

Факторный анализ

Факторный анализ: понятие и назначение

В факторном анализе предполагается, что наблюдаемые переменные являются линейной комбинацией некоторых латентных факторов. Некоторые из этих факторов допускаются общими для нескольких исходных переменных, а другие характерны для каждой переменной в отдельности. Характерные факторы ортогональны друг другу и, следовательно, не вносят вклад в ковариационную матрицу переменных. Таким образом, только общие факторы, число которых предполагается существенно меньшим, чем число исходных переменных, вносят вклад в ковариацию между ними

Принимаемая в факторном анализе линейная система такова, что структура ковариаций может быть идентифицирована без ошибок, если известна матрица нагрузок латентных факторов. При использовании постулатов и свойств данной линейной системы можно точно идентифицировать латентную факторную структуру на основе анализа исходной ковариационной матрицы.

На практике на исследуемую матрицу корреляций оказывают влияние различные случайные и неслучайные ошибки. В результате она будет отлична от корреляционной матрицы, обусловленной факторной структурой генеральной совокупности.

Таким образом, на практике невозможно получить точную структуру факторной модели, можно найти оценки параметров факторной структуры с использованием определенных критериев.

Факторный анализ – это процедура, с помощью которой большое число переменных, относящихся к имеющимся наблюдениям, сводят к меньшему количеству независимых влияющих величин, называемых факторами: - в один фактор объединяются переменные, сильно коррелирующие между собой. - переменные из разных факторов слабо коррелируют между собой

Факторы (Factors) – латентные (скрытые) переменные, конструируемые таким образом, чтобы можно было объяснить корреляцию между набором имеющихся переменных.

Цель факторного анализа – сокращение числа переменных на основе их классификации и определения структуры взаимосвязей между ними.

Благодаря сокращению числа переменных вместо исходного набора переменных появляется возможность анализировать данные по выделенным факторам, число которых значительно меньше исходного числа вза-

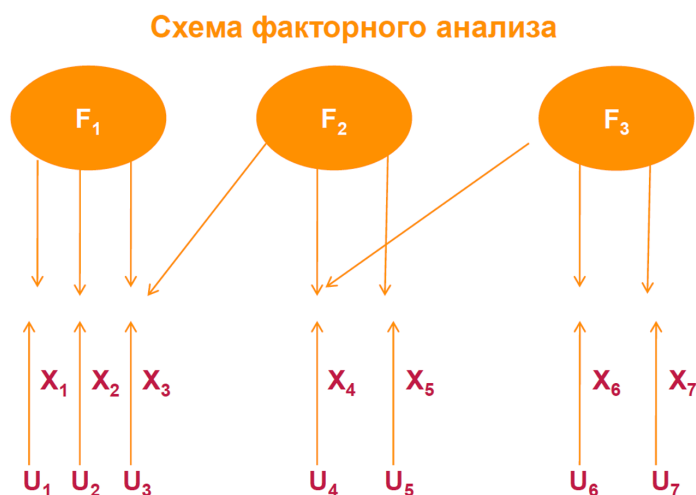


Рис. 1: Общая схема факторного анализа

имосвязанных переменных. При этом измеряемые эмпирические переменные (предикторы) считаются следствием других, глубинных, скрытых от непосредственного измерения характеристик (латентных переменных)

F –общие факторы, каждый из которых влияет на определенную совокупность переменных;

X – исходные переменные, фиксируемые на основании измерений;

U –уникальные (характерные) факторы, каждый из которых влияет только на одну переменную.

Требования (ограничения) метода факторного анализа.

- Тип шкалы исходных показателей не ниже интервального. - Все показатели должны быть количественными переменными (интервальными либо метрическими).
- Выборка должна быть однородна. - Форма связи между наблюдаемыми признаками – линейная. - Форма связи между показателями и факторами – линейная. - Требуется вращения для оптимизации интерпретации. - Форма двумерного распределения любой пары наблюдаемых признаков – нормальная. - Число наблюдений должно быть минимум в два раза больше числа переменных. - Наблюдения независимы друг от друга. - Исходные переменные должны быть распределены симметрично.

Основные этапы факторного анализа

Порядок выполнения факторного анализа

1. Подготовка корреляционной матрицы для всех переменных, участвующих в анализе. 2. Извлечение факторов. 3. Выбор факторов и вращение факторов для создания упрощенной структуры. 4. Интерпретация факторов.

