22. Метод на главните компоненти. Факторен анализ

Факторният анализ е възникнал като задача при обработката на данни от психологически анкети.

Търси се представяне на данните във формата:

$$X = FL + E$$

Тук $X \in \mathbb{R}^{n \times m}$ е матрицата на центрираните и нормирани наблюдения, а $E \in \mathbb{R}^{n \times m}$ – грешките.

Във факторния анализ са приети следните названия.

- Факторни товари се наричат коефициентите на разлагане на оригиналните променливи по фактори. те $L \in \mathbb{R}^{k \times m}$.
- Факторни стойности се наричат оценените стойности на факторите за всяко наблюдение $F \in \mathbb{R}^{n \times k}$
- Общност – това е относителната дисперсия на всяка променлива в новото ѝ описание. Т.е. общностите показват доколко стойностите на конкретната променлива могат да бъдат предсказвани или възстановени от факторните стойности.

Основните задачи на факторния анализ са:

- Да се определи бр. на факторите, достатъчни за описване на зависимостите (втората размерност на матрицата F)
- Да се намерят фактори това означава да се изчислят коефициентите на линейната функция L, представяща променливите (посредством факторите)

Корелационна матрица

Корелационна (или ковариационна) матрица на случаен нормален вектор в m-мерното пространство е симетрична и неотрицателно определена. Тя може да бъде представена:

$$S = \sum_{i=1}^{m} a_i e_i e_i' \quad \ldots (1)$$

Тук числата $\{a_1, a_2 \ldots a_m\}$ се наричат собствени числа, а векторите $\{e_i, i=1 \ldots m\}$ - собствени вектори, защото удовлетворяват условието:

$$S e_i = a_i e_i \quad \ldots (2)$$

Те са ортогонални с единична норма и образуват базис в m-мерното евклидово пространство

$$S^2 = \sum \|x_i\|^2 = \sum \|P_k x_k\|^2 + \sum \|x_i - P_k x_i\|^2$$

$S^2 = \sum \|x_i\|^2$ - представлява центрираност на данните, то естествено е да търсим това подпространство на H с размерност k (за фиксирано k, $1 \le k \le m$), за което експерименталните точки биха се преместили минимално при своето проектиране P върху него. Тогава най-малко би се изместила и S^2, което се вижда и от равенството.

Брой на факторите:

$$\sum_{i=1}^{m} \|P_k x_i\|^2 \geq 0{,}95 \sum_{i=1}^{n} \|x_i\|^2$$

Друг срещан критерий е избора на собствените вектори със собствени стойности, които са по-големи 1. Ако използваме корелационната матрица $\sum a_i = tr R = m$ - критерий на Кайзер.

Сред много методи, обясняващи интерпретацията на факторите, може би най-популярен е методът варимакс-трансформация. Идеята на варимакс-трансформацията е така да бъдат променени факторите, че те да запазят добрите си свойства (полезно на описание на индивидите) и да получат по-добри интерпретативни качества. Преразпределят се факторните тегла като големите нарастват, а малките намаляват, което води до това, че всеки фактор се обяснява от по-малко на брой променливи. Факторите вече не са така независими. Процесът се обуславя от променливите. той се състои в максимизиране на функционала:

$$\max_{K} \operatorname{var}(KL) = \max_{K} \sum_{i=1}^{k} \sigma^2(i)$$

по всевъзможните избори на факторите всъщност факторно подпространство.

Тук $\sigma^2(i)$ е „дисперсията" на квадратните факторни тегла, а $K \in K^{k \times k}$ е ортогонална матрица. Теглата на променливите в даден фактор са записани по редове в L. Максимизирането на дисперсията води до увеличаване на разликата между големите и малките тегла. Така се винаги кои променливи са ясно представени в даден фактор - те получават тегла близки до единица (по стойност) на тези тегла, съответният метод а нарича квартимакс

Общият вариант на едновременна максимизация претегленото сума на двата функционала:

- биквартимакс - два функционала са с равно тегло
- еквимакс - варимакс функционалът е с тегло $\frac{k}{2}$

и степента на тяхното определяне фактори в
отделните обекти F.
3. Да се интерпретират получените резултати в
термини на предметната област. на данните.

Методика на факторния анализ.

Ако си представим наблюдаемия като облак от точки,
в пространството, то естествена характеристика
на този облак биха били неговата форма и разс
Те се отразяват до известна степен в ковариационната
матрица корелационната матрица е обект на
класическия факторен анализ.
Съществуват обаче редица вариации на факторния
анализ, които използват други аналогични матрици,
като интерпретацията е различна.

Главните компоненти. – Метод на главните компоненти позволява да определим до колко определен брой фактори описват явлението. Тук. Главните компоненти съответстват на осите на елипсоида на разсейва-
не на точките, представящи обекти (данни) в
пространството на наблюдаваните променливи.
Техния брой е равен на броя на променливите.
Главните компоненти могат да бъдат разглеждани
като фактори, т. е като нови променливи, които
изцяло описват извадката и са независими.
Тази независимост позволява да представим общата
дисперсия на извадката като сума от дисперсиите
на новите променливи. Отстранявайки тези от
факторите които имат незначителна дисперсия
(разсейване) може да дадем описание на свойс-
твата на извадката с нямален брой нови променливи.

Когато броят на факторите е избран, решението получено по метода на главните компоненти може да бъде записано:

$$X = FL + E$$

Тук. $X \in \mathbb{R}^{n \times m}$ е матрица на центрираните и корегирани наблюдения, $F \in \mathbb{R}^{n \times k}$ – са факторните стойности $L \in \mathbb{R}^{k \times m}$ – са факторните тегла, а $E \in \mathbb{R}^{n \times m}$ – грешките.

Решението по метода на Хотелинг дава дава минимум на функционала $tr\, E'E$. Ако разгледаме следното представяне на матрицата $X = UDV$. Тук $U \in \mathbb{R}^{n \times n}$ и $V \in \mathbb{R}^{m \times m}$, са ортогонални, а $D \in \mathbb{R}^{n \times n}$ – „диагонална" с неотрицателни елементи.
$(d_i, i \geq d_{i+1, i+1})$

Тогава имаме следното представяне на ковариационната матрица

$$X'X = V'D'U'UDV = V'D'DV = \sum_{i=1}^{n} d_{i,i}^2 \, Vi\, V'i$$

Анал. е представянето на матрицата XX', но собствените вектори са в друго пространство. Имаме $d_{i\,i}^2 = \lambda$ фиксиране k и „орязване" тогава к стълба на една матрица. Получаваме $F = n^{-1/2}(UD)_k$, $L' = (V')_k$ Числата $d_i^2 i / n$ са дисперсиите на факторите, а $\sum_{i=1}^{k} l_{i,j}^2$ са т.н общности на променливите. Те обясняват каква част от дисперсията е обяснена с тази подбрани фактори.
Матрицата $L \in \mathbb{R}^{k \times m}$