БИТ-2010
Доц. д-р С. Гочева

Лекция 9. Факторен анализ

Обща характеристика

Приложението на факторния анализ (ФА) за изследване на данни е отдавна наложен и широко използван статистически подход в психологията, социалната сфера, икономиката, естествените науки и други области.

ФА е статистическа техника, предназначена за преобразуване на множество от корелиращи данни в ново множество с некорелиралищи изкуствени променливи или фактори, които обясняват възможно по-голяма част от общата вариация на изходните данни. С тази техника се постига редуциране на броя на началните променливи, чрез групирането на тези, които корелират помежду си в общ фактор и разделянето на некорелиращите в различни фактори. Математически това се постига с намаляване размерността на първоначалното пространство чрез установяване на нов базис от променливи (фактори). Факторите могат да бъдат ортогонални или наклонени.

Нека $x_1, x_2, ..., x_p$ са началните променливи, всяка с *n* наблюдения. Означаваме с $\mathbf{X} \in \mathbb{R}^{p \times n}$ съответната матрица от данни. Във ФА тази матрица се моделира като линейна комбинация на *k* ($k < p$) на брой n-мерни фактори, $F_1, F_2, ..., F_k$ плюс грешки $E_1, E_2, ..., E_p$:

$$\mathbf{X} = \mathbf{LF} + \mathbf{E}, \qquad (1)$$

където $\mathbf{F} \in \mathbb{R}^{k \times n}$ е матрицата на факторните стойности, $\mathbf{L} \in \mathbb{R}^{p \times k}$ е матрица на факторните тегла, $\mathbf{E} \in \mathbb{R}^{p \times n}$ е матрицата на грешките.

Основна задача във ФА е определянето на броя на факторите k, който зависи съществено от силата на корелационните зависимости между данните.

За по-силно разграничаване на принадлежността на изходните променливи към един или друг фактор се извършва и допълнително преобразуване на факторите с т. нар. въртене на факторите.

ФА позволява за всеки от факторите да се пресметнат факторните му стойности (factor scores), с което факторите се получават като нови латентни променливи. Тези променливи могат да бъдат използвани в последващи статистически анализи вместо първоначалните променливи. Факторните променливи са в стандартизиран вид със *z* – стойности, със средно нула и стандартно отклонение единица.

Разработени са голям брой методи за извличане на факторите: метод на главните елементи, метод на най-малките квадрати, алфа-факторинг и др. За въртене на факторите се използва най-често т.н. варимакс (Varimax) метод, но има и други методи – Quartimax, Equamax, както и такива, при които факторите са наклонени, напр. методите Oblimin, Promax и др.

Ще отбележим по-специално, че най-разпространения тип ФА е т.н. изследователски (exploratory) ФА, който се използва в този труд. Процедурите на изследователския ФА (определяне на броя на факторите, извличане на факторите, метод на въртене и др.) не са достатъчно формализирани и получаваното решение не е единствено. Счита се, че намереният факторен модел е добър, когато има достатъчно ясна практическа интерпретация в рамките на разглежданата предметна област.

На ФА е посветена огромна по количество литература, вкл. с използване на специализиран софтуер.

Основни изисквания за приложение на ФА

За получаване на адекватен факторен модел се изисква:

- Изходните данни да имат случаен характер.
- Да бъдат от интервален или относителен тип. Категорийни и номинални данни не се допускат. Наблюденията трябва да бъдат независими.

- Препоръчваният брой наблюдения е поне 50 (виж и по-долу).
- Във ФА участват корелиращи помежду си променливи. Променлива, която не корелира с останалите трябва да бъде предварително изключена от ФА (случай на unique variable). При последващи анализи такава променлива се използва заедно с получените фактори вместо изходните предиктори.
- В идеалния случай данните е добре да имат многомерно нормално разпределение. Последното се проверя с т.н. условие за адекватност на факторния модел с предварителна проверка на данните с теста на Kaiser-Mayer-Olkin или (KMO тест), чиято стойност трябва да е >0.5, както и установяване на сферичност на облака от данни с теста на Bartlett. Може да се проверяват и стойностите на MSA (Measure of Sampling Adequacy) за всяка променлива, които също е препоръчително да са над 0.5.

ФА по метода на главните елементи (МГЕ)

Ще разгледаме само най-разпространения метод на извличане на факторите по МГЕ (PCA – Principal Component Analysis).

