ЛАБОРАТОРНАЯ РАБОТА

«МЕТОД ГЛАВНЫХ КОМПОНЕНТ»

Цель работы. Практическое освоение метода главных компонент для решения задач снижения размерности.

КРАТКИЕ СВЕДЕНИЯ.

Определение главных компонент.

Пусть имеется стандартизованная матрица типа «объект – признак» $X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{Np} \end{bmatrix}$

Векторы-столбцы x^1, x^2, ..., x^p - это исходные измеряемые признаки.

$$x^j = \begin{pmatrix} x_{1j} \\ \dots \\ x_{Nj} \end{pmatrix}, \quad j = 1,\dots,p, \quad X = (x^1, x^2, \dots, x^p).$$

Главными компонентами называются новые признаки y^1, y^2, …, y^p, обладающие свойствами:

1) Главная компонента – это линейная комбинация исходных измеряемых признаков $y^j = \sum_{k=1}^{p} c_{jk} x^k, \quad j = 1,\dots, p.$

2) Главные компоненты ортогональны между собой, т.е. некоррелированы $\mathrm{cov}(y^i, y^j)=0$, если $i \neq j$.

3) Главные компоненты упорядочены по мере убывания дисперсии $D(y^1) \geq D(y^2) \geq \dots \geq D(y^p)$.

Необходимо определить такое линейное преобразование, задаваемое матрицей C, в результате действия которого исходные данные выражаются набором(матрицей) главных компонент $Y = (y^1, y^2, \dots, y^p)$, где первые p' главных компонент обеспечивают требуемую долю дисперсии γ (как правило $\gamma \geq 0{,}95$).

Задача сводится к определению матрицы C (нахождению неизвестных параметров c_{jk}, $j = 1,\dots, p$, $k = 1,\dots, p$. Для этого

используется аппарат матричной алгебры. Элементы матрицы C рассчитываются на основе корреляционной матрицы R, построенной по входным данным.

Дисперсия j-ой главной компоненты равна j-ому собственному числу λ_j корреляционной матрицы R, ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$). Координаты собственного вектора, соответствующего j-ому собственному числу матрицы R являются искомыми параметрами c_{jk}, $j = 1,..., p$, $k = 1,..., p$.

Сумма выборочных дисперсий исходных признаков равна сумме выборочных дисперсий проекций объектов на главные компоненты.

$$s^2(y^1) + s^2(y^2) + ... + s^2(y^p) = s^2(x^1) + s^2(x^2) + ... + s^2(x^p).$$

Несмотря на то, что вместо p признаков получается такое же количество главных компонент, вклад большей части главных компонент в объясняемую дисперсию оказывается небольшим. Исключают из рассмотрения те главные компоненты, вклад которых мал. При помощи p' первых (наиболее весомых) главных компонент можно объяснить основную часть суммарной дисперсии в данных.

Относительная доля разброса, приходящаяся на j-ую главную компоненту

$$\frac{s^2(y^j)}{s^2(y^1) + s^2(y^1) + ... + s^2(y^p)} = \frac{\lambda_j}{\lambda_1 + \lambda_2 + ... + \lambda_p}.$$

Относительная доля разброса, приходящаяся на p' первых компонент

$$\frac{s^2(y^1) + s^2(y^1) + ... + s^2(y^{p'})}{s^2(y^1) + s^2(y^1) + ... + s^2(y^p)} = \frac{\lambda_1 + \lambda_2 + ... + \lambda_{p'}}{\lambda_1 + \lambda_2 + ... + \lambda_p}.$$

Таким образом, метод главных компонент позволяет описать большой набор признаков p небольшим числом главных компонент p', при этом различия между объектами зависят от доли изменчивости, связанной с данной главной компонентой. Связи между признаками и главными компонентами – линейные.

Алгоритм метода главных компонент.

1) Сформировать матрицу данных

$$Z(N\times p)=\begin{pmatrix} z_{11} & z_{12} & \dots & z_{1p} \\ z_{21} & z_{22} & \dots & z_{2p} \\ \dots & \dots & \dots & \dots \\ z_{N1} & z_{N2} & \dots & z_{Np} \end{pmatrix}.$$

2) Для устранения неоднородности в исходных данных выполнить нормировку (стандартизацию) данных по столбцам

$x_{ij}=\dfrac{z_{ij}-\bar{z}^{\,j}}{s^{j}}$, где $\bar{z}^{\,j}$ и s^{j} - оценка математического ожидания и среднеквадратическое отклонение по j-ому столбцу (i=1,…,N, j=1,…,p).

