9.1. КОНЦЕПЦІЯ МЕТОДУ ГОЛОВНИХ КОМПОНЕНТ

При моделюванні складних причинних комплексів часом стикаються з проблемою надлишковості інформації, коли екзогенні змінні x_i, включені в ознаковий простір моделі, висококорельовані (мультиколінеарні). Щоб забезпечити адекватність моделі реальному процесу, вдаються до заміни такого типу ознакової множини меншою кількістю некорельованих величин, які б зберігали всю інформацію щодо причинно-наслідкового механізму формування явища (процесу) і не впливали на точність результатів аналізу. Інструментом такої заміни є *метод головних компонент*.

Основне призначення методу головних компонент — виявити приховані (латентні) першопричини, які пояснюють кореляції між ознаками і змістовно інтерпретуються. Використання методу ґрунтується на припущенні, що ознаки x_i є лише індикаторами певних існуючих властивостей явища, які безпосередньо не вимірюються. Так, хвороба людини виявляється певними симптомами, рівень життя населення — умовами праці, побуту та дозвілля. Якщо таких першопричин декілька, в ознаковому просторі X виокремлюються групи висококорельованих ознак. Скажімо, сім ознак x_i поділяються на дві групи:

Група	Ознаки	Компонента
1	$x_1\ x_2\ x_3\ x_4$	G_1
2	$x_5\ x_6\ x_7$	G_2

Першопричина кореляції ознак j-ї групи називається *компонентою* G_j. Ознаки, що належать до різних груп, некорельовані, а отже, і компоненти G_j незалежні (ортогональні). Суть методу головних компонент полягає у переході від численної множини x_i до мінімальної кількості максимально інформативних компонент G_j.

$$\underset{i=1,\,2,\,\ldots,\,m}{x_i} \quad \Rightarrow \quad \underset{j=1,\,2,\,\ldots,\,p}{G_j} \quad .$$

Основні задачі методу головних компонент:

- виокремити та ідентифікувати компоненти G_j;
- визначити рівні G_j для окремих одиниць статистичної сукупності.

Ідентифікація компонент, тобто надання їм певного змісту, залежить від ознакової множини X. Як правило, її формують на основі теоретично обґрунтованої гіпотези щодо природи латентних властивостей явища. Якщо така гіпотеза відсутня, то використовують максимальну кількість ознак, покладаючись на можливості методу виявити такі властивості. Але в такому разі інтерпретація компонент ускладнюється.

Оскільки компоненти є гіпотетичними величинами, то виміряти їх можна лише опосередковано за допомогою спеціально сконструйованих моделей. У моделі головних компонент зв'язок між первинними ознаками і компонентами описується як лінійна комбінація

$$z_i = \sum_1^m a_{ij} G_j \ ,$$

де z_i — стандартизовані значення i-ї ознаки з одиничними дисперсіями; *сумарна дисперсія* дорівнює кількості ознак m;

a_{ij} — *факторне навантаження* j-ї компоненти на i-у ознаку.

Навантаження a_{ij} характеризує щільність зв'язку між i-ю ознакою та j-ю компонентою і як будь-яка міра щільності зв'язку змінюється в межах від 0 до ± 1.

У моделі головних компонент відсутні залишки, тобто апріорі передбачається, що всі m компонент повністю пояснюють сумарну дисперсію ознакової множини. За умови ортогональності компонент квадрат факторного навантаження a_{ij}^2 характеризує внесок j-ї компоненти у варіацію i-ї ознаки. Повний внесок j-ї компоненти у сумарну дисперсію m ознак становить $\lambda_j = \sum_1^m a_{ij}^2$.

У процесі компонентного аналізу сумарна варіація m первинних ознак x_i перерозподіляється між компонентами G_j з дисперсіями λ_j . Тобто сумарну дисперсію ознакової множини X можна представити як суму дисперсій компонент $\sum_1^m \lambda_j$ або через факторні навантаження.

$$m = \sum_1^m \lambda_j = \sum_1^m \sum_1^m a_{ij}^2 \ .$$

Схема декомпозиції сумарної дисперсії ознакової множини X наведено у вигляді матриці (табл. 9.1).

