МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ
ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
"ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ"
АВТОМОБІЛЬНО-ДОРОЖНІЙ ІНСТИТУТ

Кафедра «Прикладна математика та інформатика»

МЕТОДИЧНІ ВКАЗІВКИ ДО ВИКОНАННЯ ПРАКТИЧНИХ РОБІТ З ДИСЦИПЛІНИ "БАГАТОВИМІРНИЙ СТАТИСТИЧНИЙ АНАЛІЗ" (ДЛЯ СТУДЕНТІВ НАПРЯМУ 6.030502 "ЕКОНОМІЧНА КІБЕРНЕТИКА")

22/17-2012-02

Горлівка – 2012

МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ
ДЕРЖАВНИЙ ВИЩИЙ НАВЧАЛЬНИЙ ЗАКЛАД
«ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ»
АВТОМОБІЛЬНО-ДОРОЖНІЙ ІНСТИТУТ

«ЗАТВЕРДЖУЮ»
Директор АДІ ДВНЗ «ДонНТУ»
М. М. Чальцев

Кафедра «Прикладна математика та інформатика»

МЕТОДИЧНІ ВКАЗІВКИ ДО ВИКОНАННЯ ПРАКТИЧНИХ РОБІТ З ДИСЦИПЛІНИ «БАГАТОВИМІРНИЙ СТАТИСТИЧНИЙ АНАЛІЗ» (ДЛЯ СТУДЕНТІВ НАПРЯМУ ПІДГОТОВКИ 6.030502 «ЕКОНОМІЧНА КІБЕРНЕТИКА»)

22/17-2012-02

«РЕКОМЕНДОВАНО»
навчально-методична комісія
факультету
«Економіка і управління»
протокол № 9 від 15.06.2012

«РЕКОМЕНДОВАНО»
Кафедра «Прикладна математика
та інформатика»
протокол № 8 від 27.03.2012

Горлівка – 2012

УДК 519.23(07)

Методичні вказівки до виконання практичних робіт з дисципліни «Багатовимірний статистичний аналіз» (для студентів напряму підготовки 6.030502 «Економічна кібернетика») [Електронний ресурс] / укладачі: В. Г. Хребет, Д. В. Фесенко.–Електрон. дані. – Горлівка: ДВНЗ «ДонНТУ» АДІ, 2012. – 1 електрон. опт. диск (CD-R); 12 см. – Систем. вимоги: Pentium; 32 Mb RAM; WINDOWS 98/2000/NT/XP; MS Word 2000. – Назва з титул. екрану.

Вказівки містять загальні теоретичні відомості із найбільш використовуваних методів дисципліні «Багатовимірний статистичний аналіз» та методику і матеріали до виконання практичних робіт з цієї дисципліни

Укладачі: Хребет В. Г., к.ф.-м.н., доц., Фесенко Д. В.

Відповідальний за випуск: Хребет В. Г., к.ф.-м.н., доц.

Рецензент: Ніколаєнко В. Л., к.т.н., доц

Державний вищий навчальний заклад
«Донецький національний технічний університет»
Автомобільно-дорожній інститут, 2012

ЗМІСТ

ВСТУП

Курс багатовимірного статистичного аналізу є розділом прикладної статистики, що призначений для рішення наступних задач:

1) вивчення структури та сили статистичного зв'язку між системами випадкових величин;
2) заміна вихідного набору статистичних даних великої розмірності значно меньшим масивом даних без істотної втрати інформації – зменшення розмірності;
3) розбиття сукупності статистичних величин на однорідні в деякому відношенні групи – класифікація.

Аналіз та моделювання економічних систем, які є основними цілями економічної кібернетики, неможливі без вирішення перелічених задач. Оскільки аналіз системи полягає у розкладанні ії на складові частини та встановленні структури взаємодії між ними, а моделювання - це передусім максимально точне формулювання досліджуваного процесу або явища у якомога компактному вигляді методи багатовимірного аналізу є вкрай необхідними для підготовки фахівців із економічної кібернетики.

Дані методичні вказівки призначені для формування навичок розв'язання всіх проблем, які зустрічаються у багатовимірному статистичному аналізі: перша робота присвячена методам вимірювання кореляційних зв'язків між випадковими величинами, наступні три роботи – компонентний, факторний аналіз та шкалювання розв'язують задачу зниження розмірності та останні дві задачі – дискримінантний та кластерний аналізи формують навички автоматичної класифікації.

У зв'язку з тим, що більшість методів багатовимірного статистичного аналізу є досить трудомісткими вкрай необхідним є застосування спеціалізованих математичних пакетів, в якості якого була обрана популярна програма MathCad.

Розподіл робочого часу на виконання робіт

№ з/п	Робота	Витрати часу
1	Дослідження статистичних зв’язків багатовимірних випадкових величин.	6
2	Зменшення розмірності признакового простору за допомогою метода головних компонент.	6
3	Виявлення загальних латентних факторів за допомогою методу найменших залишків.	6
4	Використання алгоритму Торгесона для поновлення геометричної структури об’єктів у шкальній системі координат	6
5	Класифікація за допомогою лінійного дискримінантного аналізу	5
6	Розбиття сукупності об'єктів на типообразуючі множини за методами кластерного аналізу	5
	Усього	34

Практична робота № 1

Дослідження статистичних зв'язків багатовимірних випадкових величин методами кореляційного аналізу

Теоретичні відомості

Однією із задач статистичного аналізу є вимірювання тісности зв`язку між випадковимим величинами. Найпоширенішим з таких вимірювачів є коефіцієнт кореляціїї. Його застосування засновано на теоремі:

Необхідною й достатньою умовою незалежності двох підмножин нормально-розподілених випадкових величин є рівність нулю всіх коваріацій між величинами з різних підмножин:

$$cov(x_i, x_j) = M((x_i - M(x_i))(x_j - M(x_j))) = 0.$$

На практиці використовують не коваріацію, а більш зручний коефіцієнт кореляції:

$$\rho = \frac{cov(x_i, x_j)}{\sqrt{cov(x_i, x_i)cov(x_j, x_j)}} = \frac{cov(x_i, x_j)}{\sigma_i \sigma_j},$$

значення якого знаходиться в інтервалі [-1;1] (тобто є безрозмірним), що дозволяє порівнювати силу зв'язку між парами змінних. Так як коефіцієнт кореляції (як і коваріація) взагалі є невідомим параметром генеральної сукупності, постає задача його оцінки. Оцінку коефіцієнта кореляції одержують за формулою:

$$r_{ij} = \frac{\sum_k (x_{ik} - x_i)(x_{jk} - x_j)}{\sqrt{\sum_k (x_{ik} - x_i)^2 \sum_k (x_{jk} - x_j)^2}}.$$

Перевірка на значущість цього показника (тобто перевірка гіпотези $H_0: \rho = 0$) може бути виконана за допомогою статистики:

$$t = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}},$$

де t – випадкова величина, що розподілена за законом Стьюдента із n-2 ступенями вільності;

n – кількість спостережень.

Гіпотеза про незначущость коефіцієнта кореляціїї (неістотності зв'язку між змінними) приймають, якщо розрахункове значення t виявляється не більше критичного.

Більш точними вимірювачами статистичного зв'язку є часткові коефіцієнти кореляції, які враховують, що на зв'язок між двома величинами можуть впливати інші змінні та виключають цей вплив. Часткові коефіцієнти можуть бути обчислені за допомогою наступних рекурентних співвідношень:

- коефіцієнти першого порядку:

$$r_{ij\bullet k} = \frac{r_{ij} - r_{ik} r_{jk}}{\sqrt{\left(1 - r_{ik}^2\right)\left(1 - r_{jk}^2\right)}}$$

- коефіцієнти другого порядку:

$$r_{ij\bullet k,l} = \frac{r_{ij\bullet k} - r_{il\bullet k} r_{jl\bullet k}}{\sqrt{\left(1 - r_{ik\bullet k}^2\right)\left(1 - r_{jl\bullet k}^2\right)}}$$

і т. д.

Де $r_{ij\bullet k}$ - частковий коефіцієнт між i-ю та j-ю змінними при фіксованій (при виключеному впливу) k-й змінній.

Частковий коефіцієнт при фіксованих всіх змінних (окрім тих, між якими він розраховується) може бути обчислений одразу:

$$r_{i,i+1\bullet i+2,\ldots p} = -\frac{R_{i,i+1}}{\sqrt{R_{i,i} R_{i+1,i+1}}},$$

де $R_{i,i+1}$ - алгебраїчне доповнення до $(i,i+1)$ елементу кореляційної матриці.

Перевірка на значущість часткових коефіцієнтів кореляції виконується за допомогою статистики:

$$t = r\frac{\sqrt{n-q}}{\sqrt{1-r^2}}$$

де q – кількість зафіксованих змінних.

