Снижение размерности признакового пространства.

Есть несколько методов решения этой задачи:

- метод главных компонент,
- факторный анализ.

Метод главных компонент.

Основная модель МГК.

Предпосылки появления:

1) Многие признаки существенно коррелированы.
2) Некоторые признаки обладают достаточно малой дисперсией, т.е. при переходе от одного объекта к другому почти не изменяются, и, поэтому, малоинформативны.
3) Возможно существуют новые признаки (может быть даже непосредственно не измеряемые).

Определение главных компонент.

Пусть дана матрица объект – признак $X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{Np} \end{bmatrix}$

Векторы x^1, x^2, ..., x^p - это измеряемые признаки. Главными компонентами называются новые признаки y^1, y^2, ..., y^p, обладающие свойствами:

1) Главная компонента - это линейная комбинация исходных измеряемых признаков $y^i = \sum_{k=1}^{p} c_{ik} X^k, \quad i = \overline{1, p}$

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

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

Вычислительная процедура МГК.

Введем вектор $C_i = \begin{bmatrix} c_{i1} \\ \dots \\ c_{ip} \end{bmatrix}$

Это вектор весов i-ой главной компоненты. Тогда i-ая главная компонента в векторной форме выглядит следующим образом:

$y^i(N \times 1) = X(N \times p) \times C_i(p \times 1)$

$y^i = XC_i$

Рассмотрим первую главную компоненту:

$y^1 = X \times C_1$

$$D(y^1) = D\left(\sum_{k=1}^{p} c_{1k} x^k\right) = \sum_{k=1}^{p}\sum_{j=1}^{p} c_{1k} c_{1j} \sigma_{kj}$$

Пусть ковариационная матрица Σ имеет вид:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1p} \\ \dots & & \\ \sigma_{p1} & \dots & \sigma_{pp} \end{bmatrix}$$, σ_{ij}- ковариация между x^i и x^j

Тогда можно показать, что:

$D(y^1) = C_1^T \Sigma C_1$

В общем виде:

$D(y^i) = C_i^T \Sigma C_i$

На вектор весов наложим ограничение, состоящее в том, что сумма квадратов весов каждой компоненты равна 1.

$$\sum_{k=1}^{p} c_{ik}^2 = 1$$

$C_1^T C_i = 1, \qquad i = 1,\dots,p.$

Задача определения первой главной компоненты.

Найти такой ненулевой вектор $C_1 = \begin{bmatrix} c_{11} \\ \dots \\ c_{1p} \end{bmatrix}$, что $D(y^1) = C_1^T \Sigma C_1$ максимальна по C_1, при условии $C_1^T C_1 = 1$.

Эта задача на условный экстремум решается с помощью метода множителей Лагранжа.

$\Gamma(\lambda_1) = C_1^T \Sigma C_1 - \lambda_1 (C_1^T C_1 - 1)$

Это выражение должно быть максимальным по C_1.

Берем производную по C_1.

$\Sigma C_1 - \lambda_1 C_1 = 0$ Отсюда следует, что $(\Sigma - \lambda_1 E) C_1 = 0$

Получили однородную систему p линейных уравнений с p неизвестными. Она имеет нетривиальное решение C_1, если $|\Sigma - \lambda_1 E| = 0$. Таким образом, λ_1- это собственное число ковариационной матрицы.

Из предыдущего уравнения получаем:

$$\Sigma C_1 = \lambda_1 C_1.$$

Обе части этого равенства умножим слева на C_1^T:

$$C_1^T \Sigma C_1 = C_1^T \lambda_1 C_1$$

$$\lambda_1 = C_1^T \Sigma C_1 = D(y^1).$$

Определение второй главной компоненты.