Таблиця 9.1

Z_i \ G_j	G_1	G_2	…	G_m	Дисперсія z_i
z_1	a_{11}^2	a_{12}^2	…	a_{1m}^2	1
z_2	a_{21}^2	a_{22}^2	…	a_{2m}^2	1
z_3	a_{31}^2	a_{32}^2	…	a_{3m}^2	1
…	…	…	…	…	…

z_m	a_{m1}^2	a_{m2}^2	…	a_{mm}^2	1
Дисперсія G_j	λ_1	λ_2	…	λ_m	m

Аналіз матриці по рядках показує, які компоненти і з якою вагою формують варіацію *i*-ї ознаки. Кожній ознаці властива своя *факторна структура*. Чим менше компонент навантажує ознаку, тим простішою вважається її факторна структура.

Аналіз матриці по стовпцях показує, які ознаки є індикаторами *j*-ї компоненти. Компоненти упорядковуються за значеннями дисперсій:

$$\lambda_1 > \lambda_2 > \lambda_3 > ... > \lambda_m .$$

Незважаючи на те, що замість *m* ознак визначається така ж кількість компонент, внесок більшості з них у сумарну варіацію виявляється незначним.

Левова частка сумарної варіації припадає на декілька перших компонент. Як показує досвід, кількість таких вагомих компонент становить 10—15% від кількості первинних ознак. Саме вони називаються *головними компонентами* і підлягають змістовній інтерпретації.

Таким чином, модель головних компонент трансформує *m*-вимірний ознаковий простір у *p*-вимірний простір компонент (*p* < *m*). Сумарна дисперсія головних компонент менша за сумарну дисперсію ознакового простору. Відношення $\frac{\sum_1^p \lambda_j}{m}$ характеризує *повноту факторизації*.

Математичною основою методу головних компонент слугує кореляційна матриця *R* з одиницями на головній діагоналі. Недіагональні елементи матриці представлені коефіцієнтами кореляції r_{ik}, які оцінюють не причинно-наслідкові, а супутні зв'язки між ознаками x_i та x_k, зумовлені наявністю спільної першопричини їх варіації.

У термінах матричної алгебри дисперсії компонент λ_j — це властиві числа кореляційної матриці *R*. Кожному з них відповідає властивий вектор *V*, який задовольняє рівняння $(R - \lambda E)V = 0$, де *E* — одинична матриця. Тобто виокремлення головних компонент є класичною задачею визначення властивих чисел λ та властивих векторів *V* кореляційної матриці *R*. Головними вважаються компоненти, для яких:

- за критерієм Кайзера $\lambda_j > 1$;
- повнота факторизації не менша, скажімо, 70%.

Наприклад, для кореляційної матриці

$$R = \begin{vmatrix} 1 & 0{,}8 & 0{,}2 \\ 0{,}8 & 1 & 0{,}6 \\ 0{,}2 & 0{,}6 & 1 \end{vmatrix}$$

властиві значення дорівнюють: $\lambda_1 = 2{,}1$; $\lambda_2 = 0{,}81$; $\lambda_3 = 0{,}09$. За критерієм Кайзера головною слід вважати першу компоненту ($\lambda_1 > 1$). Внесок цієї компоненти в сумарну варіацію трьох ознак становить 2,1 : 3 = 0,7, або 70%.

Розв'язування системи рівнянь

$$(1 - 2{,}1)V_1 + 0{,}8V_2 + 0{,}2V_3 = 0$$
$$0{,}8V_1 + (1 - 2{,}1)V_2 + 0{,}6V_3 = 0$$
$$0{,}2V_1 + 0{,}6V_2 + (1 - 2{,}1)V_3 = 0$$

дає властивий вектор $V = (1{,}213; 1{,}428; 1{,}0)$.

Щоб задовольнити умову $\lambda_j = \sum_1^m a_{ij}^2$, властивий вектор нормується

$$a_{ij} = V_{ij} \sqrt{\frac{\lambda_j}{\sum_1^m V_{ij}^2}} .$$

Отже, факторні навантаження *j*-ї компоненти є не що інше, як нормований властивий вектор матриці *R*. У розглянутому прикладі: $\sum_1^3 V_{i1}^2 = 4{,}81064$, множник $\sqrt{\frac{2{,}1}{4{,}81064}} = 0{,}668267$. Звідси факторні навантаження:

$$a_{11} = 0{,}878;\ a_{21} = 0{,}975;\ a_{31} = 0{,}683.$$

Сума квадратів факторних навантажень дорівнює значенню $\lambda_1 = 2{,}1$.