Приклад виконання роботи

$$X := \begin{pmatrix} 6.699 & -19.886 & 6.409 & 11.347 & 2.142 & -14.361 & -5.484 \\ -4.128 & 25.279 & 41.152 & -30.727 & 26.485 & 1.612 & -4.43 \\ -0.632 & 19.269 & 30.604 & -31.88 & 19.523 & 0.817 & 1.05 \\ -0.665 & 5.325 & 39.29 & -8.096 & 23.928 & -9.344 & -12.164 \\ -1.974 & -0.445 & -14.01 & 7.85 & -8.266 & 4.681 & 1.728 \\ -0.084 & -4.321 & -10.164 & 8.01 & -6.34 & 0.914 & 0.64 \\ 1.352 & -1.277 & 18.175 & -2.921 & 10.716 & -6.748 & -5.546 \\ 0.124 & 3.433 & 47.969 & -7.775 & 28.959 & -13.288 & -15.484 \\ -8.535 & 8.251 & -7.95 & 14.866 & -3.589 & 9.115 & -5.198 \\ -3.7 & 12.911 & 20.858 & -10.32 & 13.562 & 1.287 & -4.822 \end{pmatrix} \ldots$$

Задаємо розмірність вихідних даних:

$n := 15 \quad p := 10$

$i := 1, 2 .. p \quad j := 1, 2 .. p$

Матриця оцінок коефіцієнтів кореляції

$$Xsr_i := \sum_{t=1}^{n} X_{i,t}$$

$$R_{i,j} := \frac{\sum_{k=1}^{n} \left[\left(X_{i,k} - Xsr_i\right) \cdot \left(X_{j,k} - Xsr_j\right)\right]}{\sqrt{\sum_{k=1}^{n} \left(X_{i,k} - Xsr_i\right)^2 \cdot \sum_{k=1}^{n} \left(X_{j,k} - Xsr_j\right)^2}}$$

$R =$

	1	2	3	4	5	6	7	8
1	1	-0.425	-0.454	-0.268	0.281	0.397	-0.201	-0.245
2	-0.425	1	0.994	0.982	-0.983	-0.997	0.971	0.977
3	-0.454	0.994	1	0.96	-0.983	-0.997	0.95	0.954
4	-0.268	0.982	0.96	1	-0.981	-0.977	0.997	...

Для перевірки на значущість кореляцій обчислимо *t*-статистики та порівняємо їх з критичним значенням на 5% рівні значущості:

$$tk := qt(0.975, n-2)$$

$$tk = 2.16$$

$$t := \left| \begin{array}{l} \text{for } i \in 1,2..p \\ \quad \text{for } j \in 1,2..p \\ \quad\quad t_{i,j} \leftarrow R_{i,j} \dfrac{\sqrt{n-2}}{\sqrt{1-(R_{i,j})^2}} \quad \text{if } i \neq j \\ \quad\quad 0 \quad \text{otherwise} \\ t \end{array} \right.$$

$$t = \begin{pmatrix} 0 & -1.693 & -1.836 & -1.002 & 1.054 & 1.559 \\ -1.693 & 0 & 32.206 & 18.553 & -19.552 & -48.316 \\ -1.836 & 32.206 & 0 & 12.322 & -19.038 & -48.244 \\ -1.002 & 18.553 & 12.322 & 0 & -18.254 & -16.648 \\ 1.054 & -19.552 & -19.038 & -18.254 & 0 & 28.278 \\ 1.559 & -48.316 & -48.244 & -16.648 & 28.278 & 0 \\ -0.741 & 14.529 & 10.994 & 46.183 & -19.963 & -14.588 \\ -0.91 & 16.585 & 11.504 & 151.577 & -17.715 & -15.318 \\ -0.132 & -1.106 & -1.502 & -0.843 & 1.613 & 1.399 \\ -1.572 & 37.211 & 17.081 & 25.557 & -15.373 & -22.157 \end{pmatrix} \ldots$$

Для наочного представлення результатів побудуємо схему значущих зв`язків (рисунок 1.1).

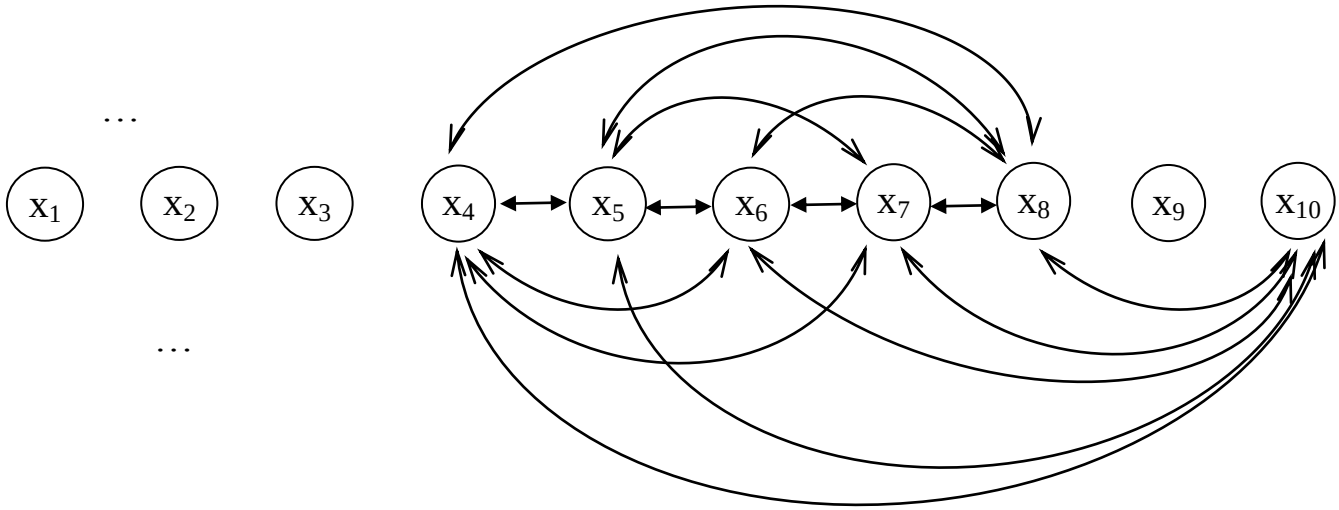

Рисунок 1.1 – Схема кореляційних зв`язків

Як видно, між більшістю змінних є тісні зв'язки. Для встановлення структури зв'язків обчислимо часткові коефіцієнти кореляції фіксуючі всі змінні. Результати наведені на рисунку 1.2.

$$r_{i,j} := -\frac{\left(R^{-1}\right)_{j,i}}{\sqrt{\left(R^{-1}\right)_{i,i}\cdot\left(R^{-1}\right)_{j,j}}}$$

$$t := \left| \begin{array}{l} \text{for } i \in 1,2..p \\ \quad \text{for } j \in 1,2..p \\ \quad\quad \left| \begin{array}{ll} t_{i,j} \leftarrow r_{i,j}\dfrac{\sqrt{n-(p-2)}}{\sqrt{1-\left(r_{i,j}\right)^2}} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{array} \right. \\ t \end{array} \right.$$

$$tk := qt[0.975, n-(p-2)]$$

$$tk = 2.365$$

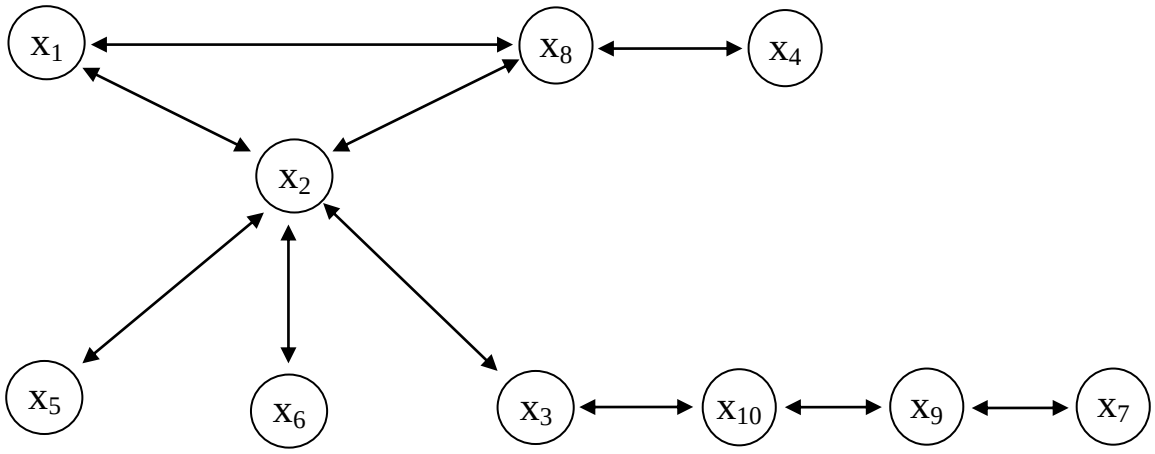

Рисунок 1.2 – Схема значущих «чистих» зв’язків

Отримана схема дає передумови зменшення розмірності насамперед за рахунок змінної x_2, яка акумулює варіацію п’яти інших змінних.