3) Построить матрицу ковариации. Ввиду произведенной стандартизации данных матрица ковариации будет корреляционной матрицей исходных данных R порядка pxp.

$$r_{ij}=\frac{1}{N}\sum_{k=1}^{N}x_{ki}x_{kj}, i=1,\dots,p,\ j=1,\dots,p.$$

4) Вычислить собственные числа и собственные векторы корреляционной матрицы, воспользовавшись алгоритмом метода Якоби с преградами (см. Приложение). Упорядочить собственные числа в порядке убывания $(\lambda_1\ge\lambda_2\ge\dots)$, а также упорядочить собственные векторы $c_1, c_2, \dots, c_p$, соответствующие этим собственным числам.

В результате получить ковариационную матрицу главных компонент

$$\begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \lambda_p \end{bmatrix}.$$

Вычислить матрицу нагрузок на главные компоненты путем нормировки собственных векторов.

5) Рассчитать проекции объектов на главные компоненты

$$y^{j}=c_{j1}x^{1}+c_{j2}x^{2}+\dots+c_{jp}x^{p}.$$

Замечание 1.

Прежде чем начинать анализ главных компонент, целесообразно проверить, значимо ли отличается от единичной матрицы корреляционная матрица исходных стандартизованных данных. В предположении, что исходные данные подчиняются многомерному нормальному распределению, можно воспользоваться статистикой

$$d = N\sum_{i,j}^{p} r_{ij}^2 ,$$

где $r_{ij}, i \neq j, i = 1,..., p, j = 1,..., p$ - недиагональные элементы корреляционной матрицы R. Статистика d подчиняется χ^2 – распределению с p(p-1)/2 степенями свободы. Если корреляционная матрица исходных данных не отличается от единичной матрицы, то есть $d \leq \chi^2$, вычисленное при заданном уровне доверительной вероятности и заданном числе степеней свободы, то применение метода главных компонент нецелесообразно.

Порядок выполнения

1. Разработать алгоритм метода главных компонент и программно его реализовать (для проверки правильности вычисления собственных чисел и собственных векторов корреляционной матрицы можно воспользоваться MATLAB).
2. Выполнить анализ экспериментальных данных методом главных компонент.
   - Загрузить данные согласно варианту. Отобразите данные на экране монитора в виде таблицы.
   - Стандартизовать исходные экспериментальные данные. Построить корреляционную матрицу.
   - Удостовериться, что корреляционная матрица значимо отличается от единичной матрицы.
   - Рассчитать проекции объектов на главные компоненты.
3. Произвести анализ результатов работы метода главных компонент.
   - Проверить равенство сумм выборочных дисперсий исходных признаков и выборочных дисперсий проекций объектов на главные компоненты.
   - Построить матрицу ковариации для проекций объектов на главные компоненты.

- Вычислить величину

$$I(p') = \frac{\sum_{i=1}^{p'} \lambda_i}{\sum_{j=1}^{p} \lambda_j}, \quad p' = 1,2,..., p,$$

- причем p' - число новых признаков, выбрать минимальным, удовлетворяющим условию $I(p') > 0.95.$
- Провести анализ и дать содержательную интерпретацию первых двух главных компонент.

Замечание 2. Как провести анализ и интерпретацию двух первых главных компонент(к последнему пункту задания)

Главная компонента – это линейная комбинация исходных измеряемых признаков $y^j = \sum_{k=1}^{p} c_{jk} x^k, \quad j = 1,..., p.$

Рассмотрим первую главную компоненту

$$y^1 = \sum_{k=1}^{p} c_{1k} x^k, \quad k = 1,..., p.$$

$$y^1 = c_{11}x^1 + c_{12}x^2 + ... + c_{1p}x^p;$$

Матрица коэффициентов С вами найдена (см. теорию).

То есть вы можете подставить вместо коэффициентов c_{1k} числа и записать первую главную компоненту в виде линейной функции от исходных признаков. Жизненный смысл этих десяти столбцов вам известен. Если c_{1k} большой по модулю, то зависимость первой главной компоненты от x^k сильная, если коэффициент положительный, то зависимость возрастающая, аналогично остальные случаи. Смотрим, от каких исходных признаков первая главная компонента более всего зависит и какая зависимость, возрастающая или убывающая.

Исходя из жизненного смысла (названий) ваших исходных признаков, наиболее влияющих на первую главную компоненту, вы можете придумать ей название или жизненный смысл.

Аналогично для второй главной компоненты.

Приложение.

1. Описание метода Якоби.

