МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ

Многомерное шкалирование (МШ)— это метод, развитый в науках о поведении и социологии с целью изучения объектов и индивидов.

Оно применялось для исследования социальной структуры организации, семантической структуры слов, логической структуры служебных обязанностей.

МШ оказалось полезным во многих областях, в том числе в антропологии, педагогике, географии, истории, психологии, социологии, науках о поведении, исследованиях маркетинга.

Как факторный и кластерный анализ, МШ используется для описания структуры. Исходные предположения МШ отличаются от исходных предположений факторного и кластерного анализа, поэтому МШ обычно позволяет получить описания, которые отличаются от описаний, полученных с помощью методов фак торного и кластерного анализа.

Кроме того, МШ применимо к данным, для которых непригодно большинство обычных методов факторизации.

Основной тип данных в МШ — меры близости между двумя объектами. Мера близости — это величина, определенная на паре объектов и измеряющая, насколько эти два объекта похожи.

Часто встречаются такие меры близости, как коэффициенты корреляции и совместные вероятности. Обозначим меру близости пары стимулов (i, j) символом δ_{ij} .

Если мера близости такова, что самые большие значения δ_{ij} соответствуют парам наиболее похожих объектов, то δ_{ij} -мера сходства.

Если же мера близости такова, что самые большие значения δ_{ij} соответствуют парам наименее похожих объектов, то δ_{ij} - мера различия.

- Согласно наиболее полному определению под МШ понимается «семейство геометрических моделей для многомерного представления данных и соответствующий набор методов для подгонки таких моделей к реальным данным».

- Под столь широкое определение подходит большинство методов многомерной статистики, в том числе факторный и кластерный анализ.

- Мы будем трактовать МШ значительно уже как набор многомерных статистических методов, предназначенных для определения соответствия данных о близости различным дистанционным пространственным моделям и для оценки параметров этих моделей.

Дистанционная модель для различий

- «Дистанционная пространственная модель» является намеком на аналогию между понятием сходства в психологии и понятием расстояния в геометрии. Строго говоря, аналогия включает не понятие сходства в психологии, а понятие различия. Параллели между различием и расстоянием просматриваются в аксиомах расстояния.
- Для того чтобы функция, определенная на парах объектов *(a, b),* была евклидовым расстоянием, она должна удовлетворять следующим четырем аксиомам:
- $d(a, b)>0$, (1.1)
- $d(a,a) = 0$, (1.2)
- $d(a,b)=d(b,a)$, (1.3)
- $d(a,b)+d(b,c)>=d(a,c)$. (1.4)
- В применении к понятию различия первая аксиома означает, что или два объекта идентичны друг другу и их различие равно 0, или они в чем-то отличны друг от друга и их различие больше 0. Вторая аксиома означает, что объект идентичен сам себе. В соответствии с третьей аксиомой объект a так же отличается от объекта b, как объект b отличается от объекта a.
- Хотя выполнение первых трех аксиом интуитивно кажется вполне возможным, никакие качества различия не дают возможности предположить, будет или не будет выполняться четвертая аксиома, называемая аксиомой треугольника (неравенством треугольника). Однако в социологии и науках о поведении выполнение трех аксиом из четырех уже неплохо, так что аналогия между различием в психологии и расстоянием в геометрии есть.

- Более формально дистанционную модель для различий можно описать следующим образом.

- Пусть δ_{ij} - мера различия между объектами *i* и j.

- Объектами могут быть автомобили, места работы, кандидаты на должности. Согласно модели меры различия функционально связаны с *K* признаками объектов.

- Если объекты - автомобили, то признаками могут быть, например, цена, расход бензина на милю, спортивность автомобиля.

- Если объекты - места работы, то признаками могут служить престижность, заработная плата, условия труда.

- Пусть x_{ik} и x_{jk} -значения признака *k* у объектов *i* и *j* соответственно. Например, если объекты- автомобили, а признак - расход бензина, то x_{ik} и x_{jk} будут означать расход бензина этих автомобилей. Или если объекты - места работы, а признак *k* - престижность, то x_{ik} и x_{jk} - престижность работы i и j соответственно.

- Согласно обычной формуле евклидова расстояния меры различия связаны со значениями признаков следующей функцией:

$$\delta_{ij} = d_{ij} = \left[\sum_{k=1}^{K} \left(x_{ik} - x_{jk} \right)^2 \right]^{1/2} .$$

- δ_{ij} -обозначает данные, величину, полученную для пары объектов *(i, j)* эмпирически, путем наблюдения. С другой стороны, d_{ij}, x_{ik} и x_{jk} — теоретические величины в статистической модели для данных о различии. Эти теоретические величины непосредственно не наблюдаемы и могут быть оценены по данным.