Завдання

1. Побудувати кореляційну матрицю семи показників та визначити пари змінних, між якими на рівні значущості 5 % є значущі зв’язки.
2. Побудувати матрицю часткових коефіцієнтів кореляції та проаналізувати зміну структури значущих зв`язків при варіації рівня значущості від 10 % до 1 %.
3. Обґрунтувати передумови зменшення розмірності ознакового простору.

Практична робота № 2

Зменшення розмірності признакового простору за допомогою метода головних компонент

Теоретичні відомості

Однією із важливих задач у статистичній практиці є представлення даних у зручній формі, яка дозволяє наочно побачити структуру даних та полегшує її аналіз та обробку. У зв`язку з тим, що досліджувані масиви змінних часто мають досить значні кореляції з`являється можливість такого перетворення змінних, при якому нові змінні не корелюють один з одним. Так як при цьому загальна варіація залишається незмінною можуть з`явитися змінні з досить невеликою дисперсією й це дає привід для виключення цих змінних із подальшого аналізу. При лінійному перетворенні нові змінні визначаються як:

$$z = X \cdot A,$$

де X – матриця, у стовпцях якої знаходяться значення вихідних змінних;

A – матриця, що складена із одиничних і ортогональних векторів (головних компонент), які відповідають новим змінним.

Головні компоненти визначають таким чином, щоб дисперсія нових змінних була найбільшою. Можна показати, що при цьому стовпці матриці A повинні бути власними векторами матриці X^TX, причому перша компонента є тим власним вектором, якому відповідає найбільше власне значення, друга – вектором з другим за величиною власним значенням і т. д. Рішення про необхідну кількість головних компонент приймають за допомогою декількох критеріїв. Згідно з першим (критерій кам'янистого осипу) виділення змінних закінчують, коли інформативність нової виділеної компоненти значно меньше попередніх. Другий критерій визначає необхідну кількість компонент, як таку, при якій виділена дисперсія досягає визначеної досить великої величини. Якщо вихідними даними є нормовані значення, то замість матриці X^TX використовують кореляційну матрицю R.

У зв`язку з труднощами точного визначення власних векторів і значень використовують чисельний алгоритм Хотеллінга. У його основу покладений той факт, що при перемноженні будь-якого вектора на матрицю той повертається таким чином, що зменьшується кут між ним та тим власним вектором, якому відповідає найбільше власне значення. Тому при достатній

кількості поворотів можна отримати скільки завгодно точне наближення до власного вектора.

Приклад виконання роботи

Кореляційну матрицю беремо з першої роботи.
Задаємо кількість спостережень, змінних та діапазони змін індексів:

$$n := 15 \qquad p := 10$$

$$i := 1,2..p \quad j := 1,2..p \quad t := 1,2..n$$

Стандартизуємо дані

$$Xsr_i := \sum_{t=1}^{n} \frac{X_{i,t}}{n} \qquad \sigma_i := \sqrt{\sum_{t=1}^{n} \left(X_{i,t} - \sum_{t=1}^{n} \frac{X_{i,t}}{n} \right)^2}$$

Переходимо до виділення першої головної компоненти

$$a_i := \sum_{j=1}^{p} R_{i,j} \qquad a := \frac{a}{\sqrt{\sum_{i=1}^{p} (a_i)^2}}$$

$a =$

	1
1	-0.038
2	0.345
3	0.323
4	0.372
5	...

$$a := R^{14} \cdot a \qquad a := \frac{a}{\sqrt{\sum_{i=1}^{p} (a_i)^2}}$$

$a =$

	1
1	-0.131
2	0.351
3	0.349
4	0.347
5	...

Виконаємо перевірку та обчислимо перше власне значення

$$\lambda_1 := \frac{(R\cdot a)_1}{a_1} \qquad \frac{(R\cdot a)_i}{a_i} =$$

8.097
8.097
8.097
...

Для отримання досить точного наближення вистачило зведення кореляційної матриці у чотирнадцяту ступінь, оскільки відношення компонент вектора *Ra* до компонент вектора *a* є постійним.
Обчислюємо залишкову матрицю та виділяємо другу та третю головні компоненти:

$$A := a$$

$$R := R - \lambda_1 \cdot a a^T$$

$$a_i := \sum_{j=1}^{p} R_{i,j} \qquad a := \frac{a}{\sqrt{\sum_{i=1}^{p} (a_i)^2}}$$

$a =$

	1
1	0.541
2	-0.006
3	-0.126
4	0.179
5	...

$$a := R^{56} \cdot a \qquad a := \frac{a}{\sqrt{\sum_{i=1}^{p} (a_i)^2}}$$

$a =$

	1
1	-0.69
2	0.061
3	0.02
4	-0.001
5	...

$$\lambda_2 := \frac{(R\cdot a)_1}{a_1} \qquad \frac{(R\cdot a)_i}{a_i} =$$

1.071
1.071
...

$$R := R - \lambda_2 \cdot a\, a^T$$

$$A := augment(A, a)$$

$$a_i := \sum_{j=1}^{p} R_{i,j} \qquad a := \frac{a}{\sqrt{\sum_{i=1}^{p} (a_i)^2}} \qquad a =$$

	1
1	0.651
2	-0.016
3	-0.13
4	0.181
5	...

$$a := R^2 \cdot a \qquad a := \frac{a}{\sqrt{\sum_{i=1}^{p} (a_i)^2}} \qquad a =$$

	1
1	0.651
2	-0.016
3	-0.13
4	0.18
5	...

$$\lambda_3 := \frac{(R \cdot a)_1}{a_1} \qquad \frac{(R \cdot a)_i}{a_i} =$$

0.832
0.832
0.832
...

$$\lambda_1 + \lambda_2 + \lambda_3 = 10$$

Так як накоплена сума власних значень практично співпадає з загальною дисперсією (рівною сумі діагональних елементів кореляційної матриці) немає необхідності у виділенні інших головних компонент.

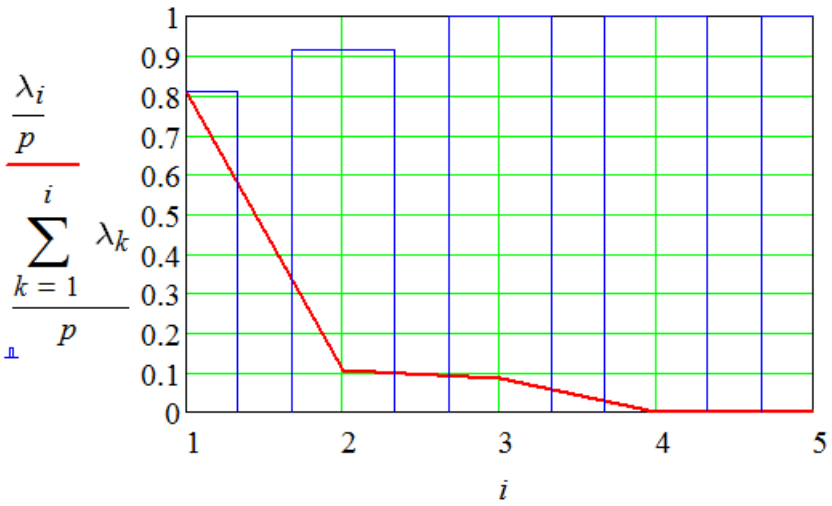

Рисунок 2.1 – Розподіл дисперсії по компонентах

Із рисунка 2.1 видно, що оптимальною кількістю головних компонент за критерієм кам`янистого осипу буде дві компоненти.

Коефіцієнти кореляції між признаками та компонентами:

$$j := 1,2..2$$

$$r_{i,j} := \frac{\sqrt{\lambda_j} \cdot A_{i,j}}{\sqrt{\sum_{t=1}^{n} (X_{i,t})^2}}$$

$r^T =$

	1	2	3	4	5	6	7	8	9	10
1	-0.372	0.998	0.993	0.986	-0.993	-0.999	0.98	0.983	-0.331	0.992
2	-0.714	0.064	0.021	-0.001	0.116	0.004	-0.077	-0.017	0.726	0.104

Виділення другої компоненти обумовлене першим та дев`ятим признаками, варіація останніх признаків досить точно пояснюється першою компонентою.

