ПАНЕЛЬ 3.
ВІЗУАЛІЗАЦІЯ ДАНИХ У СТАТИСТИЦІ

ПРИНЦИПИ ВИБОРУ ГОЛОВНИХ КОМПОНЕНТ: ОСОБЛИВОСТІ ПРИКЛАДНОГО МОДЕЛЮВАННЯ

Голубова Галина Володимирівна,
кандидат економічних наук, доцент,
доцент кафедри статистики,
Національна академія статистики, обліку та аудиту

Багатомірність соціально-економічних явищ, сьогоднішні наростаючі інформаційні потоки, в тому числі «Big data», – все це породжує значні за розмірами інформаційні бази. З метою дослідження внутрішньої структури об'єкта слід «стиснути» розмірність початкової ознакової множини, замінивши її мінімальною кількістю компонент [1]. При цьому головні компоненти зберігають всю інформацію про об'єкт дослідження. На основі головних компонент можна ранжувати та класифікувати об'єкти (країни, регіони тощо), вимірювати взаємозв'язки між первинними показниками та головними компонентами, використовувати їх для побудови регресійного аналізу та ін.

Метод головних компонент (Principal Component Analysis (PCA)) – це потужна дослідницька модель, основною задачею якої є зменшення розмірності багатовимірного простору з мінімальними втратами інформації.

Не слід плутати PCA з факторним аналізом. Факторний аналіз популярний у суспільних науках, його основою є виявлення інтерпретованих лінійних зв'язків між змінними, які називаються факторами. Натомість PCA – це корисний метод для зменшення кількості спостережуваних змінних до меншого набору незалежних компонентів. Отже, основними цілями PCA є:

1. Візуалізація даних для дослідницького аналізу, що дає змогу розкрити латентні характеристики даних та інтерпретувати компоненти.

2. Зменшення кількості предикторів для майбутнього аналізу, такого як регресія основних компонент.

У PCA використовується складна математика (лінійна алгебра) для визначення базової лінійної структури, властивої матриці даних. Основою математики в PCA є декомпозиція сингулярних значень, яка є узагальненням розкладу власного числа. Розуміння того, як працюють ці математичні комбінації, не є необхідним для освоєння PCA, однак розуміння основних

принципів щодо методу вибору головних компонент є надзвичайно необхідним при інтерпретації результатів PCA.

В ході аналізу розкрито основні класичні методи вибору головних компонент (ГК). Власне значення головної компоненти – це величина відхилення у вихідних даних, і максимізація відхилення є важливою, оскільки вона надає найбільшу інформацію про вихідні дані. Отже, одним із найпростіших методів вибору підмножини ГК є лише вибір першої *k*-кількості компонентів з найбільшими власними значеннями.

Інший класичний метод вибору ГК передбачає вивчення відсотка загальної дисперсії, що пояснюється кожною компонентою. Встановивши заздалегідь визначений поріг (як правило, 75% або 80% від загальної поясненої дисперсії), перші *k*-головних компонент, які сукупно пояснюють принаймні цю велику частину дисперсії, можна обрати як підмножину компонент. Однак, як й інші класичні методи, цей метод відбору не може повністю враховувати дисперсію даних.

Посилюючи складність відбору ГК, часто використовують метод, що передбачає збереження всіх ГК з власними значеннями більше 1. Його ще називають «правилом Кайзера», «критерієм Кайзера» або «правилом Кайзера – Гутмана». Основна ідея полягає в тому, що при стандартизованих даних дисперсія кожної з вихідних змінних дорівнює 1. Отже, головні компоненти з власним значенням більше 1 пояснюють більше дисперсії, ніж одна змінна у вихідних даних. Цей метод є логічним, але він не враховує той факт, що навіть з випадковими даними (шумами) PCA визначатиме компоненти із власними значеннями більше 1. У цих ситуаціях дисперсія, що пояснюється компонентами, насправді не є корисною, оскільки це просто дисперсія через випадкову помилку або шум.

