că vectorul w este ortogonal cu orice vector x din acest subspațiu, situație în care se spune că vectorul w este ortogonal cu întregul subspațiu L_w și se scrie:

$$w \perp L_w.$$

Având în vedere faptul că doi vectori sunt ortogonali, dacă produsul lor scalar este egal cu zero, condiția de ortogonalitate a vectorului w cu subspațiuul L_w , poate fi scrisă sub formă:

$$w \perp L_w \Leftrightarrow w^t x = 0, \quad \forall x \in L_w.$$

Din punct de vedere geometric, aceasta înseamnă că vectorul w este *perpendicular* pe suprafața liniară care trece prin extremitățile vectorilor care aparțin subspațiului L_w .

În mod similar, ortogonalitatea în raport cu un vector w , poate fi definită și în cazul hiperplanului H_w . În acest context însă, spunem că vectorul w este ortogonal cu hiperplanul H_w și scriem:

$$w \perp H_w,$$

dacă și numai dacă vectorul w este ortogonal cu subspațiuul L_w , prin căruia translație este obținut hiperplanul H_w .

Din punct de vedere geometric, ortogonalitatea dintre vectorul w și hiperplanul H_w , este reprezentată de situația în care vectorul w este *perpendicular* pe suprafața care trece prin extremitățile vectorilor care aparțin hiperplanului H_w , fără însă ca aceasta să însemne că vectorul w este ortogonal cu oricare dintre vectorii care aparțin hiperplanului H_w .

În această situație, condiția de ortogonalitate dintre vectorul w și hiperplanul H_w , poate fi scrisă sub formă:

$$w \perp H_w \Leftrightarrow w^t (z - z^{(0)}) = 0, \quad \forall z \in H_w,$$

unde $z^{(0)} \neq 0$ este vectorul de translație a subspațiului L_w , translație prin care se obține hiperplanul H_w .

După cum se poate observa, din condiția de ortogonalitate dintre vectorul w și hiperplanul H_w , rezultă că:

$$w^t z = w^t z^{(0)} \neq 0, \quad \forall z \in H_w,$$

ceea ce înseamnă că vectorul w nu este ortogonal cu nici un vector care aparține hiperplanului H_w .

În contextul prezentat anterior, ecuația:

$$w^t (z - z^{(0)}) = 0,$$

poate fi considerată ca fiind ecuația unui hiperplan H_w , care este ortogonal cu vectorul w .

În general, pentru un spațiu n -dimensional al formelor \mathbb{R}^n , subspațiuul L_w , care este ortogonal cu vectorul n -dimensional de ponderi w , are dimensiunea $n-1$, și este definit de mulțimea tuturor vectorilor $x = (x_1, x_2, \dots, x_n)^t$, care verifică ecuația:

$$w^t x = 0.$$

Aceasta înseamnă că subspațiuul L_w este format din toți vectorii $x \in \mathbb{R}^n$, ale căror coordonate verifică ecuația:

$$w_1x_1 + w_2x_2 + \dots + w_nx_n = 0,$$

respectiv relația:

$$x_n = -\frac{w_1}{w_n}x_1 - \frac{w_2}{w_n}x_2 - \dots - \frac{w_{n-1}}{w_n}x_{n-1}.$$

Această relație definește în spațiul \mathbb{R}^n o *suprafață liniară (n-1)-dimensională*, suprafață care trece prin originea spațiului \mathbb{R}^n .

O problemă foarte importantă, care apare frecvent în contextul instruirii rețelelor neuronale, este aceea a exprimării *distanței* dintre o formă și hiperplanul care separă clasele de forme, respectiv a distanței dintre un punct x din spațiul formelor și un hiperplan H_w . O astfel de distanță, este utilizată în construire criteriilor de performanță utilizate în procesul de instruire a rețelelor neuronale. De exemplu, unul dintre aceste criterii este cel bazat pe minimizarea sumei tuturor distanțelor de la formele clasificate incorrect, la hiperplanul care asigură separarea claselor.

În condițiile în care o astfel de distanță poate fi exprimată în funcție de ponderile sinaptice ale rețelei neuronale, se poate construi un algoritm de ajustare a acestor ponderi, în aşa fel încât suma distanțelor de la formele clasificate incorrect, la hiperplanul asociat cu vectorul ponderilor, să fie minimă.

Pentru a ilustra modul în care poate fi construită o astfel de distanță, vom considera situația ilustrată în cadrul figurii următoare.

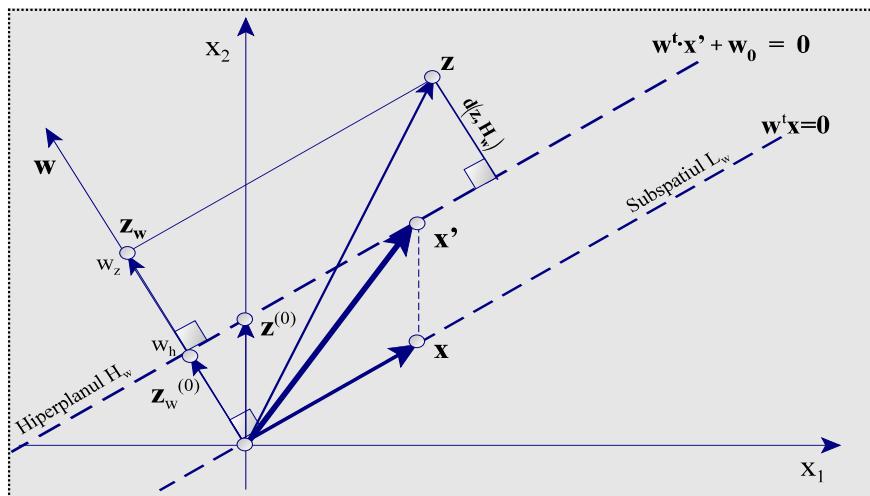


Figura 9.10: Ortogonalitatea dintre vectorul w și hiperplanul H

După cum se poate observa din această figură, distanța de la punctul z la hiperplanul H_w , poate fi exprimată ca diferență între coordonatele w_z și w_h , care reprezentă punctele ce definesc proiecțiile ortogonale ale vectorilor z și $z^{(0)}$, pe vectorul w , respectiv:

$$d(z, H_w) = |w_z - w_h| = \|z_w - z_w^{(0)}\|,$$

unde z_w și $z_w^{(0)}$ reprezintă vectorii de proiecție ai vectorilor z și $z^{(0)}$, pe vectorul w . Aceasta înseamnă că distanța de la punctul z la hiperplanul H_w , este egală cu lungimea vectorului care are originea în extremitatea vectorului $z_w^{(0)}$ și extremitatea în punctul reprezentat de vectorul z_w . În acest fel, deoarece vectorii z și $z^{(0)}$ sunt coliniari, distanța precedentă poate fi scrisă și sub forma următoare:

$$d(z, H_w) = \|\|z_w\| - \|z_w^{(0)}\|\|.$$

În orice problemă de clasificare, este recomandat să se utilizeze suprafețe liniare de tip hiperplan, și nu suprafețe liniare de tip subspațiu, deoarece hiperplanele pot conduce la o separare mai bună a claselor. În cadrul figurii următoare, este ilustrată grafic abilitatea pe care o poate avea utilizarea într-o problemă de clasificare a unei suprafețe de tip hiperplan, în comparație cu o suprafață de tip subspațiu.