В този случай по МГЕ най-напред се търси корелационната матрица

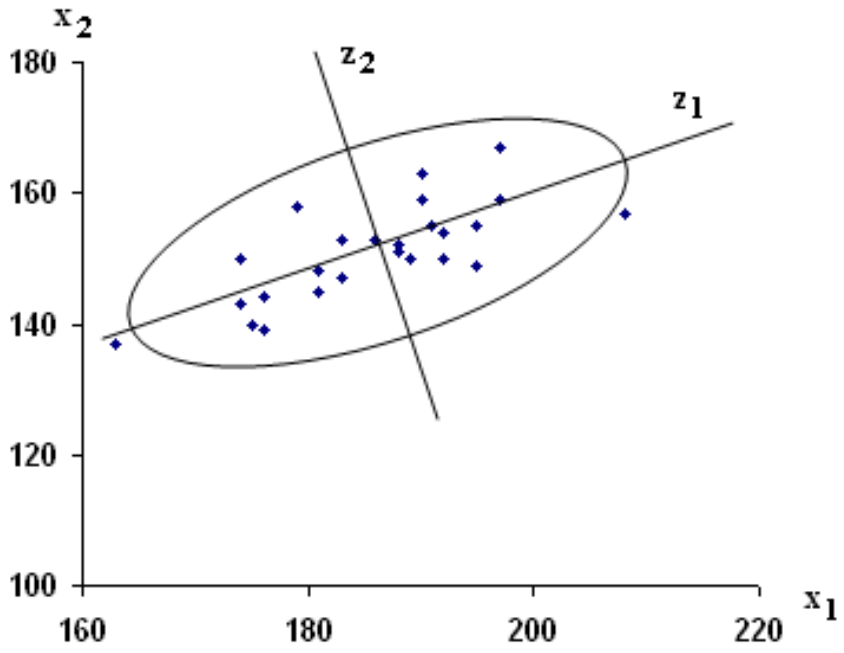
$$\mathbf{R} = \begin{pmatrix} 1 & r(x_1, x_2) & \dots & r(x_1, x_p) \\ r(x_2, x_1) & 1 & \dots & r(x_2, x_p) \\ \dots & \dots & \dots & \dots \\ r(x_p, x_1) & r(x_p, x_2) & \dots & 1 \end{pmatrix}, \qquad (2)$$

където $r(x_i, x_j)$ е корелационният коефициент за променливите x_i, x_j.

Сумата от собствените стойности на тази матрица е точно равна на броя на участващите променливи p. Относителната стойност на всяко собствено значение λ_α изразява доколко съответният й фактор F_α участва в обясняването на общата вариация на началните данни. Във ФА броят на

факторите най-често се избира да бъде равен на броя на собствените стойности на корелационната матрица, които са по-големи от единица, което е известно като правило на Кайзер. Но има примери, от които се вижда, че фактори, съответстващи на собствени стойности доста по-малки от единица също могат да оказват съществено влияние върху модела. Това налага при недостатъчно добри резултати от модела да се включват и допълнителни фактори, не отговарящи на условието на Кайзер.

На Фиг. 1 е даден пример, илюстриращ метода на главните компоненти, с който изходната ортогонална система на две променливи x_1, x_2 е трансформирана в нова ортогонална система z_1, z_2, относно която данните имат по-добро представяне. В частност се вижда, че координатните оси се завъртат и се минимизира общата сума на квадратите на разстоянията от точките до осите в новата координатна система спрямо същото за старата координатна система.



Фиг. 1. Трансформация на променливите x_1, x_2 към z_1, z_2 с метода на главните компоненти.

При определен фиксиран брой на факторите, те се извличат по метода на главните компоненти с разлагането (1) и се получава начално решение. След това се извършва процедура на въртене и окончателно се намират факторите като "завъртяно решение".

Основен момент във ФА е получаването на матрицата на факторните тегла $\mathbf{L}^{rot} = (l_{ij})$ (factor loadings) на завъртяното решение, наречена ротационна матрица. С факторните тегла се изразява връзката между факторите и изходните величини: Те са коефициентите на регресия на изходните величини върху групата от фактори, което се изразява от приближените равенства (във вид на линейни комбинации)

$$\left|\begin{array}{l} F_1 \approx l_{11}x_1 + l_{12}x_2 + \ldots + l_{1p}x_p \\ F_2 \approx l_{21}x_1 + l_{22}x_2 + \ldots + l_{2p}x_p \\ \ldots \\ F_k \approx l_{k1}x_1 + l_{k2}x_2 + \ldots + l_{kp}x_p \end{array}\right. . \qquad (3)$$

Променлива, умножена по факторно тегло, чиято абсолютна стойност е голяма (напр. над 0.5) може да бъде групирана към съответния фактор, а когато абсолютната стойност на теглото е под избраната граница, участието на променливата се пренебрегва. Успешният факторен модел изисква дадена променлива да се групира по този начин само към един фактор.