Щоб подолати цю проблему, при паралельному аналізі (Parallel Analysis (PA)) використовується багаторазове моделювання даних [2]. PA (іноді його називають «Паралельний аналіз Горна», названий на честь його творця Джона Горна) – це метод вибору головних компонент, який враховує дисперсію даних через випадкову помилку або шум [3]. Процес виконання паралельного аналізу ґрунтується на методі Монте Карло, тобто це симуляція великої кількості наборів даних (наприклад 1000 або більше), при цьому кожен імітований набір даних містить таку ж саму кількість змінних та спостережень, що і вихідні дані. Для кожної імітованої змінної дані генеруються шляхом вибірки з багатовимірного нормального розподілу, при цьому стандартне відхилення дорівнює стандартному відхиленню відповідної змінної фактичних даних. Для кожної компоненти обчислюється 95-відсотковий інтервал. Отримане власне значення ГК з вихідних даних слід порівняти з верхнім 95-м процентилем, розрахованим з модельованих наборів даних. Якщо власне значення з вихідних даних більше верхнього процентиля від модельованих даних, компонента відбирається, в іншому випадку – відкидається.

Ідея полягає в тому, що через випадкову помилку (мінливість вибірки) в даних PCA генерує деякі компоненти із власними значеннями, більшими за 1. Загалом, перші власні значення, що генеруються даними «шуму», збільшува-

тимуться зі збільшенням кількості змінних, і зменшуватимуться зі збільшенням кількості спостережень. Зберігаючи лише ті ГК з власними значеннями, що перевищують 95-й процентиль змодельованих власних значень, ви гарантуєте, що розбіжності, пояснені цими ГК, ймовірно представляють реальну дисперсію, а не дисперсію через шум. Отже, паралельний аналіз вважається більш корисним на практиці, ніж метод вибору головних компонент за правилом Кайзера або іншими методами відбору.

Розглянемо це на прикладі користувачів фейсбуку [4]. Маємо інформацію про 500 користувачів за 14 показниками: кількість щоденних публікацій, кількість постів за годину, кількість постів про особисте життя, вільний час, кількість користувачів, що підписалися на вашу сторінку, кількість людей, що вподобали вашу сторінку, користувачі, що вподобали вашу світлину, кількість коментарів, лайків, поширень та ін. Серед усіх цих показників, які можна кількісно виміряти, насправді важко виокремити ті головні компоненти, які дійсно візуалізують чи типологізують вашу активність у фейсбуці чи у якийсь спосіб здатні описати принципи взаємодії з цим соціальним медіа. Тому автором проведено аналіз PCA на основі цих показників за методом Кайзера та методом PA. Отримані власні значення наведено в табл. 1.

Таблиця 1

Власні значення головних компонент

Головні компоненти	Власні значення (вихідні дані)	Власні значення (Parallel Analysis)		
		Середнє	Верхня межа	Нижня межа
PC1	5,920	1,289	1,356	1,235
PC2	1,740	1,222	1,270	1,180
PC3	1,658	1,171	1,212	1,136
PC4	1,124	1,126	1,163	1,094
PC5	1,002	1,086	1,117	1,057
PC6	0,835	1,047	1,076	1,016
PC7	0,621	1,009	1,038	0,981
PC8	0,454	0,975	1,002	0,944
PC9	0,304	0,939	0,968	0,913
PC10	0,143	0,905	0,932	0,876
PC11	0,116	0,869	0,899	0,836
PC12	0,069	0,831	0,862	0,798
PC13	0,014	0,790	0,823	0,754
PC14	0,001	0,739	0,779	0,694

Як бачимо, за правилом Кайзера відібрано п’ять головних компонент, значення яких більше 1 (рис. 1), які пояснюють 81,2% варіації. За методом PAs виділено лише три головні компоненти, про що свідчить рис. 2, який візуалізує відсікання трьох компонент. Власне значення PC4 дорівнює 1,124, що менше за верхню межу 95-відсоткового інтервалу (1,163), це дає підстави для вилучення

цієї компоненти з подальшого аналізу, оскільки її дисперсія спричинена шумом вибірки, а не реальним процесом.