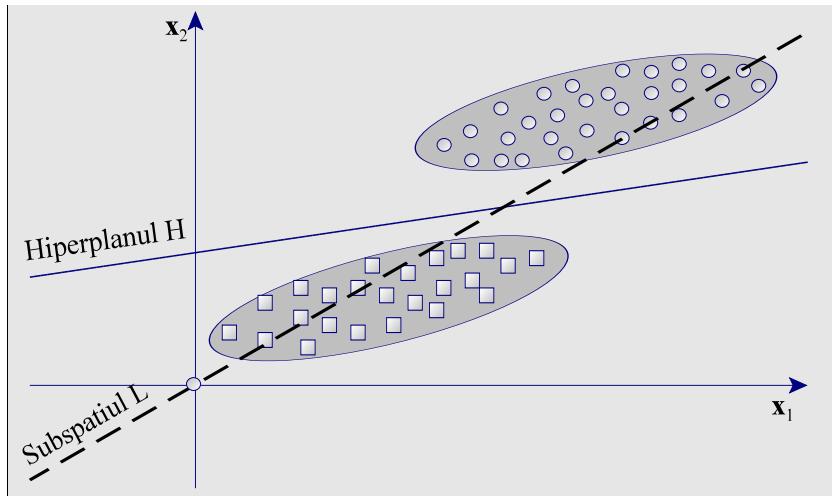


Figura 9.11: Separarea cu ajutorul unui subspațiu L și a unui hiperplan H

Natura suprafeței de separare utilizată în procesul de clasificare cu ajutorul rețelelor neuronale, este determinată în mod direct de maniera în care este definit pragul de activare al neuronilor. Astfel, considerarea unui prag de separare egal cu zero, conduce la o suprafață de separare de tip subspațiu, în timp ce considerarea unui prag de separare diferit de zero, conduce la obținerea unei suprafețe de separare de tip hiperplan.

Pentru fiecare vector de ponderi $\mathbf{w}^{(k)} = (w_1^{(k)}, w_2^{(k)})^t$, există un subspațiu unic $L_{\mathbf{w}^{(k)}}$, în raport cu care vectorul de ponderi \mathbf{w} este ortogonal. De asemenea, pentru fiecare vector de ponderi \mathbf{w} și prag de excitație w_0 , hiperplanul asociat $H_{\mathbf{w}}$ împarte întregul spațiu al formelor în două regiuni disjuncte, notează în cadrul figurii cu R_+ , respectiv cu R_- . O a treia regiune a spațiului formelor, este reprezentată de mulțimea formelor care aparțin hiperplanului $H_{\mathbf{w}}$. În condițiile în care ecuația hiperplanului $H_{\mathbf{w}}$ este de forma:

$$\mathbf{w}^t \mathbf{x} + w_0 = 0,$$

formele care aparțin regiunilor menționate verifică următoarele condiții:

a. $\mathbf{w}^t \mathbf{x} + w_0 > 0, \quad \forall \mathbf{x} \in R_+;$

b. $\mathbf{w}^t \mathbf{x} + w_0 = 0, \quad \forall \mathbf{x} \in H_{\mathbf{w}};$

c. $\mathbf{w}^t \mathbf{x} + w_0 < 0, \quad \forall \mathbf{x} \in R_-.$

Cu alte cuvinte, dacă vectorul de ponderi \mathbf{w} a fost determinat astfel încât acestuia îi corespunde acel hiperplan $H_{\mathbf{w}}$, care asigură cea mai bună separare a formelor din setul de învățare, atunci *regula* privind predicția apartenenței unei noi forme \mathbf{x} , la una dintre clasele ω_+ și ω_- , poate fi formulată sub forma următoare:

- dacă $\mathbf{w}^t \mathbf{x} + w_0 > 0$, atunci $\mathbf{x} \in \omega_+$;
- dacă $\mathbf{w}^t \mathbf{x} + w_0 < 0$, atunci $\mathbf{x} \in \omega_-$.

Dacă setul de învățare include forme ce aparțin în mod exclusiv claselor ω_+ și ω_- , iar vectorul de ponderi $\tilde{\mathbf{w}}$ și pragul de excitație \tilde{w}_0 , sunt astfel încât acestora le corespunde un hiperplan $H_{\tilde{\mathbf{w}}}$, care asigură separarea perfectă a celor două clase, atunci clasele ω_+ și ω_- se numesc clase *liniar separabile*.

Definiție: Două clase de forme ω_+ și ω_- , se numesc *liniar separabile*, dacă există un vector de ponderi $\tilde{\mathbf{w}}$ și un prag de excitație \tilde{w}_0 , astfel încât:

- a. $\tilde{\mathbf{w}}^t \mathbf{x}^+ + \tilde{w}_0 > 0, \quad \forall \mathbf{x}^+ \in \omega_+;$
- b. $\tilde{\mathbf{w}}^t \mathbf{x}^- + \tilde{w}_0 < 0, \quad \forall \mathbf{x}^- \in \omega_-.$

În condițiile în care clasele ω_+ și ω_- sunt liniar separabile, vectorul de ponderi $\tilde{\mathbf{w}}$ și pragul de activare \tilde{w}_0 determină hiperplan unic $H_{\tilde{\mathbf{w}}}$, a cărui ecuație este:

$$\tilde{\mathbf{w}}^t \mathbf{x} + \tilde{w}_0 = 0, \quad \mathbf{x} \in \mathbb{R}^n,$$

se numește **hiperplan de separare** a claselor ω_+ și ω_- , iar funcția $d : \mathbb{R}^n \rightarrow \mathbb{R}$, de forma:

$$d(\mathbf{x}) = \tilde{w}_0 + \tilde{\mathbf{w}}^t \mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^n,$$

se numește **funcție discriminant** a claselor ω_+ și ω_- .

Pe baza funcției discriminant $d(\cdot)$, regula de stabilire a apartenenței unei noi forme $\mathbf{x} \in \mathbb{R}^n$, la una dintre cele două clase ω_+ și ω_- , poate fi formulată astfel:

- | |
|--|
| <ul style="list-style-type: none"> • dacă $d(\mathbf{x}) = \tilde{w}_0 + \tilde{\mathbf{w}}^t \mathbf{x} < 0$, atunci $\mathbf{x} \in \omega_-$; • dacă $d(\mathbf{x}) = \tilde{w}_0 + \tilde{\mathbf{w}}^t \mathbf{x} > 0$, atunci $\mathbf{x} \in \omega_+$. |
|--|

În cazul în care noua formă $\mathbf{x} \in \mathbb{R}^n$, este astfel încât $\tilde{\mathbf{w}}^t \mathbf{x} + \tilde{w}_0 = 0$, adică această formă aparține zonei reprezentate de hiperplanul $H_{\tilde{\mathbf{w}}}$, numită și **zonă critică**, atunci apartenența formei nu poate fi stabilită. În această situație, se poate considera că forma respectivă poate să aparțină oricare din cele două clase.

9.9 Utilizarea perceptronului pentru clasificarea formelor

Cea mai importantă problemă a utilizării unei rețele neuronale de tip perceptron, pentru efectuarea de predicții privind apartenența unor noi forme la una dintre clasele ω_+ și ω_- , constă în determinarea vectorului de ponderi $\tilde{\mathbf{w}}$ și a pragului de excitație \tilde{w}_0 , astfel încât rețeaua neuronală să recunoască într-o măsură cât mai mare, apartenența formelor din setul de învățare. Altfel spus, determinarea vectorului de ponderi $\tilde{\mathbf{w}}$ și a pragului de excitație \tilde{w}_0 , trebuie să se facă în aşa fel încât hiperplanul $H_{\tilde{\mathbf{w}}}$, asociat acestora, să asigure o separare cât mai bună a formelor din setul de învățare.

Vom presupune în continuare că cele două clase ω_+ și ω_- sunt liniar separabile, iar $\tilde{\mathbf{w}}$ și \tilde{w}_0 reprezintă vectorul de ponderi optimale, respectiv pragul de excitație optimal, care asigură o separabilitate perfectă a formelor din setul de învățare, pe cele două clase.

