МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ФАКУЛЬТЕТ ЕКОНОМІКИ І УПРАВЛІННЯ ПІДПРИЄМНИЦТВОМ

Погорєлова Т. В.

БАГАТОВИМІРНИЙ СТАТИСТИЧНИЙ АНАЛІЗ

КОНСПЕКТ ЛЕКЦІЙ

ОДЕСА
ОНЕУ
2024

УДК 519.237:31(042.3)
П 43

Укладач:

Погорєлова Т. В. – канд. екон. наук, доцент

Рецензенти:

Янковий О. Г. – доктор екон. наук, професор (зовнішній рецензент)
Самотоєнкова О. В. – канд. екон. наук, доцент
Ольвінська Ю. О. – канд. екон.. наук, доцент

Рекомендовано до друку
Вченою радою Факультету економіки і управління підприємництвом.
Протокол № 10 від 22 квітня 2024 р.

Погорєлова Т. В.

Багатовимірний статистичний аналіз: конспект лекцій / Т. В. Погорєлова. – Одеса: ОНЕУ, ротапринт, 2024 – 115 с.

Конспект лекцій для студентів-здобувачів першого (бакалаврського) рівня освіти денної та заочної форми навчання за освітньо-професійною програмою «Бізнес-економіка та аналітика» спеціальності 051 «Економіка».

УДК 519.237:31(042.3)
П 43
©Погорєлова Т.В., 2024
©Одеській національний економічний університет, 2024

ЗМІСТ

ВСТУП

Сучасний розвиток науки та комп'ютерних технологій сприяв переосмисленню ролі кількісних статистичних методів у аналізі соціально-економічних процесів. Сучасні методи багатовимірного аналізу більш складні, мають високу наукову обґрунтованість, що впливає на точність отриманих результатів дослідження. Їх використання дозволяє виявити та оцінити приховані латентні показники, які на поверхні виявляються через фактори-симптоми, побудувати інтегральні оцінки складних багатоознакових явищ без значних втрат вихідної інформації, провести класифікацію об'єктів за всіма ознаками одночасно.

Багатовимірний статистичний аналіз – це комплексна освітня компонента програми «Бізнес-економіка та аналітика», яка орієнтована на використання статистичного інструментарію високого рівня для визначення реально існуючих закономірностей та тенденцій розвитку соціально-економічних явищ та процесів.

Мета багатовимірного статистичного аналізу полягає у формуванні системи теоретичних знань та практичних навичок з організації та методології проведення статистичних досліджень у економіці на макро- , мезо- та мікрорівнях.

Під час написання конспекту лекцій враховувалось, що студенти оволоділи такими дисциплінами, як макроекономіка, мікроекономіка, вища математика та теорія ймовірностей, статистика, бізнес-статистика, а також інших дисциплін.

У результаті вивчення освітньої компоненти «Багатовимірний статистичний аналіз»» студенти можуть оволодіти основними теоретичними та методологічними принципами економіко-статистичного дослідження соціально-економічних явищ та процесів та застосовувати ці знання на практиці.

Конспект лекцій може бути корисним під час підготовки до практичних занять, виконання індивідуальних завдань та самостійного вивчення освітньої компоненти «Багатовимірний статистичний метод».

Тема 1. МЕТОДОЛОГІЧНІ ОСНОВИ БАГАТОВИМІРНОГО СТАТИСТИЧНОГО АНАЛІЗУ

1.1. Поняття багатовимірних статистичних методів та їх роль у вивченні соціально-економічних явищ.

1.2. Аналітичні завдання, які вирішуються багатовимірними методами.

1.3. Етапи багатовимірного статистичного дослідження.

1.1. Поняття багатовимірних статистичних методів та їх роль у вивченні соціально-економічних явищ

Навколишній світ складається з різноманітних об’єктів, підприємств, регіонів, зайнятих, безробітних, кожен з яких характеризується набором якісних і кількісних ознак. Наприклад, промислові підприємства розрізняються за обсягом виробленої продукції, за розміром основних виробничих засобів, за кількістю працівників, за рівнем рентабельності, за собівартістю продукції та іншими якісними та кількісними показниками. Всього можна назвати приблизно 100 техніко-економічних показників бізнесу.

Інакше кажучи, об’єкти навколо нас відрізняються багатогранністю, тобто дозволяють описувати їх з різних боків. З цієї точку зору одновимірні об’єкти (число описових ознак дорівнює 1), які характеризує традиційна статистика є простими, частковими. Розрахунок таких показників, як середня, дисперсія, коефіцієнтів варіації та інших виконується для окремих ознак. На відміну від класичної статистики, багатовимірні статистичні методи дозволяють вивчати велику кількість показників одночасно, тобто оцінювати саме багатовимірні об’єкти. Хоча потрібно зауважити, що окремі елементи багатомірності зустрічаються у статистиці. Наприклад, комбінаційне групування. Воно використовується під час вивчення складних явищ. Як відомо, його схема передбачає послідовний розподіл вихідної сукупності об’єктів спочатку за першою ознакою, потім у середині кожної групи за другою ознакою і таке інше.

У випадку трьох і більше групувальних ознак утворюються складна громіздка таблиця, серед багаточисленних підгруп якої є незначні чи пусті. Головним недоліком такого підходу є необхідність розташовувати значні за обсягом спостереження, ймовірність отримання пустих груп, складна процедура групування.

Така проблема традиційними одномірними методами вирішується утруднено, але добре розв’язується за допомогою алгоритмів багатовимірного аналізу, насамперед, кластерного аналізу. Об'єднання об'єктів у однорідні групи відбувається за принципами їх близькості у багатовимірному просторі ознак. Розрахунок відстаней між точками простору дозволяє знайти скупчення й вирішувати завдання багатовимірного групування об’єктів незалежно від їхньої кількості.

Поряд з багатовимірним групуванням постає питання віднесення нових об’єктів до відомих, вже існуючих груп (кластерів). Наприклад, віднесення будь-якої країни до групи економічно безпечних або групи економічно небезпечних країн з метою інвестування. Це завдання також вирішується в рамках багатовимірних статистичних методів і виконується за допомогою дискримінантного аналізу.

Доволі часто на практиці необхідно оцінити деякі невимірювані внутрішні латентні якості об’єктів, які на поверхні виявляються у вигляді зовнішніх факторів-симптомів. Це так звані латентні показники. Наприклад, продуктивність праці «білих комірців» виявляється через винахідництво, раціоналізаторство, прибутковість підприємств, ритмічність праці та її оплату, впровадження нових технологій та інше.

Означені проблеми можна розв’язати за допомогою методів багатовимірного статистичного аналізу.

Багатовимірні методи мають велику кількість напрямків таких, як кластерний аналіз, дискримінанти аналіз, метод головних компонент, факторний аналіз та інші.

Основними можливостями використання багатовимірних статистичних методів є *три ключові позиції*: по-перше, одночасне використання інформаційних даних з різнорідних джерел; по-друге, просіювання даних з метою виявлення прихованих закономірностей, латентних характеристик, що сприяє підвищенню конкурентоспроможності; по-третє, новий, більш розвинутий рівень порівняльного аналізу багатовимірних об'єктів.

Отже, ***багатовимірний статистичний аналіз*** **–** це сукупність формалізованих статистичних методів, які базуються на поданні результативної інформації в багатовимірному геометричному просторі та дозволяють визначати приховані (латентні), але об'єктивно існуючі закономірності в організаційній структурі та тенденціях розвитку соціально-економічних явищ і процесів.

За змістом багатовимірний статистичний аналіз можна умовно розбити на три складові:

1. Багатовимірний статистичний аналіз спільних розподілів та їх основних характеристик.

2. Багатовимірний статистичний аналіз характеру та структури взаємозв'язків між компонентами досліджуваної багатовимірної ознаки.

3. Багатовимірний статистичний аналіз геометричної структури досліджуваної сукупності багатовимірних спостережень.

1.2. Аналітичні завдання, які вирішуються багатовимірними методами

Розвиток багатовимірного статистичного аналізу пов'язано зі сплеском розвитку науково-технічного прогресу. Спочатку методи впорядкування та класифікація багатопараметричних об'єктів та явищ застосовувалися в природничих науках – хімії, біології, ботаніці, що простежується у роботах таких учених, як:

1) Ч. Дарвін (60-ті рр. XIX ст., англійський натураліст) – використовував означені методи у селекції видів і для визначення чинників еволюції органічного світу;

2) Д. І. Менделєєв (60–70 рр. XIX ст.) – використовував для систематизації якісних характеристик хімічних елементів у розробці періодичній системі елементів;

3) А.П. Шлікевич, І.Ф. Анненський та ін. (початок XX ст.) – використовували для класифікації земельних господарств.

Самостійний науковий розвиток багатовимірні статистичні методи отримали на початку XX століття у публікаціях англійських вчених Ф. Гальтона, К. Пірсона і У. Спірмена, присвячених основам побудови алгоритмів стиснення статистичних даних. Але засновниками теорії багатовимірного аналізу вважаються такі вчені, як Л.Л. Терстоун, Л.Р. Такер, Р. Хорст, Г. Кайзер, Т. Келлі, Г. Харман, Р. Тріон, Р. Сокал, М. Жамбю, Р.В. Хемінг, Л. Заде, Р. Фішер та інші.

Поряд з проблемами багатовимірного групування та класифікацій у соціально-економічних дослідженнях виникає задача виявити та виміряти деякі внутрішні властивості об’єктів, які проявляються на поверхні у вигляді зовнішніх факторів-симптомів. Ці внутрішні (чи приховані) ознаки, які проявляються у вигляді факторів –симптомів називаються **латентними.**

Наприклад, рівень життя проявляється в середніх показниках доходу на душу населення, динаміки індексів цін споживачів товарів, тривалості життя, народжуваності, смертності, тощо. Динаміка захворювань відбувається на підставі його зовнішніх проявів – температури хворого, різноманітних аналізів, кашля. У такому випадку спочатку потрібно виявити захворювання (класифікувати), тобто провести дискримінацію, а потім оцінити латентний показник. Аналогічно про технічний рівень виробництва роблять висновки на основі таких показників, як фондоозброєність праці, енергоозброєність, механоозброєність та інші.

Отже, за допомогою багатовимірних статистичних методів можна виявити та подати у компактній формі означені приховані ознаки, тобто відтворити досліджувані латентні показники.

Під час багатовимірного аналізу виконується групування вихідної інформації за такими напрямками:

1. Виділення груп об’єктів зі схожими поєднаннями значень ознак.
2. Виділення груп ознак, які найбільше відображують образ досліджуваного латентного показника.
3. Скорочення вихідного простору ознак без значної втрати інформації.
4. Вимірювання тісноти кореляційного зв’язку між групами ознак.
5. Багатовимірне шкалювання.

Під час проведення дослідження в реальних економічних умовах означені напрямки взаємно переплітаються.

Розв’язання перших двох напрямків передбачає використовувати методи розпізнання образів. Для цього доцільно використовувати кластерний аналіз (розпізнання без вчителя) і дискримінантний аналіз (розпізнання образу з вчителем).

Завдання другого та третього напрямків, тобто скорочення вихідного простору, вирішуються за допомогою вимірювання відстані між об’єктами та еталонами чи антіеталонами. Даний підхід відомий як таксономічний аналіз. Він дозволяє у 4-5 разів скоротити кількість вихідних ознак.

Однак головним методом оцінки латентних показників, вимірювання кореляційних зв’язків між ознаками є факторний аналіз і метод канонічних кореляцій. Спочатку будуються нові змінні у вигляді лінійних комбінацій вихідних ознак, а потім між ними вимірюється тіснота кореляційних зв’язків. Такий підхід дозволяє виявити залежності між окремими групами показників. Це по суті є найвищий розвиток багатофакторної статистики.

Багатовимірне шкалювання є змішаним підходом, який поєднує проблематику розпізнання образів, скорочення простору ознак, обробки даних, які отримані на підставі експертних оцінок. У межах цього підходу розв’язуються задачі метричного та неметричного шкалювання, вибору індивідуальних думок експертів.

Оскільки статистика вивчає кількісний бік масових явищ та процесів, а багатовимірний аналіз є подальшим розвитком статистики на більш високому рівні, то будемо припускати, що всі ознаки об'єктів є кількісними, а не атрибутивними. Якщо для опису явища необхідне використання атрибутивної ознаки, то потрібно перевести їх за допомогою змінних 0 і 1 до кількісної форми.

До впровадження комп'ютерів з потужною базою для обробки великих масивів, означені питання розв'язувались теоретично, оскільки без відповідної техніки розв'язувати складні задачі неможливо. Тільки сучасні комп'ютерні технології дозволять з успіхом здійснювати обробку багатовимірних даних з метою їх класифікації, ранжування, оцінки латентних показників.

Отже, методи багатовимірного статистичного аналізу дозволяють розв'язувати різноманітні задачі дослідження багатовимірних процесів: групування, стиснення даних, моделювання складних систем, оцінювання параметрів багатовимірних сукупностей та ін. Це дозволяє використовувати отримані результати для прийняття ефективних управлінських рішень.

Зауважимо, що методи багатовимірного статистичного аналізу тісно пов'язані зі статистикою, матричною алгеброю, теорією ймовірностей, аналітичною геометрією.

1.3. Етапи багатовимірного статистичного дослідження

Обробка багатовимірних статистичних даних має певні *особливості*:

1) застосування методів багатовимірного статистичного аналізу вимагає творчого підходу до розв'язання аналітичних задач, оскільки оброблюються багатовимірні сукупності даних;

2) для багатовимірних статистичних методів характерна глибока формалізація та складна логіко-математична конструкція;

3) практичне застосування потребує використання обчислювальної техніки.

Розглянемо **етапи дослідження** за допомогою багатовимірного статистичного аналізу:

1) постановка задачі включає опис предметної області об’єктів, визначення обсягів виділених ресурсів (час, трудовитрати і т. д.);

2) визначення набору методів багатовимірного статистичного аналізу та порядку їх використання;

3) збирання вихідних даних, визначення способів збирання інформації, форм її подання;

4) аналіз даних (перевірка однорідності вибірки, відповідність законам розподілу, виявлення грубих помилок і т. д.);

5) уточнення математичної постановки задачі та визначення можливості застосування раніше відібраних методів (у разі необхідності набір методів змінюється);

6) реалізація, тобто проведення обчислень, за допомогою програмного забезпечення математичного інструментарію;

7) оцінювання адекватності моделі, визначення несуперечності математичних результатів і економічних висновків [3].

На практиці всі перелічені етапи необов'язково присутні і не розмежовані. Деякі з них можуть об'єднуватися або виключатися. Знання всіх етапів дозволяє раціонально планувати реалізацію багатовимірних методів та враховувати майбутні обсяги дослідження.

Питання для самоконтролю

1. Що таке багатовимірний статистичний аналіз?
2. Охарактеризуйте передумови розвитку багатовимірного статистичного аналізу.
3. Які існують напрямки багатовимірного статистичного досліджування?
4. Які завдання вирішують багатовимірні статистичні методи?
5. Опишіть складові багатовимірного статистичного аналізу.
6. Які існують особливості обробки багатовимірних статистичних даних?
7. Охарактеризуйте етапи багатовимірного статистичного аналізу.

Тема 2. ОСНОВИ БАГАТОВИМІРНОЇ КЛАСИФІКАЦІЇ

2.1. Математичне подання багатовимірних об’єктів.

2.2. Формування матриці вихідних даних та її попередня обробка.

2.3. Метрики відстані об’єктів у багатовимірному просторі.

2.4. Метрики подібності об’єктів у багатовимірному просторі.

2.1. Математичне подання багатовимірних об’єктів

В економічних дослідженнях оброблюються великі масиви даних, тобто число об’єктів може досягати кількох десятків чи навіть сотень. У свою чергу число ознак, що їх характеризують, також може обчислюватися десятками. Вочевидь, що безпосередній (візуальний) аналіз таких даних за великої кількості об’єктів і ознак практично малоефективний. Через це виникають завдання укрупнення, концентрації вихідних даних, аналізу структури об’єктів дослідження. Вирішення цих завдань може здійснюватися за допомогою сучасних методів багатовимірної класифікації.

Методи багатовимірної класифікації дозволяють групувати об’єкти з урахуванням усіх істотних структурно-типологічних ознак і характеру розподілу об’єктів у заданій системі ознак.

Така класифікація проводиться на основі прагнення зібрати в одну групу в деякому сенсі схожі об’єкти, причому так, щоб об’єкти з різних груп були за можливості несхожими. Наприклад, у маркетингу – це сегментація конкурентів і споживачів; у менеджменті – розбиття персоналу на різні за рівнем мотивації групи, класифікація постачальників, виявлення схожих виробничих ситуацій, за яких виникає брак; у фінансовому аналізі – для класифікації досліджуваних підприємств за рівнем фінансового стану; у макроекономічних дослідженнях – для класифікації країн чи регіонів окремої країни за рівнем життя населення, за станом трудових ресурсів, за рівнем туристичної привабливості тощо. Отже, методи багатовимірної класифікації використовуються для розділення

сукупності об’єктів на однорідні групи. Водночас кожен об’єкт характеризується великим числом різних стохастично пов’язаних ознак.

У статистичних дослідженнях використовують два підходи до класифікації об’єктів і відповідні методи класифікації (табл.2. 1).

Таблиця 2.1

Сучасні підходи до багатовимірної класифікації

Багатовимірні статистичні методи		
Мета методу	Кластерний аналіз (Розпізнавання образів без «вчителя»)	Дискримінантний аналіз (Розпізнавання образів з «вчителем»)
	Розбиття статистичної сукупності на класи (кластери)	Приписування нового об’єкта до системи кластерів

Кластерний аналіз допомагає виділити у вихідній багатовимірній сукупності однорідні об’єкти (кластери), які всередині кластера схожі між собою, а об’єкти з різних кластерів суттєво різняться між собою.

Дискримінантний аналіз дозволяє доєднати нові об’єкти до існуючих груп, методом порівняння ознак з аналогічними показниками кластерів.

Як велось раніше, багатовимірний аналіз узагальнює велику кількість методів та прийомів для обробки багатопараметричних статистичних даних, тому вибір методу залежить від подання вихідних даних. Найбільш структурованою формою подання вихідних даних про багатовимірний об’єкт є матрична. Розглянемо її.

Вихідна інформація про значення ознак завжди подається у вигляді матриці ***X***, розміром ***n***x***m***:

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2m} \\ & & \dots & & & \\ x_{i1} & x_{i2} & \dots & x_{ik} & \dots & x_{im} \\ & & \dots & & & \\ x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{nm} \end{bmatrix},$$

де ***i*** – номер об’єкта, що досліджується (*i=1,2, ..., n*);

k – номер ознаки, що вивчається (*k*=1,2, …, *m*);

x_{ik} – значення *k –оїознаки* у ***i***-го об’єкта.

Строки цієї матриці відповідають окремим об’єктам, а стовпці – окремими ознакам.

Наприклад, «нехай є дані щодо трьох промислових підприємств А, В, С (***n***=3) про величину двох економічних показників (***m***=2):

x_1 – основні виробничі засоби, млн. грн.;

x_2 – рівень рентабельності, %.

Матриця вихідної інформації матиме вигляд:

$$X=\begin{bmatrix}1200 & 8\\ 500 & 4\\ 400 & -3\end{bmatrix}.$$

Отже, досліджувані промислові підприємства А, В і С є точками двовимірного простору ознак x_1 та x_2 .» [10, с. 10].

За даними матриці вихідних ***X*** даних визначають основні статистичні характеристики: середнє значення ознаки, середнє квадратичне відхилення, коваріацію та ін. Усі розрахунки ведуться за об’єктами (стовпчиками) матриці, тобто по ***i*** (від 1 до ***n***). Більш того, можна скористатися редактором описової статистики Excel.Визначимо означені основні статистичні характеристики:

1. Середнє значення кожної ознаки:

$$\overline{x_k}=\frac{\sum x_k}{n}. \qquad (2.1)$$

Вектор середніх значень досліджуваних ознак $\bar{x}=(\bar{x}_1,\overline{x}_{2,\dots,}\bar{x}_m)$ називається **центром тяжіння** або **центроїдом** досліджуваної сукупності об’єктів.

У результаті розрахунків отримуємо середні значення $\bar{x}_1=700\ \text{тис.грн.}, \bar{x}_2=3\ \%$, тобто центр тяжіння трьох підприємств А, В і С є вектором-рядком із координатами $\bar{x}=(700,3)$. Інакше кажучи, середньостатистичне підприємство має розмір основних засобів 700 млн. грн., а рентабельності – 3%.

2. Коваріацію ***p***-ої та ***s***-ої ознаки (***p, s =1,2, …,m)***

$$\boldsymbol{\sigma_{ps}} = \frac{\sum(x_{ip}-\bar{x}_p)\cdot(x_{is}-\bar{x}_s)}{n}. \quad (2.2)$$

Якщо $\boldsymbol{p=s}$, $\boldsymbol{\sigma_{pp} = \sigma_p^2}$, $\boldsymbol{\sigma_{ss} = \sigma_s^2}$, тобто коваріація ознаки із собою дорівнює її дисперсії.

У прикладі, коваріація дорівнює $\sigma_{12} = 1366{,}667$; дисперся $\sigma_1^2 = 126666{,}667$; $\sigma_2^2 = 20{,}66667$.

3. Середнє квадратичне відхилення $\boldsymbol{k}$-ої ознаки:

$$\sigma_k = \sqrt{\frac{\sum(x_{ik}-\bar{x}_k)^2}{n}}. \quad (2.3)$$

Середнє квадратичне відхилення показує, абсолютну міру відхилення окремих ознак від його середнього значення. Отже, розмір основних засобів підприємств у середньому відхиляються від середнього на 355,9 млн.. грн. ($\sigma_1 = 355{,}9026$), а рентабельність – на 4,5 % ($\sigma_2 = 4{,}5461$).

4. Коефіцієнт парної кореляції $\boldsymbol{p}$-ої та $\boldsymbol{s}$-ої ознаки:

$$r_{ps} = \frac{\sigma_{ps}}{\sigma_p\cdot\sigma_s}. \quad (2.4)$$

Цей коефіцієнт змінюється від -1 до 1 та характеризує тісноту лінійного зв'язку між двома ознаками. Коефіцієнт парної кореляції між розміром основних засобів та рентабельністю дорівнює $r_{12} = 0{,}8447$. Його високе значення вказує на наявність прямого та тісного кореляційного зв'язку між досліджуваними ознаками.

Тісноту кореляційного зв'язку можна вимірювати за формулою (2.4) також й між об'єктами, наприклад А і С

$$r_{AC} = \frac{\sigma_{AC}}{\sigma_A\cdot\sigma_C},$$

тобто як базу розрахунків використовуються не стовпчики, а рядки матриці $\boldsymbol{X}$.

Лінійний зв'язок ознак можна подати у вигляді матриці коефіцієнтів парної кореляції:

$$r = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & 1 & r_{23} & \dots & r_{2m} \\ & & \dots\dots\dots & & \\ r_{m1} & r_{m2} & r_{m3} & \dots & 1 \end{bmatrix},$$

розміром ***mxm.*** З розглянутих формул очевидно, що $r_{11} = r_{22} = \cdots = r_{mm} = 1$ та $\boldsymbol{r_{ps} = r_{sp}}$. Отже, матриця $\boldsymbol{r}$ є симетричною, на головній діагоналі якої розташовані 1. Саме ці матриця коефіцієнтів парної кореляції є основою багатовимірних статистичних методів. Для нашого прикладу матриця коефіцієнтів парної кореляції має вигляд:

$$r = \begin{bmatrix} 1 & 0{,}8447 \\ 0{,}8447 & 1 \end{bmatrix}.$$

2.2. Формування матриці вихідних даних та її попередня обробка

Під час проведення багатовимірної класифікації приймаються дві попередні гіпотези щодо вихідної інформації:

1) відібрані ознаки для аналізу відображують головні властивості об'єктів і потребують розподіл сукупності на групи, кластери, таксони;
2) масштаб вимірювання ознак не впливає на результати багатовимірних статистичних процедур.

Перша передумова означає, що матриця $\boldsymbol{X}$ сформована до початку багатовимірного аналізу, тобто проведено якісний апріорний аналіз предмета вивчення. Зазвичай це відбувається за допомогою наукових підходів, методів та полягає в обґрунтуванні ознак, які найбільш повно характеризують результативні та факторні ознаки.

Друга гіпотеза (правильний вибір масштабів вимірювання) також грає суттєву роль у багатовимірній класифікації. Всі відібрані на апріорному етапі аналізу ознаки $\boldsymbol{x_1, x_2, ..x_m}$ мають різні одиниці вимірювання – натуральні, вартісні, трудові. Зміна масштабів вимірювання ознак не повинна впливати на результати багатовимірної класифікації. Тому всім ознакам надається один безрозмірний вид методом стандартизації. Така процедура відбувається за допомогою центрування та нормування.

Центрування – це віднімання від кожного значення ознаки за всіма об'єктами сукупності її середнього значення. При цьому середня арифметична перетворених таким чином ознак дорівнює нулю.

Нормування – це ділення вихідних значень ознаки на деяке постійне число, зазвичай на середньоквадратичне відхилення.

Отже, процедура стандартизації може бути подана таким чином:

$$z_{ik} = \frac{(x_{ik} - \bar{x}_k)}{\sigma_k}. \qquad (2.5)$$

Стандартизація ознак дозволяє позбавитися від масштабів їх вимірювання, надає всім даним один порядок. Якщо вихідна база даних підпорядковується нормальному розподілу, то діапазон зміни стандартизованих значень складає від -3 до +3. Для стандартизованих ознак справедливими є такі співвідношення:

$$\bar{z}_k = 0, \quad \sigma_k^2 = \sigma_k = 1, \quad \sigma_{ps} = r_{ps}. \qquad (2.6)$$

Таким чином, внаслідок попередньої обробки вихідних даних методом стандартизації матриця ***X*** перетворюються на матрицю ***Z:***

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z_{11}} & \boldsymbol{z_{12}} & \dots & \boldsymbol{z_{1k}} & \dots & \boldsymbol{z_{1m}} \\ \boldsymbol{z_{21}} & \boldsymbol{z_{22}} & \dots & \boldsymbol{z_{2k}} & \dots & \boldsymbol{z_{2m}} \\ & & \dots\dots & & & \\ \boldsymbol{z_{i1}} & \boldsymbol{z_{i2}} & \dots & \boldsymbol{z_{ik}} & \dots & \boldsymbol{z_{im}} \\ & & \dots\dots & & & \\ \boldsymbol{z_{n1}} & \boldsymbol{z_{n2}} & \dots & \boldsymbol{z_{nk}} & \dots & \boldsymbol{z_{nm}} \end{bmatrix}.$$

Отримана матриця має той же розмір, що й матриця ***X***, та слугує відправною точкою будь-якого багатовимірного дослідження. Як і в матриці ***X***, рядки матриці ***Z*** відповідають окремими об'єктам, а стовпчики – окремим ознакам.

Виконаєм стандартизацію ознак $\boldsymbol{x_1}$ та $\boldsymbol{x_2}$. Використаємо знайдені раніше середні значення ознак $\bar{x}_1 = 700$, $\bar{x}_2 = 3$ та середньоквадратичні значення $\sigma_1 = 355{,}9026;\ \sigma_2 = 4{,}5461$. Усі розрахунки подано у табл. 2.2.

Отже, матриця стандартизованих даних трьох промислових підприємств буде мати вигляд:

$$Z = \begin{bmatrix} 1{,}405 & 1{,}100 \\ -0{,}562 & 0{,}220 \\ -0{,}843 & -1{,}320 \end{bmatrix}.$$

Таблиця 2.2

Розрахунок стандартизованих значень

Підприємство	Основні засоби			Рентабельність		
	x_{i1}, млн. грн.	$x_{i1} - \bar{x}_1$	$\frac{x_{i1} - \bar{x}_1}{\sigma_1}$	x_{i2}, %	$x_{i2} - \bar{x}_2$	$\frac{x_{i2} - \bar{x}_2}{\sigma_2}$
1	1200	500	1,405	8	5	1,100
2	500	-200	-0,562	4	1	0,220
3	400	-300	-0,843	-3	-6	-1,320

На підґрунті математичних перетворень, можна довести, що коефіцієнт парної кореляції стандартизованих значень дорівнює:

$$r_{ps} = \frac{\sum z_{ip} z_{is}}{n} = \sigma_{ps}. \quad (2.7)$$

Ураховуючі це співвідношення, коефіцієнти парної кореляції можна визначити через стандартизовані дані, тобто кореляція співпадає з коваріацією. Звідси, матрицю коефіцієнтів парної кореляції можна знайти таким чином:

$$r = \frac{Z^T Z}{n}, \quad (2.8)$$

де Z^T – матриця, транспонована відносно матриці стандартизованих ознак. Розрахуємо матрицю коефіцієнтів парної кореляції розглянутим способом.

$$r = \begin{bmatrix} 1{,}405 & -0{,}562 & -0{,}843 \\ 1{,}100 & 0{,}220 & -1{,}320 \end{bmatrix} \cdot \begin{bmatrix} 1{,}405 & 1{,}100 \\ -0{,}562 & 0{,}220 \\ -0{,}843 & -1{,}320 \end{bmatrix} \cdot \frac{1}{3} = \begin{bmatrix} 1{,}000 & 0{,}845 \\ 0{,}845 & 1{,}000 \end{bmatrix}$$

Отже, обидва способи дозволили отримати однаковий результат.