Процедури методу головних компонент — *Principal components* — представлено в модулі *Factor Analysis* — Факторний аналіз. Інформаційною базою компонентного аналізу можуть бути як первинні ряди (*Raw data*), так і кореляційна матриця (*Correlation matrix*). Тип інформаційної бази вказується на стартовій панелі модуля (*Input file*).

Визначимо факторні навантаження за даними кореляційної матриці (табл. 9.2). Цей файл створено в модулі *Data Management* — Управління даними. Окрім коефіцієнтів кореляції він містить значення середніх і стандартних відхилень кожної ознаки, обсяг сукупності, за даними якої обчислено кореляційну матрицю, та кількість матриць (у нашому прикладі — одна).

Таблиця 9.2

DATA: FMOD1.STA 5v * 9c					
Correlation matrix					
Variable	VAR1	VAR2	VAR3	VAR4	VAR5
VAR1	1	0,839	0,927	0,871	0,753
VAR2	0,839	1	0,967	0,778	0,828
VAR3	0,927	0,967	1	0,845	0,852
VAR4	0,871	0,778	0,845	1	0,837
VAR5	0,753	0,828	0,852	0,837	1
means	112,2	59,4	76,8	107	66,8
st.dev	28,6	17,1	21	28,9	19,3
N.	12				
matrix	1				

Щільність взаємозв'язків між ознаками дає підстави зробити висновок про наявність однієї першопричини формування їх варіації. Цей висновок підтверджують розраховані факторні навантаження, наведені в табл. 9.3. Мінімальне значення — a_{15} = 0,909.

Таблиця 9.3

Factor Loadings (Unrotated) (fmod1.sta)	
Extraction: Principal components (Marked loadings are > ,700000)	
Variable	Factor 1
VAR1	0,9367
VAR2	0,9418
VAR3	0,9798
VAR4	0,9223
VAR5	0,9090
Expl.Var	4,4016
Prp.Totl	0,8803

Властиве значення кореляційної матриці *Expl.Var* становить 4,4, а ступінь факторизації *Prp.Totl* = 4.4 : 5 = 0,88.

9.2 ІДЕНТИФІКАЦІЯ ТА ВИМІРЮВАННЯ ГОЛОВНИХ КОМПОНЕНТ

У реальних багатовимірних сукупностях часто виокремлюється не одна, а декілька головних компонент, навантаження яких на окремі ознаки перетинаються. Складна факторна структура значно ускладнює ідентифікацію компонент. Пошук *простої факторної структури*, коли a_{ij} наближається до 1 або 0, здійснюється за допомогою різних процедур ортогонального чи косокутного *обертання*, в процесі якого значення одних факторних навантажень зростають, інших — зменшуються. Найчастіше використовують процедуру варімакс (*Varimax*), яка максимізує варіацію квадратів факторних навантажень для кожної компоненти, збільшуючи великі і зменшуючи малі значення a_{ij}.

В алгебраїчних термінах обертання означає перетворення матриці факторних навантажень A в матрицю простої факторної структури B. Необхідно знайти таку матрицю трансформації T, яка б забезпечила рівність $B = AT$. Матриця трансформації T залежить від кількості головних компонент і кута обертання Θ, який не повинен перевищувати 45°. Для двох компонент при обертанні за годинниковою стрілкою

$$T = \begin{vmatrix} \cos\Theta & -\sin\Theta \\ \sin\Theta & \cos\Theta \end{vmatrix}.$$