Необходимо найти вектор $C_2 = \begin{pmatrix} c_{21} \\ \dots \\ c_{2p} \end{pmatrix}$ такой, что:

1) $D(y^2) = C_2^T \Sigma C_2$
2) $C_2^T C_2 = 1$
3) $\mathrm{cov}(y^2, y^1) = 0$

Вектор C_2 является собственным вектором матрицы Σ. Он отвечает наибольшему из оставшихся собственных чисел ковариационной матрицы, т.е. собственному числу λ_2.

Аналогичен смысл векторов C_3, C_4 и т.д.

Задача определения i-ой главной компоненты:

Нужно найти вектор $C_i = \begin{pmatrix} c_{i1} \\ \dots \\ c_{ip} \end{pmatrix}$ такой, что:

1) $D(y^i) = C_i^T \Sigma C_i$,
2) $C_i^T C_i = 1$,
3) $\mathrm{cov}(y^i, y^j) = 0$, $j = \overline{1, i-1}$.

Схема МГК исследования.

1) Вычисляем ковариационную матрицу Σ, или корреляционную R.

Замечание: Главная компонента изменяется при линейном преобразовании матрицы, поэтому главные компоненты матриц Σ и R могут существенно отличаться. Решение о том, с какой матрицей работать принимает исследователь в зависимости от размерности исходных признаков. Если все признаки измеряются в одинаковых единицах, то принципиальной разницы в использовании матриц нет. Если признаки имеют различную физическую природу, то лучше работать с матрицей R.

2) Вычисляем собственные числа матрицы R и упорядочиваем их по убыванию $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, а также вычисляем собственные вектора $c_1, c_2, \ldots, c_p$, соответствующие этим собственным числам.

3) Вычисляем $I(p') = \dfrac{D(y^1) + D(y^2) + \ldots + D(y^{p'})}{D(y^1) + D(y^2) + \ldots + D(y^p)}$, $\quad где\ p' \leq p$

Замечание: На практике имеет место следующий факт: для положительно определенной симметричной матрицы сумма элементов, стоящих на главной диагонали, равна сумме всех ее собственных чисел.

$$\sum_{i=1}^{p} \mathrm{D}(\mathrm{x}^{\mathrm{i}}) = \sum_{i=1}^{p} \lambda_i = \left| \lambda_i = \mathrm{D}(\mathrm{y}^{\mathrm{i}}) \right| = \sum_{i=1}^{p} \mathrm{D}(\mathrm{y}^{\mathrm{i}})$$

Т.е. при переходе от исходной системы признаков к главным компонентам, суммарная дисперсия не изменяется.

$$I(p') = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_{p'}}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

4) Выбор числа новых признаков.

Как только I(p')>ε, то p' равняется количеству слагаемых в числителе.

5) Интерпретация и анализ полученных главных компонент.

Числовые характеристики главных компонент.

Главная компонента $y^i = \sum_{k=1}^{p} c_{ik} x^k, \quad i = \overline{1, p}$

Характеристики главных компонент:

1) $M(y^i)=0$

2) $D(y^i)=c_i^T \Sigma c_i$

3) $\mathrm{cov}\left(y^i, y^j\right) = \begin{cases} 0, & \text{если } i \neq j \\ \lambda_i, & \text{если } i = j \end{cases}$

Т.е. ковариационная матрица главной компоненты имеет вид:

$$\Sigma = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \lambda_p \end{bmatrix}$$

Модели факторного анализа.

Статистическая модель главных компонент.

Определение: основное соотношение факторного анализа:

$$x^j = \sum_{k=1}^{m} \alpha_{kj} f^k + \xi^j, \quad j = \overline{1, p},$$

x^j- стандартизованные измеряемые признаки (M=0, D=1);

f^k- стандартизованные общие факторы;

ξ^j- центрированные, но не нормированные специфические факторы (M=0, D≠1),
дополненное следующими двумя предположениями:

1) общие факторы не коррелированны между собой.

$$\left(f^i, f^j\right) = \begin{cases} 0, & i \neq j \\ N, & i = j \end{cases}$$

2) общие факторы и факторные нагрузки таковы, что суммарная дисперсия специфических факторов минимальна.