Метод Якоби с преградами применяется для отыскания всех собственных значений и собственных векторов симметричной матрицы. Его суть заключается в проведении цепочки преобразований подобия, в ходе которых из матрицы A получается некоторая диагональная матрица $A^{(k)} = T^T * A * T$, имеющая те же собственные значения, что и матрица A, в то же время известно, что собственные значения диагональной матрицы совпадают с ее диагональными элементами.

2. Входные данные процедуры Якоби.

n – размер матрицы;
ε – точность;
$A\ (n * n)$ – заданная матрица.

3. Выходные параметры процедуры Якоби.

$A\ (n * n)$ – матрица, на диагонали которой содержатся собственные числа.
$T\ (n * n)$ – матрица, столбцы которой являются собственными векторами.

4. Алгоритм реализации метода Якоби с преградами.

Шаг 1.

Задаем $T_0 = E$, где E – единичная матрица порядка n.

Шаг 2.

Вычисляем первую преграду:

$$\alpha_0 = \frac{1}{n}\sqrt{\sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_{ij}^2} = \frac{1}{n}\sqrt{2\sum_{j=2}^{n}\sum_{i=1}^{j-1} a_{ij}^2}$$

Шаг 3.

Находим наибольший по модулю внедиагональный элемент a_{pq}, превосходящий текущую преграду α_k, где $k = 0,1,2,3,\ldots$.
Если такого элемента нет, то переходим к шагу 5 алгоритма, иначе идем к шагу 4.

Шаг 4.

Анализируем найденный элемент a_{pq}. Для этого вычисляем

$$y = \frac{a_{pp} - a_{qq}}{2}$$

$$x = \begin{cases} -1, y = 0 \\ -sign(y) * \dfrac{a_{pq}}{\sqrt{a_{pq}^2 + y^2}}, y \neq 0, \end{cases}$$

$$s = \frac{x}{\sqrt{2\left(1 + \sqrt{1 - x^2}\right)}}, c = \sqrt{1 - s^2}.$$

Преобразуем строки и столбцы матрицы $A^{(k)}$ с номерами p и q следующим образом:

```
for i := 1 to n do begin
    if (i <> p) and (i<>q) then begin
        Обозначим Z1 = aip, Z2 = aiq,
        имеем:
            aqi = Z1 s + Z2 c;
            aiq = aqi;
            aip = Z1 c − Z2 s;
            api = aip;
    end;
end;
```

При преобразовании матрицы A мы не трогали элементы $a_{pp}, a_{qq}, a_{qp}, a_{pq}$.

Для них формулы преобразования выглядят следующим образом:

пусть $Z_5 = s^2, Z_6 = c^2, Z_7 = s * c, V_1 = a_{pp}, V_2 = a_{pq}, V_3 = a_{qq},$

тогда:

$a_{pp} = V_1 Z_6 + V_3 Z_5 - 2 * V_2 Z_7;$
$a_{qq} = V_1 Z_5 + V_3 Z_6 + 2 * V_2 Z_7;$
$a_{pq} = (V_1 - V_3) Z_7 + V_2 (Z_6 - Z_5);$
$a_{qp} = a_{pq}.$
В результате находим матрицу $A^{(k)}$.

Тогда столбцы матрицы T преобразуются следующим образом:

пусть $Z_3 = t_{ip}, Z_4 = t_{iq}$,

тогда $t_{iq} = Z_3 s + Z_4 c, t_{ip} = Z_3 c - Z_4 s, i = \overline{1, n}.$

Шаг 5.

Находим новую преграду $\alpha_{k+1} = \frac{\alpha_k}{n^2}$ и повторяем вычисления с шага 3 до тех пор, пока все недиагональные элементы не станут по модулю меньше числа $\varepsilon * \alpha_0$, где ε – заданная погрешность вычислений.

Результаты:

В результате, собственные значения оказываются диагональными элементами матрицы $A^{(k)}$, а собственные векторы — столбцами матрицы T.

Тестовый пример для проверки метода Якоби.

$$A = \begin{pmatrix} 1.00 & 0.42 & 0.54 & 0.66 \\ 0.42 & 1.00 & 0.32 & 0.44 \\ 0.54 & 0.32 & 1.00 & 0.22 \\ 0.66 & 0.44 & 0.22 & 1.00 \end{pmatrix}$$

$\lambda_1 = 2.3222$
$\lambda_2 = 0.7967$
$\lambda_3 = 0.2426$
$\lambda_4 = 0.6383$

$$T = \begin{pmatrix} 0.5796 & -0.05033 & -0.7188 & 0.3804 \\ 0.4600 & 0.2372 & -0.0957 & -0.8503 \\ 0.4335 & -0.8128 & 0.3874 & -0.0359 \\ 0.5143 & 0.5296 & 0.5692 & 0.3619 \end{pmatrix}$$