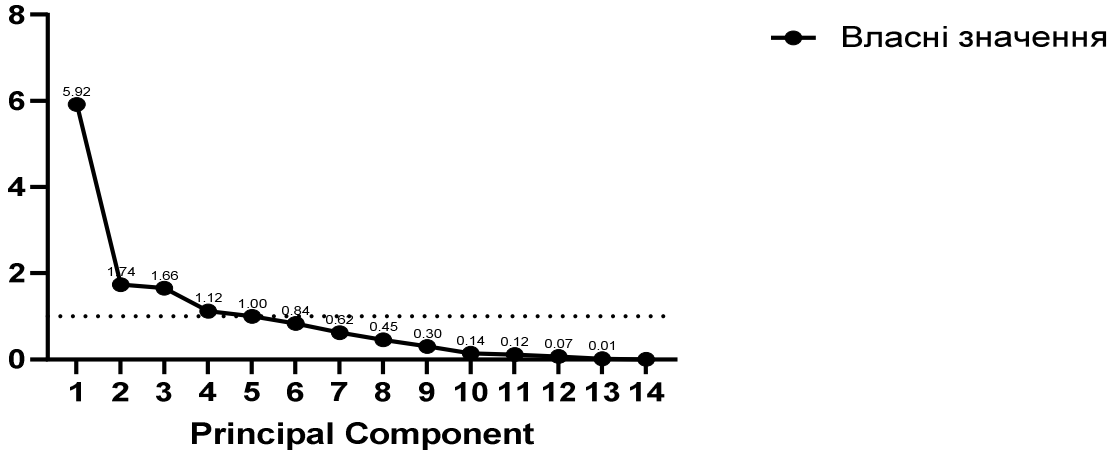

Рис. 1. Головні компоненти за критерієм Кайзера

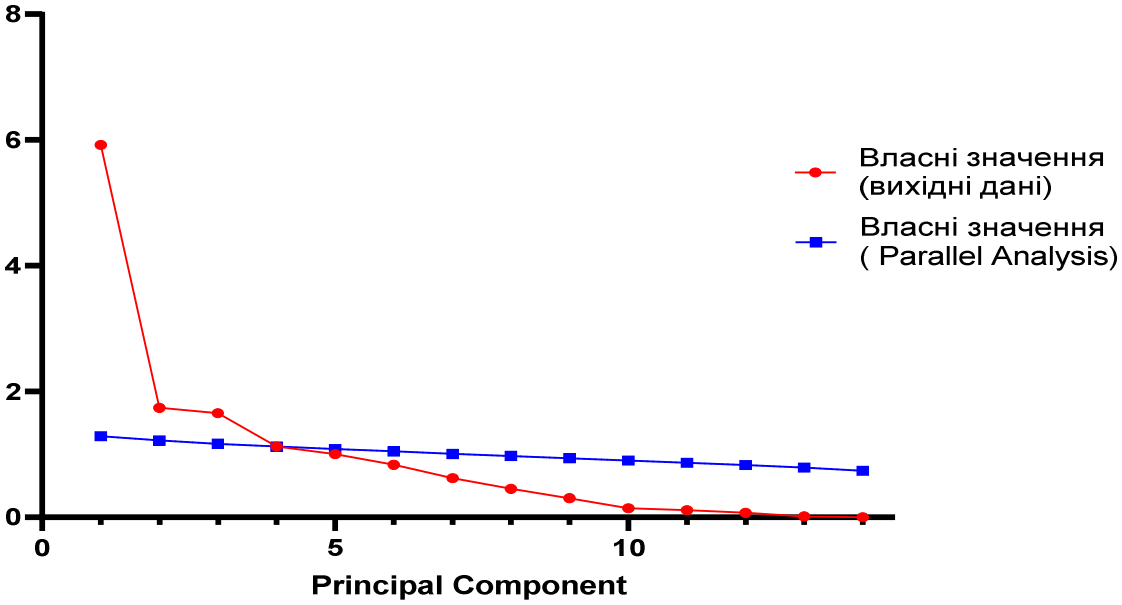

Рис. 2. Головні компоненти за методом PA

Зазвичай при виконанні PCA слід дотримуватися певних статистичних умов. Усі змінні мають бути кількісними (категорійні змінні з аналізу вилучаються), однорідними, розподіл – симетричним, при цьому кількість спостережень має переважати кількість змінних. Однак в залежності від роду дослідження можуть бути винятки, наприклад в медицині, хімії, біостатистиці та інших науках.

Розглянемо приклад генофонду людини [4] (100 ген 20 пацієнтів, тобто матриця 20 на 100). Оскільки, змінних (m) більше, ніж спостережень (n), то максимальна кількість компонент, яку можна виділити, дорівнює n-1, в нашому прикладі 19. За правилом Кайзера встановлено, що головних компонент 19,