2.3. Метрики відстані об’єктів у багатовимірному просторі

Головним поняттям багатовимірної класифікації є поняття відстані між об’єктами у просторі ознак. Воно сприяє кількісному вимірюванню близькості чи віддаленості об’єктів і встановленню однорідності чи неоднорідності досліджуваної сукупності. Розрахунки відстаней покладено в алгоритми кластерного, дискримінантного, таксономічного аналіза.

Подібність є поняттям протилежним до категорії відстані: чим далі видалені об'єкти у багатовимірному просторі ознак, тим менш вони подібні.

Визначимо що є метрикою відстані.

Для двох об’єктів ***p*** і ***s*** з сукупності ***n*** об’єктів невід’ємна речова функція $d(z_p, z_s)$ називається функцією ***відстані*** чи ***метрикою***, якщо:

1) функція невід’ємна $d(z_p, z_s) \geq 0$ для всіх z_p та z_s ;
2) відстань від об’єкта до самого себе дорівнює нулю
   $d(z_p, z_s) = 0$ при $z_p = z_s$;
3) точка відліку відстані між двома об’єктами інваріантна, тобто
   $d(z_p, z_s) = d(z_s, z_p)$;
4) спостерігається нерівність трикутника, тобто сума двох сторін трикутника завжди більша його третьої сторони
   $d(z_p, z_s) \leq d(z_h, z_s) + d(z_h, z_p)$, де z_h – довільний третій об’єкт. Зауважимо, що четверта умова не є жорсткою, тобто майже не враховується у розрахунках.

Розглянемо основні відстані, що використовуються у багатовимірному аналізі:

1. Лінійна відстань (City-blok (Мангетенська)):

$$d_1(z_p, z_s) = \sum_{k=1}^{m} |z_{pk} - z_{sk}|. \qquad (2.9)$$

Означену відстань доцільно використовувати, коли існує очікування великих скупчень точок (багатомірних об'єктів) у площині. Особливо, якщо ці скупчення розташовані на гіперплощині, ортогональні координатним осям.

2. Евклідова відстань:

$$d_2(z_p, z_s) = \sqrt{\sum_{k=1}^{m} (z_{pk} - z_{sk})^2}. \qquad (2.10)$$

Ця відстань найбільш поширена, оскільки відповідає уявленню про близькість об’єктів у трьохмірному просторі та тісно пов’язана з дисперсійними статистичним критеріями, наприклад, критерієм Фішера.

3. «Зважена» евклідова відстань:

$$d_3(z_p, z_s) = \sqrt{\sum_{k=1}^{m} (z_{pk} - z_{sk})^2 \cdot f_k}, \quad (2.11)$$

де $\mathbf{f_k}$ – статистична вага k-тої ознаки. При чому $0 \leq \mathbf{f_k} \leq 1$; $\sum_{k=1}^{m} f_k = 1$

Зважена евклідова відстань використовується, коли ознаки, що характеризують об'єкти досліджуваної сукупності нерівновагоми. Важливість, ієрархію змінних можна визначити на підставі досвіду аналогічних досліджень чи експертних опитувань спеціалістів. Статистичні ваги додаються до матриці вихідних даних ***X***. Кожній стовпчик множиться на вагу $\mathbf{f_k}$, а потім виконується стандартна процедура перетворень матриці, яку вже розглянули.

4. Відстань Махаланобіса:

$$d_M(z_p, z_s) = (z_p - z_s) r^{-1} (z_p - z_s)^T \quad (2.12)$$

де $\boldsymbol{r^{-1}}$ – матриця, звернена до кореляційної матриці; знак $^{\mathbf{T}}$ – знак транспонування матриці.

Метрика Махаланобіса є особливою, оскільки включає не тільки вектори точок, а й матрицю коефіцієнтів парної кореляції. Відстань Махаланобіса є більш загальною, порівняно з евклідовою.

Знайдені відстані між об'єктами стандартизованого простору ознак видаються у вигляді матриці відстаней:

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ d_{21} & 0 & d_{23} & \dots & d_{2n} \\ & & \dots & & \\ d_{n1} & d_{n2} & d_{n3} & \dots & 0 \end{bmatrix}.$$

У наведеній матриці підрядкові індекси позначають номери об'єктів, які змінюються ввід **1** до ***n***. Тому матриця ***D*** має розмір ***n***х***n***. Матриця ***D***відповідно визначенню функції відстані невід'ємна (умова 1), на головній діагоналі знаходяться нулі (умова 2), симетрична (умова 3). Розглянута матриця відстаней є однією з головних матриць кластерного, таксономічного, дискримінантного аналізу.

За даними нашого прикладу розрахуємо матрицю лінійних відстаней. Для цього будемо почергово розраховувати відстані між кожними двома

об’єктами. Спочатку визначимо лінійну відстань між точками z_A і z_B.Вектори-рядки стандартизованих координат підприємств A, B та C мають вигляд:

$$z_A = [1{,}405 \quad 1{,}100] \qquad z_B = [-0{,}562 \quad 0{,}220] \qquad z_C = [-0{,}843 \quad -1.320]$$

Звідси слід записати розрахунок лінійної відстані таким чином (за формулою 2.9):

$$d_1(z_A, z_B) = |1{,}405 - (-0{,}562)| + |1{,}100 - 0{,}220| = 2{,}847$$

$$d_1(z_A, z_C) = |1{,}405 - (-0{,}843)| + |1{,}100 - (-1{,}320)| = 4{,}668$$

$$d_1(z_B, z_c) = |-0{,}562 - (-0{,}843)| + |0{,}220 - (-1{,}320)| = 1{,}821$$

Матриця лінійних відстаней матиме такий вигляд:

$$D_1 = \begin{bmatrix} 0 & 2{,}847 & 4{,}668 \\ 2{,}847 & 0 & 1{,}821 \\ 4{,}668 & 1{,}821 & 0 \end{bmatrix}.$$

Як бачимо, найближче між собою розташовані підприємства B та C.

Розрахуємо за вихідними даними евклідову відстань (за формулою 2.10):

$$d_2(z_A, z_B) = \sqrt{(1{,}405 - (-0{,}562))^2 + (1{,}100 - 0{,}220)^2} = 2{,}155$$

$$d_2(z_A, z_C) = \sqrt{(1{,}405 - (-0{,}843))^2 + (1{,}100 - (-1{,}320))^2} = 3{,}303$$

$$d_2(z_B, z_C) = \sqrt{(-0{,}562 - (-{,}843))^2 + (0{,}220 - (-1{,}320))^2} = 1{,}565$$

На підставі розрахунків складемо матрицю евклідових відстаней:

$$D_2 = \begin{bmatrix} 0 & 2{,}155 & 3{,}303 \\ 2{,}155 & 0 & 1{,}565 \\ 3{,}303 & 1{,}565 & 0 \end{bmatrix}.$$

Як бачимо, із зростанням ступеня метрики відстань між точками зростає:

$$d_1(z_p, z_s) \geq d_2(z_p\,, z_s).$$

Інакше кажучи, лінійна відстань між точками z_A і z_B завжди більша, ніж евклідова відстань $d_1(z_A, z_B) \geq d_2(z_A\,, z_B)$, тобто $2{,}847 > 2{,}155$.

Як й у попередніх розрахунках, найближче між собою розташовані підприємства B та C.

2.4. Метрики подібності об’єктів у багатовимірному просторі

Як уже велось, поняття близькості, подібності у багатовимірному просторі ознак протилежне поняттю відстані.

Метрикою подібності вважається невід’ємна речова функція, обмежена зверху та знизу, яка зростає при зниженні відстані між об’єктами і навпаки.

У багатовимірному аналізі існує простий перехід від метрик відстаней до метрик подібності:

$$\mu_{ps} = \frac{1}{1+d(z_p,z_s)}, \qquad (2.13)$$

де $\mathrm{d}(\mathrm{z_p}, \mathrm{z_s})$ – деяка функція відстані (лінійна чи евклідова).

Метрика подібності змінюється в межах від нуля до одиниці. При $d(z_p, z_s) \to \infty \;\; \mu_{ps} \to 0$, а при $\; d(z_p, z_s) \to 0 \quad \mu_{ps} \to 1$. Інакше кажучи, об’єкти повністю різні, оскільки мають нульову подібність при великій відстані між ними, та повну подібність, якщо $z_p = z_s$. Метрики подібності подаються у вигляді симетричної матриці подібності:

$$\mu = \begin{bmatrix} 1 & \mu_{12} & \mu_{13} & \dots & \mu_{1n} \\ \mu_{21} & 1 & \mu_{23} & \dots & \mu_{2n} \\ & & \dots & & \\ \mu_{n1} & \mu_{n2} & \mu_{n3} & \dots & 1 \end{bmatrix}.$$

Матриця **μ** схожа на матрицю коефіцієнтів парної кореляції ***r***, які дозволяють виявити близькість між ознаками. Але потрібно розуміти, що коефіцієнт парної кореляції $\mathbf{r_{PS}}$ не є класичною метрикою подібності, оскільки знаходиться в межах від -1 до 1 та не відповідає умові $\mathbf{0 \le \mu_{PS} \le 1}$.

Зауважимо, що матриці подібності використовуються в алгоритмах оцінювання латентних показників, а також в кластерному аналізі.

За даними нашого прикладу розрахуємо матриці подібності між трьома підприємствами A, B, C. Для цього знайдемо метрики подібності між кожними двома об’єктами, тобто між A та B, A та C, B та C (за формулою 2.13) :

$$\mu_{\mathrm{A,B}} = \frac{1}{1 + d(z_{\mathrm{A}}, z_{\mathrm{B}})} = \frac{1}{1 + 2{,}847} = 0{,}260$$

$$\mu_{A,C} = \frac{1}{1 + d(z_A, z_C)} = \frac{1}{1 + 4{,}668} = 0{,}176$$

$$\mu_{B,C} = \frac{1}{1 + d(z_B, z_C)} = \frac{1}{1 + 1{,}821} = 0{,}354$$

Матриця подібності за отриманими даними на підставі лінійної відстані дорівнює:

$$\mu_1 = \begin{bmatrix} 1 & 0{,}260 & 0{,}176 \\ 0{,}260 & 1 & 0{,}354 \\ 0{,}176 & 0{,}354 & 1 \end{bmatrix}.$$

Аналогічним чином будується матриця подібності на підставі матриці евклідових відстаней:

$$\mu_2 = \begin{bmatrix} 1 & 0{,}317 & 0{,}232 \\ 0{,}317 & 1 & 0{,}390 \\ 0{,}232 & 0{,}390 & 1 \end{bmatrix}.$$

Як бачимо з обох матриць підприємства В та С більш подібні між собою.

Так як, поняття подібності протилежне поняттю відстані, то для отриманих матриць буде справедливим таке співвідношення:

$$\mu_\infty \geq \cdots \geq \mu_2 \geq \mu_1 .$$

Питання для самоконтролю

1. Опишіть сучасні підходи до методів багатовимірної класифікації.
2. Чому у багатовимірному аналізі вихідна інформація подається у матричному вигляді? Як формується матриця вихідних даних?
3. Які статистичні характеристики розраховуються на підставі матриці вихідних даних? Наведіть формули їх розрахунку.
4. Що таке центроїд?
5. Що таке стандартизація вихідних даних?
6. Охарактеризуйте поняття відстані.
7. Які метрики відстані ви знаєте?
8. Опишіть властивості метрики відстані.
9. Охарактеризуйте метрики подібності?
10. Які властивості мають метрики подібності?

Тема 3. ТАКСОНОМІЧНА ОЦІНКА ЛАТЕНТНИХ ПОКАЗНИКІВ

3.1. Сутність латентних показників.

3.2. Основні категорії таксономічного аналізу.

3.3. Алгоритми оцінки латентних показників:

а) класичний алгоритм оцінки;

б) модифікований алгоритм оцінки.

3.1. Сутність латентних показників

Термін «латентний» у перекладі з латинської мови означає прихований, недоступний. Цей науковий термін використовується для характеристики складних атрибутивних понять і оцінок, які неможливо кількісно виміряти у метричній шкалі.

Отже, такі економічні поняття, як «фінансовий стан», «інвестиційна привабливість» підприємства, «конкурентоспроможність» продукції є латентними в тому сенсі, що вони принципово не можуть бути виміряними єдиними економічним показником. Вони мають ознаки складних, прихованих властивостей виробничої системи, якою є підприємства, й виявляються на поверхні у вигляді сукупності чинників-симптомів – окремих фінансових, технічних, економічних показників та коефіцієнтів. Завдяки різноманітних обставин неможливо надати чітку, однозначну кількісну характеристику наведених показників, тому про їх рівень говорять побічно, за допомогою градацій порядкової шкали типу «краще-гірше», «більш-менш», «легше-важче».

Латентні ознаки виявляються на поверхні економічних явищ у вигляді множини чинників-симптомів, що відбувають різноманітні сторони складних атрибутивних ознак. Саме симптоми, тобто звичайні фінансово-економічні показники метричної шкали дозволяють дати уявлення про латентну ознаку, що вивчається. У таких ситуаціях є природною спроба сконцентрувати інформацію, виражаючи велику кількість чинників-симптомів за допомогою одного чи кількох більш ємних інтегральних оцінок. Справді, будь-який зацікавлений фінансовим станом підприємства суб’єкт (керівник, інвестор, аудитор) не

задовольняється простою кількісною оцінкою коефіцієнтів. Йому важливо знати, чи прийняті набутті значення, чи добрі вони і якою мірою.

Крім того, він прагне встановити логічний зв'язок кількісних значень чинників-симптомів виокремленої групи з певним комплексним показником, що характеризує фінансовий стан підприємства в цілому. Завдання ускладняється ще й тим, що фінансових коефіцієнтів мало, вони змінюються часто і різноспрямовано, тож виникає потреба об’єднати набір усіх досліджуваних окремих симптомів в один комплексний, синтетичний показник, за значенням якого можна визначити фінансовий стан підприємства, порівняти його з конкурентами.

У деяких ситуаціях виникає потреба кількісного порівняння (ранжування, групування) різних об’єктів за величиною латентного показника. Наприклад, потенційні інвестори прагнуть знайти більш привабливий об’єкт, покупці зацікавлені співвідношенням «ціна-якість» та таке інше.

Інакше кажучи, необхідно кількісно оцінити величину латентного показника у кожного з $\boldsymbol{n}$ досліджуваних об’єктів на основі $\boldsymbol{m}$ ознак-симптомів, інформація про які подана у вигляді матриці вихідних даних $\boldsymbol{X}$ чи вже стандартизованих даних $\boldsymbol{Z}$.

Отже, латентні економічні ознаки в буквальному сенсі оточують сучасного дослідника у сфері аналізу, планування та прогнозування розвитку виробничої системи. Сучасна наука має в розпорядженні досить ефективні прийоми та методи кількісної оцінки прихованих ознак. Усі вони поєднанні в багатовимірні статистичні методи і моделі.

Моделі з латентними ознаками розуміють як сукупність статистичних моделей, що конструюються за допомогою математичних методів і пояснюють дані, які спостерігаються, їх залежністю від прихованих характеристик об’єктів.

Головною метою застосування моделей з латентними ознаками є їх кількісна оцінка. При цьому потрібно розробити таку інтегральну модель, яка має кількісну загальну характеристику та сприяє обранню найперспективнішого

об’єкта інвестування, купівлі, співпраці тощо. Для вирішення такого завдання існує три основні підходи:

- кластерний, кореляційно-регресійний і дискримінантний аналіз як теоретична база моделей латентних ознак (наприклад, тестів експрес-діагностики для визначення вірогідності банкрутства підприємств);
- таксономічний аналіз об’єктів;
- методи факторизації латентних економічних ознак.

3.2. Основні категорії таксономічного аналізу

Одним з фундаментальних підходів до оцінки латентних показників є таксономічний аналіз. Його назва походить від двох грецьких слів: *таксис* (що означає розташування, порядок) і *номос* (закон, правило). Тобто таксономія – це наука про правила впорядкування і ранжування багатовимірних об’єктів. Вона ґрунтується на використанні таких понять, як стимулятор, дестимулятор, еталон, антиеталон. Розглянемо їх.

Стимулятор – чинник-симптом, зростання якого бажане з погляду латентного показника, що оцінюється. Наприклад, стимуляторами у вивченні фінансового стану об’єктів є такі ознаки господарської діяльності, як прибуток, рентабельність, дебіторська заборгованість тощо.

Дестимулятор – чинник-симптом, зростання якого не бажане з погляду латентного показника, що оцінюється. Наприклад, дестимуляторами у вивченні фінансового стану об’єктів є такі ознаки господарської діяльності, як собівартість продукції, фінансові втрати (штрафи, пені), кредиторська заборгованість тощо.

Номинатор – чинник-симптом номінальної шкали, який характеризує атрибутивні ознаки об’єктів, що досліджуються: назви підприємств, економічне призначення продукції, галузева належність тощо.

Еталон – реальна або умовна точка багатовимірного простору, чинники-симптоми якої відповідають уявленням економічної науки про найкращий рівень стимуляторів і дестимуляторів з позиції латентного показника, що

досліджується. Наприклад, як еталон для оцінки фінансового стану підприємства можна розглядати об’єкт, у котрого прибуток, рентабельність, дебіторська заборгованість та інші стимулятори – максимальні. А собівартість продукції, фінансові втрати (штрафи, пені), кредиторська заборгованість та інші дестимулятори – мінімальні на сукупності всіх підприємств.

Антиеталон – поняття, протилежне поняттю еталона, тобто реальна або умовна точка багатовимірного простору, чинники-симптоми якої відповідають уявленням економічної наук про найгірший рівень стимуляторів і дестимуляторів з позиції латентного показника, що досліджується.

Якщо еталон – це точка верхнього полюса, до якої мають прямувати всі об’єкти сукупності для досягнення максимального рівня латентного показника, то антиеталон, навпаки – точка нижнього полюса. Від неї треба триматися далі, щоб бути в лідерах.

Тоді цілком доцільно видається наступна ідея, що лежить в основі таксономічного аналізу. Метрика схожості об’єкта з еталоном чи метрика відстані кожного об’єкта від антиеталона в просторі чинників-симптомів може розглядатися як певна інтегральна (синтетична) оцінка рівня латентного показника.

Дійсно, чим ближче даний об’єкт до еталона, тим вищим є рівень прихованого латентного показника, і навпаки. Справедливим також є твердження про те, що чим далі перебуває об’єкт від антиеталона. Тим він розвинутіший щодо латентного показника і , навпаки.

Розглянемо такий умовний приклад. Нехай відомі дані про величину прибутку ($\mathbf{z}_1$) і собівартість продукції ($\mathbf{z}_2$) на чотирьох підприємствах, які потрібно дослідити на інвестиційну привабливість (латентний показник). Як антиеталон (АЕ) виберемо точку, координати якої відповідають найгіршим значенням для даної сукупності підприємств (рис.3.1). Відстані від кожної точки (підприємства) до антиеталона будуть характеризувати рівень інвестиційної привабливості об’єктів, що вивчаються.

На рис. 3.1 бачимо, що найбільш привабливим у інвестиційному плані є підприємство **A_2**, оскільки воно розташовано далі за всіх від антиеталона (має досить високий прибуток і низьку рентабельність). Це підприємство є лідером у даній сукупності об'єктів з точки зору вивчаємого латентного показника. Підприємство **A_3,** навпаки, потрібно віднести до аутсайдерів, оскілки його чинники-симптоми найближчі до антиеталона АЕ. Більш точна кількісна характеристика досліджуваної привабливості можлива за допомогою метрик відстаней.

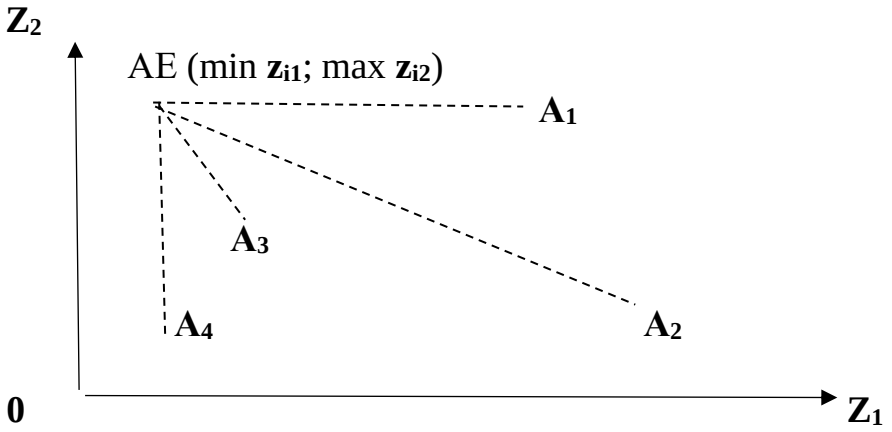

Рис. 3.1 Відстань між точками і антиеталоном у двомірному простору ознак

3.3. Алгоритми оцінки латентних показників

Відстань до антиеталона чи подібність з еталоном є синтетичними величинами, які складаються зі значень усіх чинників-симптомів. Процес (алгоритм) їх побудови передбачає проведення низки послідовних етапів. У багатовимірному аналізі існує два підходи до алгоритму таксономічного аналізу: класичний та модифікований. Розглянемо їх детальніше.

Класичний алгоритм базується на виборі еталону як точки верхнього полюса та визначенні відстані та метрики подібності від об'єкта до обраного еталону. Отримане значення метрики подібності розглядається як інтегральна синтетична величина шуканого латентного показника.

Етапи класичного алгоритму таксономічної оцінки латентного показника подано в табл. 3.1.

Таблиця 3.1

Етапи таксономічного аналізу (класичний варіант)

I етап	Відбір чинників-симптомів латентного показника та формування матриці вихідних даних $\boldsymbol{X}$
II етап	Розподіл чинників-симптомів на стимулятори і дестимулятори
III етап	Визначення статистичної ваги $\mathbf{f_k}$ для кожного чинника-симптома
IV етап	Стандартизація чинників-симптомів і побудова матриці $\boldsymbol{Z}$
V етап	Завдання еталону
VI етап	Вибір метрики відстані
VII етап	Розрахунок відстані $\boldsymbol{d_i}$ між усіма точками (об'єктами) і еталоном
VIII етап	Розрахунок мір подібності $\boldsymbol{\mu_i}$ кожного об'єкта з еталоном

Процес побудови розрахунку таксономічного показника починається з формування матриці вихідних даних $\boldsymbol{X}$. Для більшої точності та надійності проведення таксономічного аналізу необхідно ретельно відбирати змінні, які суттєво характеризують латентний показник.

Далі відбувається розподіл змінних на стимулятори і дестимулятори з точки зору латентного показника. Як уже велось, ознаки, зростання яких бажане, які позитивно впливають на досліджуваний об'єкт, потрібно віднести до стимуляторів. На відміну від них, змінні, які негативно впливають на рівень розвитку латентного показника належать до дестимуляторів. Номінатори до матриці вихідних даних $\boldsymbol{X}$ не входять.

На наступному, третьому етапі доцільно диференціювати змінні за їх роллю у розвитку латентного показника. З цією метою встановлюють коефіцієнти ієрархії або статистичні ваги кожної ознаки $\mathbf{f_k}$, які забезпечують посилення значущості окремих чинників-симптомів і послаблюють інші відповідно до їх впливу на кінцевий рівень таксономічного показника. Статистичні ваги визначаються методом експертного опитування, факторного аналізу. Необхідно пам'ятати, що сума ваг дорівнює одиниці та має знаходитися

в межах від 0 до 1. Якщо апріорна інформація про ієрархію ознак відсутня, то всі вони вважаються рівновагомими, тобто дорівнюють одиниці.

Щоб уникнути потворний вплив одиниць вимірювання на рівень таксономічного розвитку, виконується четвертий етап – стандартизація цифрової інформації:

$$\boldsymbol{z_{ik}} = \frac{(\boldsymbol{x_{ik}} - \bar{\boldsymbol{x}}_{\boldsymbol{k}})}{\boldsymbol{\sigma_k}}. \qquad (3.1)$$

На п’ятому етапі класичного таксономічного аналізу передбачено визначення еталону розвитку. Еталоном вважається реальна чи умовна точка багатовимірного простору, координати якої характеризують найкращі досягнення досліджуваних об’єктів (з урахуванням розподілу на стимулятори і дестимулятори). Еталон відбиває максимально можливий, потенційний рівень латентного показника і є орієнтиром (базою порівняння) для всіх точок досліджуваної сукупності.

Існує два підходи до завдання еталона:

1) на основі значень чинників-симптомів даної сукупності об’єктів;
2) на основі значень чинників-симптомів інших сукупностей, наприклад враховуючи досвід інших регіонів, галузей світу тощо.

У першій спосіб для стимуляторів еталонні значення визначаються таким чином:

$$\boldsymbol{z_{0k}} = \boldsymbol{max_i z_{ik}}, \qquad (3.2)$$

а для дестимуляторів

$$\boldsymbol{z_{0k}} = \boldsymbol{min_i z_{ik}} \quad \text{чи} \quad \boldsymbol{z_{0k}} = 0. \qquad (3.3)$$

У другому випадку як еталон обираються світові досягнення з точки зору досліджуваних ознак.

Далі, на шостому етапі розраховуються відстані між усіма об’єктами та еталоном. При цьому потрібно використовувати лінійні відстані:

$$d_1 = \sum_{k=1}^{m} |z_{ik} - z_{0k}|\,, \qquad (3.4)$$

або евклідові відстані:

$$d_2 = \sqrt{\sum_{k=1}^{m} (z_{ik} - z_{0k})^2}\ . \qquad (3.5)$$

Метрику Махаланобіса не рекомендується використовувати із-за мультиколінеарності змінних.

На сьомому етапі знайдені відстані $\boldsymbol{d_i}$ розглядаються як інтегральні резерви підвищення латентного показника кожного об'єкта. Для еталонної точки він дорівнює нулю.

На останньому етапі визначають міри подібності кожного об'єкта з еталоном

$$\boldsymbol{\mu_i = \frac{1}{1+d_i}}. \qquad (3.6)$$

Отримані значення знаходяться у межах від 0 до 1 та трактуються так: чим більше значення подібності, тим вищий рівень латентного показника.

Розглянемо модифікований алгоритм таксономічної оцінки латентного показника. Етапи цього алгоритму подані у табл. 3.2

Таблиця 3.2

Етапи таксономічного аналізу (модифікований варіант)

I етап	Відбір чинників-симптомів латентного показника та формування матриці вихідних даних $\boldsymbol{X}$
II етап	Розподіл чинників-симптомів на стимулятори і дестимулятори
III етап	Визначення статистичної ваги $\mathbf{f_k}$ для кожного чинника-симптома
IV етап	Перетворення чинників-симптомів дестимуляторів у стимулятори
V етап	Стандартизація чинників-симптомів і побудова матриці $\boldsymbol{Z}$
VI етап	Завдання антиеталону
VII етап	Вибір метрики відстані
VIII етап	Розрахунок відстані $\boldsymbol{d_i}$ між усіма точками (об'єктами) і антиеталоном
IX етап	Розрахунок нормованих відстаней $\boldsymbol{d_i}$ * від усіх точок (об'єктів) до антиеталона

Як бачимо, перші три етапи класичного та модифікованого алгоритмів таксономічного аналізу співпадають. Розглянемо. як передбачається завдання

антиеталону у модифікованому варіанті. Антиеталоном вважається точка нижнього полюса $\mathbf{z_0}$, яка має такі координати

$$\mathbf{z_0}(a, a, \ldots, a), \tag{3.7}$$

де a – довільне від'ємне число, що задовольняє умову

$$a \leq min_{i,k} z_{ik}. \tag{3.8}$$

Зазвичай число a приймає значення -2, -3, -4. Вихід за ці межі практично неможливий, оскільки усі стандартизовані дані знаходяться в мажах від -3 до 3.

Нижній полюс числа a є начальною точкою відліку рівня латентного показника у просторі стандартизованих чинників-симптомів, які є стимуляторами. Тому відстань від об'єкта до антиеталона у цьому просторі можна розглядати як безпосередню кількісну оцінку шуканого синтетичного показника, яка не потребує переходу до міри подібності.

Саме тому на початку таксономічного аналізу всі чинники-симптоми, які є дестимуляторами, необхідно перетворити в стимулятори таким чином:

$$x_{ik} = \frac{1}{x'_{ik}} \tag{3.9}$$

або

$$x_{ik} = 1 - x'_{ik} \tag{3.10}$$

x'_{ik} - вихідна реалізація чинника-симптома дестимулятора.

Щоб значення оцінки латентного показника знаходилися у межах від 0 до 1, необхідно провести нормування за формулою:

$$d_i^* = \frac{d_i}{-2a\sqrt{m}}. \tag{3.11}$$

Величина d_i^* інтерпретується так: високі значення d_i^* свідчать про високий рівень шуканого латентного показника і навпаки.

При порівнянні класичного алгоритму таксономічного аналізу з модифікованим, можна побачити, що вони відрізняються тільки етапами, які по'вязані з перетворенням чинників-симптомів дестимуляторів у стимулятори, завданням антиеталону та нормуванням відстаней від точки нижнього полюсу. Відмітимо, що вказані відмінності можуть призвести до різних значень оцінок латентної ознаки, що вивчається, для одних й тих самих емпіричних даних.

Розглянемо прикладні аспекти отриманих таксономічних оцінок $\boldsymbol{\mu_i}$ і $\boldsymbol{d_i^*}$.