În aceste condiții, pentru orice alt vector de ponderi $\mathbf{w} \neq \tilde{\mathbf{w}}$ și orice alt prag de excitație $w_0 \neq \tilde{w}_0$, alese în mod arbitrar, este evident că regula de clasificare menționată anterior va conduce la situația că unele dintre formele din setul de învățare vor fi clasificate în mod corect, în timp ce celelalte forme vor fi clasificate în mod incorrect.

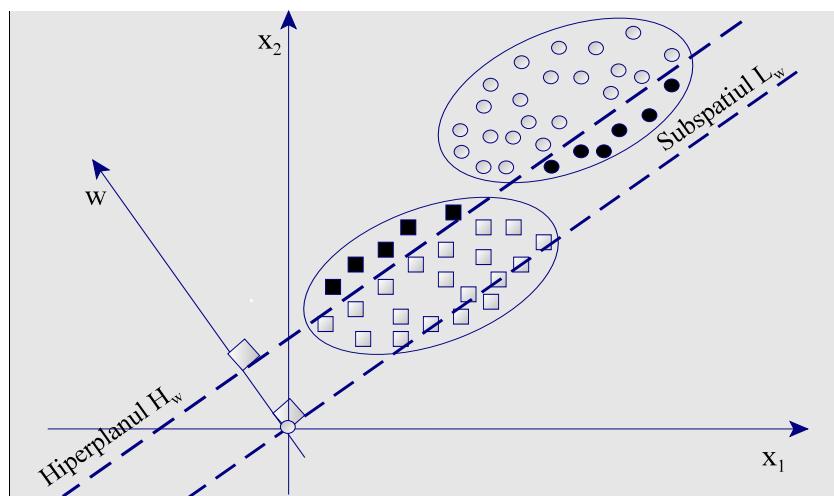


Figura 9.12: Separarea a două clase cu ajutorul unui vector de ponderi și unui prag de excitație arbitrat

În cadrul figurii precedente, poziția particulară a hiperplanului H_w , determinat de vectorul de ponderi \mathbf{w} și de pragul de excitație w_0 , este astfel încât 11 dintre forme, 5 forme aparținând clasei ω_- și 6 forme aparținând clasei ω_+ , respectiv formele a căror reprezentare este hașurată mai intens, sunt clasificate în mod incorrect, în clasa ω_+ , respectiv în clasa ω_- .

Pornind de la numărul de forme clasificate în mod corect, respectiv incorrect, de către o rețea neuronală, pentru care vectorul de ponderi este \mathbf{w} , iar pragul de excitație al neuronului de ieșire este w_0 , se poate stabili o măsură a performanței

respectivei rețele neuronale, exprimată, de exemplu, sub forma unei rate a corectitudinii clasificării.

Este evident însă că, atâtă timp cât valorile curente ale vectorului de ponderi w și pragul de excitație w_0 , sunt diferite de valorile optimale \tilde{w} și \tilde{w}_0 , performanța rețelei neuronale poate fi încă îmbunătățită, prin modificări succesive adecvate ale vectorului ponderilor w și ale pragului de excitație w_0 .

Procedura iterativă de determinare a vectorului de ponderi optimale \tilde{w} și a pragului de excitație optimal \tilde{w}_0 , pe baza informațiilor din setul de învățare și a unor valori arbitrar ale celor două entități informative, este cunoscută sub numele de *algoritm de instruire* sau *algoritm de învățare* a rețelei neuronale de tip perceptron.

În principiu, algoritmul de instruire a unei rețele neuronale de tip perceptron, constă în următoarele: inițial se pornește cu un vector de ponderi w și un prag de excitație w_0 , alese în mod arbitrar; valorile acestor mărimi sunt modificate în mod iterativ și după reguli adecvate, astfel încât să aibă loc o creștere continuă a ratei corectitudinii clasificării cu ajutorul rețelei neuronale; procesul iterativ de modificare a vectorului de ponderi w și a pragului de excitație w_0 , se oprește atunci când performanța rețelei neuronale nu mai poate fi îmbunătățită în mod semnificativ.

Cea mai importantă problemă legată de algoritmul de instruire a perceptronului, este aceea a definirii *regulilor de modificare* a vectorului de ponderi w și a pragului de excitație w_0 , astfel încât această modificare să conducă la o îmbunătățire a performanței rețelei neuronale. După cum se poate intui, definirea oricărei reguli de modificare a vectorului de ponderi w și a pragului de excitație w_0 , presupune necesitatea de a formula criteriul de performanță al rețelei neuronale, astfel încât acesta să fie exprimat în funcție de vectorul de ponderi w și de pragul de excitație w_0 .

Așa cum am arătat anterior, distanța de la orice punct x din spațiul formelor, la hiperplanul H_w , care este asociat cu vectorul de ponderi normalizat w și cu pragul de excitație w_0 , poate fi exprimată în funcție de vectorul de ponderi w , astfel:

$$d(x, H_w) = \|w^t(x - z^{(0)})\|,$$

unde $z^{(0)}$ este vectorul de translație din definirea hiperplanului, vector pe care, în acest context, este convenabil să îl considerăm a fi de forma:

$$z^{(0)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -\frac{w_0}{w_n} \end{pmatrix}.$$

În aceste condiții, cea mai elegantă modalitate de a formula criteriul de performanță a rețelei neuronale, astfel încât acesta să depindă în mod direct de vectorul de ponderi w și de pragul de excitație w_0 , este aceea de a considera că acest criteriu este reprezentat de suma distanțelor de la toate formele clasificate incorrect, la hiperplanul asociat cu vectorul de ponderi w și cu pragul de excitație w_0 .

Astfel, dacă vom nota cu $B(w_0, w) \subset \omega_+ \cup \omega_-$, mulțimea tuturor formelor din setul de învățare, care sunt clasificate *incorrect*, atunci criteriul de performanță al rețelei neuronale poate fi exprimat sub forma următoare:

$$J(w_0, w) = \sum_{u \in B(w_0, w)} d(u, H_w) = \sum_{u \in B(w_0, w)} \|w^t(u - z^{(0)})\|_{\|w\|=1},$$

unde vectorul $z^{(0)}$ este de forma:

$$z^{(0)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ -\frac{w_0}{w_n} \end{pmatrix}.$$

Este evident că, exprimată sub această formă, performanța rețelei neuronale va fi cu atât mai mare, cu cât vectorul de ponderi normalizat w și pragul de excitație w_0 sunt astfel încât, valoarea funcției $J(\cdot, \cdot)$ este cât mai mică.

Rezultă că problema determinării vectorului de ponderi w și a pragului de excitație w_0 , astfel încât performanța rețelei neuronale să fie maximă, poate fi formulată ca o problemă de extrem, de forma următoare:

$$\boxed{\min_{w_0, w} \left\{ J(w_0, w) = \sum_{u \in B(w_0, w)} \|w^t(u - z^{(0)})\|_{\|w\|=1} \right\}}$$

unde $B(w_0, w)$ este mulțimea tuturor formelor din setul de instruire, care sunt clasificate incorrect de către rețeaua neuronală care are vectorul de ponderi w și pragul de excitație a neuronului de ieșire egal cu w_0 .

Relațiile care definesc perceptronul, adică relațiile care definesc starea neuronului și output-ul acestuia sunt următoarele:

$$\begin{cases} s = w_0 + \mathbf{w}^t \mathbf{x}; \\ f(s) = \begin{cases} -1, & \text{dacă } s < 0, \text{ respectiv } \mathbf{w}^t \mathbf{x} < -w_0; \\ 1, & \text{dacă } s > 0, \text{ respectiv } \mathbf{w}^t \mathbf{x} > -w_0. \end{cases} \end{cases}$$

Din modul în care este definită ieșirea neuronului, rezultă că pragul de activare sau de excitare a neuronului este egal cu $-w_0$.