причому 75% варіації пояснюють 11 компонент, а за методом РА виділено лише одну головну компоненту (табл. 2., рис. 3–4).

Таблиця 2

Власні значення головних компонент

Головні компоненти	Власні значення (вихідні дані)	Власні значення (Parallel Analysis)		
		Середнє	Верхня межа	Нижня межа
PC1	17,291	9,749	10,606	9,013
PC2	8,457	8,808	9,449	8,282
PC3	7,587	8,131	8,650	7,678
PC4	6,924	7,575	8,011	7,156
PC5	6,425	7,058	7,477	6,647
PC6	6,157	6,589	6,986	6,229
PC7	5,983	6,151	6,489	5,807
PC8	4,816	5,730	6,050	5,400
PC9	4,524	5,344	5,653	5,019
PC10	4,472	4,969	5,287	4,666
PC11	4,140	4,621	4,924	4,334
PC12	3,983	4,278	4,567	3,984
PC13	3,723	3,954	4,239	3,661
PC14	3,282	3,640	3,920	3,360
PC15	3,117	3,318	3,600	3,037
PC16	2,745	2,989	3,284	2,712
PC17	2,452	2,686	2,964	2,389
PC18	2,201	2,365	2,655	2,046
PC19	1,720	1,990	2,324	1,651

