

помощью произвольного линейного преобразования исходных координат $x^{(1)}, x^{(2)}, \dots, x^{(p)}$, в подпространстве, натянутом на первые p' главных компонент, наименее искажается сумма квадратов расстояний между всеми возможными парами рассматриваемых точек-наблюдений.

В данном свойстве за критерий наименьшего искажения геометрической структуры совокупности исходных наблюдений $X_1, X_2, \dots, X_n$ принимается величина

$$M_p(X) = \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^T (X_i - X_j),$$

т.е. сумма квадратов евклидовых расстояний между всеми возможными парами имеющихся наблюдений.

После проектирования точек X_i в p' -мерное пространство, определенное матрицей преобразования C , мы получим точки-проекции $Z_i = C X_i$ ($i = 1, 2, \dots, n$) и соответствующую им сумму квадратов евклидовых расстояний

$$M_{p'}(Z(C)) = \sum_{i=1}^n \sum_{j=1}^n (Z_i - Z_j)^T (Z_i - Z_j).$$

Можно показать, что при $p' < p$ $M_p(X) \geq M_{p'}(Z(C))$.

Так вот, на преобразовании L , с помощью которого получают первые p' главных компонент, достигается минимум разности $M_p(X) - M_{p'}(Z(C))$, т.е.

$$M_p(X) - M_{p'}(Z(L)) = \min_C [M_p(X) - M_{p'}(Z(C))].$$

Свойство 3. Среди всех подпространств заданной размерности p' ($p' < p$), полученных из исследуемого факторного пространства с помощью произвольного линейного преобразования исходных координат $x^{(1)}, \dots, x^{(p)}$, в пространстве, натянутом на первые p' главных компонент, наименее искажаются расстояния от рассматриваемых точек-наблюдений до их общего «центра тяжести», а также углы между прямыми, соединяющими всевозможные пары точек-наблюдений с их общим «центром тяжести».

4.Б. Примеры решения типовых задач и упражнений

Пример 4.1 (задача). При формировании типобразующих признаков предприятий отрасли были обследованы 24 предприятия ($n = 24$) по трем технико-экономическим показателям: объему выпускаемой продукции $x^{(1)}$ (тыс. условных денежных единиц), основным фон-

по $z^{(2)}$ (тыс.у.д.е.) и фонду оплаты труда $x^{(3)}$ (тыс.у.д.е.). По полученным в результате обследования исходным статистическим данным $(x_i^{(1)}, x_i^{(2)}, x_i^{(3)})$, — $i = 1, 2, \dots, 24$, — были получены оценки вектора средних значений $\hat{a} = (\hat{a}^{(1)}, \hat{a}^{(2)}, \hat{a}^{(3)})^T = (420; 240; 85)^T$ и ковариационной матрицы

$$\hat{\Sigma} = \begin{pmatrix} 451,39 & 271,17 & 168,70 \\ 271,17 & 171,73 & 103,29 \\ 168,70 & 103,29 & 66,65 \end{pmatrix}$$

Требуется:

- 1) вывести уравнения для вычисления главных компонент $z^{(1)}, z^{(2)}$ и $z^{(3)}$ по заданным значениям исходных технико-экономических показателей $x^{(1)}, x^{(2)}$ и $x^{(3)}$;
- 2) определить относительные доли суммарной дисперсии, обусловленные одной и двумя главными компонентами;
- 3) используя первую главную компоненту $z^{(1)}$ как прогнозирующую (объясняющую) переменную в задаче автопрогноза исходных показателей $x^{(1)}, x^{(2)}$ и $x^{(3)}$, вычислить значение оценки суммарной относительной ошибки $\sigma_{\text{сум.отн.}}$ прогноза этих показателей по $z^{(1)}$, где под $\sigma_{\text{сум.отн.}}$ понимается величина

$$\sigma_{\text{сум.отн.}} = \frac{\sqrt{\sum_{j=1}^3 E(\tilde{x}^{(j)} - b_j z^{(1)})^2}}{\sqrt{\sum_{j=1}^3 D\tilde{x}^{(j)}}} \cdot 100\%,$$

$\tilde{x}^{(j)}$ — это центрированная переменная $x^{(j)}$ (т.е. $\tilde{x}^{(j)} = x^{(j)} - a^{(j)}$), а b_j — это коэффициент регрессии $\tilde{x}^{(j)}$ по $z^{(1)}$.

Решение

1) Для определения коэффициентов линейного преобразования (4.5), с помощью которого осуществляется переход к главным компонентам, необходимо решить вначале характеристическое уравнение (4.8), а затем использовать найденные собственные значения $\lambda_1, \lambda_2, \dots, \lambda_p$ для подстановки в системы уравнений (4.7), решения которых и дают коэффициенты $l_j = (l_{j1}, l_{j2}, \dots, l_{jp})$. В данной задаче

$$|\hat{\Sigma} - \lambda I| = \begin{vmatrix} 451,39 - \lambda & 271,17 & 168,70 \\ 271,17 & 171,73 - \lambda & 103,29 \\ 168,70 & 103,29 & 66,65 - \lambda \end{vmatrix} = 0,$$

откуда находим $\lambda_1 = 680,40$, $\lambda_2 = 6,50$, $\lambda_3 = 2,86$.