Pentru simplificare, vom considera, pentru început, că pragul de activare w_0 este egal cu zero, respectiv $w_0 = 0$, și că ponderile w_1, w_2, \dots, w_n sunt date. Atunci, din felul în care este definită funcția de activare $f(\cdot)$, rezultă următoarele trei mulțimi posibile de forme, cărora pot să le aparțină formele de intrare:

- mulțimea formelor $\mathbf{x} \in X$, pentru care:

$$s < 0, \text{ adică } \mathbf{w}^t \mathbf{x} < 0;$$

- mulțimea formelor $\mathbf{x} \in X$, pentru care:

$$s = 0, \text{ adică } \mathbf{w}^t \mathbf{x} = 0;$$

- mulțimea formelor $\mathbf{x} \in X$, pentru care:

$$s > 0, \text{ adică } \mathbf{w}^t \mathbf{x} > 0.$$

Cele trei mulțimi sunt cunoscute sub numele de *regiuni de decizie*, iar determinarea lor presupune rezolvarea ecuației:

$$\mathbf{w}^t \mathbf{x} = 0,$$

care este ecuația unui subspațiu cu dimensiunea egală cu $n-1$. Din punct de vedere geometric, subspațiu definit de ecuația precedentă este o suprafață liniară care trece prin origine.

Mulțimea soluțiilor acestei ecuații, adică mulțimea tuturor formelor \mathbf{x} , care verifică această ecuație, formează un subspațiu L_w al spațiului formelor \mathbb{R}^n , respectiv:

$$L_w = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{w}^t \mathbf{x} = 0\},$$

subspațiu care separă spațiul formelor \mathbb{R}^n în trei regiuni distincte:

- prima regiune, pe care o vom nota cu R_- , este formată din acele forme din spațiul formelor \mathbb{R}^n , pentru care mărimea s , care definește starea internă a neuronului, este *strict pozitivă*, adică:

$$\mathbf{w}^t \mathbf{x}^+ > 0;$$

- cea de-a doua regiune, pe care o vom nota cu R_0 , este formată din acele forme din spațiul formelor \mathbb{R}^n , pentru mărimea ce definește starea internă este *egală cu zero*, adică:

$$\mathbf{w}^t \mathbf{x} = 0;$$

- cea de-a treia regiune, pe care o vom nota cu R_+ , este formată din acele forme din spațiul formelor \mathbb{R}^n , pentru mărimea ce definește starea internă este *strict negativă*, adică:

$$\mathbf{w}^t \mathbf{x} < 0.$$

În condițiile în care setul de învățare este format din clasele exclusive ω_+ și ω_- , cele trei regiuni din spațiul formelor \mathbb{R}^n , delimitate cu ajutorul rețelei neuronale, pot fi considerate a fi *perfect corecte*, dacă și numai aceste regiuni sunt astfel încât regiunea R_- include absolut toate formele din clasa ω_- , iar regiunea R_+ include absolut toate formele din clasa ω_+ , adică:

$$\omega_- \subset R_- \text{ și } \omega_+ \subset R_+.$$

Această situație ideală, corespunde cazului în care rețeaua neuronală stabileste în mod perfect corect apartenența pentru toate formele din setul de învățare. O astfel de situație ideală nu poate să apară însă, decât în cazul în care clasele din setul de învățare sunt liniar separabile.

După cum se poate observa, orice vector \mathbf{x} aparținând subspațiului L_w , este ortogonal cu vectorul ponderilor \mathbf{w} , ceea ce înseamnă că vectorul de ponderi \mathbf{w} este ortogonal cu întregul subspațiu L_w . Mai mult decât atât, fiecărui vector de ponderi \mathbf{w} , i se asociază un anumit subspațiu unic L_w , în raport cu care vectorul \mathbf{w} este ortogonal. Așa cum o să vedem în continuare, problema care se pune este aceea de a determina acel vector de ponderi $\tilde{\mathbf{w}}$, căruia să i se asocieze acel subspațiu, care are proprietatea că asigură cea mai bună separare a formelor pe cele două clase.

Dacă vom considera cazul vectorilor bidimensionali, subspațiu L_w , care asigură separarea corectă a formelor, se reduce la o dreaptă în planul \mathbb{R}^2 , dreaptă care trece prin origine, iar vectorul ponderilor \mathbf{w} este ortogonal pe această dreaptă, așa

cum se poate vedea în figura următoare.

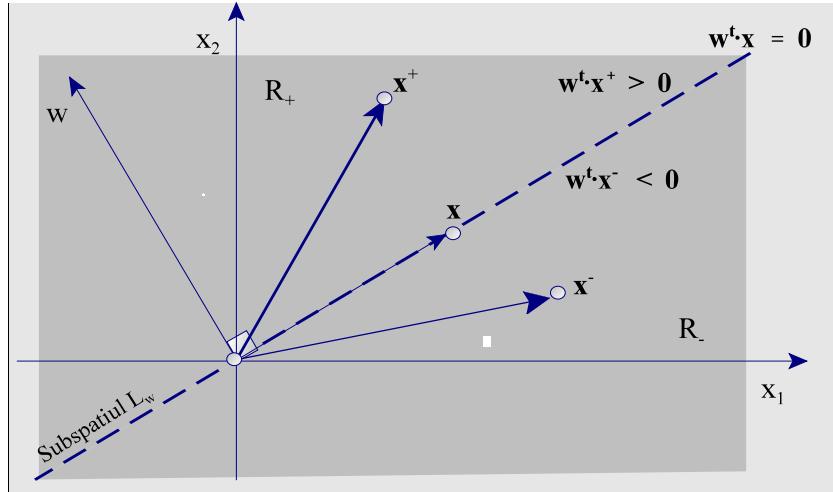


Figura 9.13: Ortogonalitatea vectorului ponderil w și hiperplanului H

Vom presupune acum că vectorul de ponderi w este la un moment dat, astfel încât acestuia îi corespunde acel subspațiu L_w , în raport cu care o formă care aparține în mod real clasei ω_+ , este plasată în mod incorrect în regiunea R_- , aşa cum este ilustrat în cadrul figurii următoare.

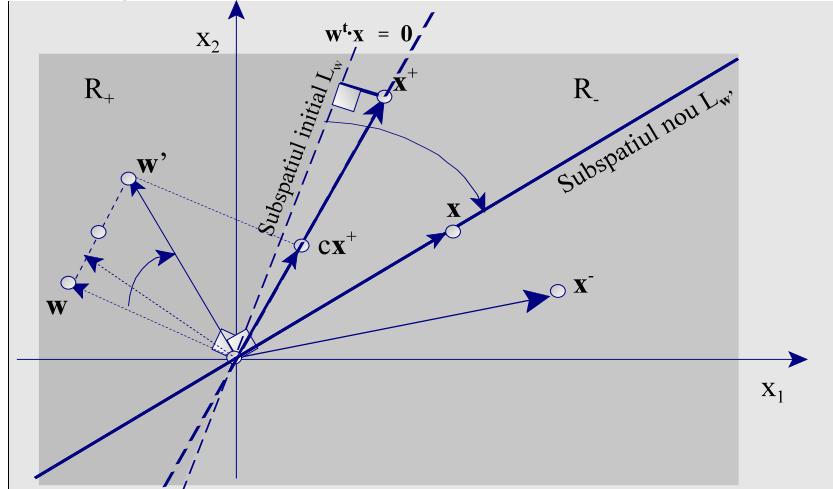


Figura 9.14: Ortogonalitatea vectorului ponderil w și hiperplanului H

Doarece în raport cu subspațiuul L_w , forma $x^+ \in \omega_+$ este situată în mod incorrect în regiunea R_- , adică ea satisface inegalitatea:

$$w^t x^+ < 0,$$

este necesar ca vectorul de ponderi w , să fie modificat în aşa fel încât noului vector de ponderi w' , să i se asocieze acel subspațiu $L_{w'}$, care aduce formă x^+ în regiunea corectă R_+ .