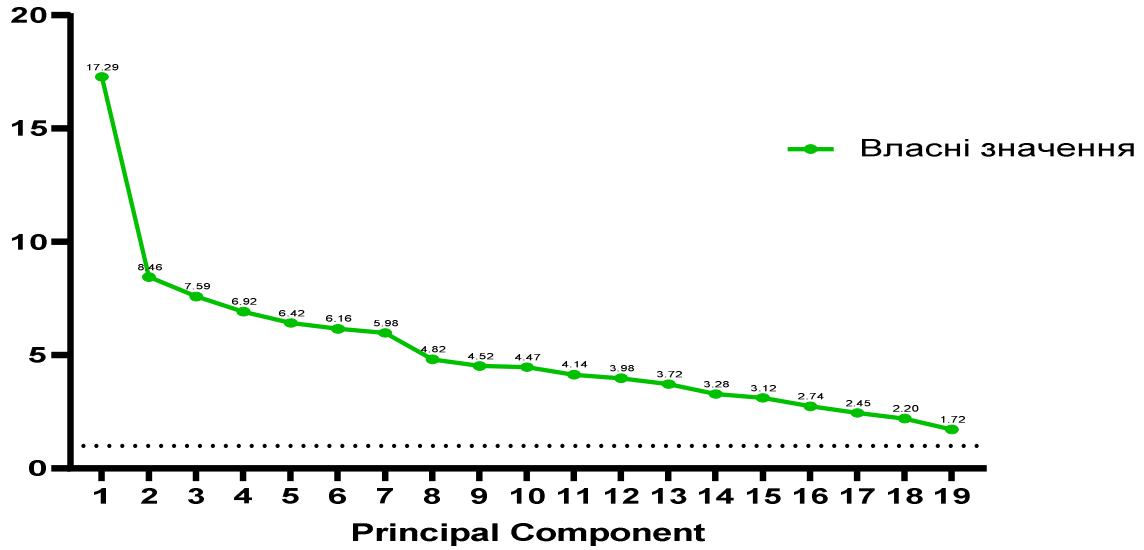

Рис. 3. Головні компоненти за критерієм Кайзера

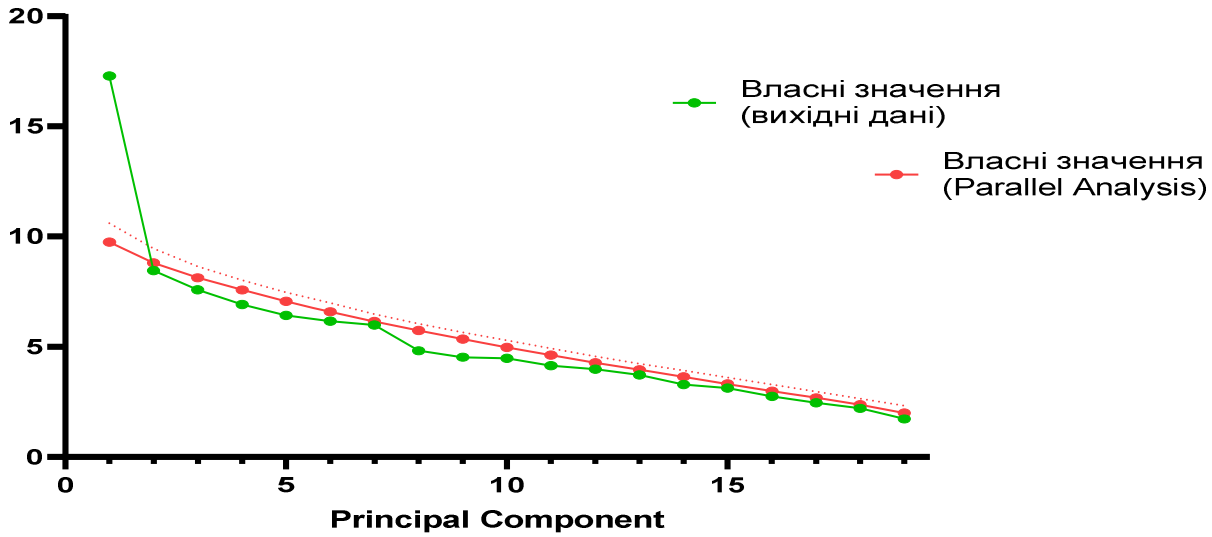

Рис. 4. Головні компоненти за методом PA

В наведених прикладах ми не інтерпретуємо виділені компоненти, оскільки це не було метою нашого дослідження і ми не є фахівцями в цій галузі. Однак радимо ретельно підходити до методу вибору головних компонент, які пояснюватимуть реальну дисперсію досліджуваного явища.

Основною перевагою Parallel Analysis вважаємо моделювання процесу вибору кількості ГК шляхом визначення точки, в якій головні компоненти неможливо відрізнити від тих, що генеруються імітованим шумом.

Список використаних джерел

1. Єріна А. М., Єрін Д. Л. Статистичне моделювання та прогнозування: підруч. К.: КНЕУ, 2014. 348 с.

2. Çokluk Ö., & Koçak D. (2016). Using Horn's parallel analysis method in exploratory factor analysis for determining the number of factors. *Educational Sciences: Theory & Practice. No 16. P. 537–551.*

3. Hayton J. C. Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis. URL: https://www.researchgate.net/publication/235726204_Factor_Retention_Decisions_in_Exploratory_Factor_Analysis_A_Tutorial_on_Parallel_Analysis/link/5582a85008ae6cf036c1a886/download.

4. Machine learning repository. URL: https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table.