По-перше, вони можуть стати основою ранжування досліджуваних об'єктів за величиною латентної ознаки. Це відбувається способом надання кожній точці рангів 1, 2, 3, …, $\boldsymbol{n}$. Такий спосіб сприяє обранню найкращого об'єкта інвестування, партнерів тощо.

По-друге, на їх основі можливе багатовимірне групування з виділенням груп аутсайдерів, лідерів. Такий спосіб дозволяє виділити конкурентів, партнерів тощо, а також провести статистичне моделювання та прогнозування.

Розглянемо приклад (дані умовні).

Визначимо таксономічний показник економічного розвитку країн за такими ознаками:

x_1 — ВВП на душу населення, тис. дол. США;

x_2 — зовнішній борг, % до ВВП;

x_3 — ступінь самозабезпеченості енергоресурсами, %.

У табл. 3.3 наведено абсолютні значення вихідних ознак.

Таблиця 3.3

Вихідні дані

Країна, $\boldsymbol{i}$	Первинні значення ознак, x_{ik}		
	$\boldsymbol{x_{1i}}$, тис. дол. США	$\boldsymbol{x_{2i}}$, %	$\boldsymbol{x_{3i}}$, %
1	5,8	14	48
2	4,7	22	55
3	3,9	28	29
У середньому на 1 країну	4,8	21,3	44

Як бачимо з умови , вихідна матриця $\boldsymbol{X}$ має розмір 3х3. Тобто перший етап подано первинною інформацією.

Далі переходимо на другий етап: розподіляємо чинники-симптоми на стимулятори і дестимулятори. Це необхідно для правильного завдання еталону розвитку. Якісний економічний аналіз дозволяє стверджувати, що $\boldsymbol{x_1}$ – ВВП на душу населення і $\boldsymbol{x_3}$ – ступінь самозабезпеченості енергоресурсами позитивно впливає на економічну міцніть країни. Це чинники-стимулятори, зростання котрих бажане. У свою чергу, $\boldsymbol{x_2}$ – зовнішній борг характеризує витратну частину

бюджету, зростання цього показника не бажане, тому його віднесемо до чинників-дестимуляторів (табл. 3.4.)

Таблиця 3.4

Класифікація вихідних ознак

Показник	Характер впливу	Група чинників
1.ВВП на душу населення	позитивний	стимулятор
2. Зовнішній борг	негативний	дестимулятор
3.Ступінь самозабезпеченості енергоресурсами	позитивний	стимулятор

Етап визначення статистичних ваг пропускаємо, оскільки апріорна інформація про важливість окремих показників відсутня. Тому вважаємо, що всі три показники рівнозначні, тобто $\mathbf{f_1=f_2=f_3=1}$.

Далі виконаємо стандартизацію чинників-симптомів. Для цього знайдемо середні значення та середньоквадратичні відхилення за кожною ознакою.

Спочатку розрахуємо середні значення кожної ознаки в цілому по трьох країнах за формулою простої арифметичної:

$$\overline{X_k} = \frac{\sum X_{ik}}{n}$$

Для першої ознаки

$$\overline{X_1} = \frac{\sum X_{ik}}{n} = \frac{14{,}4}{3} = 4{,}8 \text{ тис. дол. США}$$

Аналогічно для інших ознак (табл.3.3)

Далі розрахуємо середньоквадратичне відхилення для кожної ознаки за такою формулою

$$\sigma_k = \sqrt{\frac{\sum(X_{ik}-\bar{X}_{ik})^2}{n}}.$$

$$\sigma_1 = \sqrt{\frac{(5{,}8-4{,}8)^2+(4{,}7-4{,}8)^2+(3{,}9-4{,}8)^2}{3}} = 0{,}779 \text{ тис. дол. США}$$

$$\sigma_2 = \sqrt{\frac{(14-21{,}3)^2+(22-21{,}3)^2+(28-21{,}3)^2}{3}} = 5{,}735\ \%$$

$$\sigma_3 = \sqrt{\frac{(48-44)^2+(55-44)^2+(29-44)^2}{3}} = 10{,}985\ \%$$

Абстрагуємося від масштабів вимірювання чинників-симптомів та проведемо стандартизацію даних за допомогою формули (3.1).

Для першої ознаки:

$$z_{11} = \frac{5{,}8 - 4{,}8}{0{,}779} = 1{,}2837$$

$$z_{21} = \frac{4{,}7 - 4{,}8}{0{,}779} = -0{,}1284$$

$$z_{31} = \frac{3{,}9 - 4{,}8}{0{,}779} = -1{,}1553$$

Аналогічно для другої та третьої ознак.

Матриця стандартизованих даних буде матиме вигляд

$$Z = \begin{bmatrix} 1{,}2837 & -1{,}2729 & 0{,}3641 \\ -0{,}1284 & 0{,}1221 & 1{,}0014 \\ -1{,}1553 & 1{,}1683 & -1{,}3655 \end{bmatrix}.$$

Нагадаємо, що завдання еталона у класичному алгоритмі оцінки латентного показника виконується як визначення точки верхнього полюсу, тобто найбільш розвинутого з урахуванням розподілу на стимулятори і дистимулятори реального чи умовного об’єкта.

У даному прикладі еталон будемо задавати за максимальними та мінімальними рівнями чинників-симптомів. Для стимуляторів використовуємо формулу (3.2), а для дестимуляторів – формулу (3.3). Тому відповідно до матриці стандартизованих даних еталонними будуть максимальні значення першого и третього стовпчиків (стимулятори), а у другому – мінімальне значення (дестимулятор). Це означає, що еталонна точка матиме координати:

z_0 (1,2837, -1,2729, 1,0014).

Ця точка є умовною і виражає у стандартизованому вигляді найкращі результати країни у економічній міцності, розвитку.

Переходимо до вибору метрики відстані. Однозначної відповіді, яка метрика краще не існує. Аналітик керується власним досвідом.

Загальноприйняте використання евклідової відстані. У нашому прикладі розрахунки робляться за формулою евклідової відстані (3.5):

Наприклад, відстань між 1 країною та еталоном дорівнює

$d_{10} = \sqrt{(1{,}2837 - 1{,}2837)^2 + (-1{,}2729 - (-1{,}2729))^2 + (0{,}3641 - 1{,}0014)^2}$=0,6373.

Результати розрахунків подано у табл. 3.5.

Таблиця 3.5

Розрахунок евклідових відстаней

Країна, i	Стандартизовані значення ознак, z_{ik}			Евклідова відстань			
	Z_{i1}	Z_{i2}	Z_{i3}	$(z_{i1} - z_{01})^2$	$(z_{i2} - z_{02})^2$	$(z_{i3} - z_{03})^2$	d_i
1	1,2837	-1,2729	0,3641	0	0	0,4062	0,6373
2	-0,1284	0,1221	1,0014	1,9940	1,9460	0	1,9850
3	-1,1553	1,1683	-1,3655	5,9487	5,9595	5,6022	4,1845
Z_0	1,2837	-1,2729	1,0014	x	x	x	x

На останньому етапі знаходимо міри подібності для кожного об’єкта з еталоном за формулою (3.6). Результати розрахунку мір подібності, а також ранги, які надані країнам на основі отриманих значень подані в таблиці 3.6.

Таблиця 3.6

Таксономічні рівні розвитку країн за класичним алгоритмом

Країна	d_i	μ_i	Ранг
1	0,6373	0,6108	1
2	1,9850	0,3350	2
3	4,1845	0,1929	3

Таким чином, для трьох країн за допомогою таксономічного аналізу знайдено латентний показник «економічна міцність», за яким проаранжовано країни. Так, лідером є перша країна, яка за своїми показниками найбільш близько знаходиться до еталону ($d_{min} = d_1 = 0.6373$) і має максимальну подібність з

еталоном ($\mu_{max} = \mu_1 = 0.6408$). На другому місці розташована 2 країна, а на третьому – третя (табл.6).

Розглянемо модифікований варіант алгоритму таксономічного показника на нашому прикладі. Оскільки перші три етапи співпадають, то почнемо з четвертого етапу. Проведемо перетворення чинників-симптомів дестимуляторів у стимулятори. У нашому прикладі дестимуляторам була друга ознака x_2 – зовнішній борг.

У інформаційній базі відсутні нульові дані, можна скористатися такою формулою (3.9). Результати перетворення подані в таблиці 3.7.

Таблиця 3.7

Перетворення дестимуляторів у стимулятори

Країна	X_{i2}', %	X_{i2}
1	14	0,071
2	22	0,045
3	28	0,036

Таким чином матриця вихідних даних матиме вигляд:

$$X = \begin{bmatrix} 5,8 & 0,071 & 48 \\ 4,7 & 0,045 & 55 \\ 3,9 & 0,036 & 29 \end{bmatrix}.$$

Виконаємо перехід до стандартизованої матриці як і у класичному варіанті, за умовою, що $\bar{x}_2 = 0,051; \sigma_2 = 0,015$. На підставі розрахунків за формулою (3.1) отримаємо такі результати:

$$Z = \begin{bmatrix} 1,2837 & 1,3333 & 0,3641 \\ -\ 0,1284 & -0,4000 & 1,0014 \\ -1,1553 & -1,0000 & -1,3655 \end{bmatrix}.$$

Знаходимо антиеталон, тобто точку нижнього полюсу, яка характеризує найнижчий, начальний рівень латентного показника (3.7). Для цього з матриці стандартизованих значень обираємо від'ємне число, яке відповідає умовам (3.8). Таким числом є значення третього фактора у третій країні (-1,3655). Це значення приймаємо за антиеталон: **z_0 (-1,3655; -1,3655; -1,3655)** .

Обираємо метрику відстані. Як й у попередньому випадку звернемося до евклідової відстані (3.5). Результати розрахунків подано у табл.3.8.

Таблиця 3. 8

Розрахунок евклідових відстаней

Країна, i	Стандартизовані значення ознак, z_{ik}			Евклідова відстань			
	Z_{i1}	Z_{i2}	Z_{i3}	$(z_{i1}-z_{01})^2$	$(z_{i2}-z_{02})^2$	$(z_{i3}-z_{03})^2$	d_i
1	1,2837	1,3333	0,3641	7,0183	7,2835	2,9915	4,1585
2	-0,1284	-0,4000	1,0014	1,5304	0,9322	5,6022	2,8399
3	-1,1553	-1,0000	-1,3655	0,0442	0,1336	0	0,4217
Z_0	-1,3655	-1,3655	-1,3655	x	x	x	x

Нормуємо отримані відстані та знайдемо ранги країн. Для цього використаємо формулу (3.11). Зауважимо, що знаменник даної формули – величина постійна, у нашому прикладі вона дорівнює:

$$-2a\sqrt{m} = -2 \cdot (-1{,}3655)\sqrt{3} = 4{,}730.$$

Результати розрахунків подано у табл. 3.9.

Таблиця 3.9

Таксономічні рівні розвитку країн за модифікованим алгоритмом

Країна	d_i	d_i^*	Ранг
1	4,1585	0,8792	1
2	2,8399	0,6004	2
3	0,4217	0,0892	3

Отже, знайдені значення можна розглядати як латентний показник економічної міцності країни.

Доцільно порівняти отримані результати (табл. 3.10).

Таблиця 3.10

Порівняння результатів таксономічної оцінки економічної міцності країни

Країна	Класичний алгоритм		Модифікований алгоритм	
	μ_i	Ранг	d_i^*	Ранг
1	0,6108	1	0,8792	1
2	0,3350	2	0,6004	2
3	0,1929	3	0,0892	3

Як бачимо, результати ідентичні. На практиці при великій кількості об’єктів і ознак результати не завжди збігаються. Це можна пояснити таким чином:

1) у класичному алгоритмі спочатку визначається відстань від об’єкта до еталона, а потім відбувається перехід від відстані до міри подібності. У модифікованому алгоритмі навпаки: спочатку всі чинники-симптоми перетворюються в стимулятори, і тільки потім розраховується відстань до антиеталону;

2) правила завдання еталону в класичному алгоритмі та антиеталону в модифікованому різні;

3) перетворення чинників-дестимуляторів у стимулятори змінює роль окремих змінних у формуванні рівня інтегрального синтетичного показника.

Саме тому, необхідно досить обережно інтерпретувати знайдені оцінки із-за великого числа суб’єктивних факторів, які вносять неоднозначність у результати подібних багатовимірних процедур.

Можна припустити, що ідентичні результати отримуються, якщо

1) усі чинники-симптоми належать до одного типу, чи стимулятори, чи дестимулятори;

2) еталон та антиеталон задається однаково.

В інших ситуаціях отримані оцінки латентного показника відрізняються.

Питання для самоконтролю

1. Що означає термін «латентний»?
2. Що таке моделі з латентними ознаками?
3. Яка головна мета моделей з латентними ознаками?
4. Які існують підходи до побудови моделей з латентними ознаками?
5. Дайте визначення таксономії.
6. Що таке стимулятор? Наведіть приклад.
7. Що таке дестимулятор? Наведіть приклад.
8. Що таке номінатор?

9. Що таке еталон?
10. Що таке антиеталон?
11. У чому полягає ідея таксономічного аналізу?
12. Які алгоритми оцінки латентних показників існують?
13. Охарактеризуйте етапи класичного алгоритму таксономічного аналізу?
14. Охарактеризуйте етапи модифікованого алгоритму таксономічного аналізу?
15. У чому схожість класичного та модифікованого алгоритму таксономічного аналізу?
16. Чим відрізняється класичний алгоритм оцінювання латентних показників від модифікованого?
17. З якою метою проводять стандартизацію вихідної інформації?
18. За якою формулою виконується стандартизація даних?
19. Які існують підходи до завдання еталону в класичному алгоритмі? Опишіть їх.
20. Як розраховується міра подібності кожного об'єкта з еталоном?
21. Як перетворюються дестимулятори у стимулятори?
22. Чому проводиться нормування оцінок латентного показника?
23. За якою формулою нормуються оцінки латентного показника?
24. Чому результати, отримані за класичним алгоритмом та модифікованим не збігаються?
25. За яких умов можна отримати ідентичні результати класичного та модифікованого алгоритмів?

Тема 4. КЛАСТЕРНИЙ АНАЛІЗ

4.1. Сутність та особливості застосування методів кластерного аналізу.

4.2. Формальна постановка задачі кластерного аналізу та основні його етапи.

4.3. Класифікація алгоритмів кластерного аналізу.

4.4. Ієрархічні агломеративні процедури.

4.5. Алгоритм пошуку згущень об'єктів (алгоритм «Форель»).

4.6. Функціонали якості класифікації.

4.1. Сутність та особливості застосування методів кластерного аналізу

Однорідність сукупності є базою будь-якого дослідження. Лише в однорідній сукупності виявлені закономірності є сталими і їх можна застосовувати до всіх одиниць сукупності. Поняття *однорідності* пов'язують з наявністю в усіх одиниць сукупності таких спільних властивостей і рис, які визначають їх одноякісність, належність до одного й того ж типу. Утворення однорідних сукупностей об'єктів базується на використанні методів структурного та типологічного групування. Типологічне групування дозволяє поділити сукупність на типи, класи, наприклад, підприємства країни за видом економічної діяльності.

Структурне групування сприяє розподілу однорідних сукупностей на групи, які близькі за кількісною ознакою. Наприклад, у рослинництві господарства можна поділити на великі, середні та фермерські господарства.

Зазвичай спочатку дослідження виконують типологічне групування, а потім – структурне. Структурне групування виконується як за однією ознакою, так й за декількома ознаками. У такому випадку завдання структурного групування ускладняється послідовним розподілом сукупності на групи за окремими ознаками, на підгрупи – за іншими ознаками і т. д. Тобто, виконується комбінаційне групування.

Комбінація групувальних ознак збільшує кількість груп, одиниці яких малочисельні чи відсутні. Ці недоліки можна подолати, якщо здійснити групування за допомогою кластерного аналізу.

Термін «кластерний аналіз» запропонував К. Тріон у 1939 р. Кластер у перекладі з англійської означає згущення, гроно. Синонімом кластерного аналізу є багатовимірне групування, розпізнавання образу без вчителя.

Головна мета кластерного аналізу – виділити у багатовимірному просторі такі однорідні групи, щоб об'єкти у групах були схожі між собою, а об'єкти з різних груп – не схожі. Під «схожістю» розуміють близькість об'єктів у багатовимірному просторі ознак. Тоді завдання групування полягає у пошуку в цьому просторі природних згущень (гроно) об'єктів, які вважаються однорідними групами – кластерами.

Отже, основними задачами кластерного аналізу є:

1) проведення класифікації об'єктів з урахуванням ознак, які відображають сутність, природу об'єктів. Це приводить до поглиблення знань про сукупність класифікованих об'єктів;
2) перевірка висунутих припущень про наявність певної структури в досліджуваній сукупності об'єктів, тобто пошук існуючої структури;
3) побудова нових класифікацій для слабовивчених явищ, коли необхідно встановити наявність зв'язків у середині сукупності та спробувати привнести до неї структуру.

Кластерний аналіз застосовується у різноманітних економічних дослідженнях. Наприклад, у маркетингу це сегментація конкурентів і споживачів. У менеджменті – розбиття персоналу на різні за рівнем мотивації групи, класифікація постачальників, виявлення схожих виробничих ситуацій, за яких виникає брак. У фінансовому аналізі – для класифікації досліджуваних підприємств за рівнем фінансового стану та ін.

4.2. Формальна постановка задачі кластерного аналізу та основні його етапи

У математико-статистичному плані задача кластерного аналізу може бути сформульована таким чином. Нехай сукупність, яка складається з **n** об'єктів, кожен з яких описується множиною з **m** ознак**,** подана матрицею стандартизованих даних **Z**, розміром **n**x**m.** Визначимо основну термінологію кластерного аналізу:

L – номер групи (**L=1,2,…, R)**;

$\bar{z} = 0$ **–** центр тяжіння вихідної сукупності об'єктів;

$\bar{z}_L$ - центр тяжіння L-ої групи об'єктів;

n_L - число об'єктів у групі L;

$U_L = \sum d_{Li}^2 (z_i, \bar{z}_L)$ – сума квадратів відстаней від усіх об'єктів L-тої групи до її центра тяжіння (розкид в L-ої групі);

$U = \sum_{L=1}^{R} U_L$ - внутрішньогруповий розкид;

$B = \sum_{L=1}^{R} n_L d^2(\bar{z}_L, 0)$ – міжгруповий розкид;

$S = \sum_{s=1}^{n} d^2(z_i, 0)$ – загальний розкид.

Якщо для вимірювання відстаней між точками багатовимірного простору використовується евклідова метрика (проста чи зважена), можна записати, що:

$$S=B+U. \quad (4.1)$$

Отже, загальний розкид (**S**) дорівнює сумі міжгрупового (B) та внутрішньогрупового розкидів (**G**). Якщо поділимо кожну частину цієї рівності на загальний розкид **S,** то можна отримати частку міжгрупового розкиду у загальному розкиду таким чином:

$$G = \frac{B}{S} = 1 - \frac{U}{S}. \quad (4.2)$$

Отже, чим більше величина внутрішньогрупового розкиду **G**, тим більша частина розкиду пояснюється міжгруповим розкидом і вище якість багатовимірного групування.

Означений підхід схожий на декомпозицію дисперсії. Тому величину **G** можна інтерпретувати, як коефіцієнт детермінації. Величина **G** змінюється від 0

до 1 і слугує як показник успішності багатомірного групування. В реальних економічних дослідженнях виконується така нерівність: **0<G<1.**

Кластером називається така група об'єктів з вихідної сукупності об'єктів, для якої виконується нерівність:

$$\boldsymbol{d_L^2 = \frac{U_L}{n_L} < \frac{S}{n} = \bar{d}^2}. \quad (4.3)$$

Це означає, що середній квадрат внутрігрупової відстані від об'єктів **L**-тої групи до її центру тяжіння менш середнього квадрата відстаней від усіх об'єктів до центра тяжіння вихідної сукупності.

L-тий кластер називається **згущенням** об'єктів, якщо для нього виконується така нерівність

$$max d_{Li}^2 < \bar{d}^2, \quad (4.4)$$

тобто максимальний квадрат внутригрупової відстані від об'єктів **L**-тої групи до її центра тяжіння менш середнього квадрата відстаней від усіх об'єктів до центра тяжіння вихідної сукупності.

Зауважимо, що поняття згущення об'єктів більш вузьке. Ніж поняття кластера. Згущення – це кластер, який має підвищену щільність вхідних об'єктів.

З вище оглянутого, задача кластерного аналізу складається у пошуку та виділенню у вихідній сукупності з **n** об'єктів максимального числа кластерів і згущень, які розглядаються як кількісно однорідні групи одночасно за всіма **m** ознаками.

Рішенням задачі кластерного аналізу є таке розбиття вихідної сукупності об'єктів, яке задовольняє деякому критерію оптимальності – цільової функції, що виражає рівень бажаності різних групувань. Таким розбиттям може бути класифікація об'єктів, яка мінімізує середню з групових дисперсій.

Основними етапами кластерного аналізу є такі:

1. Відбір об'єктів для кластеризації (наявність апріорної інформації).
2. Визначення множини ознак, за якими будуть оцінюватися об'єкти.
3. Обчислення міри подібності між об'єктами відповідно до обраної метрики.

4. Групування об’єктів у кластери за допомогою тієї чи іншої процедури об’єднання.

5. Перевірка достовірності результатів кластерного аналізу.

4.3. Класифікація алгоритмів кластерного аналізу

Після визначення матриці вихідних даних **X**, стандартизації ознак і утворення матриці **Z**, а також розрахунку матриці відстаней **D** (чи матриці подібності **μ**) використовують алгоритми кластерного аналізу. На сьогодні існує більш ста алгоритмів. Усі вони групуються за трьома напрямками:

1) процедури прямої класифікації;
2) оптимізаційні алгоритми;
3) апроксимаційні підходи.

Розглянемо означені напрямки.

Процедури прямої класифікації – це перший в історії багатовимірного аналізу напрямок. Його розробляли Ф.Гейнке (нім. біолог), К. Чекановський (польський антрополог), які на початку 20 ст. запропонували компактні групи об’єктів з урахуванням множини ознак.

Суть процедур прямої класифікації полягає в чіткому формулюванні поняття кластера і утворення груп об’єктів відповідно до цього формулювання. Серед таких процедур найбільш розповсюджені ієрархічні алгоритми. Вони базуються на такому визначні кластера: всі відстані між об’єктами у середині групи мають бути менш за будь-яку відстань між об’єктами групи та рештою множини. Такий кластер називається ***компактною групою*** або кластером типу ***ядра.***

Ієрархічні алгоритми дозволяють отримати послідовне розбиття сукупності об’єктів за визначеним правилом. Усі ієрархічні алгоритми поділяються на дивізимні та агломеративні.

Дивізимні алгоритми починають опрацювання даних з розглядання вихідної сукупності як одного кластера і послідовно розбивають її на більш малі

групи аж до розбиття, коли один об’єкт вважається окремим кластером. Агломеративні алгоритми працюють у зворотному напрямку.

Ієрархічні процедури мають переваги перед іншими процедурами кластерного аналізу:

- відносна простота і змістовність;
- можливість втручання в алгоритм;
- можливість побудувати графік – дендограму (древо розбиття – визуальне разбиття на кластери);
- досить низька трудомісткість розрахунків.

Оптимізаційний напрямок кластерного аналізу сформувався у 60 рр. XX ст. Відрізняється чітким формулюванням критерія якості та послідовним досягненням його екстремума. Обґрунтування цього напрямку базується на математичних законах опису. Під час реалізації цього напрямку виконується пряма класифікація, оскільки кожна процедура доставляє екстремум якийсь функцій. І навпаки, кожному функціоналу відповідає своє визначення кластера.

Апроксимаційний напрямок кластеризації сформувався приблизно у 70 рр. XX століття. Головна ідея підходу така: відносини, що закладено у вихідних даних, потребують найкращим способом апроксимувати відношенням, що відповідає уявленням дослідника про класифікацію. Наприклад, якщо виконується класифікація без перетинання класів, то фіксується, що вихідні відношення є відношенням еквівалентності (бути рівними за випуском продукції). Третій напрямок ще не склався повністю та не мають широкого використання.

4.5. Ієрархічні агломеративні процедури

На першому етапі ієрархічних агломеративних процедур кожний об’єкт вважається окремим кластером і відбувається об’єднання (агломерація) кластерів відповідно з правилом, яке визначає послідовність (ієрархію) такого поєднання.. Алгоритми цього напрямку відрізняються критеріями, які використовуються під час об’єднання. Розглянемо головні критерії:

1. *Метод «найближчого сусіда» (одиничного зв'язку).* На кожному кроці об'єднуються кластери $\mathbf{K_p}$ та $\mathbf{K_s}$, відстань між найближчими об'єктами **p** та **s** яких мінімальна, тобто мінімізується функція:

$$\mathrm{d}(K_p, K_s) = min\left[\min d(z_p, z_s)\right]. \quad (4.5)$$

На першому кроці поєднуються два найближчих між собою об'єкти, на другому – кластери за мінімальною відстанню між двома найближчими сусідами і так далі. Потрібен тільки один мінімальний зв'язок, щоб приєднати об'єкт до кластера, оскільки враховується тільки одиничний зв'язок з однією точкою кластера.

2. *Критерій «дальнього сусіда» (повного зв'язку).* На кожному кроці об'єднуються кластери $\mathbf{K_p}$ та $\mathbf{K_s}$, відстань між найбільш віддаленими об'єктами **p** та **s** яких мінімальна, тобто мінімізується функція

$$\mathrm{d}(K_p, K_s) = min\left[\max d(z_p, z_s)\right]. \quad (4.6)$$

При використанні цього критерію на першому кроці об'єднаються два близьких між собою об'єкта, на другому – кластери за мінімальною відстанню між двома віддаленими сусідами і так далі. Потрібний один мінімальний, но повний зв'язок для приєднання до кластера.

3. *Критерій середнього зв'язку (середньої відстані).* На кожному кроці об'єднуються кластери $\mathbf{K_p}$ та $\mathbf{K_s}$, середня відстань між усіма парами об'єктів яких мінімальна:

$$\mathrm{d}(K_p, K_s) = min\left[\bar{d}(z_p, z_s)\right]. \quad (4.7)$$

Означений критерій має дві модифікації, які залежать від способу розрахунку між об'єктами кожного кластера: за простою арифметичною або за зваженою арифметичною.

Загальна схема ієрархічного агломеративного алгоритму складається з таких етапів:

1) усі об'єкти $\mathbf{z_i}$ розглядаються як **n** самостійних кластерів $\mathbf{K_1, K_2, \dots, Kn;}$
2) розраховується відстань між усіма кластерами та утворюються матриця відстаней **D**, розміром $\mathbf{n_x n}$;

3) на основі обраного критерію визначається пара найближчих кластерів, які об’єднуються у один новий кластер. Якщо відразу декілька кластерів мають мінімальну відстань між собою, то обирають будь-яку пару;
4) розраховують відстані від нового кластера до інших. Розмірність матриці **D** при цьому знижується на одиницю;
5) на наступному кроці повторюється п.п.2, 3, 4 доти не відбудеться розбиття, яке складається з одного кластера – вихідної сукупності об’єктів.

Очевидно, що доводити ієрархічний агломеративний алгоритм до кінця не має сенсу, оскільки не виконується задання багатовимірного групування. Необхідна обґрунтована зупинка процедури агломерації. Сигналом для такої зупинки слугує різке зростання мінімальної відстані між кластерами, що поєднуються. Це вказує на те, що в одну групу об’єднаються різнорідні об’єкти, ніж на попередньому кроці.

Як бачимо з загальної схеми алгоритму, для його успішного здійснення необхідно:

1) розрахувати відстань від нового (утвореного) кластера до інших кластерів;
2) своєчасно зупинити процедуру, способом обрання оптимального числа компактних груп об’єктів.

Існує загальна формула розрахунку між кластером **Kr**, який є результатом поєднання кластерів **Kp** та **Ks**, і кластером **Kg (**Ланса-Уільямса**)**:

$$d_{rg} = \alpha_p d_{pg} + \alpha_s d_{sg} + \beta d_{ps} + \gamma \left| d_{pg} - d_{sg} \right|, \qquad (4.8)$$

де d_{ps}, d_{sg}, d_{ps} – відстані між відповідними кластерами;

$\alpha_p, \alpha_s, \beta, \gamma$ – параметри, що визначаються певними критеріями об’єднання.

Значення параметрів наведеної формули для різних критеріїв об’єднання наведено у табл. 4.1.

Точна відповідь на питання про сфери використання того чи іншого алгоритму відсутня. Загальноприйняте, що критерій «найближчого сусіда» має тенденцію до виділення ланцюгово розташованих кластерів. Критерій

«дальнього сусіда» може призвести до об’єднання несхожих груп на ранніх етапах.

Таблиця 4.1

Значення параметрів під час розрахунків відстаней між об’єднаними кластерами

Критерій об’єднання	Значення параметрів			
	α_p	α_s	β	γ
1. Найближчій сусід	0,5	0,5	0	-0,5
2. Дальній сусід	0,5	0,5	0	0,5
3. Середній зв'язок (проста середня)	0,5	0,5	0	0

Розглянемо ієрархічний алгоритм на такому прикладі. Припустимо, що необхідно провести класифікацію 10 підприємств, кожне з яких характеризується двома ознаками $\mathbf{x_1}$ та $\mathbf{x_2}$, які стандартизовані (табл.4.2).