După cum se poate observa, această modificare poate fi obținută, definind noul vector de ponderi w' , sub forma sumei dintre vectorul de ponderi inițial w și un vector $c x^+$, unde $c \leq 1$, respectiv sub forma:

$$w' = w + c x^+,$$

unde constanta scalară $c \leq 1$ se numește *rată de învățare*. În aceste condiții, vom avea:

$$(w')^t x^+ > 0 \text{ și } (w')^t x^- > 0,$$

ceea ce înseamnă clasificarea corectă a celor două forme.

În mod similar, vom presupune că vectorul de ponderi w este la un moment dat, astfel încât acestuia îi corespunde acel subspațiu L_w , în raport cu care o formă care aparține în mod real clasei ω_- , este plasată în mod incorrect în regiunea R_+ , aşa cum este ilustrat în cadrul figurii următoare.

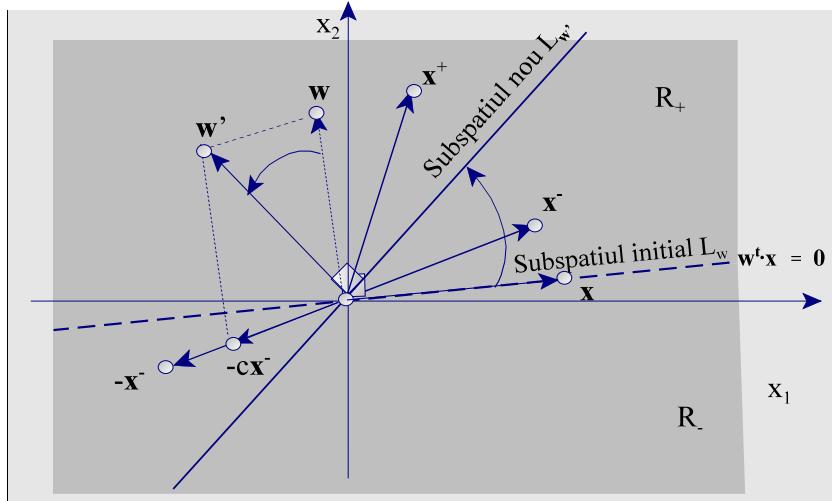


Figura 9.15: Ortogonalitatea vectorului ponderilor w și hiperplanului H

Doarece în raport cu subspațiul L_w , forma $x^- \in \omega_+$ este situată în mod incorrect în regiunea R_+ , adică ea satisfacă inegalitatea:

$$w^t x^- > 0,$$

este necesar ca vectorul de ponderi w , să fie modificat în aşa fel încât noului vector de ponderi w' , să i se asocieze acel subspațiu $L_{w'}$, care aduce forma x^- în regiunea corectă R_- .

După cum se poate observa, această modificare poate fi obținută, definind noul vector de ponderi w' , sub forma sumei dintre vectorul de ponderi inițial w și un vector $-cx^-$, unde $c \leq 1$, respectiv sub forma:

$$w' = w - cx^-,$$

unde constanta scalară $c \leq 1$ reprezintă *rata de învățare*. În aceste condiții, vom avea:

$$(w')^t x^- < 0 \text{ și } (w')^t x^+ < 0,$$

ceea ce înseamnă clasificarea corectă a celor două forme.

Pe baza celor menționate anterior, poate fi dedusă următoarea regulă generică de determinare a vectorului de ponderi w , numită și regulă de învățare: dacă x este forma prezentată la intrarea rețelei neuronale, formă care aparține fie clasei ω_- , fie clasei ω_+ , iar $w^{(k)}$ este vectorul de ponderi corespunzător pasului k de instruire a rețelei neuronale, atunci vectorul de ponderi de la pasul $k+1$ se determină după următoarea regulă:

- | |
|--|
| <ul style="list-style-type: none"> • $w^{(k+1)} = w^{(k)} - cx$, dacă $(w^{(k)})^t x > 0$ și $x \in \omega_-$; • $w^{(k+1)} = w^{(k)} + cx$, dacă $(w^{(k)})^t x < 0$ și $x \in \omega_+$. |
|--|

Prima situație, corespunde unei predicții incorecte, în sensul că o formă care aparține în mod real clasei ω_- , este predictată ca aparținând clasei ω_+ . Cea de-a doua situație, corespunde tot unei predicții incorecte, însă în sensul că o formă care aparține în mod real clasei ω_+ , este predictată ca aparținând clasei ω_- .

În toate celelalte situații, care corespund unor predicții corecte, respectiv în situațiile în care avem:

$$(w^{(k)})^t x < 0 \text{ și } x \in \omega_- \text{ sau } (w^{(k)})^t x > 0 \text{ și } x \in \omega_+,$$

vectorul ponderilor rămâne neschimbat, respectiv:

$$w^{(k+1)} = w^{(k)}.$$

După cum se poate observa, condiția de modificare ce corespunde primei situații, poate fi scrisă și sub forma echivalentă:

$$w^{(k+1)} = w^{(k)} + c(-x), \text{ dacă } (w^{(k)})^t (-x) < 0 \text{ și } x \in \omega_-.$$

Această observație este foarte utilă, deoarece oferă posibilitatea găsirii unei exprimări mai condensate a regulii de decizie anterioare, exprimare bazată pe aşa-numita *normalizare a semnului* formelor de intrare.

Normalizarea semnului formelor de intrare, presupune definirea unui formă generice z , sub forma următoare:

$$z = \begin{cases} -x, & \text{dacă } x \in \omega_-; \\ x, & \text{dacă } x \in \omega_+. \end{cases}$$

Pe baza acestei normalizări, indiferent care este forma $x \in \omega_- \cup \omega_+$, aceasta este clasificată *în mod corect* dacă:

$$\mathbf{w}^t \mathbf{z} > 0,$$

respectiv clasificată *în mod incorrect*, dacă:

$$\mathbf{w}^t \mathbf{z} \leq 0.$$

În aceste condiții, regula de instruire a rețelei neuronale, menționată anterior, poate fi reformulată sub forma următoare:

$$\mathbf{w}^{(k+1)} = \begin{cases} \mathbf{w}^{(k)}, & \text{dacă } (\mathbf{w}^{(k)})^t \mathbf{z} > 0; \\ \mathbf{w}^{(k)} + c \mathbf{z}, & \text{dacă } (\mathbf{w}^{(k)})^t \mathbf{z} \leq 0. \end{cases}$$

Vom considera că procesul de învățare începe cu un vector inițial de ponderi $\mathbf{v}^{(0)}$, ales în mod arbitrar, eventual pe calea generării aleatoare.

Deoarece vectorul inițial de ponderi $\mathbf{v}^{(0)}$ este ales în mod arbitrar, este evident că, în raport cu acest vector, unele dintre cele T forme din setul de învățare, unele vor fi clasificate în mod corect, în timp ce altele vor fi clasificate incorrect.