Для оцінки якості стиску даних обчислимо координати об`єктів у системі головних компонент та виконаємо зворотне перетворення за допомогою лише направляючих векторів головних компонент:

$$A := submatrix(A, 1, 10, 1, 2)$$

$$Z := A^T \cdot X$$

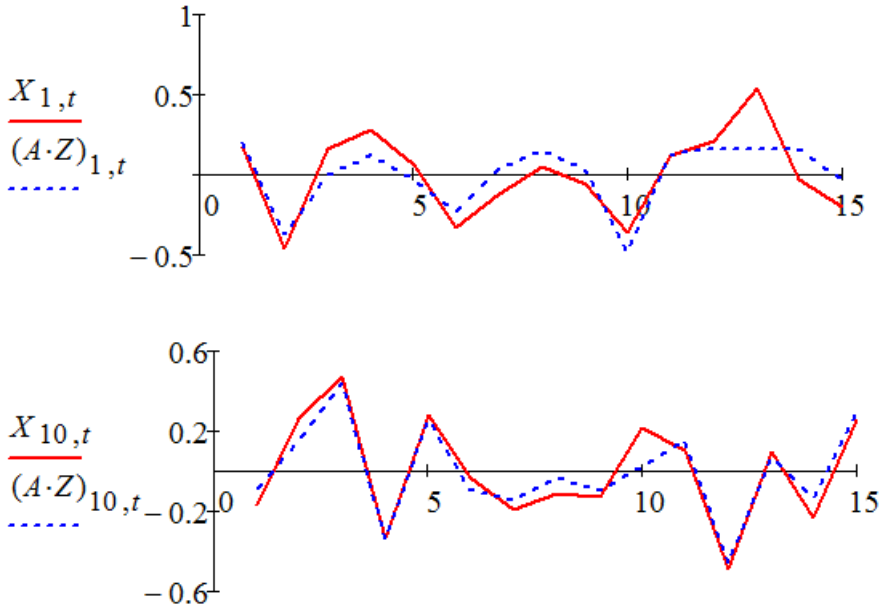

Рисунок 2.2 – Відновлення першої та десятої змінних за даними головних компонент

Як можна бачити при п’ятикратному стиску даних вдається зберегти більшість інформації, що міститься у змінних.

Завдання

1. За допомогою алгоритму Хотеллінга виділити головні компоненти системи семивимірного випадкового вектора.
2. Визначити доцільну кількість головних компонент на основі методу «кам’янистого осипу» та проаналізувати розподіл варіації змінних по компонентах.
3. Обчислити значення головних компонент у кожному спостереженні та оцініти точність самовідтворення всіх вихідних ознак.

Практична робота № 3

Виявлення загальних латентних факторів за допомогою методу послідовних замін

Теоретичні відомості

Факторний аналіз, як і компонентний аналіз відноситься до методів зменшення розмірностей, але завданням методу є не перехід до меньшої кількості нових змінних, які б акумулювали якомога більшу частину дисперсії системи, а пошук факторів, які б пояснювали взаємозв`язки. Лінійна модель факторного аналізу записується у вигляді:

$$x_{ij} = q_{i1} f_{1j} + q_{i2} f_{2j} + .. + u_{ij},$$

де x_{ij} – значення *i*-й змінной в *j*-му спостереженні;

f_{1j} – значення першого фактора в *j*-му спостереженні;

q_{i1} - коефіцієнт, що характеризує вплив 1-го фактора на *i*-у змінну;

u_{ij} - специфічність *i*-ї змінної в *j*-му спостереженні.

Наявність специфічності викликана тим, що кожній змінній властива варіація, яка не може бути пояснена тільки факторами. Для спрощення загальної моделі коваріаційну матрицю факторів вважають одиничною, а коваріаційну матрицю специфічностей вважають діагональною.
Наявність зв`язків між змінними відображається на елементах кореляційної матриці, яка при зроблених припущеннях розкладається на:

$$R = QQ^T + V$$

Згідно з цим рівнянням завданням факторного аналізу є пошук такої матриці *Q* найменьшої розмірності, яка б мінімізувала елементи матриці *V*. Ця модель є більш загальною та більш гнучкою в порівнянні з методом головних компонент, що відображається й на її складності – кількості невідомих. Найбільш точно задачу побудови такої моделі вирішує метод найменших залишків. Одночасне оцінювання всієї матриці навантажень призводить до складної системи рівнянь, тому був запропонований наближений метод послідовних замін (метод Гаусса – Зейделя). Він являє собою ітераційний процес, на кожному кроці якого знаходяться елементи одного рядка матриці навантажень *Q*, які мінімізують расходження недіагональних елементів матриць *R* та QQ^T. Цей процес повторюється для кожного рядка поки значення критерію не перестане змінюватися. Рядки

матриці навантажень, що мінімізують поточне розходження між оціненою та вихідною кореляційними матрицями обчислюють за наступною формулою:

$$\varepsilon_j = r_j^0 Q\left(\tilde{Q}^T \tilde{Q}\right)^{-1}$$

де r_j^0 – залишкові коефіцієнти кореляції в j-му рядку:

$$r_{jk}^0 = r_{jk} - \sum_t q_{jt} q_{kt}; k = 1,2..p$$

p – кількість змінних;

$\tilde{Q}$ – матриця навантажень, в якій елементи j-го рядка замінені нулями.

Недоліком цієї процедури є можливість отримати оцінені дисперсії, що перевищують 1 (так званий варіант Хевісайда). Можна показати, що у тому випадку, коли рішення без обмеження на дисперсії призводить до варіанта Хевісайда, рішення з обмеженням буде знаходитися точно на межі області $\sum_{t=1}^{p'} q_{it}^2 \le 1$ (де p' – кількість факторів), тому для корегування рішення можна пропорційно зменшити елементи i-го рядка таким чином, щоб $\sum_{t=1}^{p'} q_{it}^2 = 1$ і повторити оптимізацію інших рядків.

Приклад виконання роботи

Вихідні дані – кореляційну матрицю беремо з першої роботи. На початку задаємо розмірність та обираємо начальне наближення матриці Q таким чином, щоб матриця $\tilde{Q}^T \tilde{Q}$ була не виродженою:

$i := 1,2..10 \qquad j := 1,2..2$

$Q_{i,j} := cos(j)^i$

На першій ітерації послідовно обчислюємо елементи рядків матриці навантажень та одночасно оцінюємо розходження кореляційної матриці та її оцінки (яке повинно монотонно зменшуватися).

$$\sum_{i=1}^{9}\sum_{j=i+1}^{10}\left[\left(R - Q\cdot Q^{T}\right)_{i,j}\right]^{2} = 28.808 \qquad Q_{1,j} := 0$$

$$\left(Q_{1,1}\ \ Q_{1,2}\right) := \left(R - Q\cdot Q^{T}\right)^{\langle 1\rangle^{T}}\cdot Q\cdot\left(Q^{T}\cdot Q\right)^{-1}$$

$$\sum_{i=1}^{9}\sum_{j=i+1}^{10}\left[\left(R - Q\cdot Q^{T}\right)_{i,j}\right]^{2} = 28.257 \qquad Q_{2,j} := 0$$

$$\left(Q_{2,1}\ \ Q_{2,2}\right) := \left(R - Q\cdot Q^{T}\right)^{\langle 2\rangle^{T}}\cdot Q\cdot\left(Q^{T}\cdot Q\right)^{-1}$$

…

$$\left(Q_{10,1}\ \ Q_{10,2}\right) := \left(R - Q\cdot Q^{T}\right)^{\langle 10\rangle^{T}}\cdot Q\cdot\left(Q^{T}\cdot Q\right)^{-1}$$

$$\sum_{i=1}^{9}\sum_{j=i+1}^{10}\left[\left(R - Q\cdot Q^{T}\right)_{i,j}\right]^{2} = 18.311$$

Так як початкове значення критерію збіжності на початку ітерації не співпадає зі значенням на останньому кроці (28.808 і 18.311) переходимо до наступної ітерації, на якій в якості початкового наближення використовуємо матрицю *Q*, що отримана на попередній ітерації. Таким чином повторюємо ітерації поки критерій збіжності не стабілізується:

$Q =$

	1	2
1	-0.615	-0.897
2	0.98	-0.199
3	0.979	-0.176
4	0.932	...

$Q\cdot Q^{T} =$

	1	2	3	4
1	1.183	-0.424	-0.444	-0.274
2	-0.424	1	0.994	0.98
3	-0.444	0.994	0.989	0.971
4	-0.274	0.98	0.971	...

Серед спільності признаків – діагональних елементів матриці $\tilde{Q}^{T}\tilde{Q}$ деякі перевищують одиницю. Найбільшим серед них є перший елемент. З метою корегування рішення зменшимо довжину першого вектор-рядка

матриці Q залишаючи незмінним його напрямок та виконаємо процедуру мінімізації не перераховуючи більше елементи першого рядка:

$$\left(Q_{1,1}\ \ Q_{1,2}\right) := \frac{1}{\sqrt{\left(Q_{1,1}\right)^2 + \left(Q_{1,2}\right)^2}} \cdot \left(Q_{1,1}\ \ Q_{1,2}\right)$$

$$\sum_{i=1}^{9} \sum_{j=i+1}^{10} \left[\left(R - Q \cdot Q^T\right)_{i,j}\right]^2 = 0.062 \qquad Q_{2,j} := 0$$

$$\left(Q_{2,1}\ \ Q_{2,2}\right) := \left(R - Q \cdot Q^T\right)^{\langle 2 \rangle^T} \cdot Q \cdot \left(Q^T \cdot Q\right)^{-1}$$

$$\sum_{i=1}^{9} \sum_{j=i+1}^{10} \left[\left(R - Q \cdot Q^T\right)_{i,j}\right]^2 = 0.059 \qquad Q_{3,j} := 0$$

$$\left(Q_{3,1}\ \ Q_{3,2}\right) := \left(R - Q \cdot Q^T\right)^{\langle 3 \rangle^T} \cdot Q \cdot \left(Q^T \cdot Q\right)^{-1}$$