- М. Ричардсон предложил начать с субъективных суждений о различиях объектов в парах и получить признаки, на которых эти суждения основаны, а также значения стимулов по этим признакам. Он ввел задачу статистической оценки, откуда и появилось МШ, — задачу оценки координат стимулов x_{ik} и x_{jk} по мерам различий.

Модель Торгерсона

- В модели Торгерсона предполагается, что оценки различий равны расстояниям в многомерном евклидовом пространстве. Пусть снова δ*ij* — мера различия между объектами i и j.
- Под *xik* и x*jk (*i = 1, ..., I; j = 1, ..., J;I = J; к = 1, ..., К*)* будем понимать координаты стимулов i и j по оси k. Отметим, что число строк I в матрице различий равно числу столбцов J, так как строки и столбцы соответствуют одним и тем же стимулам. Основное предположение Торгерсона следующее:
- $$\delta_{ij} = d_{ij} = \left[\sum_k \left(x_{ik} - x_{jk} \right)^2 \right]^{1/2} \quad (1.1)$$
- Без потери общности можно предположить, что среднее значение координат стимулов по каждой оси равно нулю:
- $$\sum_i x_{ik} = \sum_j x_{jk} = 0{,}0. \quad (1.2)$$

Торгерсон начал с построения матрицы Δ^* с двойным центрированием, элементы δ_{ij}^* которой посчитаны непосредственно по матрице данных. Матрица с двойным центрированием - это матрица, у которой среднее значение элементов каждой строки и каждого столбца равно 0,0. Каждый элемент новой матрицы получается следующим образом:

- $$\delta_{ij}^* = -\frac{1}{2}\left(\delta_{ij}^2 - \delta_{i.}^2 - \delta_{.j}^2 + \delta_{..}^2\right). \quad (1.3)$$
- $$\delta_{i.}^2 = \frac{1}{j}\sum_j \delta_{ij}^2 \quad , \quad \delta_{.j}^2 = \frac{1}{i}\sum_i \delta_{ij}^2 \; , \quad \delta_{..}^2 = \frac{1}{ij}\sum_i\sum_j \delta_{ij}^2 \quad (1.4)$$
- Торгерсон показал, что если данные удовлетворяют (1.1), то каждый элемент новой матрицы будет иметь вид:
- $$\delta_{ij}^* = \sum_k x_{ik}\, x_{jk} \quad (1.5)$$
- Формула (1.5) - это основная теорема, на которой построен алгоритм Торгерсона. Матрица Δ^* часто называется *матрицей скалярных произведений.* Из формулы (1.5) видно, что каждый из ее элементов — сумма произведений скаляров *xik* и *xjk.* Уравнение (1.5) можно записать в матричном виде:
- $$\Delta^* = X\, X', \quad (1.6)$$
- где X -(I * K)-матрица координат стимулов. Найти матрицу X, удовлетворяющую (1.6), можно (если она существует) с помощью программы факторного анализа методом главных компонент.

Поворот

- Матрица X, построенная с помощью метода главных компонент, является одним из решений уравнения (1.6). Чтобы понять, почему это решение не единственно, представьте себе матрицу ортогонального преобразования T размером *(K*K).* Если X удовлетворяет (1.6), то любая матрица X* = XT тоже удовлетворяет (1.6), т. е. если
- $$\Delta^{*} = X\,X^{'}, \quad (1.7)$$
- то
- $$\Delta^{*} = X^{*}\,X^{*'}. \quad (1.8)$$
- Так как T — ортогональная матрица, TT' = I. Отсюда
- $$X^{*}\,X^{*'} = (X\,T)(X\,T)^{'} \quad (1.9)$$
- (XT)' в (1.1) равно T'X'. Подставляя этот результат в (1.9), получим
- $$X^{*}\,X^{*'} = (X\,T)(T^{'}\,X^{'}) = X\,(T\,T^{'})\,X^{'} = X\,I\,X^{'} = X\,X^{'} = \Delta^{*}. \quad (1.10)$$

- (1.10) => если X — решение (1.6), то и x* —тоже решение (1.6).
- Если размерность *K* не превышает двух, то в приложениях типа «сжатие данных» и «верификация конфигурации» этот вопрос является спорным.
- При таком небольшом числе координатных осей важные особенности конфигурации будут видны просто при ее рассмотрении, независимо от поворота.
- Однако в координатных приложениях этот вопрос бесспорный. Если координатные оси не повернуты соответствующим образом, то координаты не будут совпадать с существенными характеристиками стимулов, и интерпретировать координатные оси будет трудно.