Последовательно подставляя эти значения в систему (4.7) и решая эти системы относительно $l_j = (l_{j1}, l_{j2}, l_{j3})$, $j = 1, 2, 3$, — получаем: $l_1 = (0,813; 0,496; 0,307)$; $l_2 = (-0,545; 0,832; 0,101)$, $l_3 = (-0,205; -0,249; 0,947)$, так что уравнения для вычисления главных компонент $z^{(j)}$ ($j = 1, 2, 3$) будут иметь вид:

$$\begin{aligned} z^{(1)} &= 0,81(x^{(1)} - 420) + 0,50(x^{(2)} - 240) + 0,31(x^{(3)} - 85), \\ z^{(2)} &= -0,55(x^{(1)} - 420) + 0,83(x^{(2)} - 240) + 0,10(x^{(3)} - 85), \\ z^{(3)} &= -0,21(x^{(1)} - 420) - 0,25(x^{(2)} - 240) + 0,95(x^{(3)} - 85). \end{aligned} \quad (4.20)$$

2) В соответствии с равенствами (4.10) и (4.10') и вытекающим из них представлением $I_p(Z)$ в виде (4.4') имеем:

$$\begin{aligned} I_1(Z(X)) &= \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0,9864; \\ I_2(Z(X)) &= \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 0,9958; \end{aligned}$$

что свидетельствует о том, что почти вся информация (а именно, 98,64%) о специфике предприятия данного типа, описанной с помощью переменных $x^{(1)}$, $x^{(2)}$ и $x^{(3)}$, содержится в одной лишь первой главной компоненте $z^{(1)}$.

3) Располагая исходными статистическими данными $x_i^{(1)}, x_i^{(2)}, x_i^{(3)}$, — $i = 1, 2, \dots, 24$, — и используя уравнение (4.20), можно подсчитать значения первой главной компоненты $z_i^{(1)}$ для каждого из анализируемых предприятий. Затем, последовательно рассматривая наблюдения пар переменных $z_i^{(1)}, \tilde{x}_i^{(j)}$, — $i = 1, 2, \dots, 24$, $\tilde{x}_i^{(j)} = x_i^{(j)} - a_j$, — и интерпретируя $z^{(1)}$ как прогнозирующую (объясняющую) переменную в модели парной регрессии $\tilde{x}^{(j)}$ по $z^{(1)}$, можно вычислить (с помощью метода наименьших квадратов, см. [1], п. 15.3.1) оценки коэффициентов регрессии b_j в прогнозных моделях вида $\tilde{z}^{(j)} = b_j z^{(1)} + \varepsilon_j$, где ε_j — ошибка прогноза $x^{(j)}$ по значениям $z^{(1)}$. Далее, воспользовавшись результатами и обозначениями

Главные ком
Собственные
Вклад i -й гл
в суммарную
Суммарный
главных ком

При расче
что $\sum_{i=1}^p \lambda_i =$
компоненты (p
Для интер
ных нагрузок

4.18)~(4.18'), получаем:

$$\sigma_{\text{сум.отн.}} = \frac{\sqrt{\sum_{j=1}^3 E(\tilde{x}^{(j)} - b_j z^{(1)})^2}}{\sqrt{\sum_{j=1}^3 D\tilde{x}^{(j)}}} \cdot 100\% = \frac{\sqrt{\sum_{j=1}^3 \Delta_{jj}^2}}{\sqrt{\sum_{j=1}^3 D\tilde{x}_j}} \cdot 100\%$$

$$\approx \sqrt{\frac{\lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3}} \cdot 100\% = \sqrt{\frac{9,36}{689,76}} \cdot 100\% = 11,65\%$$

(точный знак равенства заменен на приближенный в силу того, что теоретические значения Δ_{jj}^2 и $D\tilde{x}_j$ заменены на их выборочные аналоги).

Пример 4.2 (упражнение). Компонентный анализ проведен по данным двадцати сельскохозяйственных районов ($n = 20$) области, которые содержат результаты измерений следующих показателей: $x^{(1)}$ — число колесных тракторов на 100 га; $x^{(2)}$ — число зерноуборочных комбайнов на 100 га; $x^{(3)}$ — число орудий поверхностной обработки почвы на 100 га; $x^{(4)}$ — количество удобрений, расходуемых на гектар; $x^{(5)}$ — количество средств защиты растений, расходуемых на гектар (исходные статистические данные $x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, x_i^{(4)}, x_i^{(5)}$, — $i = 1, 2, \dots, 20$, — приведены в табл. П2.1 Приложения 2).