…

За результатами мінімізації оцінюємо матрицю кореляцій:

$$Q \cdot Q^T = \begin{pmatrix} 1 & -0.42 & -0.44 & -0.27 & 0.27 & 0.39 & -0.2 & -0.25 & -0.01 & -0.41 \\ -0.42 & 1 & 1 & 0.98 & -0.99 & -1 & 0.97 & 0.97 & -0.29 & 0.99 \\ -0.44 & 1 & 0.99 & 0.97 & -0.98 & -0.99 & 0.96 & 0.97 & -0.28 & 0.99 \\ -0.27 & 0.98 & 0.97 & 0.98 & -0.99 & -0.98 & 0.98 & 0.98 & -0.3 & 0.97 \\ 0.27 & -0.99 & -0.98 & -0.99 & 1 & 0.99 & -0.99 & -0.99 & 0.31 & -0.98 \\ 0.39 & -1 & -0.99 & -0.98 & 0.99 & 1 & -0.98 & -0.98 & 0.29 & -0.99 \\ -0.2 & 0.97 & 0.96 & 0.98 & -0.99 & -0.98 & 0.99 & 0.98 & -0.31 & 0.97 \\ -0.25 & 0.97 & 0.97 & 0.98 & -0.99 & -0.98 & 0.98 & 0.98 & -0.31 & 0.97 \\ -0.01 & -0.29 & -0.28 & -0.3 & 0.31 & 0.29 & -0.31 & -0.31 & 0.1 & -0.29 \\ -0.41 & 0.99 & 0.99 & 0.97 & -0.98 & -0.99 & 0.97 & 0.97 & -0.29 & 0.99 \end{pmatrix}$$

Порівняємо результати компонентного аналізу (для двох компонент) та факторного аналізу й побудуємо діаграму векторів-признаків:

$$\sum_{i=1}^{9}\sum_{j=i+1}^{10}\left[\left(R - A\cdot diag(\lambda)\cdot A^{T}\right)_{i,j}\right]^2 = 0.216$$

$$\sum_{i=1}^{9}\sum_{j=i+1}^{10}\left[\left(R - Q\cdot Q^{T}\right)_{i,j}\right]^2 = 0.049$$

$k := 1,2..20 \qquad Ql_{2\cdot i,j} := Q_{i,j}$

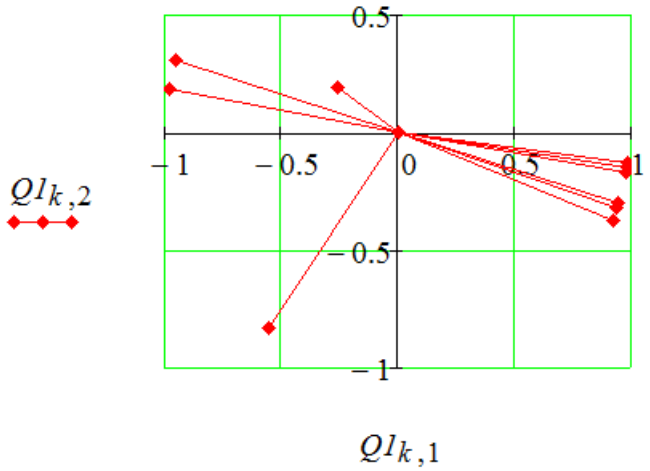

Рисунок 3.1 – Діаграма розподілу вектор-признаків

Можна зробити висновок, що метод мінімальних залишків майже в 4 рази краще пояснює кореляції й показує, що за допомогою двох факторів дуже добре моделюються зв`язки між дев'ятьма признаками – признаки з другого по восьмий тісно пов`язані з першим фактором, а перший – з другим. Варіацію дев`ятого признака в рамках двохфакторної моделі пояснити не вдається.

Завдання

1. Оцінити матрицю навантажень однофакторної моделі методом послідовних замін. В якості початкового наближення обрати першу головну компоненту із практичної роботи № 2.
2. Скорегувати рішення у випадку, якщо оцінені дисперсії перевищують одиницю.

3. Повторити пункти 1 і 2 для двофакторної моделі та обрати найбільш придатну.
4. Проаналізувати розподіл варіації признаків та порівняти результати аналізу з результатами метода головних компонент.

Практична робота № 4

Використання алгоритму Торгесона для поновлення геометричної структури об'єктів у шкальній системі координат

Теоретичні відомості

Методи шкалювання використовують у ситуаціях, коли вихідні дані подані не у вигляді таблиці «спостереження–змінна», а як міри розбіжності між об’єктами – спостереженнями. Завданням шкалювання у цьому випадку є встановлення змінних (шкал), на основі яких можуть бути пояснені зареєстровані розбіжності спостережень. Якщо розбіжності досить точно підпорядковані декартовій метриці

$$\delta_{ij} = \sqrt{\sum_t (x_{it} - x_{jt})^2},$$

то основою для шкалювання є теорема Торгесона:

Якщо

$$\Delta_{ij} = -\frac{1}{2}\left(\delta_{ij}^2 - \delta_{\cdot j}^2 - \delta_{i\cdot}^2 + \delta_{\cdot\cdot}^2\right)$$

$$\delta_{\cdot j}^2 = \frac{1}{n}\sum_i \delta_{ij}^2,\ \delta_{i\cdot}^2 = \frac{1}{n}\sum_j \delta_{ij}^2,\ \delta_{\cdot\cdot}^2 = \frac{1}{n^2}\sum_{i,j} \delta_{ij}^2,$$

де n – кількість об’єктів,
то

$$\Delta = X \cdot X^T$$

Так як будь-яка симетрична матриця може бути представлена у вигляді:

$$\Delta = C\Lambda C^T = \left(C\Lambda^{1/2}\right)\cdot\left(C\Lambda^{1/2}\right)^T,$$

то для вирішення задачі шкалювання необхідно знайти власні вектори та власні числа матриці Δ, що можна зробити за допомогою алгоритма Хотеллінга.

Приклад виконання роботи:

$$\delta := \begin{pmatrix} 0 & 3.62 & 1.11 & 12.59 & 5.48 \\ 3.62 & 0 & 5.5 & 13.2 & 7.94 \\ 1.11 & 5.5 & 0 & 12.31 & 5.78 \\ 12.59 & 13.2 & 12.31 & 0 & 18.03 \\ 5.48 & 7.94 & 5.78 & 18.03 & 0 \end{pmatrix}$$

Перевіримо виконання третьої аксіоми відстаней

$i := 1,2..5 \qquad j := 1,2..5 \quad k := 1,2..5$

$w1_{i,j} := \left(\delta_{i,j} - \delta_{i,1} - \delta_{1,j}\right) \qquad w2_{i,j} := \left(\delta_{i,j} - \delta_{i,2} - \delta_{2,j}\right)$

$w4_{i,j} := \left(\delta_{i,j} - \delta_{i,4} - \delta_{4,j}\right) \qquad w5_{i,j} := \left(\delta_{i,j} - \delta_{i,5} - \delta_{5,j}\right)$

$w3_{i,j} := \left(\delta_{i,j} - \delta_{i,3} - \delta_{3,j}\right)$

$max(stack(w1, w2, w3, w4, w5)) = 0.77$

Так як найбільша різниця позитивна коректуємо матрицю розрізнень:

$\delta := \delta + 0.77 \qquad \delta_{i,i} := 0$

Обчислимо матрицю із подвійним центруванням:

$$\Delta_{i,j} := \frac{-1}{2} \cdot \left[\left(\delta_{i,j}\right)^2 - \frac{1}{5} \cdot \sum_{k=1}^{5} \left(\delta_{i,k}\right)^2 - \frac{1}{5} \cdot \sum_{k=1}^{5} \left(\delta_{k,j}\right)^2 + \frac{1}{25} \cdot \sum_{k=1}^{5} \sum_{t=1}^{5} \left(\delta_{k,t}\right)^2 \right]$$

Перевірка:

$$\sum_{j=1}^{5} \Delta_{i,j} \qquad \sum_{i=1}^{5} \Delta_{i,j}$$

0	0
0	0
0	0
0	0
0	0

За результатами алгоритму Хотеллінга отримали наступні результати:

$$A = \begin{pmatrix} -0.122 & -0.02 & -0.327 & -0.823 & 0.447 \\ -0.099 & -0.85 & 0.095 & 0.241 & 0.447 \\ -0.095 & 0.317 & -0.656 & 0.51 & 0.447 \\ 0.835 & 0.171 & 0.272 & 0.007 & 0.447 \\ -0.519 & 0.383 & 0.617 & 0.066 & 0.447 \end{pmatrix} \qquad \lambda = \begin{pmatrix} 191.869 \\ 26.678 \\ 5.198 \\ -0.12 \\ 0 \end{pmatrix}$$

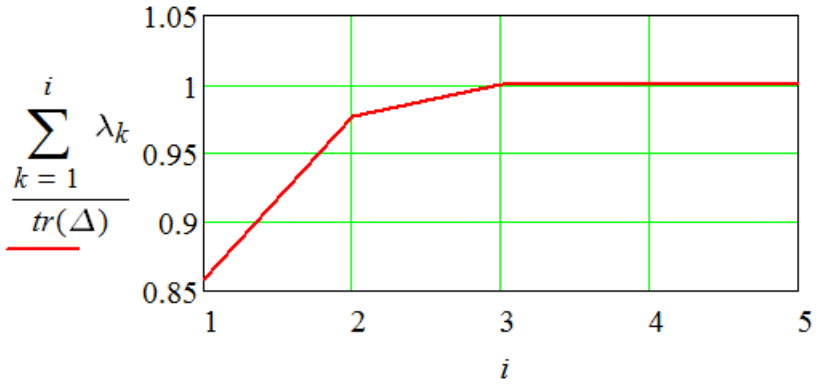

Рисунок 3.1 – Накопичена дисперсія

Видно, що серед власних значень є одно негативне, тому виділити можна не більше трьох шкал. Оптимальною кількістю шкал буде 2, тому що вони пояснюють більше 95 % варіації.