Первый этап достаточно очевиден. Рассмотрим второй этап.

Извлечение факторов.

Извлечение фактора начинается с подсчета суммарного разброса значений всех участвующих в анализе переменных. Для этого «суммарного разброса» непросто подобрать логическую интерпретацию, однако он является вполне строго определенной математической величиной.

1. Извлечение факторов заключается в выборе взаимодействующих переменных, чья взаимная корреляция обуславливает наибольшую долю общей дисперсии. Эти переменные образуют первый фактор. 2. Затем первый фактор исключается и из оставшегося множества переменных снова выбираются те, чье взаимодействие определяет наибольшую долю оставшейся общей дисперсии. Эти переменные образуют второй фактор. 3. Процедура извлечения факторов продолжается до тех пор, пока не будет исчерпана вся общая дисперсия переменных. 4. После того как процедура извлекает каждый фактор, при этом определяется соответствующее собственное значение, которое пропорционально доле общей дисперсии, определяемой данным фактором.

Извлечение факторов может производиться несколькими методами. Методы факторного анализа различаются в зависимости от подходов для нахождения коэффициентов значения факторов.

Методы факторного анализа

Метод главных факторов

Наиболее ранним методом факторного анализа является метод главных факторов, здесь методика РСА используется применительно к редуцированной корреляционной матрице, на главной диагонали которой располагаются общности. Обычно для их оценки пользуются квадратом множественного коэффициента корреляции между соответствующей переменной и всеми остальными переменными. После размещения на главной диагонали корреляционной матрицы оценок общностей для выделения факторов используют те же способы что и при РСА. То есть составляется и решается соответствующее характеристическое уравнение.

Метод наименьших квадратов.

МНК в факторном анализе сводится к минимизации остаточной корреляции после выделения определенного числа факторов и оценивания степени соответствия вычисленных и наблюдаемых коэффициентов корреляции (расчет суммы квадратов отклонений). Расхождение уменьшается при увеличении предполагаемых факторов.

Алгоритм

1. Предполагается число факторов k (можно задать $k = 1$) 2. Производится оценка общностей (оценка квадрат коэффициента множественной корреляции между данной переменной и остальными) 3. Выделяется k факторов, для которых вычисленные коэффициенты корреляции наилучшим образом приближают наблюдаемые корреляции. Здесь решается соответствующее характеристическое уравнение и находятся соответствующие собственные числа. 4. Снова производится оценка общностей, но для матрицы факторного отображения, полученной на предыдущем шаге.

Процесс продолжается до тех пор пока дальнейшее улучшение станет невозможным. Данный алгоритм называется "Метод главных факторов с итерациями по общностям". Кроме того используется "метод минимальных остатков". Он аналогичен выше рассмотренному алгоритму, но при его реализации для проверки окончания алгоритма используется критерий хи-квадрат.

Метод максимального правдоподобия.

Метод максимального правдоподобия имеет своей задачей найти факторное решение, которое наилучшим образом объясняет наблюдаемые корреляции. Предполагается, что наблюдаемые данные — это выборка из генеральной совокупности, которая точно соответствует k факторной модели. Совместное распределение переменных (включая факторы) предполагается многомерным нормальным. Задача сводится к оцениванию значений латентных переменных (нагрузок) генеральной совокупности, при которых функция правдоподобия для распределения элементов корреляционной матрицы максимальна.

Можно использовать иной критерий при котором осуществляется поиск факторных нагрузок, обеспечивающих максимизацию коэффициента корреляции между общими факторами и наблюдаемыми переменными.

Третий критерий сводится к определению нагрузок, при которых детерминант матрицы остаточных корреляций максимален. Для нахождения решения используется для всех критериев соответствующие итераци-

онные схемы. В принципе все варианты метода максимального правдоподобия сводятся к решению характеристического уравнения вида:

$$\det(\mathbf{R}_2 - \lambda \mathbf{I}) = 0 \quad (1)$$

Где $\mathbf{R}_2$ определяется соотношением:

$$\mathbf{R}_2 = \mathbf{U}^{-1}(\mathbf{R} - \mathbf{U}^2)\mathbf{U}^{-1} = \quad (2)$$

$$=\mathbf{U}^{-1}\mathbf{R}_1\mathbf{U}^{-1} \quad (3)$$

Здесь $\mathbf{U}^2$ - оценка дисперсии характерных параметров, $\mathbf{R}_2$ - редуцированная корреляционная матрица

В отличии от МНК в вычисляемую на каждом шаге оценку общности с большим весом входят корреляции с переменными, имеющими меньшую характерность. В методе максимального правдоподобия характерность играет роль "квази-ошибки": больший вес имеют переменные с максимальной общностью. Те менее точные наблюдения учитываются с меньшим весом.