Deoarece formele clasificate corect nu determină modificarea vectorului de ponderi, vom presupune că în procesul iterativ de învățare sunt luate în considerare doar acele forme din setul de învățare, care sunt clasificate în mod incorrect de către vectorul de ponderi $\mathbf{v}^{(0)}$, adică formele \mathbf{z} pentru care:

$$(\mathbf{v}^{(0)})^t \mathbf{z} \leq 0,$$

și vom presupune că formele care verifică această condiție sunt elemente ale mulțimii:

$$B(\mathbf{v}^{(0)}) = \left\{ \mathbf{z}^{(t_1)}, \mathbf{z}^{(t_2)}, \dots, \mathbf{z}^{(t_m)} \right\}.$$

În acest context, procesul de învățare va fi reprezentat de următoarea succesiune a vectorilor de ponderi:

$$\begin{aligned} \mathbf{v}^{(0)} &= \text{dat}; \\ \mathbf{v}^{(1)} &= \mathbf{v}^{(0)} + c \mathbf{z}^{(t_1)}; \\ \mathbf{v}^{(2)} &= \mathbf{v}^{(1)} + c \mathbf{z}^{(t_2)} = \mathbf{v}^{(0)} + c \left(\mathbf{z}^{(t_1)} + \mathbf{z}^{(t_2)} \right); \\ &\dots \\ \mathbf{v}^{(m)} &= \mathbf{v}^{(m-1)} + c \mathbf{z}^{(t_{m-1})} = \mathbf{v}^{(0)} + c \left(\mathbf{z}^{(t_1)} + \mathbf{z}^{(t_2)} + \dots + \mathbf{z}^{(t_m)} \right). \end{aligned}$$

Din analiza procesului iterativ anterior, desfășurat peste mulțimea de forme clasificate incorrect $B(\mathbf{v}^{(0)})$, se desprind o serie de concluzii foarte importante pentru rafinarea procesului de instruire a rețelei neuronale.

În primul rând, ca urmare a faptului că rata de învățare c este aleasă în mod arbitrar în intervalul $(0, 1]$, modificarea la care este supus vectorul de ponderi $\mathbf{v}^{(k)}$, la cel de-al k -lea pas al schemei iterative, modificare dată de relația:

$$\mathbf{v}^{(k)} = \mathbf{v}^{(0)} + c \left(\mathbf{z}^{(t_1)} + \mathbf{z}^{(t_2)} + \dots + \mathbf{z}^{(t_k)} \right), \quad k=1,2,\dots,m,$$

nu asigură certitudinea că noul vector de ponderi $\mathbf{v}^{(k)}$, va clasifica, dintr-o dată și în mod corect forma $\mathbf{z}^{(t_k)}$, deoarece clasificarea corectă a formei $\mathbf{z}^{(t_k)}$ este asigurată numai pentru o valoare optimală \tilde{c} a ratei de învățare c . Cu toate acestea, problema poate fi însă rezolvată, determinând de la început chiar valoarea optimală \tilde{c} a constantei de învățare c , astfel încât vectorul de ponderi $\mathbf{v}^{(k)}$ să asigure în mod direct clasificarea corectă a formei $\mathbf{z}^{(t_k)}$. Determinarea directă a constantei de învățare \tilde{c} , ar putea fi privită ca fiind asociată cu un proces de *învățarea rapidă*, în contextul căreia rețeaua neuronală este instruită într-un număr mai mic de iterații, respectiv cu o viteză mai mare.

În al doilea rând, chiar dacă o formă $\mathbf{z}^{(k)}$ este clasificată în mod corect de către vectorul de ponderi $\mathbf{v}^{(k)}$, este posibil ca anumite forme care au fost clasificate anterior în mod corect, prin succesiunea de vectori $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(k-1)}$, să apară ca forme clasificate în mod incorrect prin intermediul noului vector de ponderi $\mathbf{v}^{(k)}$, care asigură clasificarea corectă a formei $\mathbf{z}^{(k)}$.

O soluție rezonabilă pentru rezolvarea acestei probleme, poate fi aceea a sortării formelor din mulțimea $B(\mathbf{v}^{(k-1)})$, în ordinea descrescătoare a distanțelor față de spațiul asociat cu vectorul de ponderi $\mathbf{v}^{(k)}$, și modificarea succesivă a vectorului $\mathbf{v}^{(k)}$, în funcție de formele ordonate din acest punct de vedere.

O modalitate elegantă de rezolvare simultană a celor două categorii de probleme menționate anterior, este aceea în care modificarea succesivă a vectorului de ponderi $\mathbf{v}^{(k)}$, se face peste un șir infinit de forme, șir format dintr-o succesiune infinită de subșiruri identice, în care fiecare subșir este reprezentat de succesiunea de forme normalize din setul de instruire, respectiv de mulțimea $\{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(t)}, \dots, \mathbf{z}^{(T)}\}$, unde:

$$\mathbf{z}^{(t)} = \begin{cases} -\mathbf{x}^{(t)}, & \text{dacă } \mathbf{x}^{(t)} \in \omega_-; \\ \mathbf{x}^{(t)}, & \text{dacă } \mathbf{x}^{(t)} \in \omega_+. \end{cases}$$

În acest context, eșantionul de forme este utilizat pentru instruirea rețelei neuronale într-o manieră *ciclică, repetată indefinit*, în care șirul de forme normalize utilizate pentru instruirea rețelei neuronale este de forma:

$$\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(T)}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(T)}, \dots, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(T)}, \dots$$

Formele din acest sir de forme normalize, vor fi utilizate in mod succesiv, incepand cu prima forma din sir, pentru a determina modificarile vectorului de ponderi v , astfel incat reteaua neuronală să fie configurată astfel incat să recunoască in mod corect apartenența tuturor formelor din setul de învățare.

Vom considera că modificarea vectorului de ponderi $v^{(k)}$, de la o formă la alta, în cadrul același subșir, definește o *subiterație*, iar modificarea vectorului de ponderi $v^{(k)}$ odată cu trecerea de la un subșir la altul, definește o *iterație*. Cu alte cuvinte, o iterație este formată dintr-un număr de T subiterații, motiv pentru care o astfel de iterație, este cunoscută în literatura de specialitate a domeniului sub numele de *epochă*.

În aceste condiții, regula de instruire a rețelei neuronale poate fi formulată sub forma următoare:

$$v^{(k+1)} = \begin{cases} v^{(k)}, & \text{dacă } (v^{(k)})^T z^{(k \bmod T)} > 0; \\ v^{(k)} + c z^{(k \bmod T)}, & \text{dacă } (v^{(k)})^T z^{(k \bmod T)} \leq 0, \end{cases}$$

unde $k = 1, 2, \dots, T, T+1, T+2, \dots, 2T, 2T+1, 2T+2, \dots, 3T, \dots$, iar " $k \bmod T$ " este valoarea lui " k modulo T " sau restul împărțirii întregi a lui k prin T , respectiv:

$$k \bmod T \in \{1, 2, \dots, T\}.$$

Cu alte cuvinte, dacă vectorul de ponderi $v^{(k)}$, obținut la cea de-a k -a iterație, rămâne neschimbat numai dacă al clasifică în mod corect forma $z^{(k \bmod T)}$. În caz contrar, vectorul de ponderi $v^{(k)}$ este modificat conform relației menționate.

În ceea ce privește criteriul de oprire a algoritmului perceptronului, facem precizarea că acesta se referă la acea iterație k^* , pentru care s-a obținut acel vector de ponderi $v^{(k^*)}$, care asigură clasificarea corectă a tuturor celor T forme din setul de instruire. Cu alte cuvinte, iterația k^* , la care algoritmul se oprește, nu poate fi decât un multiplu de numărul de forme T , respectiv:

$$k^* = qT, \quad q \in \{1, 2, 3, \dots\},$$

adică iterația pentru care are loc egalitatea:

$$v^{(qT)} = v^{((q-1)T)}.$$

Vom numi procesul de instruire descris de schema iterativă precedentă, proces de instruire *formă cu formă*. Acest proces este similar proceselor de memorare bazate pe repetarea cunoștințelor, cunoștințe reprezentate, în acest caz, de formele din setul de instruire, forme a căror apartenență este cunoscută.

În sfârșit, în al treilea rând, din analiza schemei iterative precedente, rezultă că modificarea vectorului de ponderi $v^{(k)}$, se poate face de fiecare dată când este considerată o formă $z^{(t_k)}$, clasificată în mod *incorrect* de către vectorul de ponderi $v^{(k)}$, ci global, dintr-o dată, pentru toate formele clasificate în mod incorrect. În acest fel, toate cele T iterații care vizează modificarea succesivă a vectorului de ponderi $v^{(0)}$, pentru fiecare formă clasificată în mod incorrect, pot fi reduse la o singură iterație, efectuată peste mulțimea tuturor formelor clasificate incorrect la pasul respectiv.