Координати об`єктів у шкальному просторі:

$$X := augment\left(\sqrt{\lambda_1} \cdot A^{\langle 1 \rangle}, \sqrt{\lambda_2} \cdot A^{\langle 2 \rangle}\right)$$

Оцінені розрізнення

$$\delta 1_{i,j} := \sqrt{\sum_{k=1}^{2} \left(X_{i,k} - X_{j,k}\right)^2}$$

$$\delta 1 = \begin{pmatrix} 0 & 4.299 & 1.784 & 13.293 & 5.876 \\ 4.299 & 0 & 6.032 & 13.964 & 8.629 \\ 1.784 & 6.032 & 0 & 12.909 & 5.874 \\ 13.293 & 13.964 & 12.909 & 0 & 18.784 \\ 5.876 & 8.629 & 5.874 & 18.784 & 0 \end{pmatrix}$$

Порівнюючи оцінені та зареєстровані розрізнення видно, що шкалювання виконано досить точно.

Завдання

1. Перевірити виконання аксіом відстаней та у разі необхідності скорегувати елементи матриці розрізнень.
2. Побудувати матрицю з подвійним центруванням та знайти координати об’єктів у шкальному просторі.
3. Визначити доцільну кількість шкал та оцінити точність шкалювання.

Практична робота № 5

Класифікація за допомогою лінійного дискримінантного аналізу

Теоретичні відомості

Дискримінантний аналіз відноситься до однієї із двох великих груп методів класифікації – класифікації з навчанням. Призначення цих методів - розбиття усієї сукупності даних на однорідні множини. Методи дискримінантного аналізу базуються на теоретико –імовірнісному або геометричному підходах. Критерієм розбиття у першому випадку є мінімальне математичне очікування витрат від помилкової класифікації. Для двох класів правило класифікації (дискримінант не правило) виглядає як:

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)p_2}{c(2|1)p_1};$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)p_2}{c(2|1)p_1},$$

де R_1 , R_2 – множини, до яких відносять спостереження з першого та другого класів;

f_1, f_2 – щільності розподілу в класах;

$c(1|2), c(2|1)$ – штрафи за помилкову класифікацію спостереження з другого класу, як спостереження з першого та на навпаки;

p_1, p_2 – імовірність отримати спостереження з першого або другого класу.

Критерієм розбиття при використанні геометричного підходу є відношення внутрішньокласової дисперсії до міжкласової. Якщо обидва класи розподілені за нормальним законом і відрізняються тільки математичним очікуванням, то дискримінант виявляється однаковим при геометричному і імовірнісному підходах:

$$R_1 : x'\delta \geq \frac{1}{2}(\mu_1 + \mu_2)'\delta$$

$$R_1 : x'\delta < \frac{1}{2}(\mu_1 + \mu_2)'\delta$$

$$\delta = \Sigma^{-1}(\mu_1 - \mu_2)$$

де Σ – коваріаційна матриця;

μ_1, μ_2 – математичні очікування в класах.

Для оцінки параметрів генеральних сукупностей μ_1, μ_2 та Σ використовують допоміжну вибірку, яку називають навчальною.

Якщо крім навчальної вибірки є екзаменуюча, то помилку класифікації можна оцінити як:

$$p_1 \frac{n_1}{m_1} + p_2 \frac{n_2}{m_2}$$

Приклад виконання роботи

Вибірка з першого класу Вибірка з другого класу

$$X1 := \begin{pmatrix} 0.03 & 0.32 & 2.62 \\ 0.74 & 0.74 & 1.82 \\ 1.98 & 0.25 & 1.37 \\ 0.86 & 0.21 & 2.3 \\ 1.58 & 0.77 & 1.48 \\ 1.27 & 0.5 & 1.32 \\ 1.46 & 0.45 & 2.22 \\ 1.6 & 1.31 & 1.4 \\ 0.79 & 0.56 & 1.6 \\ 0.86 & 0.62 & 1.09 \end{pmatrix} \qquad X2 := \begin{pmatrix} 0.45 & 0.66 & 1.43 \\ 0.33 & 0.15 & 1.51 \\ 0.68 & 1.2 & 0.68 \\ 0.34 & 1.76 & 1.53 \\ 0.66 & 0.41 & 0.62 \end{pmatrix}$$

$i := 1, 2 .. 3$

$\mu 1_i := mean\left(X1^{\langle i \rangle}\right) \qquad \mu 2_i := mean\left(X2^{\langle i \rangle}\right)$

Центровані спостереження

$$k1 := 1,2..rows(X1) \qquad k2 := 1,2..rows(X2)$$

$$X1_{k1,i} := X1_{k1,i} - \mu1_i \quad X2_{k2,i} := X2_{k2,i} - \mu2_i$$

$$S := \frac{1}{15-2}\cdot\left(X1^T X1 + X2^T X2\right)$$

$$\delta := S^{-1}\cdot(\mu1 - \mu2)$$

Екзаменуюча вибірка

$$X1 := \begin{pmatrix} 0.43 & 0.18 & 1.65 \\ 0.88 & 0.15 & 1.91 \\ 0.57 & 0.34 & 1.68 \\ 1.72 & 0.38 & 1.94 \\ 1.7 & 0.09 & 1.89 \\ 0.84 & 0.14 & 1.94 \\ 0.6 & 0.21 & 2.06 \end{pmatrix} \qquad X2 := \begin{pmatrix} 0.24 & 0.23 & 1.54 \\ 0.67 & 1.1 & 1.07 \\ 0.54 & 0.68 & 1.22 \end{pmatrix}$$

Прокласифікуємо елементи екзаменуючої вибірки:

$$X1\cdot\delta - \frac{1}{2}\cdot(\mu1 + \mu2)^T\delta = \begin{pmatrix} -0.522 \\ 3.64 \\ 0.313 \\ 8.597 \\ 8.509 \\ 3.576 \\ 2.725 \end{pmatrix} \qquad X2\cdot\delta - \frac{1}{2}\cdot(\mu1 + \mu2)^T\delta = \begin{pmatrix} -2.32 \\ -3.252 \\ -2.761 \end{pmatrix}$$

На об`єктах з першого класу дискримінаційна функція повинна приймати позитивні значення, а на об`єктах з другого – негативні. Тому при рівних імовірностях отримати спостереження з різних класів імовірність помилкової класифікації дорівнює:

$$0.5\cdot\frac{1}{7} + 0.5\cdot\frac{0}{3} = 0.071$$

Завдання

1. По заданій навчальній вибірці семи показників побудувати лінійне дискримінантне правило та оцінити його точність.
2. Повторити п.1, використовуючи в якості вихідних даних координати об’єктів у системі головних компонент.

Практична робота № 6

Розбиття сукупності об'єктів на типообразуючі множини за методами кластерного аналізу

Теоретичні відомості

До кластерного аналізу відносять методи класифікації без наявності навчаючої вибірки. Як правило, кількість класів (кластерів) з'ясовується також під час класифікації. Існує велика кількість методів кластерного аналізу, яку умовно можна поділити на такі підгрупи:

- Ієрархічні – на кожному кроці відбувається злиття двох найбільш близьких кластерів (агломеративні методи) або поділ кластера на два найбільш несхожих (дивізімні методи). Класифікація завершується, коли поєднуються два дуже несхожих кластери (або кластер розбивається на два дуже близьких кластери).
- Ітеративні методи. Спочатку вся множина об'єктів розбивається на деяку кількість кластерів, для кожного з яких обчислюється їхній центр ваги. На другому кроці об'єкти перерозподіляються по кластерах з урахуванням їх близькості до центрів ваг, що служить базою для визначення нових центрів. Такий перерозподіл виконується до стабілізації центрів.
- Методи пошуку модальних значень щільності націлені на відшукання таких областей, в яких щільність об'єктів максимальна.
- Методи, що використовують теорію графів.