Таблиця 4.2

Вихідні стандартизовані дані промислових підприємств

Номер підприємства	$\mathbf{Z_1}$	$\mathbf{Z_2}$		Номер підприємства	$\mathbf{Z_1}$	$\mathbf{Z_2}$
1	1,7	2,9		6	4,8	3,2
2	1,9	3,1		7	5,1	3,8
3	2,5	1,5		8	5,3	4,1
4	3,6	1,8		9	3,5	5,4
5	4,2	2,3		10	3,7	5,8

Для спрощення розрахунків візьмемо лінійну відстань (2.9). Відповідно до неї відстань між 1 і 2 об’єктом дорівнює:

$$d_{12} = |1{,}7 - 1{,}9| + |2{,}9 - 3{,}1| = 0{,}4$$

між 1 і 3 об’єктами

$$d_{13} = |1{,}7 - 2{,}5| + |2{,}9 - 1{,}5| = 2{,}2$$

Між 1і 4 об’єктами

$$d_{14} = |1{,}7 - 3{,}6| + |2{,}9 - 1{,}8| = 3{,}0$$ і так далі.

За отриманими розрахунками сформуємо матрицю лінійних відстаней $\mathbf{D_0}$ (табл.4.3). На її головній діагоналі знаходяться 0, вона симетрична.

Таблиця 4.3

Матриця лінійних відстаней

Об'єкти	1	2	3	4	5	6	7	8	9	10
1	0,0	0,4	2,2	3,0	3,1	3,4	4,3	4,8	4,3	4,9
2		0,0	2,2	3,0	3,1	3,0	3,9	4,4	3,9	4,8
3			0,0	1,4	2,5	4,0	4,9	5,4	4,9	5,5
4				0,0	1,1	2,6	3,5	4,0	3,7	4,1
5					0,0	1,5	2,4	2,9	3,8	4,0
6						0,0	0,9	1,4	3,5	3,7
7							0,0	0,5	3,2	3,4
8								0,0	3,1	3,3
9									0,0	0,6
10										0,0

Спочатку кожний об'єкт приймемо за окремий кластер. З матриці відстаней видно, що найбільш близькі об'єкти 1 і 2, тому що відстань між ними мінімальна $\mathbf{d_{12}}$=0,4. Тому на першому кроці класифікації поєднаємо їх у один кластер. Після об'єднання будується нова матриця відстаней $\mathbf{D_1}$, розміром 9х9. Вочевидь, що вісім рядків нової матриці $\mathbf{D_1}$ можна перенести зі старої матриці $\mathbf{D_0}$ (3-10), оскільки відстані між 3 і наступними об'єктами залишаються незмінними. Перший рядок (тобто результат об'єднання 1і 2 об'єктів) перераховують заново. Отже, надалі виконується об'єднання кластерів відповідно до критерію простої середньої відстані.

Відповідно до формули Ланса-Уільямса (4.8) відстань між кластерами розраховується таким чином:

$$d_{rg} = 0{,}5d_{pg} + 0{,}5d_{sg}. \qquad (4.9)$$

Отже, відстань між новим кластером (1,2) і кластером 3 відповідно формули (4.9) дорівнює:

$$d_{(1,2)3} = 0{,}5d_{1,3} + 0{,}5d_{2,3} = 0{,}5(2{,}2 + 2{,}2) = 2{,}2;$$

відстань між новим кластером (1,2) і кластером 4 дорівнює:

$$d_{(1,2)4} = 0{,}5d_{1,4} + 0{,}5d_{2,4} = 0{,}5(3{,}0 + 3{,}0) = 3{,}0;$$

відстань між новим кластером і кластером 5 дорівнює:

$$d_{(1,2)5} = 0{,}5d_{1,5} + 0{,}5d_{2,5} = 0{,}5(3{,}1 + 3{,}1) = 3{,}1;$$

відстань між новим кластером (1,2) і кластером 6 дорівнює:

$$d_{(1,2)6} = 0{,}5d_{1,6} + 0{,}5d_{2,6} = 0{,}5(3{,}4 + 3{,}0) = 3{,}2$$ і так далі.

Щоб не переписувати нову матрицю відстаней **D,** наведемо результати першого кроку класифікації тільки одного перерахованого рядка (табл. 4.4), тобто наведемо відстані від кластера (1,2) до інших об'єктів (стовпчик 1). При чому у першій клітинці повинно розташувати 0, але напишемо значення відстані, що стало базою для об'єднання кластерів на першому кроці (min $\mathbf{d_{1,2}}$=0,4) і надалі такі значення позначимо дужками.

Таблиця 4.4

Відстані між кластерами під час об'єднання

Номер об'єкта	Номер кроку								
	1	2	3	4	5	6	7	8	9
1	(0,4)	4,35	4,40	2,60	3,15	3,75	3,175		
2									
3	2,20	5,15	5,20	(1,4)	2,55	2,225			
4	3,00	3,75	3,90					(3,175)	
5	3,10	2,65	3,90	1,80	(1,5)	(1,90)	(2,225)		
6	3,20	1,15	3,60	3,30					(4,213)
7	4,10	(0,5)	3,25	4,45	1,90				
8	4,00								
9	4,10	3,15	(0,6)	4,55	3,75	3,50	4,025	4,213	
10	4,70	3,35							
Кількість кластерів	9	8	7	6	5	4	3	2	1

На другому кроці знаходимо мінімальне значення матриці відстаней (рядки 3-10) табл.4.3 і стовпець 2 табл.4.4. Воно дорівнює $\mathbf{d_{7,8}}$=0,5, тому об’єднуємо 7 і 8 об’єкти. У результаті отримуємо 8 кластерів. Нова матриця відстаней буде мати розмір 8х8. Перераховуємо тільки один рядок, значення якого наведено у 2 стовпчику табл.4. Потім процедура повторюється. До 6 кроку відбувається поєднання тільки первинних об’єктів, а починаючи з 6 кроку об’єднуються два кластери (5,6) та (7,8), відстань між якими складає 1,9. Знову перераховуємо матрицю відстаней між кластерами (5,6,7,8) та іншими об’єктами та отримуємо 4 кластери. Деякі кластери поєднувались два рази.

Таким чином, послідовно доходимо до одного кластера. Розглянуту процедуру можна відобразити графічно, за допомогою дендограми. Зазвичай при її побудові по горизонталі вказують номери об’єктів, а по вертикалі – значення між кластерними відстанями.

Якщо ми проаналізуємо значення міжкластерних відстаней, то побачимо, що різке зростання відбулося на 4 кроці (з 0,6 до 1,4). Це вказує на необхідність зупинки ієрархічної процедури групування на третьому кроці. Отже, у результаті багатовимірного групування , дійшли до висновку, що досліджувану сукупність доцільно розбити на 7 груп(класів) – (1,2); (7,8); (9,10); (3); (4); (5); (6). Їх подальше об’єднання призведе до виникнення неоднорідних груп.

4.5. Алгоритм пошуку згущень об'єктів (Алгоритм «Форель»)

Головна мета даного методу – групування багатовимірних об’єктів на основі пошуку згущень точок у просторі ознак. Алгоритм використовує плаваючу сферу, яка зупиняється під час реалізації процедури у місці локального скупчення точок.

Для виконання алгоритму пошуку згущень необхідно задати величину **T** – радіус сфери у просторі ознак, яка виконує роль еталона. Нижньою межею **T** є мінімальна відстань між об’єктами вихідної сукупності, а верхньою – половина відстані між найбільш віддаленими точками. У першому випадку число

утворених кластерів дорівнює числу об'єктів **n**, а у другому – єдина сфера охоплює всю сукупність об'єктів, що вивчаються.

Робота алгоритму починається з вибору довільної точки $\mathbf{Z_0}$ $\mathbf{(z_{01}, z_{02}, .., z_{0m})}$. Вона є центром сфери заданого радіусу **T**. Потім визначається сукупність точок, які потрапили до сфери. Далі розраховуються координати центру тяжіння утвореної сукупності точок $\bar{\mathbf{Z}}_\mathbf{0}$ $(\bar{\mathbf{z}}_{\mathbf{01}}, \bar{\mathbf{z}}_{\mathbf{02}}, \dots, \bar{\mathbf{z}}_{\mathbf{0m}})$. У цьому випадку мають справу не з координатами об'єкта, а з вектором середніх значень.

Точка $\bar{\mathbf{Z}}_\mathbf{0}$ вважається новим центром сфери і усі об'єкти, які потрапили до нею підлягають відбору. Потім знову розраховується центр тяжіння $\bar{\mathbf{Z}}_\mathbf{0}$ нової сфери за всіма ознаками об'єктів, що потрапили до неї. Точка $\bar{\mathbf{Z}}_\mathbf{0}$ з новим координами приймається за новий центр сфери і так далі.

При цьому склад точок, які потрапили до сфери, може відрізнятися від сукупності точок, які належать попередній сфері, що вказує на відмінність від нуля відстані між центром сфери і центром тяжіння утвореної сукупності об'єктів. Математично це можна виразити так: $\boldsymbol{d(z_0, \bar{z}_1) \neq 0.}$

Повтор даної процедури сприяє до руху центру сфери в область згущень точок. Плавання сфери продовжується доки перерахунок координат центра тяжіння не дасть той самий результат, що й на попередньому кроці. Тобто $\boldsymbol{d(\bar{z}_i, \bar{z}_{i-1}) = 0}$.

Точки, які потрапили у сферу, приймаються за окремий кластер і з подальшого аналізу виключаються. Серед об'єктів, що залишилися, знову довільно обирається нова точка і алгоритм повторюються з початку. Тобто, з випадкового об'єкта – центра радіусу **T**.

Таким чином, процедура алгоритму «Форель» завершується за кінчене число кроків, і усі точки розподіляються за кластерами.

Вочевидь, що результат роботи алгоритмів даного типу, насамперед, число утворених кластерів **R,** залежить від вибору величини **T**. Для отриманого оптимального числа кластерів усю процедуру потрібно повторювати при різних значеннях **T**, кожного разу змінюючи довжину радіусу на невелику величину.

Сигналом стійкості розбиття об’єктів є утворення постійного числа кластерів для низки послідовних значень **T** і достатньо різка зміна **R** на попередньому та послідовному кроках.

Розглянемо приклад, нехай 12 підприємств характеризуються двома ознаками $\mathbf{x_1}$ і $\mathbf{x_2}$ (табл.4.5).

Таблиця 4.5

Вихідні стандартизовані дані підприємств

Номер підприємства	$\mathbf{Z_1}$	$\mathbf{Z_2}$		Номер підприємства	$\mathbf{Z_1}$	$\mathbf{Z_2}$
1	1,75	5,25		7	2,82	4,22
2	2,65	5,50		8	2,53	4,07
3	1,80	4,47		9	2,25	4,04
4	2,50	4,75		10	2,06	3,95
5	3,00	5,00		11	2,75	3,75
6	3,54	4,71		12	3,24	3,93

За існуючими даними необхідно за допомогою алгоритму пошуку згущень об’єктів виділити групу однорідних підприємств. Спочатку знаходимо матрицю лінійних відстань. Її значення потрібні для вибору радіусу.

Матриця лінійних відстаней D:

0	1,15	0,83	1,25	1,5	2,33	2,1	1,81	1,71	1,61	2,50	2,81
	0	1,88	0,90	0,85	1,68	1,45	1,55	1,86	2,14	1,85	2,16
		0	0,98	1,75	1,98	1,27	1,13	0,88	0,78	1,67	1,98
			0	0,75	1,08	0,85	0,71	0,96	1,24	1,25	1,56
				0	0,83	0,96	1,40	1,71	1,99	1,50	1,37
					0	1,21	1,65	1,96	2,24	1,75	1,08
						0	0,44	0,75	1,03	0,54	0,71
							0	0,31	0,59	0,54	0,85
								0	0,28	0,79	1,10
									0	0,89	1,20
										0	0,67
											0

Як нижню межу радіусу приймемо мінімальну відстань між вихідними об'єктами, тобто $\mathbf{d_{min}=d_{9;10}}$=0,28. Верхня межа радіусу дорівнює половині відстані між найбільш віддаленими точками, тобто 1,41($\mathbf{d_{max}=d_{1;12}}$=2,81). Це означає, що гіпосфера радіусом T=0,28 дозволяє виділити рівно 12 кластерів, кожен з яких містить тільки один об'єкт, а гіпосфера радіусом 1,41 охопить усю вихідну сукупність, створить один великий кластер з 12 об'єктів. Тому якісь тривіальні рішення у такому випадку не мають практичного сенсу. Оберемо радіус гіперсфери довільно з інтервалу [0,28;1,41], наприклад **T=1,10**.

На першому кроці відповідно до алгоритму пошуку об'єктів, початковою точкою обираємо, наприклад об'єкт №2, тобто $\mathbf{z_0}$ (2,65; 5,50). Скористаємось метрикою лінійної відстані і визначимо сукупність точок, які потрапили у гіперсферу з центром в точці $\mathbf{z_0}$ (2,65; 5,50). Раніше розрахована матриця відстаней **D** містить результати цих розрахунків у другій строчці. Отримані значення можна перенести у перший стовпчик табл. 4.6.

Дані табл. 4.6 вказують, що на першому кроці в круг потрапили три об'єкти: №№ 2; 4; 5, оскільки їх значення менші за 1,10.

За даними матриці стандартизованих даних (табл.4.5) визначимо за допомогою простої арифметичної координати тяжіння цих точок:

$$\bar{z}_{11} = \frac{2,65+2,50+3,00}{3} = 2,72,$$

$$\bar{z}_{12} = \frac{5,50+4,75+5,00}{3} = 5,08.$$

Наразі центр тяжіння матиме такі координати $\bar{\boldsymbol{z}}_{\mathbf{1}}$ (2,72; 5,08). Порівняння двох точок $\boldsymbol{z_0}$ та $\bar{\boldsymbol{z}}_{\mathbf{1}}$ вказує, що у гіперсфері, радіусом **T**=1,10, відбувся незначний рух. Тому нову точку $\bar{\boldsymbol{z}}_{\mathbf{1}}$ вважатиме за новий центр гіперсфери і на другому кроці розрахуємо відстані від цього центру до усіх 12 об'єктів. У табл. 4.6 наведені ці відстані тільки для об'єктів, що потрапили до сфери, і до них відносяться об'єкти під номерами 2,4,5,7. Далі розраховуємо для означених об'єктів наступний центр тяжіння:

$$\bar{z}_{21} = \frac{2,65+2,50+3,00+2,82}{4} = 2,74,$$

$$\bar{z}_{22} = \frac{5{,}50+4{,}75+5{,}00+4.22}{4} = 4{,}87.$$

Таблиця 4.6

Відстані між центрами тяжіння та об’єктами сукупності

Номер об’єкта	Номер кроку									
	1	2	3	4	5	6	7	8	9	10
1	1,15									1,84
2	(0,0)	0,49	0,72	0,98						1,51
3	1,88								1,10	(1,01)
4	(0,90)	0,55	0,6	0,38	0,51	0,81	0,77	0,73	0,68	(0,61)
5	(0,85)	0,36	0,39	0,45	0,60	0,94	0,86	1,02		1,36
6	1,68		0,86	0,70	0,85	0,99				1,61
7	1,45	0,96	0,73	0,51	0,36	0,22	0,10	0,12	0,19	(0,40)
8	1,55		1,01	0,95	0,80	0,66	0,54	0,38	0,25	(0,12)
9	1,86					0,97	0,85	0,69	0,56	(0,35)
10	2,14							0,97	0,84	(0,63)
11	1,85			1,05	0,90	0,76	0,63	0,52	0,57	(0,66)
12	2,16				1,05	0,75	0,79	0,53	0,88	(0,97)
Координати центра тяжіння	2,72	2,74	2,84	2,83	2,91	2,83	2,73	2,64	2,49	2,49
	5,08	4,87	4,71	4,57	4,35	4,31	4,25	4,21	4,15	4,15

Знову приймаємо точку $\bar{\boldsymbol{z}}_{\boldsymbol{2}}$ з координатами (2,74; 4,87) за новий центр тяжіння сфери. І так далі. Весь ітераційний процес виділення першого кластера подано у табл. 4.6. На десятому кроці сукупність об’єктів залишається незмінною а центр згущення співпадає з центром попередньої гіпосфери. $\bar{\boldsymbol{z}}_{\boldsymbol{9}} = \bar{\boldsymbol{z}}_{\boldsymbol{10}}$ і відстань ($d(\bar{\boldsymbol{z}}_{\boldsymbol{9}}; \bar{\boldsymbol{z}}_{\boldsymbol{10}})=0$). Це означає, що склад сукупності такий, як й був. Таким чином, у перший кластер можна об’єднати об’єкти 3, 4, 7, 8, 9, 10, 11, 12 і ці об’єкти з подальшого аналізу вилучаються.

З числа залишившихся об’єктів №№ 1, 2, 5, 6 знову вибирається новий, стандартизовані значення якого вважаються початковою точкою гіперсфери. Відносно неї розраховуються та аналізуються нові відстані, які потрібно

порівняти з радіусом T=1,10. Досліджується рух гіперсфери, розраховуються нові центри тяжіння, поки всі об’єкти не розносяться за класами (таксонами).

4.6. Функціонали якості класифікації

Досліджувану сукупність можна розбити на кластери різними способами. При заданій кількості кластерів **R** таких груп може бути велика кількість. Отже, необхідно порівняти їх якість. Для цього використовується функціонал якості ***Q(R)*** Найкращим за обраним функціоналом вважають таку класифікацію об’єктів, в якій досягається екстремальне (максимальне або мінімальне) значення функціоналу якості.

Загальна кількість функціоналів дуже велика, тому розглянемо найважливіші:

1. Загальна внутрікластерна дисперсія за всіма ознаками:

$$Q(R) = \sum_{L=1}^{R} \sum_{k=1}^{m} \sigma_{Lk}^2 \rightarrow min\ , \qquad (4.10)$$

де R – кількість кластерів.

При даному функціоналі вибирають таке розбиття, щоб функціонал прагнув до мінімуму. Означений функціонал покладено в оптимізаційний алгоритм:

1) задається початкове розбиття на $\mathbf{R_0}$ кластерів;

2) у кожному кластері визначаються свої центри тяжіння, тобто умовні точки, координати яких дорівнюють середнім арифметичним значенням ознак. Наприклад, k – ая координата центру тяжіння дорівнює:

$$\bar{x}_{Rk} = \frac{1}{n_R} \sum_{L=1}^{R} x_{Lk} \qquad (4.11)$$

де $\boldsymbol{n_R}$ – кількість об’єктів у R-му кластері.

3) усі вихідні об’єкти знову розподіляють за **R** кластерами відповідно принципу: і –й об’єкт належить кластеру **L**, якщо відстань від нього до центра тяжіння даного кластера мінімальна порівняно з відстанями до центрів тяжіння інших кластерів;

4) якщо отримане розбиття сукупності точок відрізняється від початкового, то пп. 2 і 3 повторюються. Інакше робота оптимізаційного алгоритму закінчується.

Наведена процедура кластерного аналізу побудована так, щоб величина функціоналу якості на кожному кроці зменшувалась

2. Загальна сума квадратів попарних внутрікласових відстаней:

$$Q(R) = \sum_{L=1}^{R} \sum_{i,j} d_{ij}^2 \to min, \ \ (i,j = 1,2, \dots, n) \qquad (4.12)$$

Мета класифікації при даному випадку функціоналу якості – отримати такі **R** кластерів, для яких загальна сума квадратів відстаней між всіма об'єктами була би мінімальною.

Питання для самоконтролю

1. Розкрийте поняття однорідності.
2. У чому сутність структурного та комбінаційного групування?
3. У чому головна мета та завдання кластерного аналізу?
4. Чому дорівнює загальний розкид?
5. Що таке кластер? Що таке згущення об'єктів? Як вони пов'язані?
6. Опишіть основні етапи кластерного аналізу.
7. Охарактеризуйте напрямки алгоритмів кластерного аналізу.
8. Опишіть процедуру прямої класифікації.
9. Що таке ієрархічні алгоритми кластерного аналізу?
10. Що таке дивізимні алгоритми кластерного аналізу?
11. Охарактеризуйте метод «найближчого сусіда».
12. Охарактеризуйте метод «дальнього сусіда».
13. Охарактеризуйте критерій середнього зв'язку.
14. Опишіть етапи алгоритму «Форель».
15. Як задається радіус гіперсфери?
16. Який існує критерій стійкого розбиття на кластери методом «Форель»?
17. Наведіть та розкрийте сутність функціоналів якості кластерного аналізу.

Тема 5. ДИСКРИМІНАНТНИЙ АНАЛІЗ

5.1. Сутність методів дискримінантного аналізу.

5.2. Теоретичні основи класифікації методами дискримінантного аналізу.

5.3. Приклад використання дискримінантного аналізу для двох кластерів.

5.4. Особливості застосування дискримінантного аналізу в практиці соціально-економічних досліджень.

5.1. Сутність методів дискримінантного аналізу

У соціально-економічних дослідженнях значну роль грає класифікація нових об'єктів, тобто віднесення їх до вже відомих кластерів, груп об'єктів. Термін «новий об'єкт» розглядається не в абсолютному , а у відносному сенсі – порівняно з описаними раніше класами об'єктів. У теорії багатовимірних статистичних методів означена проблема отримала назву задачі дискримінантного аналізу чи дискримінації, класифікації з навчанням.

Дискримінантний аналіз – це розділ багатовимірного аналізу, зміст якого полягає у розробці методів розв'язання задач класифікації нових об'єктів, методом порівняння величини їх ознак з аналогічними показниками кластерів, що досліджені раніше.

Отже, кластерний аналіз передує дикримінантному аналізу і складає його базу (навчальну вибірку чи вчитель).

Методи дискримінантного аналізу мають широкий спектр використання у суспільстві, насамперед, медицині, соціології, інженерно-технічних дослідженнях, економіці. Наприклад, за допомогою оцінки показників виробничо-господарчої діяльності (випуск продукції, прибуток, рентабельність тощо) банк класифікує клієнтів за кредитоспроможністю на дві категорії – надійних та ненадійних. Задача дискримінантного аналізу у даній ситуації визначити таке правило, яке дозволить віднести нового клієнта до першої чи другої групи позичальників з мінімальною вірогідністю помилитися у класифікації.

Розглянемо основні *практичні ситуації використання методів дискримінантного аналізу*:

1. Втрачена (чи відсутня) інформація за окремими об'єктами і досліднику необхідно встановити належність таких об'єктів до певного типу, класу. Описана ситуація найбільш властива для археології, геології, криміналістиці. Наприклад, при визначенні століття, культури, до яких належить знайдений предмет старовини.

2. Недосяжна інформація, коли досліднику потрібно розпізнати внутрішні (латентні) характеристики досліджуваного об'єкта за зовнішнім проявом факторів-симптомів. Наприклад, під час проведення технічної, економічної експертизі.

3. Прогнозна інформація, коли аналітику на базі ретроспекції (досвіду поточних та попередніх періодів часу) необхідно моделювати типи поведінки різних систем і зробити спробу передбачити поведінку нової системи у майбутньому. Наприклад, прогнозування дохідності чи збитковості нового введеного в експлуатацію підприємства за проектними техніко-економічними показниками.

Методи дискримінантного аналізу розроблялися у середині 30-х років XX століття Р. Фішером, П.Махаланобісом, Г. Хотеллингом.

Вирішальне правило, за яким виконується багатовимірна класифікація нових об'єктів, ґрунтується на спеціальних математичних функціях, які називаються дискримінантними.

Дискримінантні функції визначаються таким чином, щоб мінімізувати загальні втрати при помилкових класифікаціях нових об'єктів.

За статистичною формою задача дискримінантного аналізу формулюється таким чином. Нехай кожний з n об'єктів описується m ознаками та вся інформація подана матрицею даних **X**. Окрім того, відома додаткова характеристика **S** (**S=1, 2, …, L, …, R**), яка вказує на належність об'єкта до певного кластера.

Сукупність об’єктів з відомими параметрами **X** і **S** використовується для виявлення зв’язку між значеннями цих характеристик та називається навчальною вибіркою (вчителем). Тоді задача полягає у побудові дискримінантної функції ***f(x)*** за допомогою вчителя і на її основі віднесенні нового об’єкта до одного з **R** кластерів $\mathbf{W_1, W_2, \ldots, W_R}$.

Вочевидь, що проблема дискримінації відрізняється від задачі кластерного аналізу наявністю навчальної вибірки (вчителя). Тому іноді її називають задачею розпізнання образів з вчителем на відміну від задачі кластеризації, яку зазвичай називають задачею розпізнавання образів без учителя.

Дискримінантний аналіз має дві проблеми:

1) вибір виду функції ***f(x)***;
2) визначення невідомих коефіцієнтів дискримінантної функції ***f(x).***

У статистичній практиці найбільш обґрунтованою є лінійна дискримінантна функція.

Стосовно визначення невідомих коефіцієнтів дискримінантної функції ***f(x)***, то зараз відомі два підходи до цієї проблеми. Перший (класичний) підхід базується на розрахунках відстаней Махаланобіса між центрами тяжіння кластерів з наступним визначенням коефіцієнтів функції ***f(x).***

Другий підхід ґрунтується на використанні кореляційно-регресійного аналізу. При цьому як результативна ознака береться змінна **S**, а в якості факторів – змінні $x_1, x_2, \ldots, x_m$.

Дискримінантна функція матиме вигляд:

$$S = f(\mathrm{x}) + \varepsilon, \qquad (5.1)$$

коефіцієнти якої визначаються за допомогою статистичного апарату кореляційно-регресійного аналізу.

5.2.Теоретичні основи класифікації методами дискримінантного аналізу

Розглянемо класичний підхід до вирішення задачі дискримінації, який сформульовано для двох кластерів – $\mathbf{W_1}$ і $\mathbf{W_2}$ (**R=2**).

Стандартна процедура дискримінантного аналізу передбачає побудову лінійної комбінації вихідних даних – дискримінантної функції (5.1). Так, об'єкти, які належать двом кластерам, доцільно їх розділити прямою лінією, математичне рівняння якої (рис.5.1):

$$f(\mathrm{x}) = a_1\mathrm{x}_1 + a_2\mathrm{x}_2 = c \qquad (5.2)$$

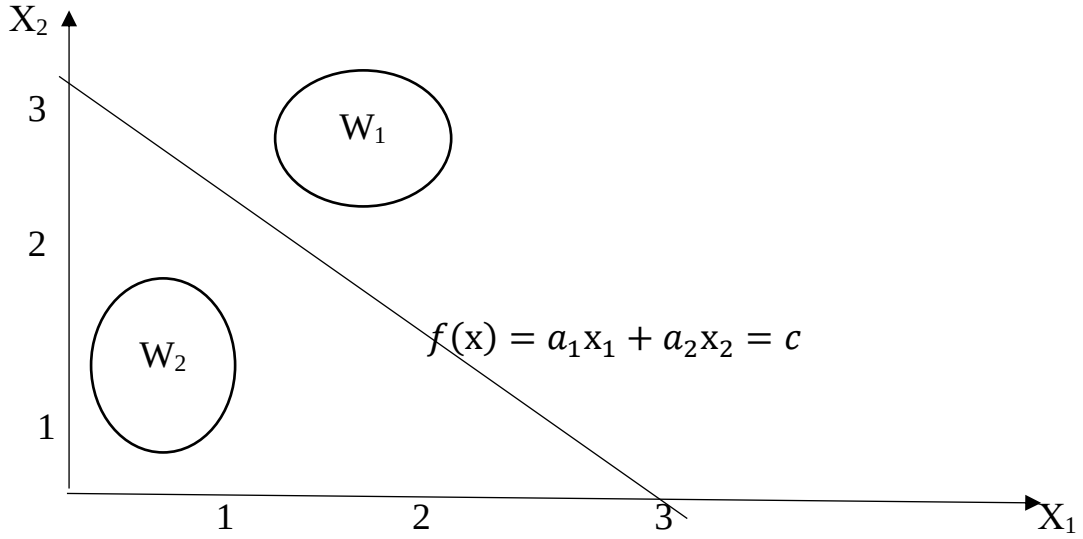

Рис. 5.1 Геометрична інтерпретація дискримінантного аналізу (**m=2; R=2**)

У загальному випадку, коли кількість ознак **m=3** – це площина у трьохмірному просторі, при **m=4** – це гіперплощина у чотирьохмірному просторі і так далі.

Вочевидь, що підстановка у таку функцію координат будь-якого об'єкта кластера $\mathbf{W_1}$, надає значення, яке більше за **C**. Значення , які менш за **C**, функція ***f(x)*** прийме після підстановки координат об'єктів кластера $\mathbf{W_2}$. Тому функцію ***f(x)*** (формула 5.2) можна використовувати як вирішальне правило під час класифікації невідомого об'єкта, тобто можна стверджувати, що цей об'єкт належить кластеру $\mathbf{W_1}$, якщо $\boldsymbol{f}(\mathrm{x}) > \boldsymbol{c}$, і кластеру $\mathbf{W_2}$, якщо $\boldsymbol{f}(\mathrm{x}) \leq \boldsymbol{c}$.

Зауважимо, що такий підхід можна використовувати при більшій, ніж 2, кількості кластерів.

Перейдемо до проблеми розрахунку коефіцієнтів дискримінантної функції ***f(x)*** (формула 5.2). При цьому невідомі коефіцієнти необхідно визначати так, щоб

розділ кластерів $\mathbf{W_1}$ і $\mathbf{W_2}$ був найкращим. Це означає, що розділ буде тим кращим, чим далі будуть знаходитися центри тяжіння кластерів $\mathbf{W_1}$ і $\mathbf{W_2}$ при вимірюванні відстані між ними в одиницях дисперсії.