Având în vedere cele menționate anterior, rezultă că pentru fiecare epocă p , putem defini un vector de ponderi $w^{(p)}$, sub forma:

$$\begin{aligned} w^{(0)} &= v^{(0)}; \\ w^{(1)} &= v^{(T)} = v^{(0)} + c \left(z^{(t_{0,1})} + z^{(t_{0,2})} + \dots + z^{(t_{0,k_0})} \right) = w^{(0)} + c \sum_{z \in B(w^{(0)})} z; \\ w^{(2)} &= v^{(2T)} = v^{(T)} + c \left(z^{(t_{1,1})} + z^{(t_{1,2})} + \dots + z^{(t_{1,k_1})} \right) = w^{(1)} + c \sum_{z \in B(w^{(1)})} z; \\ &\dots \\ w^{(p)} &= v^{(pT)} = v^{((p-1)T)} + c \left(z^{(t_{p-1,1})} + z^{(t_{p-1,2})} + \dots + z^{(t_{p-1,k_{p-1}})} \right) = w^{(p-1)} + c \sum_{z \in B(w^{(p-1)})} z; \\ w^{(p+1)} &= v^{((p+1)T)} = v^{(pT)} + c \left(z^{(t_{p,1})} + z^{(t_{p,2})} + \dots + z^{(t_{p,k_p})} \right) = w^{(p)} + c \sum_{z \in B(w^{(p)})} z, \end{aligned}$$

unde:

$$\left\{ z^{(t_{q,1})} + z^{(t_{q,2})} + \dots + z^{(t_{q,k_q})} \right\} = B(w^{(q)}), \quad q = 0, 1, 2, \dots, p,$$

reprezintă mulțimea formelor clasificate incorrect prin intermediul vectorului de ponderi $w^{(q)}$, iar k_q reprezintă numărul de forme clasificate incorrect prin intermediul acestui vector.

În aceste condiții, dacă vectorul de ponderi $w^{(p)}$, corespunzător iterației p , nu asigură clasificarea corectă a tuturor formelor din setul de instruire, adică $B(w^{(p)}) \neq \emptyset$, atunci el trebuie modificat în continuare, după schema iterativă:

$$w^{(p+1)} = w^{(p)} + c \sum_{z \in B(w^{(p)})} z.$$

În ceea ce privește criteriul de oprire specific acestei scheme iterative, este evident că dacă, la sfârșitul unei epoci p^* ,

toate formele din eșantionul de învățare sunt clasificate în mod corect, adică vectorul de ponderi $\mathbf{w}^{(p*)}$ verifică relația:

$$(\mathbf{w}^{(p*)})^t \mathbf{z} > 0,$$

oricare ar fi forma normalizată \mathbf{z} , atunci procesul de învățare se oprește, iar vectorul de ponderi care asigură separabilitatea claselor este:

$$\tilde{\mathbf{w}} = \mathbf{w}^{(p*)}.$$

După cum se poate constata, criteriul de oprire menționat anterior este satisfăcut, în mod implicit, la iterația p^* , în cazul în care:

$$\mathbf{w}^{(p*)} = \mathbf{w}^{(p*-1)}.$$

Indiferent de tipul său particular, respectiv *formă cu formă* sau *epochă cu epochă*, în condițiile în care clasele considerate sunt liniar separabile, procesul de învățare este un proces convergent.

Penrtru a descrie și justifica procesul de convergență al algoritmului perceptronului, vom considera mai întâi cazul particular al învățării *formă cu formă*. În acest context, vom presupune că șirul de învățare este reprezentat de următoarea succesiune infinită de forme normalize:

$$\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(T)}, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(T)}, \dots, \mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(T)}, \dots,$$

iar schema iterativă a procesului de instruire este de forma:

$$\mathbf{v}^{(k+1)} = \begin{cases} \mathbf{v}^{(k)}, & \text{dacă } (\mathbf{v}^{(k)})^t \mathbf{z}^{(k \bmod T)} > 0; \\ \mathbf{v}^{(k)} + c \mathbf{z}^{(k \bmod T)}, & \text{dacă } (\mathbf{v}^{(k)})^t \mathbf{z}^{(k \bmod T)} \leq 0. \end{cases}$$

Vom presupune că ω_- și ω_+ sunt două clase de forme liniar separabile, astfel încât există vectorul de ponderi $\tilde{\mathbf{v}}$, care verifică relația:

$$\tilde{\mathbf{v}}^t \mathbf{z}^{(t)} = 0, \quad \forall t > 1, 2, \dots, T,$$

ceea ce înseamnă că hiperplanul $H_{\tilde{\mathbf{v}}}$, care este ortogonal cu vectorul de ponderi $\tilde{\mathbf{v}}$, asigură separarea liniară a celor două clase.

Teorema următoare, justifică convergența șirului de vectori $\mathbf{v}^{(0)}, \mathbf{v}^{(1)}, \dots, \mathbf{v}^{(k)}, \dots$, și corespunzător procesului de instruire formă cu formă, către vectorul de ponderi $\tilde{\mathbf{v}}$, care asigură separabilitatea liniară a celor două clase considerate.

Teoremă: Dacă $\tilde{\mathbf{v}}$ este vectorul de ponderi care asigură separabilitatea liniară a claselor ω_- și ω_+ , iar $\mathbf{v}^{(0)}$ este un vector de ponderi arbitrar, atunci șirul de vectori $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \dots, \mathbf{v}^{(k)}, \dots$, definit de recurență următoare:

$$\mathbf{v}^{(k+1)} = \begin{cases} \mathbf{v}^{(k)}, & \text{dacă } (\mathbf{v}^{(k)})^t \mathbf{z}^{(k \bmod T)} > 0; \\ \mathbf{v}^{(k)} + c \mathbf{z}^{(k \bmod T)}, & \text{dacă } (\mathbf{v}^{(k)})^t \mathbf{z}^{(k \bmod T)} \leq 0, \end{cases}$$

converge către vectorul $\tilde{\mathbf{v}}$, adică:

$$\lim_{k \rightarrow \infty} \mathbf{v}^{(k)} = \tilde{\mathbf{v}}.$$

Vom considera acum cazul particular al celui de-al doilea proces de instruire, respectiv procesul de instruire epochă cu epochă, pentru care schema iterativă este de forma:

$$\mathbf{w}^{(p+1)} = \mathbf{w}^{(p)} + c \sum_{\mathbf{z} \in B(\mathbf{w}^{(p)})} \mathbf{z}.$$

Teorema următoare, justifică convergența șirului de vectori $\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \dots, \mathbf{w}^{(p)}, \dots$, și corespunzător procesului de instruire epochă cu epochă, către vectorul de ponderi $\tilde{\mathbf{w}}$, care asigură separabilitatea liniară a celor două clase considerate.

Teoremă: Dacă $\tilde{\mathbf{w}}$ este vectorul de ponderi care asigură separabilitatea liniară a claselor ω_- și ω_+ , iar $\mathbf{w}^{(0)}$ este un vector de ponderi arbitrar, atunci șirul de vectori $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \dots, \mathbf{w}^{(p)}, \dots$, definit de recurență următoare:

$$\mathbf{w}^{(p+1)} = \mathbf{w}^{(p)} + c \sum_{\mathbf{z} \in B(\mathbf{w}^{(p)})} \mathbf{z},$$

unde $B(\mathbf{w}^{(p)})$ este mulțimea formelor clasificate incorrect de către vectorul de ponderi $\mathbf{w}^{(p)}$, converge către vectorul $\tilde{\mathbf{w}}$, adică:

$$\lim_{p \rightarrow \infty} \mathbf{w}^{(p)} = \tilde{\mathbf{w}}.$$

Deoarece vectorul de ponderi $\mathbf{w}^{(k)}$, obținut după parcurgerea întregii mulțimi de forme clasificate incorrect, poate să nu conducă la situația în care toate formele din setul de învățare să fie clasificate în mod corect, procesul de învățare este reluat, cu luarea în considerare a formelor clasificate incorrect după parcurgerea întregii mulțimi de forme clasificate incorrect.