Використання кластерного аналізу потребує оцінки міри розходження кластерів та об`єктів (метрик). Використовують наступні види метрик об'єктів:

а) коефіцієнт кореляції – оцінює форму (характер зміни показника), але не чутливий до розрізнень у величинах;

б) міри відстаней – чутливі до абсолютних розрізнень та схильні нехтувати динамікою. Використовують такі міри відстаней:

1. Евклідова

$$d_{ij} = \sqrt{\sum_k \left(x_{ik} - x_{jk} \right)^2}$$

2. Манхетенська

$$d_{ij} = \sum_k \left| x_{ik} - x_{jk} \right|$$

3. Мінковського – є узагальненням більшості мір відстаней:

$$d^r{}_{ij} = \left(\sum_k \left(x_{ik} - x_{jk} \right)^r \right)^{1/r}$$

в) коефіцієнти асоціативності – призначені для даних, що мають дискретну природу.

Найбільш поширеними на сьогодні є ієрархічні агломеративні методи, які відрізняються між собою мірою розбіжності кластерів. У залежності від неї можуть використовуватися методи одиночного зв`язку, повного зв`язку та середнього зв`язку. У методі одиночного зв`язку розбіжність між кластерами оцінюється як відстань між найближчими його представниками, в методі повного зв`язку – як відстань між найвіддаленішими й в методі середнього зв`язку – середня відстань між усіма парами об`єктів із різних кластерів. Перевагою методу одиночного зв`язку є стійкість його результатів до вихідних даних, недоліком – схильність до утворювання одного довгого кластера, що ускладнює визначення кількості кластерів. Результати методу повного зв`язку більш наочно відображають кластерну структуру, але не є інваріантними до перетворень матриці розбіжностей. Метод середнього зв`язку сполучає позитивні сторони обох методів.

Приклад виконання роботи

$$\delta = \begin{pmatrix} 0 & 5 & 9 & 5 & 7 \\ 5 & 0 & 7 & 9 & 4 \\ 9 & 7 & 0 & 13 & 5 \\ 5 & 9 & 13 & 0 & 11 \\ 7 & 4 & 5 & 11 & 0 \end{pmatrix}$$

Мінімальною є відстань між другим та п`ятим об`єктами, тому об`єднуємо їх в один кластер та розраховуємо відстані до нового кластера, як середню відстань, наприклад відстань між кластерами (2,5) і 1 є середнє між $\delta_{2,1}$ і $\delta_{5,1}$. Інші відстані переносимо без змін. В отриманій матриці знову знаходимо мінімальний елемент – це відстань між об`єктами 1 і 4, тому об`єднуємо їх та перераховуємо відстані. На останньому кроці об`єднуємо кластер (2,5) з об`єктом 3 та обчислюємо середню відстань між

елементами кластерів (2,3,5) і (1,4). За результатами побудуємо дендограму, з якої можна зробити висновок, що оптимальним буде розбиття на два кластери.

	Кластери			
	(2,5)	1	3	4
(2,5)	0	6	6	9
1	6	0	9	5
3	6	9	0	13
4	9	15	13	0

	Кластери		
	(2,5)	(1,4)	3
(2,5)	0	7.5	6
(1,4)	6	0	11
3	6	9	0

	Кластери	
	(2,5,3)	(1,4)
(2,5,3)	0	8.25
(1,4)	8.25	0

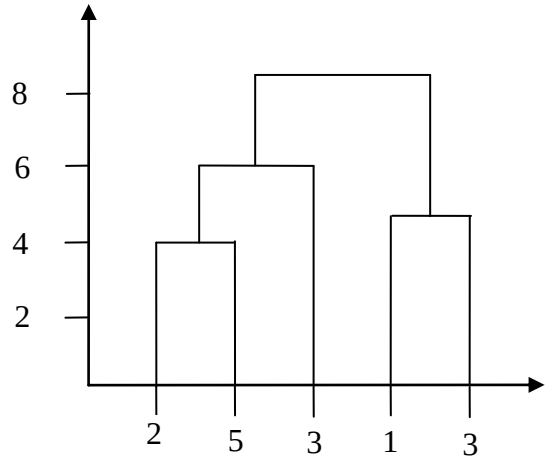

Рисунок 6.1 – Дендограма розподілу по кластерах

Завдання

1. Виконати ієрархічну агломеративну кластеризацію п’ятнадцяти об’єктів, використовуючи в якості метрики розходження кластерів відстань середнього зв’язку.

2. Повторити кластеризацію, використовуючи метрики одиночного та повного зв’язку й порівняти результати.

Перелік рекомендованих літературних джерел

1. Айвазян С.А. Прикладная статистика. Основы эконометрики: учебник для вузов: в 2 т. 2-е изд., испр. /С. А. Айвазян, В. С. Мхитарян Теория вероятностей и прикладная статистика. Т. 1-М.: ЮНИТИ-ДАНА, 2001.– 656 с.

2. Дронов С. В. Многомерный статистический анализ: учебное пособие. /С. В. Дронов.– Барнаул: изд-во Алт. Гос. Ун-та, 2003.– 213 с.

3. Дубров А. М. Многомерные статистические методы: учебник/А. М.Дубров, В. С. Мхитарян, Л. И. Трошин.М.: Финансы и статистика. 2000.–352 с.

Додаткова:

1. Дейвисон М. Многомерное шкалирование: Методы наглядного представления данных;/ М. Дейвисон; пер. с англ. В. С. Каменского.–М.: Финансы и статистика, 1988.–254 с.

2. Джонстон Дж. Эконометрические методы/ Дж. Джонстон; пер. с англ. А.А. Равкина.–М.: Статистика, 1980.–444 с.

3. Андерсон Т. Введение в многомерный статистический анализ/Т. Андерсон.– М.: Физматгиз, 1963. 500 с.

4. Факторный, дискриминантный и кластерный анализ: пер. с англ./Дж.-О. Ким, Ч. У. Мюллер, У. Р. Клекка и др.; под ред. И. С. Енюкова. – Финансы и статистика, 1989.–215 с.

ДОДАТОК А
Варіанти завдань

Номера варіантів співпадають із номером по журналу. Вихідні дані до першої, другої та третьої робіт отримують із таблиці 1. Номера признаків обирають згідно з таблицею 2. Вихідні дані на п`яту роботу співпадають з даними на першу роботу. При цьому перші 10 спостережень відносять до навчальної вибірки, а останні – до екзаменуючої. До першого класу відносять ті спостереження, всі признаки в якому від`ємні.

Таблиця 1 – Вихідні дані

	Номера признаків									
	1	2	3	4	5	6	7	8	9	10
1	-3.17	-33.85	-217.04	-54.14	77.43	-18.95	44.66	-50.3	-110.15	102.35
2	-21.81	-30.9	-143.17	-11.91	85.32	-14.56	24.06	-51.62	-74.61	119.91
3	6.73	1.73	9.21	-14.99	-21.66	2.15	-5.19	48.69	24.92	-38.86
4	-15.32	-13.97	-32.46	0.94	30.78	-3.76	-2.24	17.04	0.71	38.58
5	-33.06	-27.19	-60.04	9.95	66.99	-7.63	-4.06	16.23	-4.42	88.48
6	-12.85	-1.57	19.6	24.71	18.76	0.06	-5.47	-24.86	0.35	32.88
7	-11.42	5.32	69.15	32.03	-3.25	4.75	-17.75	5.36	34.33	0
8	17.34	9.42	7	-16.03	-30.18	2.2	4.93	4.65	-1.1	-43.58
9	23.59	14.41	-20.92	-7.65	-14.45	-1.18	25.28	-96.85	-63.74	-5.52
10	12.35	-2.19	-69.42	-23.17	9.18	-5.38	23.55	-48.06	-55.49	15.95
11	36.15	7.6	-64.82	-55.2	-37.03	-2.18	26.62	-5.19	-41.3	-57.53
12	-1.71	-4.07	-49.32	3.61	30.23	-5.44	17.92	-76.07	-56.04	52.7
13	11.84	4.97	-20.98	-9.13	-5.62	-1.32	13.52	-40.11	-32.06	-2.76
14	28.71	20.09	50.33	-20.93	-66.33	7.3	-2.55	36.64	27.73	-97.55
15	3.03	-1.73	10.64	-17.65	-21.28	2.41	-10.6	78.04	41.05	-43.33

	11	12	13	14	15	16	17	18	19	20
1	98.32	83.23	1.02	87.12	-8.46	-60.93	-27.54	19.34	102.76	122.88
2	62.8	79.06	35.68	31.57	-12.51	-57.54	-46.79	11.11	41.38	98.52
3	38.93	-30.16	-39.04	14.31	13.81	10.99	5.95	15.54	40.67	19.41
4	53.53	12.99	0.54	2.43	4.12	-19.7	-30.24	16.7	29.95	60.53
5	87.1	34.8	16.72	-4.23	2.96	-41.29	-61.49	26.23	39.32	107.94
6	-28.22	16.45	34.8	-29.09	-8.57	-7.16	-14.56	-9.54	-42.22	-12.37
7	-31.61	-11.73	21.08	-43.85	-1.4	8.54	-9.38	-6.83	-48.52	-27.42
8	-15.12	-17.79	-21.78	16.8	2.72	16.28	27.46	-4.3	8.04	-29.1
9	-108.12	31.25	34.93	13.32	-23.94	7.79	45.62	-40.88	-60.59	-95.58
10	-13.86	30.67	6.65	35.12	-10.27	-10.6	15.87	-10.1	10.9	-6.07
11	7.12	-9.18	-48.75	68.98	3.65	12.83	48.58	-0.85	58.08	-14.91
12	-44.67	49.38	45.03	3.76	-19.83	-18.82	-0.13	-20.62	-35.08	-19.26
13	-38.17	14.56	10.03	13.59	-9.4	1.79	20.5	-15.35	-15.58	-34.03
14	-21.13	-53.03	-49.09	16.26	11.46	38.66	47.13	-2.61	12.56	-54.17
15	72.01	-41.07	-54.4	16.37	21.51	10.21	-3.12	27.81	64.14	46.4

Вихідні дані на шосту роботу – елементи матриці відстаней обчислюють, як евклідові відстані між спостереженнями, що увійшли до вибірки в першому завданні.