Знайдемо значення дискримінантної функції в центрах тяжіння кожного з двох кластерів:

$$\boldsymbol{f'} = \sum_{k=1}^{m} \boldsymbol{a_k}\bar{\boldsymbol{x}}_{\boldsymbol{1k}} \ , \tag{5.3}$$

$$\boldsymbol{f''} = \sum_{k=1}^{m} \boldsymbol{a_k}\bar{\boldsymbol{x}}_{\boldsymbol{2k}}. \tag{5.4}$$

Коефіцієнти $\boldsymbol{a_k}$ обираємо таким чином, щоб значення $\boldsymbol{f'}$ та $\boldsymbol{f''}$ були якнайдалі віддалені один від одного. Для цього введемо відстань Махаланобіса між одномірними точками $\boldsymbol{f'}$ та $\boldsymbol{f''}$:

$$\boldsymbol{d_{|M}(f', f'')} = (\boldsymbol{f'} - \boldsymbol{f''})^{\boldsymbol{T}}\boldsymbol{W^{-1}}(\boldsymbol{f'} - \boldsymbol{f''}) = \frac{(f'-f'')^2}{\sigma_f^2}, \tag{5.5}$$

де $\boldsymbol{W^{-1}} = \frac{\mathbf{1}}{\boldsymbol{\sigma_f^2}}$ – матриця, зворотна до матриці розсіювання, що дорівнює одиниці, поділеної на дисперсію всіх значень $\boldsymbol{f_i}$.

Отже, необхідно знайти коефіцієнти $\boldsymbol{a_1}$, $\boldsymbol{a_2}$, .., $\boldsymbol{a_m}$, які максимізують відстань Махалонобіса $\boldsymbol{d_M(f', f'')}$ (формула 5.5). У теорії математичної статистики доведено, що шукані $\boldsymbol{a_k}$ є результатом рішення такої системи лінійних рівнянь:

$$\begin{cases} a_1\sigma_{11} + a_2\sigma_{12} + \cdots + a_m\sigma_{1m} = \bar{x}_{11} - \bar{x}_{21} \\ a_1\sigma_{21} + a_2\sigma_{22} + \cdots + a_m\sigma_{2m} = \bar{x}_{12} - \bar{x}_{22} \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_1\sigma_{m1} + a_2\sigma_{m2} + \cdots + a_m\sigma_{mm} = \bar{x}_{1m} - \bar{x}_{2m} \end{cases} \tag{5.6}$$

де $\boldsymbol{\sigma_{kp}}$ – коваріація ознак $\mathbf{x_k}$ і $\mathbf{x_p}$, яка визначається за такою формулою:

$$\sigma_{kp} = \frac{\sum_{k=1}^{m}(z_k - \bar{z}_k)(z_p - \bar{z}_p)}{n}, \tag{5.7}$$

де $\boldsymbol{n}$ – обсяг сукупності.

При цьому нагадаємо властивість дисперсій-коваріацій:

$\boldsymbol{\sigma_{kk}^2} = \boldsymbol{\sigma_k^2}$ – дисперсія $\boldsymbol{k}$-ої ознаки.

Після визначення всіх дисперсій-коваріацій і розв'язання системи рівнянь, отримуємо коефіцієнти a_k, на підставі яких розраховуємо значення дискримінантної функції $f(x)$ для кожного об'єкта двох кластерів. Щоб мінімізувати ймовірність похибки класифікації, значення константи **С** визначаються за формулою простої середньої арифметичної:

$$C = \frac{f' + f''}{2}. \qquad (5.8)$$

Отже, процедура класифікації складається з визначення параметрів a_k, оцінки f' та f'', розрахунку константи **С**. Для нового вектора спостережень (об'єкта) розраховують значення дискримінантної функції $f(x_0)$ і об'єкт відносять до кластера W_1, якщо $f(x_0) > c$, інакше до кластера W_2.

5.3. Приклад використання дискримінантного аналізу для двох кластерів

Розглянемо використання методів дискримінантного аналізу на такому прикладі. Нехай 10 підприємствами промисловості характеризуються такими показниками (табл.5.1):

X_1 – продуктивність праці 1 робітника, тис. грн .

X_2 – собівартість одиниці продукції , грн.

По кожному підприємству відома загальна оцінка результатів господарчої діяльності – рівень рентабельності. До кластера W_1 **(**прибутковість**)** належать підприємства 1-5 **(S=1),** до кластера W_2 (збитковість) належать підприємства 6-10 (S=2).

Таблиця 5.1

Вихідні дані

№№ підприємства	1	2	3	4	5	6	7	8	9	10
X_1	12	13	15	14	18	6	8	5	7	12
X_2	25	28	24	26	20	45	48	46	44	54

За наведеними даними необхідно побудувати дискримінантну функцію і виконати прогноз прибутковості нового збудованого підприємства (X_0), яке має проектні значення X_1=10 тис. грн.; X_2 = 30 грн.

Для визначення параметрів лінійної дискримінантної функції ***f(x)*** запишемо систему з двох рівнянь (відповідно до системи рівнянь 5.6):

$$\begin{cases} a_1\sigma_{11} + a_2\sigma_{12} = \bar{x}_{11} - \bar{x}_{21} \\ a_1\sigma_{21} + a_2\sigma_{22} = \bar{x}_{12} - \bar{x}_{22} \end{cases} . \quad (5.9)$$

Розрахуємо загальні середні значення ознак за правилом простої арифметичної: $\overline{X}_1$= 11,0 тис. грн.; $\overline{X}_2$= 36,00 грн.

Таблиця 5.2

Розрахункові дані

№ підпр.	X_1	X_2	$X_1 - \overline{X}_1$	$X_2 - \overline{X}_2$	$(X_1 - \overline{X}_1) \cdot (X_2 - \overline{X}_2)$	$(X_1 - \overline{X}_1)^2$	$(X_2 - \overline{X}_2)^2$	f_i	$(f_i - \bar{f})^2$
А	1	2	3	4	5	6	7	8	9
1	12	25	1	-11	-11	1	121	-2,042	2,506
2	13	28	2	-8	-16	4	64	-2,335	1,664
3	15	24	4	-12	-48	16	144	-1,581	4,178
4	14	26	3	-10	-30	9	100	-1,958	2,779
5	18	20	7	-16	-112	49	256	-0,718	8,451
6	6	45	-5	9	-45	25	81	-5,376	3,066
7	8	48	-3	12	-36	9	144	-5,560	3,744
8	5	46	-6	10	-60	23	100	-5,619	3,976
9	7	44	-4	8	-32	16	64	-5,133	2,274
10	12	54	1	18	18	1	324	-5,928	5,304
Усього	110	360	x	x	-372	166	1398	-36,250	37,942

Знайдемо дисперсії-коваріації за формулою (5.7):

$$\sigma_{12} = \frac{\sum(X_1 - \bar{X}_1)\cdot(X_2 - \bar{X}_2)}{n} = \frac{-372}{10} = -37{,}2.$$

Перейдемо до розрахунків дисперсії кожної ознаки:

$$\sigma_{11} = \sigma_1^2 = \frac{\sum(X_1 - \overline{X}_1)^2}{n} = \frac{166}{10} = 16{,}6.$$

$$\sigma_{22} = \sigma_2^2 = \frac{\sum(X_2 - \overline{X}_2)^2}{n} = \frac{1398}{10} = 139{,}8.$$

Знайдемо середні значення ознак у кожному кластері. Середні значення за ознакою X_1 (продуктивність праці робітника) у першому кластері $\mathbf{W_1}$ дорівнює

$\bar{X}_{11} = \frac{12+13+15+14+18}{5} = \frac{72}{5} = 14{,}4$ тис. грн., а у другому кластері $\mathbf{W_2}$ – $\bar{X}_{21} = \frac{39}{5} = 7{,}6$ тис. грн.

Аналогічні розрахунки проводимо за другою ознакою (собівартість одиниці продукції):

$$\bar{X}_{12} = \frac{123}{5} = 24{,}6 \text{ грн.} \quad ; \quad \bar{X}_{22} = \frac{237}{5} = 47{,}4 \text{ грн.}$$

Звідси система лінійних рівнянь (5.9) для визначення a_1 і a_2 приймає такий вигляд:

$$\begin{cases} 16{,}6a_1 - 37{,}2a_2 = 14{,}4 - 7{,}6 \\ -37{,}2a_1 + 139{,}8a_2 = 24{,}6 - 47{,}4 \end{cases}$$

Розв'язання отриманої системи дозволяє визначити параметри

a_1=0,109 і a_2= – 0,134.

Визначимо значення дискримінантної функції ***f(z)*** в центрах тяжіння обох кластерів (формули 5.3-5.4):

$$f' = \sum_{k=1}^{m} a_k \bar{x}_{1k} = a_1\bar{X}_{11} + a_2\bar{X}_{12} = 0{,}109 \cdot 14{,}4 - 0{,}134 \cdot 24{,}6 = -1{,}727$$

$$f'' = \sum_{k=1}^{m} a_k \bar{x}_{2k} = a_1\bar{X}_{21} + a_2\bar{X}_{22} = 0{,}109 \cdot 7{,}6 - 0{,}134 \cdot 47{,}4 = -5{,}523.$$

Знайдемо значення константи **C** (формула 5.8), щоб мінімізувати ймовірність похибки:

$$C = \frac{f' + f''}{2} = \frac{-1{,}727 - 5{,}523}{2} = -3{,}625.$$

Отже, дискримінантна функція (5.2) може бути записана таким чином:

$$f(x) = 0{,}109x_1 - 0{,}134x_2 = -3{,}625 \,. \qquad (5.10)$$

Об'єкти, для яких $f(x) > -3{,}625$ віднесемо до кластера $\mathbf{W_1}$ (прибутковість підприємств). Якщо же $f(x) \leq -3{,}625$, то об'єкті належать до кластера $\mathbf{W_2}$ (збитковість).

Розрахуємо значення дискримінантної функції для всіх досліджуваних підприємств (табл.5.2, стовпчик 8). Можна побачити, що перші п'ять значень більш за -3,625 ($f(x) > -3{,}625$) і тому склали кластер $\mathbf{W_1}$, а наступні п'ять – менш за -3,625 ($f(x) \leq -3{,}625$) і склали кластер $\mathbf{W_2}$.

Визначимо відстань Махаланобіса між центрами тяжіння обох кластерів. Для цього спочатку знайдемо значення функції (5.10) для кожного об'єкта (табл. 5.2, гр.8):

f(1)=0,109·12+(-0,134)·25=-2,042;

f(2)=0,109·13+(-0,134)·28=-2,335 і т.д.

На підставі отриманих даних розраховуємо дисперсію значень дискримінантної функції f_i (табл. 5.2 (гр.9)).:

$$\sigma_f^2 = \frac{37,942}{10} = 3,794.$$

Відстань Махаланобіса розрахуємо за формулою (5.5):

$$d_{|M}(f', f'') = \frac{(f'-f'')^2}{\sigma_f^2} = \frac{(-1,727+5,523)^2}{3,794} = 3,798.$$

Здійснимо прогноз прибутковості нового промислового підприємства X_0 з очікуваними характеристиками: виробіток продукції 1 робітника за проектом складе 10 тис. грн., а собівартість продукції – 30 грн.

Підставимо ці дані у отриману дискримінантну функцію (5.10):

$$f(x_0) = 0,109 \cdot 10 - 0,134 \cdot 30 = -2,930, \ \text{-2,930>-3,625}$$

Так як, $f(x_0) > C$, то даний об'єкт віднесемо до кластера $\mathbf{W_1}$, тобто новий об'єкт, виходячи з проектних даних, буде прибутковим.

5.4. Особливості застосування дискримінантного аналізу в практиці соціально-економічних досліджень

Дискримінантну функцію ***f(x)*** (5.1)можна подати таким чином:

$$f(x) = a_1x_1 + a_2x_2 + \cdots + a_mx_m + a_{m+1}. \qquad (5.11)$$

Таке подання ідентично тому, що розглянуто раніше, якщо прийняти $a_{m+1} = -C$. Тоді розподіл вихідних даних на два кластери визначається за такими властивостями дискримінантної функції:

$$f(x) = \begin{cases} >0, \text{ якщо } x \in \mathbf{W_1} \\ \\ \leq 0, \text{ якщо } x \in \mathbf{W_2} \end{cases} \qquad (5.12)$$

Узагальнимо ці властивості на випадок ***i*** – кластерів, **W*i*** (***i=1, 2, …,R***), тобто **R>2.**

Виділимо кожний кластер від будь-якого іншого, який виділяється індивідуальною гіперплощиною. Приклад для трьох кластерів (R=3) і двох ознак (m=2) (рис.5.2):

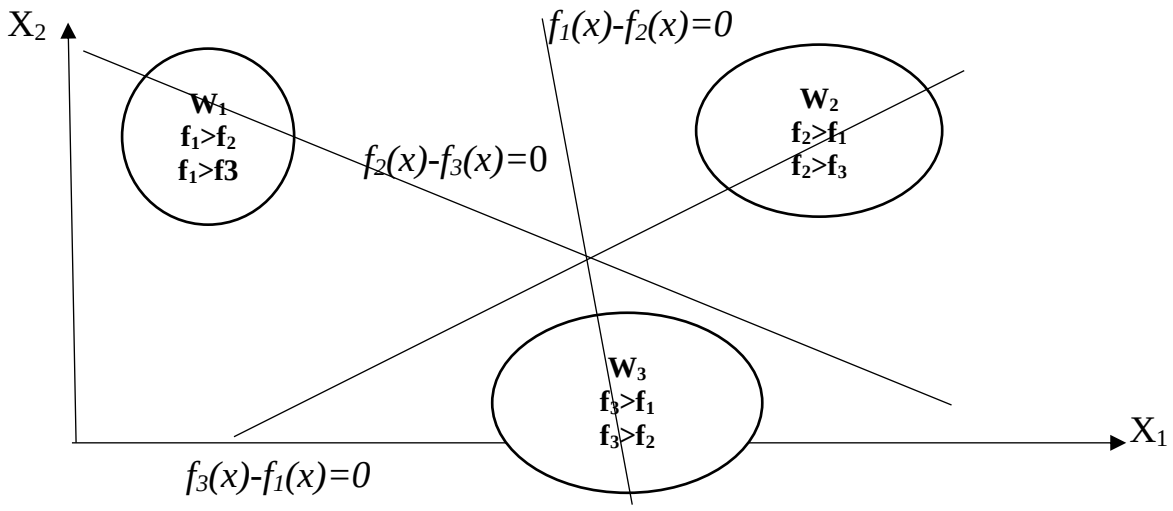

Рис. 5.2 Розподіл трьох кластерів трьома прямими

Очевидно, що ні один кластер неможливо відділити від інших за допомогою єдиної лінії гіперплощини. Кожна з наведених на рис. 2. меж кластерів забезпечує розподіл тільки двох класів. У цьому випадку існує **R** дискримінантних функцій:

$$f_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{im}x_m + a_{im+1}$$

таких, як що об'єкт належить до кластера **W*i*** , то ***fi(x)*** > ***fj(x)*** для всіх ***j≠i*** .

Межа між кластерами **W*i*** та **W*j*** визначається такими значеннями ***x,*** при яких виконується рівність $f_i(x)=f_j(x)$ або теж саме, що й $f_i(x)-f_j(x)= 0$. Отже, під час виділення рівнянь розділяючих гіперплощин кластерів **W*i*** та **W*j*** значення дискримінантної функції розглядаються разом. Для випадку, поданого на рис. 2, для об'єктів кластера $\mathbf{W_1}$ повинні виконуватися умови $f_1(x)>f_2(x)$ та $f_1(x)>f_3(x)$. Це еквівалентно тому, що об'єкти, які ввійшли до даного кластера, розташовані у додатних зонах на півплощинах (полуосях) $f_1(x)-f_2(x)=0$ та $f_1(x)-f_3(x)=0$. Процедура визначення невідомих параметрів a_{ij} дискримінантної функції $f_j(x)$

така сама, як й у випадку двох кластерів (**R=2**). Такі розрахунки досить трудомісткі і тому їх виконують за допомогою спеціальних комп'ютерних програм.

Під час проведення дискримінантного аналізу виникає низка проблем, розв'язання яких впливає на успіх статистичної процедури. Перш за все, це проблема правильного вибору діагностичних ознак, які викликають суттєві відмінності між кластерами. Рішенням даної проблеми стає практичний досвід аналітика, а також якісний економічний аналіз явища, що досліджується.

Залежно від специфіки задач використовуються різні типи ознак. Деякі з них легко інтерпретуються. Наприклад, при класифікації студентів на групи успішних та неуспішних такими ознаками становяться оцінки (бали) за окремими дисциплінами. Якщо використовуються складні (багаточисельні) ознаки, наприклад, підприємства за рівнем рентабельності, працівників за ступнем задовільності працею, то спочатку необхідно виділити найбільш суттєві за допомогою метода головних компонент, факторного, таксономічного аналізу. При виборі вихідних ознак потрібно звернути увагу на ті змінні, які спричиняють суттєві розбіжності між кластерами. І тільки після вибору і обґрунтування ознак, починається процедура дискримінантного аналізу.

Відмітимо, що у практичних дослідженнях вихідна сукупність поділяється на окремі кластери не чітко, як у прикладах. Завжди існує зона ризику (рис. 5.3):

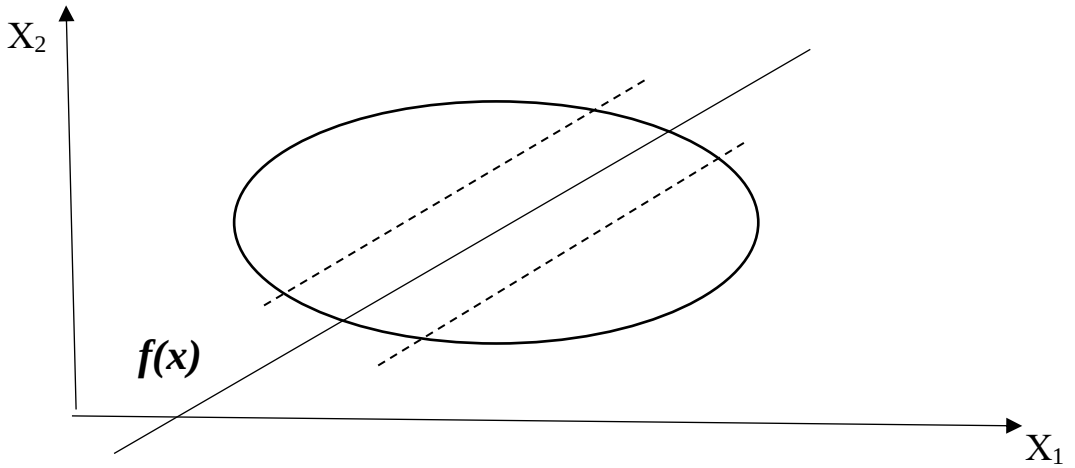

Рис.5. 3 Дискримінація у випадку розмитих меж кластерів (m=2; R=2)

Тому важливим поняттям дискримінантного аналізу виступає відмова від розпізнання образу для нових об’єктів, коли вони розташовані у зоні високого ризику, тобто біля дискримінантної функції.

У цілому, при виборі ознак, що сприяють відмінностям між кластерами, дискримінантний аналіз добре відображує соціально-економічну реальність.

Питання для самоконтролю

1. Що таке дискримінантний аналіз?
2. У яких практичних ситуаціях можливо застосувати дискримінантний аналі?
3. Як сформулювати основне завдання дискримінантного аналізу?
4. Які вимоги існують для визначення дискримінантної функції?
5. Які проблеми виникають при дискримінантному аналізі?
6. Як визначаються невідомі коефіцієнти дискримінантної функції?
7. Запишіть дискримінантну функцію.
8. Опишіть геометричну інтерпретацію дискримінантного аналізу.
9. За якою формулою розраховується відстань Махаланобіса?
10. Наведіть систему лінійних рівнянь для визначення коефіцієнтів дискримінантної функції.
11. Наведіть формулу дисперсії-коваріації.
12. За якою формулою визначається значення константи С у дискримінантної функції?
13. Які властивості має дискримінантна функція?
14. Що таке зона ризику?

Тема 6. МЕТОД ГОЛОВНИХ КОМПОНЕНТ

6.1. Постановка задачі і математична модель методу головних компонент.

6.2. Найважливіші етапи методу головних компонент.

6.3. Приклад визначення матриці факторних навантажень.

6.4. Визначення, якісне інтерпретування та вимірювання головних компонент.

6.5. Геометрична інтерпретація методу головних компонент.

6.1. Постановка задачі і математична модель методу головних компонент

Одним з ефективних способів зниження розмірності простору ознак є метод головних компонент (МГК). Його сутність полягає у тому, що вихідні ознаки (фактори-симптоми), які безпосередньо спостерігаються під час дослідження, замінюються деякими новими, синтетичними змінними, які є лінійними комбінаціями вихідних факторів-симптомів. Ці нові синтетичні, штучні змінні називаються головними компонентами та мають важливу властивість: вони лінійно незалежні між собою, тобто є ортогональними векторами. У тих випадках, коли можна якісно інтерпретувати головні компоненти, метод головних компонент являє собою міцний інструмент вивчення внутрішніх властивостей соціально-економічних явищ та процесів. Наприклад, такі фактори виробництва, як рівень механізації праці, ступінь автоматизації, витрати на впровадження нової техніки можливо об’єднати та виразити їх вплив через одну головну компоненту. Ця компонента розглядається як узагальнюючий фактор, що характеризує рівень науково-технічного прогресу на виробництві.

Метод головних компонент запропонував Карл Пірсон у 1901 році як окремий метод головних осей, який було розвинуто та доопрацьовано такими вченими, як Х. Хотелінг, Г. Харман, С. Рао. Цей метод досить довго розглядався

як самостійний напрям багатовимірних статистичних методів. У дійсності він є одним з різновидів факторного аналізу. Саме так трактується метод головних компонент (чи компонентний аналіз) сучасними світовими спеціалістами.

Незважаючи на наукове обґрунтування методу головних компонент, його широке використання стримувалось наявністю двох проблем:

1) проблеми розрахункового характеру;

2) складності інтерпретації результатів.

З появою комп'ютерів першу проблему повністю подолано. Дійсно, за допомогою сучасних комп'ютерних програм можна виконати будь-які складні розрахунки. Друга проблема залишається перешкодою під час застосування даного методу.

Загальна ідея МГК полягає у виявленні головних компонент, які описують максимально можливу частку варіації вихідних факторів-симптомів, але при цьому число компонент має бути меншим, ніж число вихідних ознак.

Якщо кількість вихідних ознак **m,** то загальна кількість компонент, яку можна виділити **m.** Але немає необхідності використовувати всі **m** головних компонент. Достатньо розглянути тільки **p** перших компонент, які пояснюють більшу частину дисперсії вихідних факторів-симптомів. Інші **m-p** компонент виключаються з подальшого аналізу як несуттєві, такі, що не містять корисної інформації про досліджувані об'єкти. На практиці число **p** приблизно у 4-5 раз менше за число **m**.

Для запису математичної моделі МГК введемо такі позначення:

i – номер досліджуваного об'єкта (**i**=1,2, …, **n**);

k – номер вихідного фактора –симптома **(k =** 1, 2, 3, …, **m)**

L – номер головної компоненти (**L= 1, 2, …, m).**

Головне рівняння МГК має вигляд:

$$\boldsymbol{Z} = \boldsymbol{A} \cdot \boldsymbol{F}, \qquad (6.1)$$

де $\boldsymbol{Z}$ – матриця стандартизованих значень вихідних факторів-симптомів, розміром **m**x**n;**

$\boldsymbol{F}$ – матриця головних компонент, розміром **m**x**n;**

A – матриця факторних навантажень, розміром **m**x**m.**

Записане матричне рівняння (6.1) можна розглядати як деяке лінійне перетворення змінних $\mathbf{z_{1i}}$,$\mathbf{z_{2i}}$, …, $\mathbf{z_{mi}}$ у змінні $\mathbf{f_{1i}}$,$\mathbf{f_{2i}}$,…, $\mathbf{f_{mi}}$ і, навпаки.

Факторні навантаження, тобто елементи матриці ***A*** позначаються через $\boldsymbol{a_{kL}}$ являють собою звичайні коефіцієнти парної кореляції , які відображують зв'язок між **k**-фактором $\boldsymbol{x_k}$ та **L**-тою головною компонентою $\mathbf{f_L}$.Отже, можна записати, що $\boldsymbol{a_{kL} = r_{kL}}$, (нагадаємо, що $\boldsymbol{r_{kL}}$ – коефіцієнт парної кореляції, який змінюється від -1 до 1). Так само змінюється і коефіцієнти $\boldsymbol{a_{kL}}$. Якщо $\boldsymbol{a_{kL} = 0}$, то зв'язок між $\mathbf{x_k}$ та $\mathbf{x_L}$ відсутній.

Якщо значення $\boldsymbol{a_{kL}}$ за модулем близьке до 1(зазвичай $|a_{kL}| \geq 0{,}7$, зв'язок між **k**-тим фактором і **L**-тою компонентою вважається тісним. У даному випадку вважають, що **k**-тий фактор навантажує **L**-тую компоненту, надає їй якісний зміст.

Систему рівнянь компонентної моделі для **m**–ознак називають **факторним відображенням.**

Наприклад, для m=3 факторне відображення, отримане з рівняння МГК матиме такий вигляд:

$$\begin{cases} z_1 = a_{11}f_1 + a_{12}f_2 + a_{13}f_3 \\ z_2 = a_{21}f_1 + a_{22}f_2 + a_{23}f_3 \\ z_3 = a_{31}f_1 + a_{32}f_2 + a_{33}f_3 \end{cases} \qquad (6.2)$$

У цій системі матриці **Z** і **F** розглядаються як вектори-стовпці, тобто відбувається абстрагування від значень окремих об’єктів (за індексом **i**).

Факторне відображення спроситься, якщо не враховувати навантаження, які незначно відрізняються від нуля. Розглянемо схему спрощеного факторного відображення (табл. 6.1).

Кожній ознаці-симптому відповідно до наведеної схеми властива своя факторна структура – рядки табл. 6.1. Число факторних навантажень, для яких виконується умова $|a_{kL}| \geq 0{,}7$, визначає складність даної ознаки. Так, перший вихідний показник у схемі має складність 3, другий – 2, третій – 1.

Таблиця 6.1

Схема факторного відображення

(х – високі за абсолютною величиною факторні навантаження)

Ознаки-симптоми	Головні компоненти – $\boldsymbol{F_L}$		
	F_1	F_2	F_3
Z_1	х	х	х
Z_2	х	х	
Z_3	х		

У розрізі стовпців таблиці 1 такий підхід дозволяє класифікувати компоненти таким чином. Головна компонента називається **генеральною** , якщо вона тісно пов'язана зі всіма факторами-симптомами. Компонента називається **груповою,** якщо вона тісно пов'язана з більш, ніж однією ознакою. Наприклад, у табл. 6.1 перша головна компонента F_1 називається генеральною, а F_2 – груповою.

Незважаючи, що замість **m** ознак, розглядається **p** головних компонент (тобто **p< m),** досвід практичних досліджень вказує, що цього цілком достатньо для подальшого аналізу внутрішніх загальних факторів. Втрати інформації незначні, а стиск простору досягається приблизно у 4-5 раз.

6.2. Найважливіші етапи методу головних компонент

Процедура компонентного аналізу складається з п'яти важливих етапів:

1. Формування матриці **X** вихідних ознак-симптомів
2. Стандартизація ознак, побудова матриці **Z**
3. Розрахунок кореляційної матриці **r,** яка відображує зв'язки між вихідними даними
4. Визначення матриці факторних навантажень **A**
5. Виділення, вимірювання та інтерпретація головних компонент

Розглянемо зміст наведених етапів МГК.

Початковим і найважливішим етапом процедури є перший етап, формування матриці вихідних даних. Основою формування такої матриці слугує

теоретичний якісний аналіз сутності явищ, які вивчаються. Інформація про всі значення змінних подається у вигляді матриці **X**, розміром **mxn.** Означена матриця X має **m** рядків (за числом ознак) та **n** стовпців (за кількістю об'єктів). Елемент матриці $\mathbf{x_{ki}}$ – це значення **k**-ої ознаки у **i**-ого елемента сукупності.

Оскільки у соціально-економічних дослідженнях ознаки $\mathbf{x_{ki}}$ неоднорідні, мають різні одиниці вимірювання, то для проведення розрахунків їх необхідно стандартизувати. Стандартизація полягає у центруванні та нормуванні вихідних даних за формулою (2.5).

Нагадаємо, що розмірність матриці **Z** така сама, як і в матриці **X**, тобто **mxn.**

У статистиці можна довести, що для стандартизованих даних виконуються співвідношення (2.6), тобто :

$$\bar{z}_k = 0 \ ; \ \sigma^2_{z_k} = 1 \ . \qquad (6.3)$$

Розглянемо розкладання загальної дисперсії факторів-симптомів на доданки. З цією метою розглянемо дисперсію **k**-ої стандартизованої ознаки. Відповідно до визначення дисперсії та формули (6.3) можна записати, що:

$$\sigma^2_{z_k} = \frac{\sum(z_{ki}-\bar{z}_k)^2}{n} = \frac{1}{n}\sum z^2_{ki}=1 \ . \qquad (6.4)$$

Відповідно до формули факторного відображення (6.2), а також беручи до уваги, що головні компоненти некорельовані, нормовані величини, для яких справедливі рівняння:

$$\frac{1}{n}\sum f^2_{ij} = 1 \ \ \frac{1}{n}\sum f_{ik}f_{iL} = r_{fkf_L} = 0,$$

формулу (6.4) можна записати таким чином:

$$\sigma^2_{z_k} = a^2_{k1} + a^2_{k2} + \cdots + a^2_{kL} + \cdots + a^2_{km} = 1 \ . \qquad (6.5)$$

Ліворуч і праворуч у співвідношенні (6.5) записана повна дисперсія **k**-ої стандартизованої ознаки, а у середині – частки повної дисперсії, що пояснюються головними компонентами. З цього співввідношення бачимо, що

дисперсію кожної ознаки можна розкласти на складові, які являють собою частки повної дисперсії, що належать до відповідних головних компонент.