Формула для вычисления статистики χ^2 показывает, что ее значение определяется объемом выборки, а число степеней свободы от нее не зависит:

$$\mathbf{U}_k = N\{\ln |\mathbf{C}| - \ln |\mathbf{R}| + tr(\mathbf{R}\mathbf{C}^{-1})\} - n \quad (4)$$

где $\ln$ - натуральный логарифм; tr - след матрицы; N - объем выборки; n - число переменных; R - матрица ковариаций; $\mathbf{C} = \mathbf{F}\mathbf{F}^T + \mathbf{U}^2$; $\mathbf{F}$ - матрица факторных нагрузок; $\mathbf{U}^2$ - характерности.

При фиксированной корреляционной матрице, величина $\mathbf{U}_k$ пропорциональна объему выборки N . Соответствующее число степеней свободы равно:

$$df_k = 1/2[(n - k)^2 - (n + k)], \quad (5)$$

где k - число гипотетических факторов; n - число переменных.

Существенное преимущество метода максимального правдоподобия состоит в том, что для большой выборки он позволяет получить критерий значимости. Если критерий χ^2 показывает значимое отклонение наблюдений от k факторной модели, то в рассмотрение вводится модель $k + 1$ факторами. В разведочном анализе обычно начинают с одного

фактора, а заканчивают, когда отклонение наблюдений от модели становится статистически незначимым.

Альфа-факторный анализ.

В альфа-факторном анализе используемые переменные считаются выборкой из некоторой совокупности переменных, о которой можно судить наблюдаемой совокупности объектов. Таким образом, в альфа-факторном анализе выводы носят психометрический, а не статистический характер (альфа-факторный метод был разработан для обработки данных в психологии).

Утверждается, что метод основан на выделении таких факторов, которые имеют максимальные корреляции с соответствующими факторами генеральной совокупности переменных. Характерные факторы при этом, рассматриваются как ошибки, обусловленные психометрической выборкой переменных. Следовательно, оценки общностей имеют смысл "надежностей".

На первом шаге образуется подправленная корреляционная матрица вида:

$$\mathbf{R}_3 = \mathbf{H}^{-1}(\mathbf{R} - \mathbf{V}^2)\mathbf{H}^{-1} \quad (6)$$

где $\mathbf{V}^2$ и $\mathbf{H}^2$ - диагональные матрицы характеристик и общностей соответственно, а $\mathbf{H}^{-1}$ - диагональная матрица, элементами которой являются обратные величины к квадратным корням из общностей.

Тогда характеристическое уравнение, связанное с подправленной матрицей, представляется следующим образом:

$$\det(\mathbf{R}_3 - \lambda \mathbf{I}) = 0 \quad (7)$$

В методе максимального правдоподобия матрица инормируется с помощью характеристик, а в альфа-факторном анализе - дисперсией общностей. То есть в первом случае больший вес имеют переменные с большей общностью, а во втором наоборот, с меньшей.

В альфа-факторном анализе число выделяемых факторов определяется с помощью величины собственных значений, которая должна быть больше единицы. Это эквивалентно использованию коэффициента обобщенности α (квадрат коэффициента корреляции данного фактора с соответствующими факторами, взятыми из генеральной совокупности). Выделяются факторы, для которых коэффициент α положителен.

Анализ образов.

В анализе образов определение общей и характерной части переменной отличается от принятого в классическом факторном анализе. Под общей частью переменной подразумевается та ее составляющая, которая выражается через линейную комбинацию других переменных. Эта доля переменной называется "образ-переменной". Вторая часть, независимая от остальных переменных, называется "антиобразом".

В анализе образов предполагается, что потенциальное множество переменных бесконечно. При этом средний квадрат образа равен общности переменной, определяемой в факторном анализе, а средний квадрат "антиобраза" равен общности данной переменной.

Образы и антиобразы, определяемые для некоторого набора наблюдаемых переменных, называются соответственно частными образами и частными антиобразами. Они полностью задаются наблюдаемыми переменными. В этом смысле анализ образов в корне отличается от классического факторного анализа, в котором общая часть переменной является линейной комбинацией гипотетических факторов и не может быть явной функцией наблюдаемых переменных.