Таблиця 2 – Розподіл признаків по варіантах

Номер варіанта	Номера признаків
1	2,8,11,15,17,18,20
2	3,6,9,13,15,18,19
3	1,2,5,9,10,15,17
4	1,3,5,6,9,12,18
5	4,6,8,9,13,15,19
6	3,6,8,10,16,18,20
7	5,7,11,15,18,19,20
8	2,5,7,9,11,16,20
9	3,4,8,10,15,17,19
10	2,6,10,13,14,16,18

Вихідні дані для четвертої роботи:

Варіант 1 Варіант 2 Варіант 3

$$\delta = \begin{pmatrix} 0 & 4 & 7 & 5 & 6 & 6 & 9 \\ 4 & 0 & 6 & 6 & 3 & 4 & 7 \\ 7 & 6 & 0 & 7 & 7 & 4 & 7 \\ 5 & 6 & 7 & 0 & 5 & 8 & 5 \\ 6 & 3 & 7 & 5 & 0 & 6 & 6 \\ 6 & 4 & 4 & 8 & 6 & 0 & 8 \\ 9 & 7 & 7 & 5 & 6 & 8 & 0 \end{pmatrix}, \quad \delta = \begin{pmatrix} 0 & 2 & 3 & 6 & 5 & 6 & 3 \\ 2 & 0 & 7 & 3 & 6 & 8 & 6 \\ 3 & 7 & 0 & 7 & 9 & 9 & 2 \\ 6 & 3 & 7 & 0 & 9 & 7 & 6 \\ 5 & 6 & 9 & 9 & 0 & 8 & 4 \\ 6 & 8 & 9 & 7 & 8 & 0 & 8 \\ 3 & 6 & 2 & 6 & 4 & 8 & 0 \end{pmatrix}, \quad \delta = \begin{pmatrix} 0 & 5 & 7 & 6 & 4 & 5 & 2 \\ 5 & 0 & 3 & 7 & 7 & 2 & 5 \\ 7 & 3 & 0 & 5 & 8 & 4 & 7 \\ 6 & 7 & 5 & 0 & 5 & 5 & 8 \\ 4 & 7 & 8 & 5 & 0 & 6 & 5 \\ 5 & 2 & 4 & 5 & 6 & 0 & 6 \\ 2 & 5 & 7 & 8 & 5 & 6 & 0 \end{pmatrix}.$$

Варіант 4

$$\delta = \begin{pmatrix} 0 & 6 & 6 & 5 & 3 & 6 & 3 \\ 6 & 0 & 4 & 6 & 3 & 6 & 6 \\ 6 & 4 & 0 & 8 & 4 & 4 & 5 \\ 5 & 6 & 8 & 0 & 4 & 9 & 3 \\ 3 & 3 & 4 & 4 & 0 & 5 & 7 \\ 6 & 6 & 4 & 9 & 5 & 0 & 9 \\ 3 & 6 & 5 & 3 & 7 & 9 & 0 \end{pmatrix}.$$

Варіант 5

$$\delta = \begin{pmatrix} 0 & 4 & 7 & 5 & 4 & 7 & 7 \\ 4 & 0 & 5 & 4 & 2 & 6 & 6 \\ 7 & 5 & 0 & 8 & 4 & 4 & 5 \\ 5 & 4 & 8 & 0 & 7 & 4 & 9 \\ 4 & 2 & 4 & 7 & 0 & 7 & 3 \\ 7 & 6 & 4 & 4 & 7 & 0 & 6 \\ 7 & 6 & 5 & 9 & 3 & 6 & 0 \end{pmatrix}.$$

Варіант 6

$$\delta = \begin{pmatrix} 0 & 5 & 8 & 4 & 4 & 6 & 5 \\ 5 & 0 & 5 & 7 & 9 & 5 & 6 \\ 8 & 5 & 0 & 7 & 4 & 9 & 8 \\ 4 & 7 & 7 & 0 & 6 & 8 & 5 \\ 4 & 9 & 4 & 6 & 0 & 8 & 6 \\ 6 & 5 & 9 & 8 & 8 & 0 & 4 \\ 5 & 6 & 8 & 5 & 6 & 4 & 0 \end{pmatrix}.$$

Варіант 7

$$\delta = \begin{pmatrix} 0 & 6 & 6 & 6 & 2 & 5 & 2 \\ 6 & 0 & 7 & 6 & 5 & 5 & 6 \\ 6 & 7 & 0 & 9 & 8 & 8 & 3 \\ 6 & 6 & 9 & 0 & 3 & 5 & 6 \\ 2 & 5 & 8 & 3 & 0 & 2 & 3 \\ 5 & 5 & 8 & 5 & 2 & 0 & 7 \\ 2 & 6 & 3 & 6 & 3 & 7 & 0 \end{pmatrix}.$$

Варіант 8

$$\delta = \begin{pmatrix} 0 & 7 & 7 & 5 & 5 & 6 & 1 \\ 7 & 0 & 8 & 8 & 7 & 6 & 8 \\ 7 & 8 & 0 & 5 & 6 & 6 & 3 \\ 5 & 8 & 5 & 0 & 6 & 3 & 4 \\ 5 & 7 & 6 & 6 & 0 & 4 & 8 \\ 6 & 6 & 6 & 3 & 4 & 0 & 3 \\ 1 & 8 & 3 & 4 & 8 & 3 & 0 \end{pmatrix}.$$

Варіант 9

$$\delta = \begin{pmatrix} 0 & 4 & 7 & 5 & 3 & 7 & 5 \\ 4 & 0 & 4 & 6 & 7 & 6 & 6 \\ 7 & 4 & 0 & 7 & 7 & 7 & 3 \\ 5 & 6 & 7 & 0 & 4 & 8 & 9 \\ 3 & 7 & 7 & 4 & 0 & 2 & 8 \\ 7 & 6 & 7 & 8 & 2 & 0 & 7 \\ 5 & 6 & 3 & 9 & 8 & 7 & 0 \end{pmatrix}.$$

Варіант 10

$$\delta = \begin{pmatrix} 0 & 6 & 5 & 7 & 4 & 7 & 6 \\ 6 & 0 & 7 & 5 & 2 & 3 & 7 \\ 5 & 7 & 0 & 3 & 2 & 2 & 7 \\ 7 & 5 & 3 & 0 & 7 & 6 & 7 \\ 4 & 2 & 2 & 7 & 0 & 9 & 8 \\ 7 & 3 & 2 & 6 & 9 & 0 & 2 \\ 6 & 7 & 7 & 7 & 8 & 2 & 0 \end{pmatrix}.$$

ЕЛЕКТРОННЕ НАВЧАЛЬНО-МЕТОДИЧНЕ ВИДАННЯ

Хребет Валерій Григорович
Фесенко Дмитро Володимирович

МЕТОДИЧНІ ВКАЗІВКИ ДО ВИКОНАННЯ ПРАКТИЧНИХ РОБІТ З ДИСЦИПЛІНИ "ЕКОНОМЕТРИКА" (ДЛЯ СТУДЕНТІВ НАПРЯМУ 6.030502 "ЕКОНОМІЧНА КІБЕРНЕТИКА" ТА 6.030601 "МЕНЕДЖМЕНТ ОРГАНІЗАЦІЙ")

Підписано до випуску 2012 р. Гарнітура Times New.
Умов. друк. арк. 3,0. Зам. № 94
Державний вищий навчальний заклад
" Донецький національний технічний університет "
Автомобільно-Дорожній інститут
84646, м. Горлівка, вул. Кірова, 51
E-mail:
Редакційно – видавничий відділ
Свідоцтво про внесення до Державного реєстру видавців, виготовників і розповсюджувачів видавничої продукції ДК № 2982 від 21.09.2007 р.