$$\sum_{i=1}^{p} s^2\left(\xi^i\right) \to \min_{\substack{\alpha_{kj} \\ f_k}}, \quad k = 1,\dots,m; \quad j = 1,\dots,p$$

-называется статистической моделью главных компонент (СМГК).

$$\sum_{i=1}^{p} s^2\left(\xi^i\right) = \frac{1}{N}\sum_{i=1}^{p}\left(\xi^i, \xi^i\right)$$

Таким образом, близость между совокупностью измеряемых признаков и совокупностью общих факторов в СМГК понимается в смысле суммы квадратов евклидовых расстояний между соответствующими векторами.

Смысл факторных нагрузок.

Специфический фактор $\xi^j = -\sum_{k=1}^{m}\alpha_{kj} f^k + x^j$

Общие факторы и факторные нагрузки выбираются из условия:

$$\frac{1}{N}\left(x^j - \sum_{k=1}^{m}\alpha_{kj} f^k, x^j - \sum_{k=1}^{m}\alpha_{kj} f^k\right) \to \min_{\substack{\alpha_{kj} \\ f^k}}, \quad k = 1,\dots,m, \quad j = 1,\dots,p$$

Найдем нагрузку j-го параметра на s-ый вектор.

Пусть имеется вектор $a(x)=(a_1(x),..,a_n(x))$

$$(a(x),a(x))'_x = \left(\sum_{k=1}^{n} a_k^2(x)\right)'_x = 2\sum_{k=1}^{n} a_k(x) a'_k(x) = 2\left(a(x), a'(x)\right)$$

$$\frac{1}{N}\left(x^j - \sum_{k=1}^{m}\alpha_{kj} f^k, x^j - \sum_{k=1}^{m}\alpha_{kj} f^k\right)'_{\alpha_{sj}} = 0$$

$$\left(x^j-\sum_{k=1}^{m}\alpha_{kj}f^k, f^s\right)=0;$$

$(\xi^j, f^s)=0$, $j=1,\dots,p$; $s=1,\dots,m$

В СМГК специфические факторы и общие факторы не коррелируют между собой.

$$\left(x^j, f^s\right)-\sum_{k=1}^{m}\alpha_{kj}\left(f^k, f^s\right)=0;$$

$(x^j, f^s)-\alpha_{sj}N=0$

$$\alpha_{sj}=\frac{1}{N}\left(x^j,f^s\right)=r(x^j,f^s)$$

То есть $|\alpha_{sj}|\le 1$.

<u>Следствие:</u> Специфический фактор и вычисленный признак $\hat{x}_j=\sum_{k=1}^{m}\alpha_{kj}f^k$ **не коррелируют между собой.**

Доказательство:

$$\left(\hat{x}_j,\xi_j\right)=\left(\sum_{k=1}^{m}\alpha_{kj}f^k,\xi^j\right)=\sum_{k=1}^{m}\alpha_{kj}\left(f^k,\xi^j\right)=0$$

Дисперсия измеряемого признака.

$$s^2\left(x^j\right)=\frac{1}{N}\left(x^j,x^j\right)=\frac{1}{N}\left(\hat{x}^j+\xi^j,\hat{x}^j+\xi^j\right)=\frac{1}{N}\left(\hat{x}^j,\hat{x}^j\right)+\frac{2}{N}\left(\hat{x}^j,\xi^j\right)+\frac{1}{N}\left(\xi^j,\xi^j\right)=$$
$$=\frac{1}{N}\left(\hat{x}^j,\hat{x}^j\right)+\frac{1}{N}\left(\xi^j,\xi^j\right).$$

Так как x^j – стандартизованные признаки, то

$$1=\frac{1}{N}\left(\hat{x}^j,\hat{x}^j\right)+\frac{1}{N}\left(\xi^j,\xi^j\right)$$
$$1=s^2\left(\hat{x}^j\right)+s^2\left(\xi^j\right)$$
$$s^2\left(\hat{x}^j\right)=s^2\left(\sum_{k=1}^{m}\alpha_{kj}f^k\right)=\sum_{k=1}^{m}\alpha_{kj}^2 s^2\left(f^k\right)=\sum_{k=1}^{m}\alpha_{kj}^2$$

$$1=\underbrace{\alpha_{1j}^2+...+\alpha_{mj}^2}_{\text{общность}}+\underbrace{s^2\left(\xi^j\right)}_{\text{специфично сть}}$$

Нужно, чтобы общность давала больший вклад, а специфичность меньший.