Підсумуємо дисперсії за всіма ознаками:

$$\sum_{k=1}^{m} \sigma_{z_k}^2 = \sum_{k=1}^{m} a_{k1}^2 + \sum_{k=1}^{m} a_{k2}^2 + \cdots + \sum_{k=1}^{m} a_{kL}^2 + \cdots + \sum_{k=1}^{m} a_{km}^2 = m\,. \quad (6.6)$$

Таким чином, загальна дисперсія всіх ознак, що дорівнює **m,** може бути розкладена на частки, обумовлені кожною головною компонентою. Згідно до вираження (6.6) абсолютний внесок кожної компоненти, наприклад **L**-тої, у пояснення дисперсії всіх факторних ознак розраховується так:

$$v_L = \sum_{k=1}^{m} a_{kL}^2. \quad (6.7)$$

Співвідношення (6.6) та (6.7) показують, що сума внесків усіх головних компонент дорівнює сумі дисперсій всіх ознак, тобто:

$$\sum_{L=1}^{m} v_L = \sum_{L=1}^{m} \sum_{k=1}^{m} a_{kL}^2 = m. \quad (6.8)$$

Охарактеризуємо третій етап процедури МГК – побудову кореляційної матриці. Велике значення у МГК під час розрахунків факторних навантажень відіграє матриця парних кореляцій. Розглянемо її основні властивості.

Елементами кореляційної матриці є коефіцієнти парної кореляції між вихідними факторними ознаками:

$$r_{kL} = \frac{\sum_i^n (x_{ki} - \bar{x}_k)(x_{Li} - \bar{x}_L)}{n\sigma_k \sigma_L}\,, \quad (6.9)$$

де $\bar{x}_k$, $\bar{x}_L$ – середні значення **k** – го та **L** – го факторів;

σ_k , σ_L – середньоквадратичні значення **k** – го та **L** – го факторів.

Враховуючи властивості стандартизованих даних (6.3), можна довести, що

$$r_{kL} = \frac{1}{n} \sum_i^n z_{ki}\, z_{Li}.$$

Звідси у матричному запису

$$r_x = \frac{1}{n} Z Z^T \quad (6.10)$$

де Z^T - матриця, транспонована до матриці Z.

Оскільки коефіцієнт кореляції кожної ознаки з самим з собою дорівнює одиниці ($r_{kk} = r_{LL} = 1$), та $r_{kL} = r_{Lk}$, то кореляційна матриця є симетричною матрицею, на головній діагоналі якої розташовані 1. Матриця $\mathbf{r_X}$ має розмірність **mx m**, тобто являє квадратну симетричну матрицю (на основі властивостей транспонованої матриці та добутку матриць).

Після стандартизації ознак та побудови кореляційної матриці, переходять до четвертого етапу МГК – визначенню матриці факторних навантажень **A**.

Основну модель МГК (6.1) можна перетворити до такого виду:

$$\boldsymbol{F=A^{T}\cdot Z} \quad . \tag{6.11}$$

Отже, матрицю факторних навантажень можна розглядати як матрицю, зворотного лінійного перетворення змінних $\mathbf{z_1, z_2, \dots, z_m}$ у змінні $\mathbf{f_1, f_2, \dots, f_m}$. Вона визначається на підставі таких вимог:

1) усі головні компоненти лінійно незалежні (ортогональні);
2) стандартизовані;
3) перша компонента має пояснювати максимальну частку дисперсії вихідних змінних, друга компонента – максимальну частку дисперсії вихідних змінних, які залишилися після першої компоненти і т.д.

Відповідно до теорії матричної алгебри таким вимогам відповідає матриця **A**, стовпці якої визначаються таким чином:

$$\boldsymbol{a_L = \sqrt{\lambda_L}\cdot u_L,} \tag{6.12}$$

де $\boldsymbol{\lambda_L}$ – L-оє власне (характеристичне) значення матриці $\mathbf{r_X}$**;**

$\mathbf{u_L}$ – L-ий власний вектор кореляційної матриці $\mathbf{r_X}$, який має одиничну довжину.

Для знаходження власних значень та відповідних до них власних векторів матриці $\mathbf{r_X}$ необхідно розв'язати відносно **U** матричне рівняння:

$$\boldsymbol{(r_X - \lambda E)\cdot U = 0\ .} \tag{6.13}$$

Означене рівняння називається **характеристичним**. Воно однорідне і має нульове рішення тільки тоді, коли характеристичний визначник дорівнює нулю, тобто:

$$|r_X - \lambda E| = 0\ , \tag{6.14}$$

де **E –** одинична матриця, розміром **mx m**.

Запишемо характеристичну матрицю у розгорнутому вигляді:

$$\left|r_X - \lambda \cdot E\right| = \begin{pmatrix} 1-\lambda & r_{12} & \dots & r_{1m} \\ r_{21} & 1-\lambda & \dots & r_{2m} \\ & \dots\dots\dots\dots & & \\ r_{m1} & r_{m2} & \dots & 1-\lambda \end{pmatrix}. \tag{6.15}$$

Визначник даної матриці є многочленом ступеня **m** відносно **λ.** Враховуючи формулу (6.14) характеристичний многочлен визначається так:

$$|r_X - \lambda \cdot E| = (-1)^m \lambda^m + (-1)^{m-1} g_1 \lambda_1^{m-1} + \dots + g_m = 0, \tag{6.16}$$

де $g_1 = t_r(r_X) = m$ – слід матриці $\mathbf{r_X}$, тобто сума її діагональних елементів;

$g_m = |r_X|$ – визначник матриці $\mathbf{r_X}$;

λ – корінь поліному (характеристичний корінь).

Відомо, що характеристичні корні (чи власні значення) мають такі властивості:

1) $\sum_L^m \lambda_L = t_r(r_X) = g_1 = m$
2) $\lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \dots \cdot \lambda_m = |r_X| = g_m$.

Оскільки власні вектори $\mathbf{U_L}$ матриці $\mathbf{r_X}$ визначаються з точністю до постійного множника **C**, то для забезпечення одиничної довжини кожного власного вектора $\mathbf{U_L}$ нормований множник **C** доцільно прийняти таким, що дорівнює:

$$C = \frac{1}{\sqrt{\sum_1^m u_{kL}^2}}. \tag{6.17}$$

Дійсно, під час нормування $\mathbf{U_L}$ за формулою (6.17) скалярний добуток (тобто сума квадратів елементів) дорівнює:

$$U_L^T U_L = \sum_{L=1}^{m} (C \cdot U_{kL})^2 = C^2 \sum_{L=1}^{m} U_{kL}^2 = \frac{\sum_{L=1}^{m} U_{kL}^2}{\sum_{k=1}^{m} U_{kL}^2} = 1. \qquad (6.18)$$

Звідси норма власного вектора: $|U_L| = 1$.

З урахуванням вимоги (6.12), отримуємо таку формулу для розрахунку факторних навантажень за значеннями власних векторів матриці $\mathbf{r_X}$:

$$a_{kL} = U_{kL} \sqrt{\frac{\lambda_L}{\sum U_{kL}^2}}. \qquad (6.19)$$

Сукупність значень $\boldsymbol{a_{kL}}$ формує матрицю факторних навантажень:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}. \qquad (6.20)$$

Означена матриця має розмір **mxm** і використовується у основному рівнянні методу головних компонент (6.1).

Якщо через $\boldsymbol{\Lambda^{1/2}}$ позначити діагональну матрицю, розміром **mxm,**

$$\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & \sqrt{\lambda_m} \end{bmatrix}, \qquad (6.21)$$

а через **U** – матрицю власних векторів кореляційної матриці $\mathbf{r_X}$, яка має одиничну довжину, то з формули (6.19) слідкує матричне рівняння:

$$\boldsymbol{A = U \cdot \Lambda^{1/2}.} \qquad (6.22)$$

Наведена формула дозволяє виявити властивості матриці факторних навантажень **A**. Для цього розглянемо добуток $A^T A$:

$$A^T \cdot A = (U \cdot \Lambda^{1/2})^T \cdot U \cdot \Lambda^{1/2} = \Lambda^{1/2} \cdot U^T \cdot U \cdot \Lambda^{1/2} = \Lambda^{1/2} \Lambda^{1/2} = \Lambda. \quad (6.23)$$

Це означає, що сума квадратів елементів **L**-го стовпця матриці **A** дорівнює відповідному даному стовпцю власному значенню $\boldsymbol{\lambda_L}$, тобто

$$\lambda_L = \sum_k^m a_{kL}^2. \qquad (6.24)$$

У свою чергу, сума добутків елементів двох різних стовпців, наприклад, **k** і **L** дорівнює нулю. Як було доведено раніше (формула (6.7)), вираз праворуч

формули (6.24) визначає внесок **L**-ої головної компоненти у пояснення варіації вихідних факторів. Отже, величина власних значень матриці $\mathbf{r_X}$ і визначає важливість кожної головної компоненти. При цьому має місце співвідношення:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_m. \quad (6.25)$$

Зауважимо, що завжди найважливішою є перша головна компонента, потім друга і так далі. Останні головні компоненти незначущі та виключаються з аналізу.

Надалі розглянемо добуток $\mathbf{A \cdot A^T}$:

$$A \cdot A^T = U \cdot \Lambda^{1/2} \cdot (U \cdot \Lambda^{1/2})^T = U \cdot \Lambda^{1/2} \cdot \Lambda^{1/2} \cdot U^T = U \cdot \Lambda \cdot U^T. \quad (6.26)$$

З теорії матричної алгебри відомо, що будь яку симетричну, у тому числі й матрицю $\mathbf{r_X}$, можна привести до діагонального вигляду за допомогою матриць власних векторів. Як наслідок, діагональну матрицю $\mathbf{\Lambda}$ можна подати таким чином:

$$\Lambda = U^T \cdot r_X \cdot U. \quad (6.27)$$

Якщо підставимо співвідношення (6.27) у формулу (6.26), то отримуємо:

$$A \cdot A^T = U \cdot \Lambda \cdot U^T = U \cdot U^T \cdot r_X \cdot U \cdot U^T = r_X\,. \quad (6.28)$$

Це означає, що сума квадратів елементів будь якого **i** рядка дорівнює 1, що узгоджується з формулою (6.5).

Сума добутків **k**-го та **L**-го рядків дорівнює парному коефіцієнту кореляції між **k**-ою та **L**-ою вихідними ознаками – $\mathbf{r_X}$.

Співвідношення (6.28) вказує, як на основі матриці факторних навантажень **A** відтворити кореляційну матрицю $\mathbf{r_X}$.

6.3. Приклад визначення матриці факторних навантажень

Оскільки перші три етапи МГК (формування матриці вихідних даних, стандартизація, побудова кореляційної матриці) відповідно до статистичної точки зору особливих труднощів не мають, то зосередимся на четвертому етапі, коли необхідно побудувати матрицю факторних навантажень. З цією метою

розглянемо приклад, який демонструє розрахунок цієї матриці на підставі заданої кореляційної матриці **r_X:**

$$r_X = \begin{bmatrix} 1 & 0{,}8 & 0{,}2 \\ 0{,}8 & 1 & 0{,}6 \\ 0{,}2 & 0{,}6 & 1 \end{bmatrix}.$$

Це симетрична матриця, на головній діагоналі якої розташовані 1; 0,8 – коефіцієнт парної кореляції між 1 і 2 ознаками; 0,2 – коефіцієнт парної кореляції між 1 і 3 ознаками; 0,6 – коефіцієнт парної кореляції між 2 і 3 ознаками .

Розглянемо задачу виділення трьох головних компонент за трьома факторними ознаками (m=3), тому матриця $\mathbf{r_X}$ має розмір 3x3.

Для того, щоб розрахувати факторні навантаження за формулою (6.19) необхідно знайти власні числа матриці $\mathbf{r_X}$. Знаходження власних чисел $\boldsymbol{\lambda_L}$ потребує запису характеристичного визначника:

$$|r_X - \lambda E| = \begin{vmatrix} 1-\lambda & 0{,}8 & 0{,}2 \\ 0{,}8 & 1-\lambda & 0{,}6 \\ 0{,}2 & 0{,}6 & 1-\lambda \end{vmatrix} = 0.$$

Нагадаємо, що $\boldsymbol{E}$ – одинична матриця, розміром 3x3, у якої змінюються тільки діагональні елементи, а інша залишаються без змін у порівнянні з кореляційною матрицею.

Для отримання характеристичного многочлена (формула 6.16) можна використати правило трикутника (методом розкриття визначника 3 порядку):

$$(1-\lambda)^3 + 0{,}8 \cdot 0{,}6 \cdot 0{,}2 + 0{,}8 \cdot 0{,}6 \cdot 0{,}2 - 0{,}2 \cdot 0{,}2 \cdot (1-\lambda) - (1-\lambda) \cdot 0{,}6 \cdot 0{,}6 - 0{,}8 \cdot 0{,}8 \cdot (1-\boldsymbol{\lambda}) = 0$$

Якщо розкрити дужки та привести подібні члени, можна отримати кубічне рівняння відносно $\boldsymbol{\lambda}$:

$$-\lambda^3 + 3\lambda^2 + 1{,}96\lambda + 0{,}152 = 0$$

Знайдене характеристичне рівняння надає три значення $\boldsymbol{\lambda}$. Треба зауважити, що у практиці рівняння вищі за другий ступінь розв'язуються за допомогою спеціальних комп'ютерних програм.

Корнями даного рівняння є такі значення:

$$\lambda_1 = 2{,}10\,;\ \ \lambda_2 = 0{,}81;\ \lambda_3 = 0{,}09$$

Отримані власні числа матриці $\mathbf{r_X}$ мають дві властивості, про які вже велось раніше. Перевіримо їх.

По-перше, сума власних чисел дорівнює сліду матриці $\mathbf{r_X}$:

$$\lambda_1 + \lambda_2 + \lambda_3 = 2{,}10 + 0{,}81 + 0{,}09 = t_r(r_X) = g_1 = m = 3,$$

де t_r – слід матриці, тобто сума діагональних елементів.

По-друге, добуток власних чисел дорівнює визначнику матриці $\mathbf{r_X}$ або вільному члену характеристичного рівняння (6.16):

$$\lambda_1 \cdot \lambda_2 \cdot \lambda_3 = 2{,}10 \cdot 0{,}81 \cdot 0{,}09 = |r_X| = g_3 \approx 0{,}152$$

Для того, щоб розрахувати факторні навантаження за формулою (6.19), необхідно знати відповідні власні вектори. Для цього скористуємось формулою (6.13). Замість $\boldsymbol{\lambda}$ вписуємо певне значення, тобто наприклад, $\boldsymbol{\lambda}_1$, потім $\boldsymbol{\lambda}_2$, $\boldsymbol{\lambda}_3$.Розрахуємо власний вектор, який відповідає першому власному значенню $\boldsymbol{\lambda}_1$=2,10. Отримаємо таку систему рівнянь:

$$\begin{cases} (1\text{-}2{,}1)\cdot U_1 + 0{,}8\cdot U_2 + 0{,}2\cdot U_3 = 0 \\ 0{,}8\cdot U_1 + (1\text{-}2{,}1)\cdot U_2 + 0{,}6\cdot U_3 = 0 \\ 0{,}2\cdot U_1 + 0{,}6\cdot U_2 + (1\text{-}2{,}1)\cdot U_3 = 0 \end{cases}.$$

З метою спрощення розрахунків приймемо U_3=1. Така дія припустима, оскільки власні вектори визначаються з точністю до постійного множника **C** (який є визначником системи, що прирівняна до нуля). В результаті отримуємо таку систему рівнянь:

$$\begin{cases} 0{,}8U_1 - 1{,}1\cdot U_2 = -0{,}6 \\ 0{,}2\cdot U_1 + 0{,}6\cdot U_2 = 1{,}1. \end{cases}$$

Щоб її розв’язати, друге рівняння множимо на -4, та додаємо обидва рівняння:

$$-3{,}5\ U_2 = -5{,}0.$$

Звідси, U_2=1,429, а U_1=1,213: тобто розв’язуємо перше рівняння системи ($0{,}8U_1 = 1{,}1\cdot 1{,}429 - 0{,}6 = 0{,}971$.).

Таким чином, отримано координати власного вектора, який відповідає першому власному значенню $\lambda_1 = 2{,}10$.

Аналогічно визначаються власні вектори, які відповідають власним числам $\boldsymbol{\lambda}_2$ та $\boldsymbol{\lambda}_3$.

Розрахуємо факторні навантаження першої головної компоненти за формулою (6.19). Всі розрахунки подано у табл. 6. 2.

Зауважимо, що величина $\sqrt{\frac{\lambda_1}{\sum U_{k1}^2}} = \sqrt{\frac{2{,}100}{4{,}510}} = 0{,}6827$ – постійна для першої головної компоненти. Якщо її помножити на числа з першого стовпця, отримуємо факторні навантаження:

1,213x0,6827=0,829; 1,429x0,6827=0,975

Таблиця 6.2

Розрахунок факторних навантажень першої головної компоненти

U_{k1}	U_{k1}^2	$a_{k1} = U_{k1}\sqrt{\frac{\lambda_1}{\sum U_{k1}^2}}$	a_{k1}^2
1,213	1,471	0,829	0,685
1,429	2,039	0,975	0,951
1,000	1,000	0,683	0,464
Усього	4,510	-	2,100

У останньому стовпці табл.6.2 наведені суми квадратів факторних навантажень першої головної компоненти. Як бачимо, їх сума дорівнює першому власному значенню $\lambda_1 = 2{,}100$ (формула 6.23).

Аналогічно розраховуються факторні навантаження на другу та третю компоненти. Вони подані у вигляді матриці факторних навантажень **A**, розміром 3x3.

$$\mathrm{A} = \begin{bmatrix} 0{,}829 & -0{,}532 & 0{,}170 \\ 0{,}975 & -0{,}052 & -0{,}221 \\ 0{,}683 & 0{,}723 & 0{,}109 \end{bmatrix}$$

Наведена матриця приблизно задовольняє властивостям, які виражені формулами (6.23) та (6.28). Доведемо це. Для цього знайдемо добуток

транспонованої матриці факторних навантажень та звичайної матриці факторних навантажень (формула 6.23):

$$A^T \cdot A = \begin{bmatrix} 0{,}829 & 0{,}975 & 0{,}683 \\ -0{,}532 & -0{,}052 & 0{,}723 \\ 0{,}170 & -0{,}221 & 0{,}109 \end{bmatrix} \cdot \begin{bmatrix} 0{,}829 & -0{,}532 & 0{,}170 \\ 0{,}975 & -0{,}052 & -0{,}221 \\ 0{,}683 & 0{,}723 & 0{,}109 \end{bmatrix} \approx \begin{bmatrix} 2{,}1 & 0 & 0 \\ 0 & 0{,}81 & 0 \\ 0 & 0 & 0{,}09 \end{bmatrix}.$$

За рахунок арифметичних округлень існують деякі розбіжності.

Розглянемо добуток матриці факторних навантажень та транспонованої матриці факторних навантажень, тобто перевіряємо формулу (6.28):

$$\mathrm{A} \cdot \mathrm{A}^{\mathrm{T}} = \begin{bmatrix} 0{,}829 & -0{,}532 & 0{,}170 \\ 0{,}975 & -0{,}052 & -0{,}221 \\ 0{,}683 & 0{,}723 & 0{,}109 \end{bmatrix} \cdot \begin{bmatrix} 0{,}829 & 0{,}975 & 0{,}683 \\ -0{,}532 & -0{,}052 & 0{,}723 \\ 0{,}170 & -0{,}221 & 0{,}109 \end{bmatrix} \approx \begin{bmatrix} 1 & 0{,}8 & 0{,}2 \\ 0{,}8 & 1 & 0{,}6 \\ 0{,}2 & 0{,}6 & 1 \end{bmatrix}.$$

Співвідношення (6.28) наочно демонструє як за матрицею факторних навантажень А відтворити кореляційну матрицю $\mathbf{r_X}$.

Отже, на 4 етапі побудовано матрицю факторних навантажень та перевірено її властивості.

Елементи матриці А дозволяють проаналізувати факторну структуру кожного фактора-симптома і кожної головної компоненти. Розглянемо рядки матриці **А**: 1 р. – 2 (0,829; -0,532)

2 р. – 1 (0,975)

3 р. – 2 (0,683; 0,723).

Складність першого та третього факторів-симптомів дорівнює 2, а другого - один (визначено за високими факторними навантаженнями).

Аналіз стовпців довів, що перша головна компонента є генеральною, оскільки вона тісно пов’язана зі всіма ознаками; друга компонента – групова, оскільки тісно пов’язана з першим та третім факторами. Третя компонента слабо пов’язана з факторам, тому її можна виключити з подальшого аналізу.

6.4. Визначення, якісне інтерпретування та вимірювання головних компонент

Останнім, п’ятим етапом процедури МГК, є виділення, інтерпретація та вимірювання головних компонент. На цьому етапі виділяють найбільш суттєві

компоненти, як правило перші. Для цього потрібно проаналізувати величину дисперсії, яка пояснюється даною компонентою, а також накопичену варіацію, яка описується декількома компонентами. Важливість чи внесок **L**-ої компоненти визначається відношенням:

$$d_L = \frac{\lambda_L}{m} \cdot (100). \qquad (6.29)$$

Записане співвідношення виражається у % та показує частку варіації всіх факторів-симптомів, яка описується певною компонентою. Щоб знайти внесок перших ***p***-компонент, підраховують суму:

$$\sum_{L-1}^{p} d_L = \sum_{L=1}^{p} \frac{\lambda_L}{m}. \qquad (6.30)$$

Досвід практичних розрахунків вказує, що для подальшого дослідження доцільно використовувати ті головні компоненти, для яких $\mathbf{d_L > 10\ \%}$. Знайдемо внесок компонент за нашим прикладом:

$$d_1 = \frac{\lambda_1}{m} = \frac{2{,}100}{3} = 0{,}700$$

$$d_2 = \frac{\lambda_2}{m} = \frac{0{,}810}{3} = 0{,}270$$

$$d_3 = \frac{\lambda_3}{m} = \frac{0{,}090}{3} = 0{,}030.$$

Як бачимо, сума всіх внесків складає 100 %. Це означає, що всі **m** – головних компонент описують всю варіацію вихідних факторів-симптомів. Вочевидь, що немає сенсу використовувати всі три компоненти, оскільки остання пояснює менш 10 % варіації. У даному прикладі можна обмежитися тільки першими двома компонентами, що пояснюють 97 % варіації вихідних ознак.

Більш складним та відповідальним моментом всього методу є якісна інтерпретація виділених головних компонент. Вона базується на дослідженні матриці факторних ознак..

Розглянемо інтерпретацію головних компонент на такому прикладі. Нехай для 7 показників технічного прогресу виділено три головні компоненти, які мають факторні навантаження, що наведені у таблиці 6.3.

Таблиця 6.3

Факторні навантаження головних компонент

Показник (фактор-x_k)	a_{k1}	a_{k2}	a_{k3}
1.Фондоозброєність праці	0,84	0,09	0,04
2. Рівень механізації	0,92	0,15	0,11
2. Рівень автоматизації	0,87	0,26	0,13
3. Впровадження нової технології	0,62	0,38	0,76
5. Впровадження комп'ютерних технологій	0,46	0,65	0,24
6. Амортизація обладнання	0,72	0,88	0,31
7. Впровадження нової продукції	0,48	0,21	0,68

Для економічної інтерпретації головних компонент необхідно звернути увагу на високі факторні навантаження. За їх значеннями першої головної компоненти, її можна інтерпретувати як синтетичну змінну, яка відображує технічний рівень виробництва.

Друга головна компонента відображує вдосконалення засобів виробництва.

Третя головна компонента – відображує удосконалення технологічних процесів.

Логічно, що різні дослідники надають різну інтерпретацію отриманих головних компонент, що вносить елементи суб'єктивізму у цей метод.

Головні компоненти можуть поєднувати різні фактори, які можуть бути економічно не пов'язані між собою, але тісно переплітатися з факторами-симптомами. У таких випадках інтерпретація головних компонент дуже ускладнюється.

Вимірювання головних компонент для окремих об'єктів дослідження можна виконувати відповідно до основного рівняння методу (формула 6.1). При цьому вектор **F** виражається через зворотну матрицю (рівняння 6.11).

Матриця **A** для всіх **m**–головних компонент зворотна. На практиці зазвичай розраховують не всі, тільки ***p***-перших компонент. Тому після

нескладних перетворень формул (6.11) та (6.23), отримуємо такий вираз матриці головних компонент:

$$\boldsymbol{F = \Lambda^{-1} A^T Z}\ , \qquad (6.31)$$

де **Λ** – діагональна матриця власних значень кореляційної матриці $\mathbf{r_X}$;

A – матриця факторних навантажень;

Z – матриця стандартизованих факторів.

Значення будь-якої головної компоненти, наприклад, **L**-ої для **i**-об'єкта, може бути розраховане за наступною формулою, яка випливає з вираження (6.31):

$$f_{Li} = \sum_{k=1}^{m} \frac{a_{kL}}{\lambda_L} \cdot z_{ki}\ , \qquad (6.32)$$

де **k** та **L** змінюються у межах від 1 до **m**, а i –від 1 до **n.**

За даними табл. 6.3 та заданими значеннями стандартизованих ознак, визначимо значення першої головної компоненти для двох підприємств, яка характеризує технічний рівень виробництва.

У нашому прикладі $\lambda_1 = 3{,}62$.

Таблиця 6.4

Розрахунок значень першої головної компоненти для двох підприємств

$\mathbf{X_i}$	$\frac{a_{k1}}{\lambda_1}$	Підприємство 1		Підприємство 2	
		z_{k1}	$\frac{a_{k1} z_{k1}}{\lambda_1}$	z_{k2}	$\frac{a_{k1} z_{k2}}{\lambda_1}$
X_1	0,232	0,846	0,196	-0,113	-0,026
X_2	0,254	0,428	0,109	0,326	0,083
X_3	0,240	0,242	0,058	0,553	0,133
X_4	0,171	0,289	0,049	1,067	0,182
X_5	0,127	-0,150	-0,019	0,111	0,014
X_6	0,199	0,211	0,042	-0,065	-0,013
X_7	0,133	-0,407	-0,054	0,455	0,061
Усього	-	-	0,381	-	0,434

Аналіз даних табл.6.4 вказує, що значення першої головної компоненти на другому підприємстві вище за перше підприємство(оскільки 0,434>0,381). Тому

можна вважати, що друге підприємство має більш високий рівень технічного розвитку. Аналогічно можна виміряти значення першої компоненти на всіх об'єктах дослідження, порівняти їх.

Саме так вимірюються значення другої та третьої головних компонент. Це дозволяє дослідити та кількісно охарактеризувати інші напрямки науково-технічного прогресу, удосконалити засоби виробництва (2 головна компонента) та удосконалити технологічні процеси (3 головна компонента).

6.5. Геометрична інтерпретація методу головних компонент

Геометрична інтерпретація головних компонент базується на їх властивостях. Нагадаємо ці властивості: по-перше, головні компоненти є лінійними комбінаціями вихідних факторів–симптомів; по-друге, ортогональні (взаємонезалежні); по-третє, перша головна компонента пояснює максимум загальної дисперсії всіх ознак, друга – максимум з того, що залишилося після першої і т.д.

Перетворення вихідних факторів у головні компоненти зводиться до повороту координат осей таким чином, щоб нова координатна система мала такі ж самі властивості, як й попередня і про які вже йшла мова. Якщо факторні ознаки, відповідні до окремих об'єктів, зобразити у вигляді кореляційної хмари, то перетворення вихідних факторів у головні компоненти буде відповідати такому повороту системи координат, коли нові осі, орієнтовані по великій і малій осях еліпса кореляційної хмари.

У простішому випадку, коли розглядаються тільки два фактори-симптоми (m=2) всю процедуру методу головних компонент можна зобразити графічно (рис. 6.1).

У системі координат $\mathbf{X_1 0 X_2}$ еліпс об'єднує об'єкти, що вивчаються. Стандартизація вихідних даних (2 етап процедури) являє собою перенесення початку координат у центр кореляційного еліпсу (центрування) та деякий стиск (нормування). Лінійне перетворення стандартизованих даних у головні компоненти (етапи 3 і 4) полягає у повороті координатних осей на кут $\boldsymbol{\alpha}$, при

якому нова координатна система має вищеназвані властивості. При цьому кореляційна матриця $\mathbf{r_x}$ за допомогою ортогонального перетворення приводиться у діагональну матрицю **Λ.** Це означає, що нова вісь абсцис, яка відповідає першій головній компоненті ($\boldsymbol{\lambda_1}$), орієнтована по першій головній осі кореляційного еліпса, а вісь ординат, яка відповідає другій головній компоненті ($\boldsymbol{\lambda_2}$) – перпендикулярна до неї і орієнтована по другій головній осі еліпсу. Вочевидь, що максимальний розкид точок (дисперсія) спостерігається по першій головній осі еліпсу, а мінімальний – по другій головній осі ($\boldsymbol{\lambda_1 > \lambda_2}$). При чому осі ортогональні між собою. Вимірювання головних компонент (5 етап процедури) є вираженням координат усіх точок у новій системі координат $\mathbf{f_1 0 f_2}$.