Очевидно, що проста факторна структура недосяжна, але наближення до неї все ж спрощує ідентифікацію компонент. Наприклад, трансформуємо матрицю A з кутом обертання Θ = 30° (sin 30° = 0,500; cos 30° = 0,866):

$$B = AT = \begin{vmatrix} 0{,}60 & 0{,}40 \\ 0{,}40 & 0{,}50 \\ -0{,}30 & 0{,}60 \\ -0{,}20 & 0{,}80 \\ -0{,}10 & 0{,}70 \end{vmatrix} \cdot \begin{vmatrix} 0{,}866 & -0{,}500 \\ 0{,}500 & 0{,}866 \end{vmatrix} = \begin{vmatrix} 0{,}72 & 0{,}05 \\ 0{,}60 & 0{,}23 \\ 0{,}04 & 0{,}67 \\ 0{,}24 & 0{,}79 \\ 0{,}26 & 0{,}91 \end{vmatrix}.$$

На основі факторних навантажень матриці B можна зробити висновок, що перша компонента навантажує ознаки x_1 та x_2, друга — решту ознак. Зміст кожної компоненти визначається змістом ознак, які її представляють.

Отже, побудова моделі головних компонент здійснюється в три етапи:

- розрахунок кореляційної матриці *R*;
- виокремлення головних компонент і розрахунок факторних навантажень;
- ідентифікація головних компонент.

Розглянемо аналітичні можливості моделі за даними файла *faktor sta*. (піддиректорія *Examples*), в якому наведено результати соціологічного опитування 100 респондентів щодо ступеня задоволеності їх життям. Ініціювавши кнопку *Variables*, сформуємо ознакову множину моделі, надавши кожній ознаці соціально-економічний зміст:

1 — самооцінка професійного статусу респондента;
2 — оцінка умов праці;
3 — оцінка рейтингу компанії;
4 — оцінка можливостей самореалізації поза роботою;
5 — ефективність відпочинку;
6 — оцінка матеріального добробуту сім'ї;
7 — задоволеність соціальним статусом сім'ї;
8 — оцінка навколишнього середовища.

Після команди *OK* з'являється вікно *Define Method of Factor Extraction* — Визначити метод виокремлення факторів. У функціональній його частині з-поміж запропонованих методів вибираємо *Principal components* — Головні компоненти. Праворуч розміщено поля для установки параметрів моделі: *Maximum no. of factors* — максимальне число факторів і *Minimum eigenvalue* — мінімальне властиве число. За умовчування ці параметри становлять відповідно 2 і 1.

За командою на виконання програми з'являється вікно *Factor Analysis Results* — Результати факторного аналізу, в інформаційній частині якого вказується кількість ознак, метод аналізу, десятковий логарифм детермінанта кореляційної матриці, число виокремлених факторів і властиві значення матриці λ_j. Для детальнішого аналізу результатів скористаємося опціями функціональної частини вікна. Скажімо, для візуальної оцінки виокремлення головних компонент можна скористатися графічним критерієм «кам'янистий обвал» — *Scree plot* (рис. 9.1). Значення властивих чисел кореляційної матриці представлено на осі ординат. Як бачимо, ці значення стрімко зменшуються і лише два перших більші за одиницю.

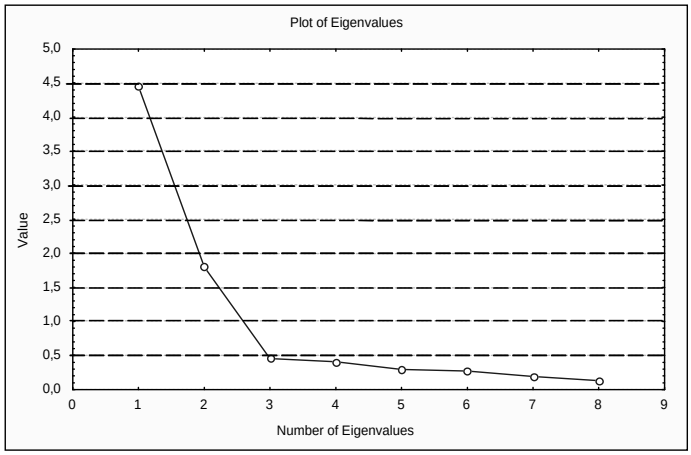

Рис. 9.1. Властиві числа кореляційної матриці

За установкою *Eigenvalues* система видає таблицю значень властивих чисел, які є дисперсіями головних компонент, а також внесок кожної з них у сумарну варіацію ознакової множини — % *total Variance* (табл. 9.4). Внесок першої компоненти в сумарну дисперсію ознакової множини становить 55,6%, другої — 22,5%. Разом (*Cumul.%*) дві компоненти пояснюють 78,1% сумарної варіації, що свідчить про високий ступінь факторизації.