Другая формулировка СМГК.

Равенство $1=s^2(\hat{x}^j)+s^2(\xi^j)$ просуммируем по j.
Получим

$$p=\sum_{j=1}^{p}\underset{\substack{\Downarrow\\ \max}}{s^2\left(\hat{x}^j\right)}+\sum_{j=1}^{p}\underset{\substack{\Downarrow\\ \min}}{s^2\left(\xi^j\right)}$$

$$\sum_{j=1}^{p}s^2\left(\hat{x}^j\right)\to\max$$

В силу предположения 2 общие факторы и факторные нагрузки выбираются из условия: суммарная дисперсия вычисленных признаков должна быть максимальна.
поэтому можно сделать вывод:

$$\sum_{j=1}^{p}s^2\left(\xi^j\right)\to\min\Leftrightarrow\left(f^k,\xi^j\right)=0\Leftrightarrow\alpha_{kj}=r\left(f^k,x^j\right)\Leftrightarrow\sum_{i=1}^{p}s^2\left(\hat{x}^j\right)\to\max.$$

Это эквивалентно условию:

$$\sum_{j=1}^{p}\sum_{k=1}^{m}r^2\left(x^j,f^k\right)\to\max \qquad \left(2^*\right)$$

Вывод условия 2*:

$$\sum_{j=1}^{p}s^2\left(\hat{x}^j\right)=\sum_{j=1}^{p}s^2\left(\sum_{k=1}^{m}\alpha_{kj}f^k\right)=\sum_{j=1}^{p}\sum_{k=1}^{m}\alpha_{kj}^2=\sum_{j=1}^{p}\sum_{k=1}^{m}r^2\left(x^j,f^k\right)\to\max$$

Основное соотношение факторного анализа, дополненное предположением (1) и предположением (2^*) называется статистической моделью главных компонент (СМГК).

При этом факторные нагрузки выбираются из условия $(f^k,\xi^j)=0$ или $\alpha_{kj}=r(x^j,f^k)$. Таким образом, в СМГК близость между совокупностью измеряемых признаков и совокупностью общих факторов может пониматься в смысле суммы квадратов парных коэффициентов корреляции.

Коэффициент корреляции измеряемых признаков.

$x^i = \hat{x}^i + \xi^i$ -измеряемый признак.

$$r\left(x^i, x^j\right) = r\left(\hat{x}^i + \xi^i, \hat{x}^j + \xi^j\right) = \frac{1}{N}\left(\hat{x}^i + \xi^i, \hat{x}^j + \xi^j\right) = \frac{1}{N}\left(\hat{x}^i, \hat{x}^j\right) + \frac{1}{N}\left(\xi^i, \hat{x}^j\right) + \frac{1}{N}\left(\hat{x}^i, \xi\right.$$

$$= \frac{1}{N}\left(\hat{x}^i, \hat{x}^j\right) + \frac{1}{N}\left(\xi^i, \xi^j\right) = \operatorname{cov}\left(\hat{x}^i, \hat{x}^j\right) + \operatorname{cov}\left(\xi^i, \xi^j\right)$$

Недостатком СМГК является то, что в последнем соотношении, несмотря на минимизацию суммарной дисперсии специфических признаков, ковариации могут быть большими. Т.е. вычисляемые признаки хорошо объясняют суммарную дисперсию исходных признаков, но плохо - корреляцию между ними.