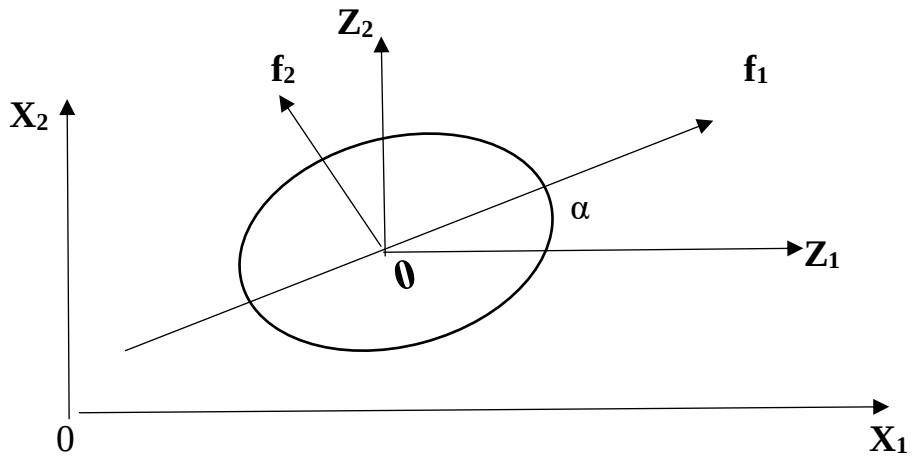

Рис. 6.1 Геометрична інтерпретація методу головних компонент

Відмітимо, що МГК – це один з найбільш доведених та математично обґрунтованих методів, який добре описує дисперсію вихідних ознак за допомогою нових штучних змінних. Його використання дозволяє отримати стислий опис досліджуваних об’єктів без значних втрат інформації про їх властивості. При цьому побудовані змінні – головні компоненти, ортогональні між собою, що сприяє їх подальшому використанню у статистичному моделюванні, наприклад, як фактори регресійної моделі у регресійно-кореляційному аналізі. Успіх подібних досліджень у соціально-економічній сфері залежить від можливостей чіткої та якісно з’ясованої інтерпретації виділених головних компонент.

Питання для самоконтролю

1. У чому головна ідея методу головних компонент?
2. Які проблеми розвитку має метод головних компонент?
3. Запишіть та охарактеризуйте математичну модель методу головних компонент.
4. Що таке факторне відображення?
5. Як розраховуються факторні навантаження?
6. Яка головна компонента називається генеральною?
7. Яка головна компонента називається груповою?
8. З яких етапів складається процедура компонентного аналізу?
9. Як відбувається стандартизація вихідних даних?
10. Які властивості мають стандартизовані дані?
11. Розкладіть загальну дисперсію факторів-симптомів на доданки.
12. Як будується кореляційна матриця у факторному аналізі?
13. Як визначити матрицю факторних навантажень?
14. Які вимоги ставлять до матриці факторних навантажень?
15. Запишіть та охарактеризуйте характеристичне рівняння.
16. Які властивості мають характеристичні корні?
17. Запишіть формулу нормованого множника.
18. Як можна відтворити кореляційну матрицю на основі матриці факторних навантажень?
19. Як оцінюється внесок компоненти у поясненні вихідної інформації?
20. Запишіть рівняння матриці головних компонент.
21. За якою формулою розраховується значення головної компоненти?
22. Охарактеризуйте властивості головних компонент.
23. Як надати геометричну інтерпретацію методу головних компонент?

Тема 7. ФАКТОРНИЙ АНАЛІЗ

7.1. Математична модель факторного аналізу.

7.2.Розкладання дисперсії у факторному аналізі.

7.3. Основні етапи факторного аналізу.

7.4. Методи визначення факторного рішення.

7.5. Інтерпретація факторів і обертання системи координат.

7.6. Вимірювання факторів.

7.1. Математична модель факторного аналізу

Засновником факторного аналізу вважається Ф. Гальтон та Ч. Спирмен. Їх праці розвинули С. Барт, К. Пирсон, Г. Томсон та інші.

Факторний аналіз розглядається у широкому та вузькому сенсі.

У широкому сенсі факторний аналіз – це сукупність методів і моделей, які орієнтовані на виявлення, оцінювання, аналіз та використання латентних показників за інформацією про їх зовнішній прояв. Такими методами є метод головних компонент, багатовимірне шкалювання, канонічний факторний аналіз та безпосередньо факторний аналіз у вузькому сенсі.

В узькому сенсі факторний аналіз охоплює методи виявлення, оцінки та інтерпретації латентних показників, які пояснюють кореляційні зв’язки між факторами-симптомами.

Головна ідея саме факторного аналізу полягає в описанні деякої множини непрямих, безпосередньо невимірюваних ознак (факторів-симптомів) меншим числом максимально інформативних змінних, які відображають сутність досліджуваного явища чи об’єкта. Отримані змінні є лінійними ортогональними комбінаціями вихідних ознак, що дуже схоже з методом головних компонент. Потрібно відмітити, що ортогональність – це частковий випадок взаємного розташування векторів і тому факторне рішення може відрізнятися від ортогонального. Це так зване ***косокутне рішення***, яке спостерігається у випадку корельованих факторів. Вочевидь, що корельованість факторів краще відповідає

реальній економічній дійсності, оскільки на практиці важко виділити абсолютно незалежні ознаки. Наприклад, технічний рівень виробництва та рівень концентрації виробництва неможливо вважати повністю незалежними. Тому підходи щодо факторного аналізу мають більш загальний характер порівняно з методом головних компонент, хоча класична теорія розроблена саме для ортогональних векторів.

Зауважимо, що загальними для обох методів (факторного та компонентного) є методи обробки вихідної інформації, які ґрунтуються на розв'язанні характеристичних рівнянь і знаходженні власних значень та власних векторів кореляційної матриці. Однак на відміну від методу головних компонент, у факторному аналізі потрібно розрахувати редуковану кореляційну матрицю, на головній діагоналі якої розташовані значення спільностей замість одиниць.

Принциповою відмінністю методу головних компонет та факторного аналізу є те, що МГК має виділити компоненти, що пояснюють максимальну варіацію вихідних ознак, а факторний аналіз має визначити такі лінійні комбінації (загальні фактори), які максимально відтворюють кореляційні зв'язки між вихідними ознаками. Останнє можливе через обертання осей і подання знайденого факторного рішення в нових факторних навантаженнях. Якщо метод головних компонент має одне обертання, то факторний аналіз має безліч обертань залежно від мети дослідження.

Основним принципом обертання є принцип досягнення простої структури, яка дозволяє надати зрозумілу якісну інтерпретацію отриманим загальним факторам.

У факторному аналізі оцінюється вплив не тільки загальних факторів, але й вплив специфічних характерних причин. Наприклад, якість та рівень заглибленості вугілля в промисловості.

Для запису математичної моделі факторного аналізу введемо позначення:

i – номер досліджуваного об'єкта (**i**=1,2, …, **n**);

k – номер вихідного фактора –симптома **(k =** 1, 2, 3, …, **m);**

L – номер загального фактора (**L= 1, 2, …, m)**

$\mathbf{y_{ki}}$ – значення **k**-го фактора у **i**-го об'єкта;

$\mathbf{u_k}$ **–** коефіцієнт парної кореляції між **k** –ою досліджуваною ознакою та **k**-м характерним фактором.

Головне рівняння факторного аналізу має вигляд:

$$\boldsymbol{Z = A \cdot F + UY}, \tag{7.1}$$

де $\boldsymbol{Z}$ – матриця стандартизованих значень вихідних факторів-симптомів, розміром **mxn;**

$\boldsymbol{F}$ – матриця загальних факторів, розміром **mxn;**

$\boldsymbol{A}$ – матриця факторних навантажень, розміром **mxm;**

$\boldsymbol{Y}$ **–** матриця характерних факторів, розміром **mxm;**

$\boldsymbol{U}$ **–** діагональна матриця коефіцієнтів при характерних факторах, розміром **mxm.**

Припускається, що загальні фактори $\boldsymbol{F_1, F_2 ,\ldots,F_p}$ в моделі (7.1) лінійно незалежні, тобто ортогональні.

Як і в методі головних компонент елементи матриці **A** є звичайними коефіцієнтами парної кореляції між загальними факторами та вихідними даними, тобто $\boldsymbol{a_{kL}=r_{kL}}$, тому вірна нерівність $-1 \leq \boldsymbol{a_{kL}} \leq \mathbf{1}$.

Якщо значення факторного навантаження за модулем близьке до 1 ($|\boldsymbol{a_{kL}}| \geq 0{,}7$), то зв'язок між **k**–ою ознакою і **L**-м фактором тісний. В такому випадку вважають, що **k**–я ознака навантажує **L**-й загальний фактор. Інакше кажучи, надає йому якісний зміст. У протилежному випадку, якщо $|\boldsymbol{a_{kL}}| \approx 0$, **k**–я ознака не пов'язана з **L**-м фактором.

Система рівнянь факторної моделі (7.1) для **m** досліджуваних змінних називається факторним відображенням. Для **m=3** факторне відображення має вигляд:

$$\begin{cases} z_1 = a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + u_1y_1 \\ z_2 = a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + u_2y_2 \\ z_3 = a_{31}F_1 + a_{32}F_2 + a_{33}F_3 + u_3y_3. \end{cases} \tag{7.2}$$

У даній системі (7.2) матриці ***Z, F, Y*** розглядаються як вектори-стовпці, тобто абстрагуємося від значень окремих об'єктів (**i**). Загальні фактори F_1, F_2 та F_3 можуть бути як ортогональними, так і корельованими між собою. Характерні фактори Y_1, Y_2 та Y_3 є як некорельовані між собою, так і з загальними факторами.

Факторне відображення (7.2) значно спрощується, якщо не враховувати факторні навантаження, які незначно відрізняються від нуля. У табл. 7.1 наведена схема спрощеного факторного відображення.

Таблиця 7.1

Схема факторного відображення

(х – високі за абсолютною величиною факторні навантаження)

Ознаки-симптоми	Загальні фактори - F_L			Характерні фактори - y_k		
	F_1	F_2	F_3	Y_1	Y_2	Y_3
Z_1	х	х	х	х		
Z_2	х	х			х	
Z_3	х					х

Класифікація ознак-симптомів виконується аналогічно компонентному аналізу: за рядками табл. 7.1 – складність фактора; за стовпцями табл. 7.1 – на генеральні та групові. Отже, фактор F_1 є генеральним (пов'язаний з усіма ознаками), фактор F_2 – груповим, оскільки пов'язаний з двома ознаками.

Якщо число вихідних ознак **m**, то число загальних факторів, які виділяє аналіз, дорівнює **p (p< m).** У практичних дослідженнях отриманих **p** загальних факторів достатньо для описання проблеми, оскільки вони пояснюють значущу частку варіації ознак. Останні **p-m** загальних факторів потрібно вилучити з подальшого аналізу як несуттєві. Втрати інформації при цьому незначні, а стиск простору досягається приблизно у 3-4 рази.

Отже, ***головна мета*** факторного аналізу з математичної точки зору зводиться до розв'язання таких задач:

1) визначення найпростішої факторної структури, яка сприяє якісної інтерпретації загальних факторів;

2) оцінка коефіцієнтів при характерних факторах;
3) оцінка значень факторів для всіх об’єктів, що вивчаються;
4) якісне тлумачення отриманого факторного рішення та його використання на практиці.

Далі розглянемо простіший випадок ортогональних векторів.

7.2. Розкладання дисперсії у факторному аналізі

З метою кращого розуміння математичної моделі факторного аналізу розглянемо дисперсію **k**-ої стандартизованої ознаки докладніше. Відповідно до статистичного визначення дисперсії та властивостей стандартизованих ознак

$$\bar{z}_k = 0 \ ; \ \sigma_{z_k}^2 = 1 \qquad (7.3)$$

можна записати, що:

$$\sigma_{z_k}^2 = \frac{\sum(z_{ki}-\bar{z}_k)^2}{n} = \frac{1}{n}\sum z_{ki}^2 = 1 \ . \qquad (7.4)$$

Враховуючи факторне навантаження (7.2), а також, що загальні фактори ортогональні та нормовані величини, дисперсію (7.4) можна записати так:

$$\sigma_{z_k}^2 = a_{k1}^2 + a_{k2}^2 + \cdots + a_{kL}^2 + \cdots + a_{km}^2 + u_k^2 = 1 \ . \qquad (7.5)$$

Отриманий вираз (7.5) визначає повну дисперсію **k**-ої стандартизованої ознаки. Члени у центральній частині є частками дисперсії **k**-ої вихідної ознаки, яка припадає на відповідні фактори. Наприклад, $\boldsymbol{a_{k1}^2}$ – частка дисперсії **k**-ої ознаки, яка пояснюється першим загальним фактором $\boldsymbol{F_1}$.

Абсолютний вклад **L**-го загального фактора $\boldsymbol{F_L}$ у пояснення варіації всіх чинників-симптомів складає:

$$v_L = \sum_{k=1}^{p} a_{kL}^2 . \qquad (7.6)$$

Повний абсолютний вклад усіх $\boldsymbol{p}$ загальних факторів у сумарну дисперсію усіх **m** вихідних ознак визначається так:

$$V = \sum_{L=1}^{m} v_L . \qquad (7.7)$$

Нагадаємо, що у методі головних компонент цей сумарний вклад дорівнював **m**, тобто загальній дисперсії усіх вихідних ознак. Таке відбувалося тому, що вплив характерних факторів у методі головних компонент не

враховується. На відміну від нього, у факторному аналізі повний вклад усіх загальних факторів буде меншим за число вихідних ознак **m**, що безпосередньо випливає з підсумування формули (7.5) за всіма вихідними факторами-симптомами. В результаті такого підсумування отримуємо:

$$V = \sum_{L=1}^{m} v_L = m - \sum_{k=1}^{m} u_k^2 \,. \qquad (7.8)$$

Отримане рівняння означає, що у факторному аналізі можна виділити частку дисперсії вихідних ознак, яка пояснюється загальними факторами, і частку дисперсії, яка пояснюється характерними факторами.

Отже, сума квадратів факторних навантажень вихідної стандартизованої ознаки $\mathbf{z_k}$ називається ***спільністю*** та позначається через $\boldsymbol{h_k^2}$:

$$h_k^2 = a_{k1}^2 + a_{k2}^2 + \cdots + a_{kL}^2 + \cdots + a_{km}^2. \qquad (7.9)$$

Частина дисперсії **k**-ої ознаки, яка обумовлена дією характерного фактора, називається ***характерністю*** і позначається через $\boldsymbol{u_k^2}$**:**

$$u_k^2 = 1 - h_k^2. \qquad (7.10)$$

Характерність є тою частиною повної дисперсії стандартизованої ознаки, яка не пов’язана з дією загальних факторів.

Звідси слідкує, що у факторному аналізі дисперсія будь-якої ознаки може бути подана як сума двох складових: спільності та характерності. Тому вираз (7.5) можна записати таким чином:

$$\sigma_{Z_k}^2 = h_k^2 + u_k^2. \qquad (7.11)$$

Якщо спільність $\boldsymbol{h_k^2}$ характеризує частку повної дисперсії **k**-ої стандартизованої ознаки, яка обумовлена дією всіх загальних факторів, то сумарна спільність відображує частку повної дисперсії всіх стандартизованих ознак, які обумовлені дією всіх загальних факторів:

$$\sum_{k=1}^{m} h_k^2 = V = \sum_{L=1}^{m} v_L. \qquad (7.12)$$

Якщо характерність $\boldsymbol{u_k^2}$ виражає частину повної дисперсії **k**-ої стандартизованої ознаки, яка обумовлена дією **k**-ого характерного фактора, то сумарна характерність відображує частку повної дисперсії всіх стандартизованих ознак, яка обумовлена дією всіх характерних факторів:

$$\sum_{k=1}^{m} u_k^2 = 1 - V = 1 - \sum_{L=1}^{m} v_L. \quad (7.13)$$

Як наслідок сумарна дисперсія всіх вихідних ознак, складається із сумарної спільності (7.12) та сумарної характерності (7.13):

$$\sum_{k=1}^{m} \sigma_{Z_k}^2 = \sum_{k=1}^{m} h_k^2 + \sum_{k=1}^{m} u_k^2 = m. \quad (7.14)$$

Наведена формула (7.14) повністю узгоджується з формулою (7.11).

Сумарні спільність та характерність використовуються при визначенні вкладу окремих факторів у варіацію вихідних ознак.

7.3. Основні етапи факторного аналізу

Статистичний алгоритм виділення загальних факторів передбачає проведення низки послідовних етапів.

Таблиця 7.2

Етапи факторного аналізу

I етап	Відбір чинників-симптомів латентного показника та формування матриці вихідних даних ***X***
II етап	Стандартизація чинників-симптомів і побудова матриці ***Z***
III етап	Побудова кореляційної матриці **r**
IV етап	Побудова редукованої кореляційної матриці $\mathbf{r_h}$
V етап	Побудова матриці факторних навантажень ***A***
VI етап	Виділення, інтерпретація загальних факторів та обертання системи координат
VII етап	Вимірювання загальних факторів

Сутність кожного етапу розглянемо на прикладі продуктивності праці в сільському господарстві на 10 об'єктах. Попередній апріорний аналіз чинників-симптомів дозволів виявити такі фактори:

x_1 – обсяг тракторних робіт на 100 га ріллі, сотень еталонних га;

x_2 – капіталовіддача, грн.;

x_3 – активна частина основних засобів на 1 га сільськогосподарських угідь, тис. грн.

Матриця вихідних даних ***X***, розміром 3x10:

$$\mathbf{X}=\begin{bmatrix} 8{,}40 & 5{,}57 & 4{,}87 & 5{,}66 & 4{,}23 & 5{,}09 & 8{,}14 & 3{,}46 & 5{,}73 & 6{,}83 \\ 5{,}79 & 4{,}69 & 5{,}60 & 3{,}65 & 4{,}53 & 4{,}29 & 5{,}05 & 4{,}10 & 7{,}12 & 7{,}49 \\ 100{,}8 & 84{,}4 & 85{,}6 & 62{,}3 & 59{,}1 & 82{,}0 & 68{,}3 & 66{,}4 & 143{,}9 & 126{,}1 \end{bmatrix}.$$

За формулою (2.5) проведено стандартизацію вихідних даних:

$$\boldsymbol{Z}=\begin{bmatrix} 1{,}726 & -0{,}153 & -0{,}618 & -0{,}093 & -1{,}043 & -0{,}471 & 1{,}554 & -1{,}554 & -0{,}047 & 0{,}684 \\ 0{,}464 & -0{,}447 & 0{,}306 & -1{,}308 & -0{,}580 & -0{,}778 & -0{,}149 & -0{,}935 & 1{,}565 & 1{,}871 \\ 0{,}483 & -0{,}131 & -0{,}086 & -0{,}958 & -1{,}078 & -0{,}221 & -0{,}783 & -0{,}805 & 2{,}095 & 1{,}429 \end{bmatrix}$$

Розмірність матриці ***Z*** така сама, як і в матриці ***X***, тобто **3x10.**

Також доведено виконання таких співвідношень (7.3) для стандартизованих даних.

На третьому етапі будується кореляційна матриця **r**, елементами якої виступають коефіцієнти парної кореляції між ознаками, розраховані за формулою:

$$r_{kL} = \frac{1}{n}\sum_i^n z_{ki}\, z_{Li}. \qquad (7.15)$$

Оскільки $\mathbf{r_{kk}} = \mathbf{r_{LL}} = \mathbf{1}$ та $\mathbf{r_{kL}} = \mathbf{r_{Lk}}$,то кореляційна матриця є симетричною матрицею, на головній діагоналі якої розташовані 1. Кореляційна матриця **r** має розмір **mxm**, тобто вона є квадратною симетричною матрицею.

Для нашого прикладу елементи матриці **r** розраховані за формулою (15):

$$\mathbf{r} = \begin{bmatrix} 1{,}000 & 0{,}458 & 0{,}321 \\ 0{,}458 & 1{,}000 & 0{,}912 \\ 0{,}321 & 0{,}912 & 1{,}000 \end{bmatrix}.$$

Отже, перші три етапи такі самі, як й у метода головних компонент: формування матриці вихідних даних, їх стандартизація та розрахунок кореляційної матриці. Означені етапи стандартні, тому сконцентруємо увагу на етапах 4-7.

У факторному аналізі увага звертається на загальні фактори, а отримана матриця **r** відображує вплив як загальних, так й специфічних факторів. З метою уникнення впливу специфічних факторів необхідно у матриці **r** ввести на головну діагональ спільності. Перетворена таким чином кореляційна матриця

називається редукованою кореляційною матрицею – $\mathbf{r_h}$. При чому коефіцієнти кореляції між двома різними ознаками залишаються такими самими, як й в матриці **r**.

Отже, ***редукованою кореляційною матрицею*** називається матриця парних коефіцієнтів кореляції, на головній діагоналі якої замість одиниць розташовані оцінки спільностей: $\boldsymbol{h_k^2 \neq 1.}$

Як велось раніше, загальність **k**-ої ознаки відображує частину дисперсії, яка обумовлена дією всіх факторів-симптомів. До початку розрахунків необхідно приблизно оцінити цю величину, щоб побудувати редуковану матрицю $\mathbf{r_h}$. Існує декілька способів оцінки спільностей:

1) за максимальним коефіцієнтом парної кореляції;
2) за середнім коефіцієнтом парної кореляції;
3) спосіб тріад.

Розглянемо їх детальніше.

Перший спосіб максимальної кореляції добре себе зарекомендував на практиці. На головній діагоналі записується з позитивним знаком найбільший за абсолютною величиною коефіцієнт парної кореляції даної строки (стовпця) матриці **r**. Ніяких додаткових розрахунків не потрібно. Цей спосіб краще використовувати при великій кількості вихідних ознак (m>10).

Другим способом оцінки спільностей є усереднення кореляцій. Зазвичай цей спосіб використовується при незначній кількості вихідних ознак **m** і заключається у находженні середньої арифметичної з коефіцієнтів парної кореляції відповідної строки чи стовпця матриці **r**.

Третій спосіб – спосіб тріад, надає добрі результати. В цьому випадку спільність розраховується так: у **k**-му рядку матриці **r** знаходять два найбільших за абсолютною величиною коефіцієнта парної кореляції $\mathbf{r_{ks}}$ та $\mathbf{r_{kg}}$. Їх значення разом з коефіцієнтом парної кореляції між **s**-ою і **g**-ою ознаками підставляють у формулу:

$$h_k^2 = \frac{|r_{ks}| \cdot |r_{kg}|}{|r_{sg}|}. \qquad (7.16)$$

Такий розрахунок спільностей дозволяє підсилити вплив найбільших коефіцієнтів парної кореляції.

Відмітимо, що думки вчених щодо способів оцінки спільностей базуються на різних концепціях і суттєво різняться. Сьогодні існує багато розрахункових процедур побудови редукованої матриці. Вибір того чи іншого алгоритму залежить від мети та завдань дослідження.

Розрахунок редукованої кореляційної матриці для нашого прикладу доцільно виконати у спосіб усереднення кореляцій, оскільки число вихідних ознак невелике (m=3). В якості оцінки спільностей приймаємо середнє квадратичне за строкою кореляційної матриці **r**:

$$h_1^2 = \frac{1{,}000+0{,}458+0{,}321}{3} = \frac{1{,}779}{3}=0{,}593;$$

$$h_2^2 = \frac{0.458+1{,}000+0{,}912}{3} = \frac{2{,}370}{3}=0{,}790;$$

$$h_3^2 = \frac{0{,}321+0{,}912+1{,}000}{3} = \frac{2{,}233}{3}=0{,}744.$$

Отже, будуємо редуковану кореляційну матрицю, розміром 3x3:

$$\mathbf{r_h}=\begin{bmatrix} 0{,}593 & 0{,}458 & 0{,}321 \\ 0{,}458 & 0{,}790 & 0{,}912 \\ 0{,}321 & 0{,}912 & 0{,}744 \end{bmatrix}.$$

7.4. Методи визначення факторного рішення

У практичній діяльності є різні способи розрахунків факторних навантажень. Їх використання залежить від наявної інформації, від мети дослідження. Класичним методом серед них є метод головних компонент. Саме цей метод розглянемо у факторному аналізі.

Після того, як редукована матриця $\mathbf{r_h}$ побудована, вона підлягає тим самим статистичним процедурам, що й при визначенні головних компонент: будується та розв'язується характеристичне рівняння, визначаються власні значення і власні вектори редукованої кореляційної матриці $\mathbf{r_h}$, що дозволяє розрахувати факторні навантаження a_{kL}. Зауважимо, що властивості матриці факторних навантажень однакові для факторного та компонентного методів.

Для знаходження власних чисел $\boldsymbol{\lambda}$ запишемо характеристичний визначник та прирівняємо його до нуля:

$$|r_h - \lambda \cdot E| = \begin{bmatrix} 0{,}593-\lambda & 0{,}458 & 0{,}321 \\ 0{,}458 & 0{,}790-\lambda & 0{,}912 \\ 0{,}321 & 0{,}912 & 0{,}744-\lambda \end{bmatrix} = 0,$$

де Е – одинична матриця розміру 3x3.

Щоб отримати характеристичний многочлен, розкриймо визначник третього порядку за правилом трикутника:

$$(0{,}593-\boldsymbol{\lambda})\cdot(0{,}790-\boldsymbol{\lambda})\cdot(0{,}744-\boldsymbol{\lambda})+0{,}458\cdot0{,}912\cdot0{,}321+0{,}458\cdot0{,}912+0{,}321-0{,}321^2\cdot$$

$$\cdot(0{,}790-\boldsymbol{\lambda})-0{,}912^2\cdot(0{,}593-\boldsymbol{\lambda})-0{,}458^2\cdot(0{,}744-\boldsymbol{\lambda})=0.$$

Після математичних перетворень можна отримати таке кубічне рівняння відносно $\boldsymbol{\lambda}$:

$$-\boldsymbol{\lambda}^3+2{,}127\cdot\boldsymbol{\lambda}^2-0{,}353\cdot\boldsymbol{\lambda}-0{,}114=0.$$

Знайдене характеристичне рівняння надає три значення $\boldsymbol{\lambda_k}$. Нагадаємо, що рівняння вищі за другий порядок розв'язуються за допомогою спеціальних ітеративних процедур.

Для нашого прикладу, коренями даного рівняння є такі значення:

$$\boldsymbol{\lambda_1}=1{,}911; \quad \boldsymbol{\lambda_2}=0{,}375; \quad \boldsymbol{\lambda_3}=-0{,}159.$$

Розв'язання характеристичного рівняння дозволило отримати як позитивні ($\boldsymbol{\lambda_1}$ та $\boldsymbol{\lambda_2}$), так і від'ємні ($\boldsymbol{\lambda_3}$) власні значення. Це слідкує з того факту, що слід редукованої матриці $\mathbf{r_h}$, менший за кількість вихідних ознак **m.** Тому сума позитивних власних значень більше суми спільностей, тобто:

$$\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2=1{,}911+0{,}375=2{,}286.$$

Сума спільностей при цьому буде дорівнювати:

$$h_1^2 + h_2^2 + h_3^2 = 0{,}593 + 0{,}790 + 0{,}744 = 2{,}127\ .$$

І як наслідок, 2,286>2,127. При врахуванні від'ємних власних значень наведені суми повністю співпадають, тобто

$$\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2+\boldsymbol{\lambda}_3=1{,}911+0{,}375-0{,}159=2{,}127; \quad 2{,}127=2{,}127.$$

Але на практиці від'ємні власні значення не мають сенсу, оскільки повинна виконуватися така властивість матриці факторних навантажень: $\sum_{k=1}^{m} a_{kL}^2 = \lambda_k$.

Отже, на підставі вище викладеного та відповідно до формули (7.12) можна записати таку властивість характеристичних коренів матриці $\mathbf{r_h}$:

$$\sum_{k=1}^{m} \lambda_k = tr(r_h) = \sum_{k=1}^{m} h_k^2 = g_m < m. \quad (7.17)$$

Що повністю узгоджується з визначенням спільності ($h_k^2 < 1$).

Тому з **m** можливих характеристичних коренів реальний інтерес мають тільки **s** (**s<m**) позитивні, яким відповідають дійсні вектори та дійсні факторні навантаження. Процес виділення загальних факторів зупиняється на тому кроці, коли сума власних значень редукованої матриці $\mathbf{r_h}$ дорівнює сумарній спільності, тобто:

$$\sum_{k=1}^{p} \lambda_k = \sum_{k=1}^{m} h_k^2 \ , (p < s < m). \quad (7.18)$$

Відповідно до наведеного критерію (7.18) у прикладі потрібно виділити два загальних фактори.