Facem observația că normalizarea de semn a formelor din clasele ω_- și ω_+ , nu modifică vectorul de ponderi w , deci nu modifică nici subspațiul L_w , asociat acestuia.

Pentru a ilustra efectul normalizării de semn a formelor, vom considera situația din figura următoare.

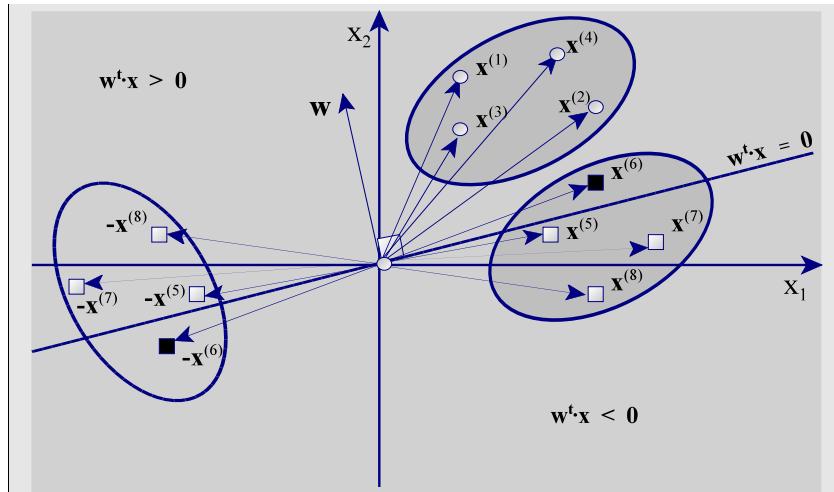


Figura 9.16: Efectul normalizării de semn a formelor

După cum se poate observa, singura formă normalizată ca semn, care nu verifică inegalitatea:

$$w^t z > 0,$$

adică este clasificată incorect cu vectorul de ponderi w , este forma $z = -x^{(6)}$. Pentru toate celelalte forme, are loc verificarea inegalității precedente, ceea ce înseamnă că sunt clasificate în mod corect cu vectorul de ponderi w .

Considerând situația în care formele sunt normalizate ca semn, mulțimea formelor clasificate incorect cu vectorul de ponderi w , poate fi scrisă sub forma:

$$B(w) = \{z \mid w^t z \leq 0\},$$

unde u este o formă normalizată. În acest caz, suma distanțelor de la formele clasificate incorect, la subspațiul L_w , ortogonal cu vectorul w , capătă forma următoare:

$$J(w) = \sum_{z \in B(w)} d(z, L_w) = \sum_{z \mid w^t z \leq 0} \|w^t z\|_{\|w\|=1}.$$

Având în vedere că sumarea se face după toate formele clasificate incorect, adică după formele z , pentru care $w^t z \leq 0$, rezultă că suma distanțelor de la formele clasificate incorect, la subspațiul L_w , poate fi scrisă sub forma:

$$J(w) = \sum_{z \mid w^t z \leq 0; \|w\|=1} -w^t z = - \sum_{z \mid w^t z \leq 0; \|w\|=1} w^t z = \left(-w^t \sum_{z \mid w^t z \leq 0} z \right)_{\|w\|=1},$$

iar performanța rețelei neuronale poate fi scrisă sub forma:

$$\min_{w \mid \|w\|=1} \left\{ J(w) = -w^t \sum_{z \mid w^t z \leq 0} z \right\}.$$

Deoarece, în general, $\min_x f(x) = -\max_x \{-f(x)\}$, performanța rețelei neuronale poate fi rescrisă sub forma:

$$\boxed{\max_{w \mid \|w\|=1} \left\{ -J(w) = w^t \sum_{z \mid w^t z \leq 0} z \right\}.}$$

Având în vedere că gradientul funcției $J(\cdot)$, notat cu $\nabla J(\cdot)$, este un vector care definește direcția în care trebuie să se avanseze, astfel încât să se obțină o creștere a funcției $J(\cdot)$, orice schema iterativă bazată pe evaluarea gradientului, care asigură maximizarea funcției $J(\cdot)$, este de forma:

$$\boxed{w^{(k+1)} = \frac{1}{\|w^{(k)} + h_k \nabla J(w^{(k)})\|} (w^{(k)} + h_k \nabla J(w^{(k)})),}$$

unde punctul de start, reprezentat de vectorul normalizat $w^{(0)}$, este dat, h_k este o constantă scalară pozitivă, aleasă în mod convenabil, care determină lungimea pasului care se face la iterația k , iar $\nabla J(w^{(k)})$ reprezintă evaluarea gradientului funcției $J(\cdot)$ în punctul $w^{(k)}$, respectiv:

$$\nabla J(\mathbf{w}^{(k)}) = \left. \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \right|_{\mathbf{w}=\mathbf{w}^{(k)}}.$$

Având în vedere că derivarea unui produs scalar de forma $\mathbf{w}^t \mathbf{x}$, în raport cu vectorul \mathbf{w}^t , se face după regula:

$$\frac{\partial(\mathbf{w}^t \mathbf{x})}{\partial \mathbf{w}^t} = \mathbf{x},$$

rezultă că gradientul funcției $J(\cdot)$ este de forma:

$$\nabla J(\mathbf{w}) = \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\mathbf{w}^t \sum_{\mathbf{z} | \mathbf{w}^t \mathbf{z} \leq 0} \mathbf{z} \right) = \sum_{\mathbf{z} | \mathbf{w}^t \mathbf{z} \leq 0} \mathbf{z},$$

iar schema iterativă care produce maximul funcției, are forma:

$$\boxed{\mathbf{w}^{(k+1)} = \frac{1}{\|\mathbf{w}^{(k)} + h_k \sum_{\mathbf{z} | \mathbf{w}^t \mathbf{z} \leq 0} \mathbf{z}\|} \left(\mathbf{w}^{(k)} + h_k \sum_{\mathbf{z} | \mathbf{w}^t \mathbf{z} \leq 0} \mathbf{z} \right)}.$$

Având în vedere cele menționate anterior, algoritmul perceptronului poate fi descris prin intermediul următoarei succesiuni de pași:

Se consideră existența esantionului de forme n -dimensionale $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)}\}$, formă a că

Datorită faptului că pentru orice formă $\mathbf{x} \in \omega_-$ este verificată, prin definiția clasei ω_- , condiția:

$$\mathbf{w}^t \cdot (-\mathbf{x}) > 0,$$

vom defini o formă generică \mathbf{z} , numită *formă normalizată ca semn*, și definită sub forma următoare:

$$\mathbf{z} = \begin{cases} \mathbf{x}, & \text{dacă } \mathbf{x} \in \omega_+ \\ -\mathbf{x}, & \text{dacă } \mathbf{x} \in \omega_-, \end{cases}$$

Cu această notație, separabilitatea liniară a celor două clase poate fi definită sub o formă mai sintetică, respectiv sub forma următoare:

Definiție: Două clase ω_1 și ω_2 se numesc **liniar separabile**, dacă și numai dacă
are loc inegalitatea:

$$\mathbf{w}^t \cdot \mathbf{z} > 0, \quad \forall \mathbf{z} \in \omega_1 \cup \omega_2,$$

 unde \mathbf{z} este o formă normalizată ca semn.

În concluzie, se poate spune că, în cazul în care clasele sunt liniar separabile, există un hiperplan care delimită cele două clase, hiperplan care poate fi determinat prin metode matematice specifice.