Зависимост между размера на извадката и факторното тегло

Няма установено правило за това, какъв трябва да бъде размерът на извадката, за да се провежда ФА. Битува мнението, че за всяка променлива трябва да има поне 10 или дори 20 наблюдения. Например, при 8 променливи по това правило извадката трябва да е с обем поне 80.

Друг основен момент е съотношението между обема на извадката и големината на факторното тегло, за да се приеме, че факторното тегло е значимо и променливи с факторни тегла над посоченото се групират със

съответния фактор. Обикновено се приема, че границата за значимост е 0.5. По-прецизни резултати при ниво на значимост $\alpha = 0.05$ са приведени в табл. 1. От таблицата следва, че например при извадка с размер $n = 100$, само променливи с факторни тегла над 0.55 трябва да се считат за статистически значими за съответния фактор.

Табл. 1. Зависимост между обема на извадката и нивото на значимост на факторното тегло.

Размер на извадката, необходим за значимост (ниво 0.05)	**Факторните тегла трябва да са не по-малки от**
350	0.30
250	0.35
200	0.40
150	0.50
100	0.55
85	0.60
70	0.65
60	0.70
50	0.75

Алгоритъм на процедурата на ФА

- Нормализиране на всички променливи чрез трансформация към z-стойности
- Изчисляване на корелационната матрица и съответните статистики
- Проверка за адекватност на ФА
- Извличане на факторите по на метода на главните елементи или друга техника за изчисляване на комуналите (натрупваните вариации) и разпределение на общата вариация
- Избор на броя на факторите
- Получаване на начално факторно решение
- Въртене на факторите и получаване на ротационната матрица
- Преценка за правилно групиране на променливите по фактори

- Изчисляване и запомняне на факторните стойности за по-нататъшни статистически анализи
- Интерпретация на факторите и резултатите от ФА.

Валидация на резултатите от ФА

Макар ФА да не е добре формализирана техника, отделните му етапи, описани по-горе трябва да удовлетворяват известен брой тестове и изисквания, които служат за валидация на получавания факторен модел. Основните от тях са:

- Използваните променливи да корелират помежду си и със зависимите променливи, т.е. бивариантните корелационни коефициенти трябва да са високи. В противен случай ФА не се препоръчва
- Статистическите тестове за адекватност на ФА за дадената извадка да са изпълнени: KMO тестът трябва да е >0.5, Бартлет тестът за сферичност, включващ и χ^2 теста за многомерно нормално разпределение трябва да е статистически значим, т.е. Sig. < 0.05.
- Валидиране на избора на броя на факторите по някакъв метод, напр. с изчисляване на репродуцираните корелационни остатъци. За целта по пресметнатите фактори и факторни тегла се получават приближенията на променливите по формула (1), изчислява се корелационната им матрица, наречена репродуцирана корелационна матрица и тя се сравнява с първоначалната корелационна матрица. Резидиумите трябва да са достатъчно малки, в рамките на 0.05.
- Групирането във факторите в ротационната матрица трябва да бъде коректно, т.е. дадена променлива може да участва само в един фактор (да корелира силно с него), а с другите фактори да има слаба корелация (виж Табл. 1).

Кога да използваме ФА

Фа дава възможност за построяване на модели на базата на експерименталните данни, когато прякото прилагане на регресионни техники е невъзможно или силно затруднено.

С ФА се решават следните видове задачи:

- Класификация на изследваните независими величини чрез групирането им във фактори на базата на взаимната им корелация
- Отхвърляне на несъстоятелните входни величини, т.е. тези, които нямат реално влияние върху изходните променливи
- Получаване на адекватни на данните мултифакторни модели, които обясняват в много голям процент (желателно над 70%) изменчивостта на приближаваните данни
- Получаване на факторни променливи, които са взаимно независими и подходящи за по-нататъшни статистически обработки от типа на параметрични и непараметрични регресионни анализи.