Розрахунок факторних навантажень потребує визначення власних векторів. Для цього необхідно розв'язати характеристичне рівняння. Тому розрахуємо власний вектор, який відповідає першому власному значенню $\boldsymbol{\lambda_1}$ =1,911, на підставі такої системи рівнянь:

$$\begin{cases} (0{,}593\text{-}1{,}911)\cdot U_1+0{,}458\cdot U_2+0{,}321\cdot U_3=0 \\ 0{,}458\cdot U_1+(0{,}790\text{-}1{,}911)\cdot U_2+0{,}912\cdot U_3=0 \\ 0{,}321\cdot U_1+0{,}912\cdot U_2+(0{,}744\text{-}1{,}911)\cdot U_3=0 \end{cases}$$

З метою спрощення розрахунків приймемо, що U_3=1. Така дія можливо тому, що власні вектори знаходяться з точністю до постійного множника С (тобто визначник дорівнюємо нулю). В результаті таких перетворень отримуємо таку систему рівнянь:

$$\begin{cases} 0{,}458\cdot U_1\text{-}1{,}121\cdot U_2=\text{-}0{,}912 \\ 0{,}321\cdot U_1+0{,}912\cdot U_2=1{,}167 \end{cases}$$

Розв'язання такої системи дозволило отримати такі значення:

U_1=0,613 ; U_2=1,064; U_3=1,000.

Вони є координатами власного вектора, які відповідають власному значенню $\boldsymbol{\lambda_1}$=1,911.

Аналогічно визначається власний вектор,. Який відповідає другому власному числу $\boldsymbol{\lambda_2}$=0,375:

$$\begin{cases} (0{,}593-0{,}375)\cdot U_1+0{,}458\cdot U_2+0{,}321\cdot U_3=0 \\ 0{,}458\cdot U_1+(0{,}790-0{,}375)\cdot U_2+0{,}912\cdot U_3=0 \\ 0{,}321\cdot U_1+0{,}912\cdot U_2+(0{,}744-0{,}375)\cdot U_3=0 \end{cases}$$

Якщо прийняти, U_3=1, то система рівнянь спрощується:

$$\begin{cases} 0{,}458\cdot U_1+0{,}415\cdot U_2=-0{,}912 \\ 0{,}321\cdot U_1+0{,}912\cdot U_2=-0{,}369. \end{cases}$$

Рішенням такої системи будуть координати власного вектора, який відповідає другому власному значенню $\boldsymbol{\lambda_2}$=0,375:

U_1=-2,385; U_2=0,435; U_3=1,000.

Отримані власні вектори та власні значення дозволяють розрахувати факторні навантаження першого та другого загальних факторів за формулою:

$$a_{kL} = U_{kL}\sqrt{\frac{\lambda_k}{\sum U_{kL}^2}}. \qquad (7.19)$$

Усі розрахунки подані у табл.7.3 і табл. 7.4.

Таблиця 7.3

Розрахунок факторних навантажень першого загального фактора

U_{k1}	U_{k1}^2	$a_{k1} = U_{k1}\sqrt{\frac{\lambda_1}{\sum U_{k1}^2}}$	a_{k1}^2
0,613	0,376	0,535	0,286
1,064	1,132	0,929	0,863
1,000	1,000	0,873	0,762
Усього	2,508	-	1,911

Зауважимо, що величина $\sqrt{\frac{\lambda_1}{\sum U_{k1}^2}} = \sqrt{\frac{1{,}911}{2{,}508}} = 0{,}873$ – постійна для першого загального фактора.

Таблиця 7.4

Розрахунок факторних навантажень другого загального фактора

U_{k2}	U_{k2}^2	$a_{k2} = U_{k2}\sqrt{\frac{\lambda_2}{\sum U_{k2}^2}}$	a_{k2}^2
-2,385	5,688	-0,557	0,310
0,435	0,119	0,102	0,010
1,000	1,000	0,234	0,055
Усього	6,877	-	0,375

Зауважимо, що величина $\sqrt{\frac{\lambda_2}{\sum U_{k2}^2}} = \sqrt{\frac{0{,}375}{6{,}877}} = 0{,}234$ – постійна для другого загального фактора.

Останні стовпці табл.7.3 і табл.7.4 містять суми квадратів факторних навантажень. Їх сума дорівнює власним числам $\boldsymbol{\lambda}_1$=1,911 та $\boldsymbol{\lambda}_2$=0,375 відповідно.

Сукупність значень факторних навантажень утворює матрицю **A**, розміру **m**x**p**, причому **p<m**, оскільки третій фактор в процесі розрахунків було усунено. Отже, матриця **A** розміру 3x2 має такий вигляд:

$$\mathrm{A} = \begin{bmatrix} 0{,}535 & -0{,}557 \\ 0{,}929 & 0{,}102 \\ 0{,}873 & 0{,}234 \end{bmatrix}.$$

Знайдемо спільність та характерність кожної ознаки за формулами (7.9) та (7.10). Розрахунки наведено у табл.7. 5.

Таблиця 7.5

Розрахунок факторних навантажень першої головної компоненти

Ознаки	Квадрати факторних навантажень		Спільності h_k^2	Характерності u_k^2
	a_{k1}^2	a_{k2}^2		
X_1	0,286	0,310	0,596	0,404
X_2	0,863	0,010	0,873	0,127
X_3	0,762	0,055	0,817	0,183
Усього	1,911	0,375	2,286	0,714

Далі оцінимо вклад факторів у сумарну дисперсію. Як велось раніше сумарна дисперсія всіх вихідних ознак дорівнює m=3. Тому відповідно до співвідношення (7.14):

$$\sum_{k=1}^{m} \sigma_{Z_k}^2 = 3 = 2{,}286 + 0{,}714 \, .$$

Знайдемо частку сумарної спільності у сумарній дисперсії всіх ознак:

$$\frac{\sum_{k=1}^{m} h_k^2}{m} = \frac{2{,}286}{3} = 0{,}762 \text{ або } 76{,}2\%.$$

Знайдемо частку сумарної характерності у сумарній дисперсії всіх ознак:

$$\frac{\sum_{k=1}^{m} u_k^2}{m} = \frac{0{,}714}{3} = 0{,}238 \text{ або } 23{,}8\%.$$

Отже, два знайдених загальних фактори пояснюють 76,2 % варіації вихідних ознак і 23,6% обумовлено дією характерних факторів, котрі акумулюють усі інші, не враховані у дослідженні фактори.

Можна також оцінити вклад кожного фактора у сумарну дисперсію. Розрахуємо вклад першого загального фактора:

$$\frac{v_1}{m} = \frac{\sum_{k=1}^{m} a_{k1}^2}{m} = \frac{1{,}911}{3} = 0{,}637 \text{ або } 63{,}7\%.$$

Перший загальний фактор пояснює 63,7% варіації вихідних чинників-симптомів.

Вклад другого фактора в пояснення варіації вихідних ознак:

$$\frac{v_2}{m} = \frac{\sum_{k=1}^{m} a_{k2}^2}{m} = \frac{0{,}375}{3} = 0{,}125 \text{ або } 12{,}5\%.$$

Другий загальний фактор пояснює тільки 12,5 % варіації вихідної інформації.

Сума вкладів обох факторів наддасть сумарну спільність:

$$63{,}7\%+12{,}5\%=76{,}2\%$$

7.4. Інтерпретація факторів і обертання системи координат

Отримані навантаження a_{kL} під час факторного рішення варіюють у визначених межах. Це ускладнює аналіз факторної структури кожної вихідної ознаки та кожного загального фактора. Особливо це стосується першої ознаки (a_{11}=0,535 та a_{21}=-0,557). Визиває деякі труднощі й економічна інтерпретація

знайдених факторів. З метою запобігання означених проблем необхідно здійснити обертання системи координат.

Нагадаємо, що стовпці матриці $\boldsymbol{A}$ ортогональні та займають довільну позицію відносно змінних. Можливе існування великої кількості матриць, які будуть однаково добре відтворювати редуковану кореляційну матрицю і задовольняти такій формулі:

$$\boldsymbol{A \cdot A^T = r_h} . \qquad (7.20)$$

Виділимо у просторі загальних факторів систему координат, яку можна отримати виходячи з вихідної, через обертання на кут $\boldsymbol{\alpha}$. Таке відбувається методом добутку матриці $\boldsymbol{A}$ на деяку невідому матрицю ортогонального перетворення $\boldsymbol{T}$.

Отже нова матриця факторних навантажень буде матиме такий вигляд:

$$\boldsymbol{B = A \cdot T}. \qquad (7.21)$$

До визначення ортогональності, матриця $\boldsymbol{T \cdot T^T}$ надає одиничну матрицю. Враховуючи означене, можна записати, що

$$\boldsymbol{B \cdot B^T = A \cdot T \cdot (A \cdot T)^T = A \cdot T \cdot T^T \cdot A^T = A \cdot A^T} \quad . \qquad (7.22)$$

Тому матриця $\boldsymbol{B}$ задовольняє відношенню (7.20). Нова матриця факторних навантажень $\boldsymbol{B}$ розглядається як один з варіантів факторного рішення. Інакше кажучи, шукана матриця факторних навантажень може бути визначена тільки з точністю до довільного ортогонального перетворення.

Обертання матриці потрібне для виявлення у просторі загальних факторів однієї з можливих систем координат. При цьому ця система координат має накладатися на конфігурацію векторів, щоб отримати просту факторну структуру. Конфігурація – це система векторів, співвідношення між якими постійні.

Геометрично це означає, що матриця факторних навантажень знаходить у просторі стандартизованих ознак підпростір загальних факторів, але не фіксує там систему координат.

Така невизначеність факторного рішення дозволяє висунути нові вимоги до матриці факторних навантажень, які її частково обмежують. Такою вимогою

є вимога на просту структуру, тобто факторні навантаження a_{kL} прагнуть або до 0, або до 1. Означену вимогу сформулював Л. Терстоун.

Проілюструємо наведені теоретичні викладки (рис. 7.1). На цьому рисунку зображено конфігурацію трьох векторів (ознак) для двох допустимих випадків розташування системи координат, що створені двома загальними факторами.

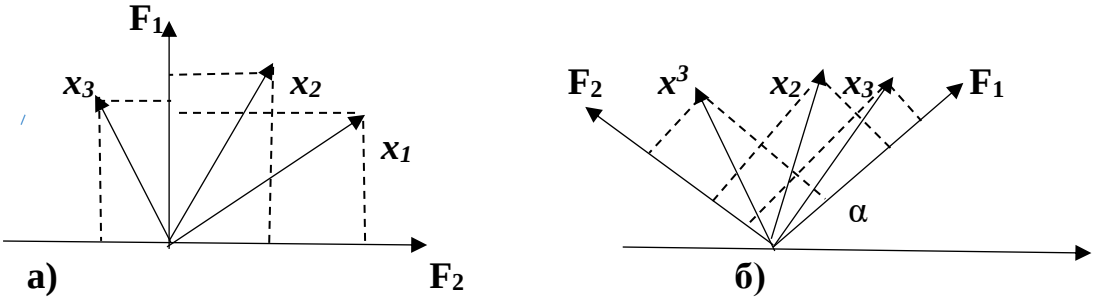

Рис. 7.1 Обертання системи координат:

а) вихідне факторне рішення;

б) перетворене факторне рішення

Матрицю ортогонального перетворення T з формули (7.21) можна записати так:

$$T = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}. \quad (7.23)$$

Наведена формула (7.23) характеризує попарне обертання системи координат за годинною стрілкою на кут **α** .

Під час обертання проти годинної стрілки формула (7.23) набуває такого виду:

$$T = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}. \quad (7.24)$$

За умови, що **p**>2 повна матриця перетворень буде визначатися як добуток матриць перетворень для всіх парних комбінацій факторів. Розглянемо таку матрицю для **p**=3:

$$\boldsymbol{T = T_{12} \cdot T_{13} \cdot T_{23}}\ , \quad (7.25)$$

де

$$T_{12} = \begin{bmatrix} \cos\alpha_1 & -\sin\alpha_1 & 0 \\ \sin\alpha_1 & \cos\alpha_1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad T_{13} = \begin{bmatrix} \cos\alpha_{12} & 0 & -\sin\alpha_2 \\ 0 & 1 & 0 \\ \sin\alpha_2 & 0 & \cos\alpha_2 \end{bmatrix},$$

$$T_{23} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_3 & -\sin\alpha_3 \\ 0 & \sin\alpha_3 & \cos\alpha_3 \end{bmatrix}.$$

У кожній з наведених матриць діагональні елементи, які відповідають номеру необернених осей, приймають значення 1, а недіагональні – 0. Оскільки кути $\boldsymbol{\alpha_1}$, $\boldsymbol{\alpha_2}$, $\boldsymbol{\alpha_3}$ заздалегідь невідомі, обертання виконується послідовно за допомогою матриць $\mathbf{T_{12}}$, $\mathbf{T_{13}}$, $\mathbf{T_{23}}$. Виникає питання , як знайти кут повороту? У теорії багатовимірного аналізу є низка методів находження кута $\boldsymbol{\alpha}$, які визначаються метою дослідження (метод квартимакс, метод варимакс та інші).

Розглянемо метод квартимакс, як найбільш доведений. Його обґрунтовували такі вчені, як Дж. Керол, Ш. Ріглі, Д. Саундерс, Дж. Нейгауз та інші. Метою методу квартимакс є отримання однофакторного рішення, складність ознак в якому дорівнювала 1.

З математики відомо, що обертання кута $\boldsymbol{\alpha}$ визначається за формулою:

$$tg4\alpha = \frac{2\cdot\sum_{k=1}^{m}(2a_{k1}\cdot a_{k2})\cdot(a_{k1}^2 - a_{k2}^2)}{\sum_{k=1}^{m}\left[(a_{k1}^2 - a_{k2}^2)^2 - (2\cdot a_{k1}\cdot a_{k2}\right]} = \frac{c}{w}. \qquad (7.26)$$

У наведеній формулі (7.26) будь-якому знаку чисельника відповідає позитивне або від’ємне значення знаменника. Тому можливі чотири ситуації, кожна з яких має свій кут обертання (табл.7.6).

Таблиця 7.6

Визначення кута обертання за методом квартимакс

Знаки формули			Знак	Квадрант кута	Інтервал
Чисельник ***c***	Знаменник ***w***	***tg4α***	***cos4α***	***4α***	***α***
+	+	+	+	$0^\circ<4\alpha<90^\circ$	$0^\circ – 22{,}5^\circ$
+	-	-	-	$90^\circ<4\alpha<180^\circ$	$22{,}5^\circ – 45^\circ$
-	-	+	-	$-180^\circ<4\alpha<-90^\circ$	$-45^\circ – -22{,}5^\circ$
-	+	-	+	$-90^\circ<4\alpha<0^\circ$	$-22{,}5^\circ – 0^\circ$

Розглянемо на прикладі розрахункову процедуру методу квартимакс. Спочатку за формулою (7.26) проведемо розрахунок кута обертання (табл.7.7).

Таблиця 7.7

Квартимакс-рішення для двох загальних факторів

№ ознаки	Навантаження		Квадрат навантаження		Добуток $2 \cdot a_{k1} \cdot a_{k2}$	Різниця квадратів $a_{k1}^2 - a_{k2}^2$	$(2a_{k1}a_{k2}) \cdot (a_{k1}^2 - a_{k2}^2)$
	a_{k1}	a_{k2}	a_{k1}^2	a_{k2}^2			
1	0,535	-0,557	0,2863	0,3101	-0,5960	0,0238	-0,0142
2	0,929	0,102	0,8626	0,0103	0,1887	0,8523	0,1608
3	0,873	0,234	0,7620	0,0545	0,4076	0,7075	0,2884
Сума квад-ратів	1,911	0,375	1,4067	0,0992	0,5570	1,2275	0,1092

Відповідно до теорії чисельник є подвійною сумою добутку трьох значень, тобто

$$c = 2 \cdot (-0{,}0142 + 0{,}111608 + 0{,}2884) = 2 \cdot 0{,}4350 = 0{,}870.$$

Отримання знаменника формули (7.26) відбувається методом суми квадратів добутку та різниці квадратів (останній рядок табл. 7.7):

w=1,2275-0,5570=0,6705.

Підставляємо отримані значення у формулу (7.26):

$$tg4\alpha = \frac{0{,}870}{0{,}6705} = 1{,}2975.$$

Оскільки і чисельник, і знаменник позитивні, кут **α** потрапляє у першій квадрант (табл. 7.6). Знаходимо, що 4·α=52˚23' і як наслідок, кут обертання буде дорівнювати α=13˚3'. Отриманий результат потрапляє в інтервал від 0˚ до 22,5˚ (табл. 7.6). Тому кут α=13˚3' – це той кут, який шукали. Тоді можна записати, що sin13˚3' = 0,2258; cos13˚3' =0,9705 . Обертання виконувалось проти годинної стрілки. Відповідно до формули (7.24) матрицю перетворень записуємо так:

$$T = \begin{bmatrix} 0{,}9705 & -0{,}2258 \\ 0{,}2258 & 0{,}9705 \end{bmatrix}.$$

Помножимо матрицю ***A*** на матрицю перетворень ***T*** за формулою (7.21):

$$B = A \cdot T = \begin{bmatrix} 0{,}535 & -0{,}557 \\ 0{,}929 & 0{,}102 \\ 0{,}873 & 0{,}234 \end{bmatrix} \cdot \begin{bmatrix} 0{,}9705 & -0{,}2258 \\ 0{,}2258 & 0{,}9705 \end{bmatrix} = \begin{bmatrix} 0394 & -0{,}661 \\ 0{,}924 & -0{,}111 \\ 0{,}900 & 0{,}030 \end{bmatrix}.$$

Порівняємо вихідну матрицю А та нову матрицю В, яку отримано методом обертання. Як бачимо, структура факторного навантаження спростилася. Наразі кожна ознака має складність 1; перший фактор тісно пов'язано з другою (0,924) та третьою ознаками (0,900), а другий загальний фактор – тільки з першою ознакою (-0,661). Отримане рішення спрощує інтерпретацію знайдених загальних факторів.

Отже, можна сказати, що перший загальний фактор $\mathbf{F_1}$ – технічна оснащеність фермерських господарств, а другий фактор F_2 – обсяг тракторних робіт на 100 га ріллі.

Перевіримо властивості матриці факторних навантажень за підставі формул (7.20) і (7.22):

$$A \cdot \mathrm{A}^{\mathrm{T}} = \begin{bmatrix} 0{,}535 & -0{,}557 \\ 0{,}929 & 0{,}102 \\ 0{,}873 & 0{,}234 \end{bmatrix} \cdot \begin{bmatrix} 0{,}535 & 0{,}929 & 0{,}873 \\ -0{,}557 & 0{,}102 & 0{,}234 \end{bmatrix} \approx \begin{bmatrix} 0{,}593 & 0{,}458 & 0{,}321 \\ 0{,}458 & 0{,}790 & 0{,}912 \\ 0{,}321 & 0{,}912 & 0{,}744 \end{bmatrix};$$

$$\mathrm{B} \cdot \mathrm{B}^{\mathrm{T}} = \begin{bmatrix} 0{,}394 & -0{,}661 \\ 0{,}9240 & -0{,}111 \\ 0{,}900 & 0{,}030 \end{bmatrix} \cdot \begin{bmatrix} 0{,}394 & 0{,}924 & 0{,}900 \\ -0{,}661 & -0{,}111 & 0{,}030 \end{bmatrix} \approx \begin{bmatrix} 0{,}593 & 0{,}458 & 0{,}321 \\ 0{,}458 & 0{,}790 & 0{,}912 \\ 0{,}321 & 0{,}912 & 0{,}744 \end{bmatrix}.$$

Як бачимо всі властивості виконуються, деякі незбіжності виникли під час округлень.

7.6. Вимірювання факторів

Визначення загальних факторів порівняно з розрахунком головних компонент ускладняється двома обставинами. Перш за все, у факторній моделі (7.1) поряд з загальними враховуються характерні фактори значення яких не можливо розрахувати поки невідома повна дисперсія змінної. По-друге, усі фактори обертаються. Тому точно виміряти значення факторів неможливо. Щоб позбавитися цих проблем у багатофакторному аналізі використовується метод найменших квадратів у межах багатофакторного кореляційно-регресійного аналізу. Як результативна ознака в цьому випадку розглядаються латентні ознаки. Визначення матриці **А** (матриці коефіцієнтів парної кореляції між незалежними ознаками-симптомами та залежними загальними факторами), а

також перехід до стандартизованої форми рівняння множинної регресії дозволяє розв'язати означені проблеми. Рівняння регресії, яке пов'язує будь-який загальний фактор $\mathbf{F_{Li}}$ з вихідними ознаками має такий вид:

$$\widehat{F_{Li}} = \beta_{L1} \cdot z_{1i} + \beta_{L2} \cdot z_{2i} + \cdots + \beta_{Lm} \cdot z_{mi}, \quad (7.27)$$

де $\widehat{F_{Li}}$ – розрахункове значення **L**-го загального фактора для **i**-го об'єкта;

β_{Lk} – **k**-тий стандартизований коефіцієнт в **L**-му рівнянні регресії (**β**-коефіцієнт).

Щоб знайти вектор-стовбець невідомих **β**-коефіцієнтів використовують матричну формулу з кореляційно-регресійного аналізу:

$$\beta_L = r^{-1} \cdot a_L , \quad (7.28)$$

де r^{-1} – матриця, зворотна кореляційній матриці, розміром mxm;

a_L – **L**-тий вектор-стовпець факторних навантажень.

Матриця, зворотна кореляційній матриці має ряд властивостей, насамперед:

$$(r^{-1})^T = (r^T)^{-1} = r^{-1}.$$

Враховуючи означену властивість та формулу (7.28), вираз (7.27) приймає вид:

$$\widehat{F}_L = \beta_L^T \cdot Z = (r^{-1} \cdot a_L)^T \cdot Z = a_L^T \cdot r^{-1} \cdot Z. \quad (7.29)$$

За допомогою отриманого рівняння можна розрахувати значення **L**-го загального фактора для всіх **n** об'єктів.

Перезапишемо рівняння (7.29) у матричному вигляді. Для цього всі загальні фактори, розташовані зліва, необхідно записати як матрицю, розміру **p**x**n,** а справа замість вектора-строки a_L^T – записати транспоновану матрицю факторних навантажень $\mathbf{A^T}$, розміру **p**x**m**.Таким чином оцінки виділених загальних факторів відбуваються за матричним рівнянням:

$$\widehat{F} = A^T \cdot r^{-1} \cdot Z. \quad (7.30)$$

У наведеному рівнянні матриця **A** означає або початкову матрицю факторних навантажень, або вторинну, нову матрицю, яку отримано під час

обертання за формулою (7.21). Інакше кажучи, матриця **A** – це деяке факторне рішення, не обов'язково початкове.

Зауважимо, що рівняння (7.30) виражає шукані значення у випадку ортогональності загальних факторів. Використання методу найменших квадратів і стандартних процедур кореляційно-регресійного аналізу для знаходження значень виділених загальних факторів за формулами (7.27) – (7.30) свідчить, що точно визначити значення невідомих параметрів неможливо. Є тільки приблизна їх оцінка на базі рівняння (7.27).

Розглянемо заключний етап факторного аналізу продуктивності праці, тобто оцінимо два виділених загальних фактора. Спочатку знайдемо матрицю, зворотну кореляційній, r^{-1}:

$$r^{-1} = \begin{bmatrix} 1{,}3612 & -1{,}3368 & 0{,}7823 \\ -1{,}3368 & 7{,}2563 & -1886 \\ 0{,}7823 & -6{,}1886 & 6{,}3929 \end{bmatrix}.$$

Отримана матриця дозволяє за формулою (7.30) оцінити загальні фактори:

$$\widehat{\boldsymbol{F}} = \boldsymbol{B^T} \cdot \boldsymbol{r^{-1}} \cdot \boldsymbol{Z} = \begin{bmatrix} 0{,}394 & 0{,}924 & 0{,}900 \\ -0{,}661 & -0{,}111 & 0{,}030 \end{bmatrix} \cdot \begin{bmatrix} 1{,}3612 & -1{,}3368 & 0{,}7823 \\ -1{,}3268 & 7{,}2563 & -6{,}1886 \\ 0{,}7823 & -6{,}1886 & 6{,}3929 \end{bmatrix} \cdot$$

$$\cdot \begin{bmatrix} 1{,}726 & 0{,}153 & -0{,}618 & -0{,}093 & -0{,}043 & -0{,}471 & 1{,}554 & -1{,}554 & -0{,}047 & 0{,}684 \\ 0{,}464 & -0{,}447 & 0{,}306 & -1{,}308 & -0{,}580 & -0{,}778 & -0{,}149 & -0{,}935 & 1{,}565 & 1{,}871 \\ 0{,}483 & -0{,}131 & -0{,}086 & -0{,}958 & -1{,}078 & -0{,}221 & -0{,}733 & 0{,}805 & 2{,}095 & 1{,}429 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}455 & -0{,}319 & 0{,}155 & -1{,}127 & -0{,}726 & -0{,}553 & 0{,}335 & -0{,}853 & 1{,}671 & 1{,}634 \\ -1{,}133 & 0{,}111 & 0{,}87 & -0{,}139 & 0{,}434 & 0{,}346 & -1{,}380 & 0{,}941 & 0{,}622 & -0{,}184 \end{bmatrix}.$$

Як бачимо, самий високий рівень технічної оснащеності на дев'ятому об'єкті (1,671), а самий низький – на четвертому (-1,127). У той же час обсяг тракторних робіт на 100 га угідь найбільший на 8 господарстві (0.941), та самий низький – на 7 (-1,380).

У подальших дослідженнях загальні фактори доцільно використовувати як аргументи у регресійних моделях продуктивності праці. В результаті факторного аналізу замість трьох ознак, будуть використані два загальних фактори, які є ортогональними. Отже, спостерігається редукція вихідних даних без значних втрат інформації про продуктивність праці.

Питання для самоконтролю

1. Дайте визначення факторному аналізу у широкому сенсі.
2. Дайте визначення факторному аналізу у вузькому сенсі.
3. У чому полягає головна ідея факторного аналізу?
4. Запишіть математичну модель факторного аналізу.
5. Дайте визначення загальним факторам.
6. Що таке факторне відображення?
7. Які основні етапи має факторний аналіз? Опишіть їх.
8. Як розрахувати повний вклад загального фактора у сумарну дисперсію вихідних ознак?
9. Що таке спільність?
10. Як розраховується спільність?
11. Що таке характерність?
12. Як розрахувати характерність?
13. Чому дорівнює дисперсія ознаки у факторному аналізі?
14. Наведіть формулу розрахунку сумарної дисперсії.
15. Що таке редукована матриця?
16. Охарактеризуйте перетворення кореляційної матриці в редуковану.
17. Як розрахувати факторні навантаження?
18. З якою метою виконується обертання системи координат загальних факторів?
19. Як визначається нова матриця факторних навантажень після обертання?
20. Охарактеризуйте властивості матриці факторних навантажень.
21. Що таке вимога простої структури факторних навантажень?
22. Запишіть матрицю ортогональних перетворень?
23. Як знайти кут обертання?
24. У чому полягає сутність методу квартимакс?
25. Як оцінюється значення загальних факторів?
26. Запишіть та охарактеризуйте матричне рівняння значень загальних факторів.

СПИСОК ЛІТЕРАТУРИ

1. Бахрушин В.Є. Методи аналізу даних: навч. посіб. для студентів. Запоріжжя: КПУ, 2011. 268 с.
2. Барковський В.В., Барковський Н.Р., Лаптін О.К. Теорія ймовірностей та математична статистика: навч. посіб.. К.: ЦНЛ, 2019. 424 с.
3. Бізнес-аналітика багатовимірних процесів: мультимедійний навч. посіб. Т.С. Клебанова, Л.С. Гур'янова, Л.О. Чаговець та ін. Харків, 2020.URL: http://ebooks.git-elt.hneu.edu.ua/babap/literature.html
4. Вітлінський В.В. Економіко-математичні методи та моделі: оптимізація: навч. посібник. К.: КНЕУ, 2016. 303 с.
5. Кузьмичов А.І. Ймовірне та статистичне моделювання в EXCEL для прийняття рішень. Навч.посіб. К.: Видавництво Ліра-К., 2020.с. 200.
6. Основи статистичного моделювання: навч. посіб. С.В. Чугаєвська, Н.В. Ковтун та ін. Житомир: Видавництво ПП "Рута", 2022. 604 с.
7. Пістунов І.М., Антонюк О.П., Турчанінова І.Ю. Кластерний аналіз в економіці: навч. посіб. Дніпропетровськ: Національний гірничий університет, 2008. 84 с.
8. Пономаренко В.С., Малярець Л.М. Багатовимірний аналіз соціально-економічних систем: навч. посіб. Харків: Вид. ХНЕУ, 2009. 384 с.
9. Статистичне дослідження соціально-економічних і демографічних процесів: моногр. за заг. ред. Ю.О. Ольвінської. Київ: ФОП Гуляєва В.М., 2021. 207 с.
10. Янковий О.Г. Багатовимірний аналіз у системі STATISTICA: моногр. Вип. 1. Одеса: Оптимум, 2001. 216 с.
11. Янковий О.Г. Багатовимірний аналіз у системі STATISTICA: моногр. Вип. 2. Одеса: Оптимум, 2002. 325 с.
12. Яровий А.Т., Страхов Є.М. Багатовимірний статистичний аналіз: навч.-метод. посіб. для студентів математичних та економічних фахів. Одеса: Астропринт, 2015. 132 с.

БАГАТОВИМІРНИЙ СТАТИСТИЧНИЙ АНАЛІЗ

КОНСПЕКТ ЛЕКЦІЙ

Укладач: Тетяна Валеріївна Погорєлова

Коректор: Алла Олексіївна Коовальова

Обсяг 4,4 др. арк.