Таблиця 9.4

Eigenvalues (factor.sta)				
Continue…	Extraction: Principal components			
Value	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	4,56	55,6	4,56	55,6
2	1,80	22,5	6,36	78,1

З-поміж процедур обертання факторів — *Factor rotation* вибираємо *Varimax normalized* — Варімакс нормалізований. За опцією *Factor loadings* маємо таблицю факторних навантажень, значення яких наближаються до 1 або до 0 (табл. 9.5). Ознаки, які навантажує кожна компонента, виділено.

Перша компонента зв'язана з ознаками 1—5, її можна ідентифікувати як ступінь задоволеності роботою і дозвіллям; друга компонента навантажує ознаки 6—8, які характеризують матеріальний добробут і соціальний статус сімей респондентів. Наведені в останніх рядках таблиці властиві значення і внесок окремих компонент у сумарну дисперсію визначені за трансформованими факторними навантаженнями, а тому відрізняються від первинних, проте сумарний їх внесок процедура обертання не змінює: *Prp.Totl.* = 0,418 + 0,363 = 0,781.

Поглиблений факторний аналіз складних соціально-економічних явищ передбачає вимірювання головних компонент для окремих одиниць сукупності. Процедура, за якою *h*-й одиниці сукупності надається певна оцінка латентної величини *G*,

називають *факторним шкалюванням*. Значення компонент можна визначити, спираючись на зв'язок їх з первинними ознаками $Z = AG$, звідки

$$G = A^{-1}Z,$$

Таблиця 9.6

Factor Scores (factor.sta)		
Rotation: Unrotated Extraction: Principal components		
Variable	Factor 1	Factor 2
1	–1,442	–1,243
2	1,007	0,773
3	0,200	1,612
4	–0,717	0,063
5	–0,007	–0,162
6	0,401	1,354
7	–2,556	0,859

де A^{-1} — обернена матриця факторних навантажень m компонент.

Враховуючи, що в процесі факторного аналізу виокремлюється p головних компонент ($p < m$), вимірюванню підлягають саме ці компоненти:

$$G = \lambda^{-1}A'Z,$$

де λ^{-1} — дисперсії головних компонент.

Алгебраїчно ця процедура зводиться до підсумовування значень ознак x_i (у стандартизованому масштабі) з вагами, пропорційними факторним навантаженням (до обертання):

$$G_h = \sum_{i=1}^{m}\left(\frac{a_{ij}}{\lambda_j} z_{hi}\right).$$

Ділення факторних навантажень на λ_j забезпечує нульове математичне сподівання та одиничну дисперсію оцінок G. Знаки (+, –) свідчать про те, що рівень компоненти у h-ї одиниці сукупності вищий або нижчий за середній. Обчислені за даними файла *faktor sta*. значення обох головних компонент для семи респондентів наведено у табл. 9.6. Згідно з даними в одних респондентів обидві компоненти додатні, у других — обидві компоненти від'ємні, у третіх — знаки оцінок компонент протилежні.

Оцінки головних компонент застосовують при ранжуванні та типології одиниць сукупності, при вивченні закономірностей динаміки, при вимірюванні взаємозв'язків. У системах одночасних рівнянь, коли коефіцієнти регресії визначаються двокроковим МНК, головні компоненти використовуються на першому кроці як визначені наперед змінні приведеної форми моделі. Такий підхід значно спрощує розрахунки, не впливаючи на точність результатів аналізу.

Таблиця 9.5

Factor Loadings (Varimax normalized) (factor.sta)		
Continue …	Extraction: Principal components (Marked loadings are > ,700000)	
Variable	Factor 1	Factor 2
WORK_1	0,8425	0,0196
WORK_2	0,9023	0,0958
WORK_3	0,8700	0,1185
HOBBY_1	0,7109	0,6075
HOBBY_2	0,7182	0,5165
HOME_1	0,0834	0,8438
HOME_2	0,1213	0,8971
HOME_3	0,1415	0,8538
Expl.Var	3,3438	2,9056
Prp.Totl	0,4180	0,3632