Расчеты проводились по *нормированным данным* вида (4.1') и представлены в следующей таблице:

Главные компоненты $z^{(i)}$	$z^{(1)}$	$z^{(2)}$	$z^{(3)}$	$z^{(4)}$	$z^{(5)}$
Собственные значения λ_i	3,04	1,41	0,43	0,10	0,02
Вклад i -й главной компоненты (%) в суммарную дисперсию	60,8	28,2	8,6	2,0	0,4
Суммарный вклад первых главных компонент (%)	60,8	89,0	97,6	99,6	100,0

При расчете относительного вклада главных компонент учитывалось, что $\sum_{i=1}^p \lambda_i = p = 5$. Для анализа были оставлены две первые главные компоненты ($p' = 2$), на которые приходится 89% суммарной вариации.

Для интерпретации главных компонент построена матрица факторных нагрузок

$$A = \begin{pmatrix} 0,95^* & 0,97^* & 0,94^* & 0,24 & 0,56 \\ -0,19 & -0,17 & -0,28 & 0,88^* & 0,67^* \end{pmatrix}^T$$

Звездочкой (*) отмечены элементы, удовлетворяющие условию $|a_{ij}| > 0.1$, т.е. те, которые следует учитывать при интерпретации главных компонент $z^{(1)}$ и $z^{(2)}$.

Требуется:

дать содержательную интерпретацию первых двух главных компонент.

Решение

Из вида матрицы A следует, что первая главная компонента наиболее тесно связана с показателями: $x^{(1)}$ — число колесных тракторов ($a_{11} = r(x^{(1)}, z^{(1)}) = 0,95$); $x^{(2)}$ — число зерноуборочных комбайнов ($a_{21} = r(x^{(2)}, z^{(1)}) = 0,97$); $x^{(3)}$ — число орудий поверхностей обработки почвы на 100 га ($a_{31} = r(x^{(3)}, z^{(1)}) = 0,94$). Поэтому первая главная компонента $z^{(1)}$ интерпретируется как *уровень механизации работ*.

Вторая главная компонента $z^{(2)}$ тесно связана с количествами удобрений ($x^{(4)}$) и средств защиты растений ($x^{(5)}$), расходуемых на гектар ($a_{42} = r(x^{(4)}, z^{(2)}) = 0,88$; $a_{52} = r(x^{(5)}, z^{(2)}) = 0,67$). Соответственно $z^{(2)}$ интерпретируется как *уровень химизации растениеводства*.

Пример 4.3 (задача). В табл. П2.2 Приложения 2 приведены значения десяти характеристик $x_i^{(1)} \sim x_i^{(10)}$ качества жизни населения по 15-ти странам мира ($i = 1, 2, \dots, 15$).

Требуется:

- 1) предложить и обосновать метод вычисления интегрального индикатора качества жизни (ИИКЖ), основанного на знании значений $x_i^{(j)}$ этих десяти частных (статистически регистрируемых) критериев;
- 2) численно реализовать этот метод (с помощью подходящей компьютерной программы) и оценить степень информативности ИИКЖ;
- 3) проранжировать заданные характеристики (частные критерии) в порядке убывания удельного веса их влияния на ИИКЖ.

Решение

1) Определим в качестве ИИКЖ такую переменную, зная значения которой можно с наибольшей точностью восстановить значения всех десяти анализируемых частных критериев в виде линейной аппроксимации от этой переменной. Именно этим свойством (см. в п. 4.А «Свойство наименьшей ошибки автопрогноза», соотношения (4.18') и (4.19')) обладает 1-я главная компонента.

Таким образом, в качестве интегрального индикатора качества жизни населения предлагается использовать линейную комбинацию десяти

$$z^{(1)} = l_{11}(x^{(1)} - a^{(1)})$$

$$a^{(j)} = \frac{1}{15} \sum_{i=1}^{15} x_i^{(j)}$$

и l_{1j} определяются

из матрицы $\hat{\Sigma}$ наибольшем собственном

2) Численная реализация

данной задачи сводится

а) оценка $\hat{a} = (\hat{a}^{(1)}, \hat{a}^{(2)}, \dots, \hat{a}^{(10)})^T$ по формуле

$$\hat{a}^{(j)} = \frac{1}{15} \sum_{i=1}^{15} (x_i^{(j)} - \hat{a}^{(j)})(x_i^{(1)} - \hat{a}^{(1)})$$

б) оценка $\hat{\Sigma} = (\hat{\sigma}_{jl})$ ковариационной матрицы

$$\hat{\sigma}_{jl} = \frac{1}{15} \sum_{i=1}^{15} (x_i^{(j)} - \hat{a}^{(j)})(x_i^{(l)} - \hat{a}^{(l)})$$

в) решение характеристического уравнения

наибольшего корня

г) определение коэффициентов преобразования (4.21)

3) Сравнительную оценку

характеристик $x^{(1)} \sim x^{(10)}$ на

абсолютных величинах ко

целью вычислим матрицу

сколько $r(x^{(j)}, z^{(1)}) = r(z^{(1)}, x^{(j)})$

$x^{(j_0)}$ будет обладать наибольшим

$$|r(x^{(j_0)}, z^{(1)})| = \max_{1 \leq j \leq 10} |r(x^{(j)}, z^{(1)})|$$

Пример 4.4 (упражнение). Пусть $x^{(1)}$ и $x^{(2)}$ — две главные компоненты