Методика анализа образов предполагает введение матрицы ковариации частных образов:

$$\mathbf{R}_4 = (\mathbf{R} - \mathbf{S}^2)\mathbf{R}^{-1}(\mathbf{R} - \mathbf{S}^2) \quad (8)$$

где $\mathbf{R}$, а $\mathbf{S}^2$ -диагональная матрица элементами которой являются доли дисперсий антиобразов. Получение матрицы (8) сводится, во-первых, к замене диагональных элементов матрицы $\mathbf{R}$ на квадраты множественного корреляции каждой переменной с совокупностью всех остальных переменных, и, во-вторых к преобразованию недиагональных для получения матрицы Грама (если определитель Грама равен нулю, то соответствующие вектора линейно зависимы). Характеристическое уравнение для этой матрицы имеет вид:

$$\det(\mathbf{R}_4 - \lambda \mathbf{I}) = 0 \quad (9)$$

Число выделяемых факторов определяется количеством собственных чисел, больших единицы для матрицы $\mathbf{S}^{-1}\mathbf{R}\mathbf{S}^{-1}$. Обычно число выделяемых таким образом факторов велико, приблизительно половина числа

исходных параметров. При вращении можно обрасывать незначительные и неинтерпретируемые факторы.

Методы вращения.

На первом этапе анализа определяется минимальное число факторов, адекватно воспроизводящих наблюдаемые корреляции, а также значения общностей каждой переменной. Следующий шаг состоит в нахождении легко интерпретируемых факторов с помощью процедуры вращения. При этом число факторов и значения общностей переменных фиксируются.

Существуют три различных подхода к проблеме вращения:

1. Графический подход при котором вращение заключается в проведении новых осей, которые соответствуют некоторому критерию простоты, легко интерпретируемой структуры. Если в пространстве факторов есть явные кластеры точек легко отделяемые друг от друга, простая структура получается в том случае, когда оси проведены через эти скопления.

2. Подход, связанный с аналитическими методами. Здесь выбирается некоторый объективный критерий, которым надо руководствоваться при вращении. В рамках этого подхода различают два вида вращения- ортогональное и косоугольное.

3. Подход на основе априорного задания целевой матрицы. Цель вращения найти такое факторное отображение, которое максимально близко к некоторой заданной матрице. Целевая матрица задается исходя из определенных предположений о факторной структуре.

Геометрический метод вращения.

Геометрический метод практически не применим, когда скопления точек неразделимы или когда число факторов больше двух. Целью вращения является получение простой факторной структуры, но не существует единых формальных критериев для оценки простоты факторной структуры.

Попытаемся определить минимальные требования к простой структуре. Если взять число факторов r и число переменных n , то факторная структура является наипростейшей, когда каждая переменная имеет ненулевую только на один общий фактор. Если число факторов два и больше, то это означает, что в наиболее простой матрице факторной структуры будут выполняться следующие условия:

- каждая строка будет содержать только один ненулевой элемент;
- каждый столбец будет иметь несколько нулей;

- для каждой пары столбцов нулевые элементы не совпадают

Для реальных данных такая простая структура недостижима. Следовательно, задача состоит в том чтобы определить факторную структуру, которая наиболее близка к простой структуре. Эта задача решается путем вращения первичных осей и перехода к вторичным осям для получения матрицы факторного отображения, наиболее близкой к простейшей идеальной структуре.

Методы ортогонального вращения: Квартимакс, Варимакс, Эквимакс.

Рассмотренный выше способ оценки степени сложности факторной структуры для численного использования неудобен. Одной из возможных мер сложности модели является вариация квадрата факторной нагрузки для каждой строки (для каждой переменной). Следовательно, дисперсия квадратов факторных нагрузок переменной есть мера факторной сложности этой переменной. После некоторых преобразований выражение для этой меры можно представить следующим образом:

$$q_i = \frac{\sum_{j=1}^r (b_{ij}^4) - (\sum_{j=1}^r b_{ij}^2)^2}{r^2} \quad (10)$$

где r - число столбцов факторной матрицы; b_{ij} - факторная нагрузка j -го фактора на i -ю переменную.

Число факторов r и общности каждой переменной считаются известными в результате решения задачи выделения первоначальных факторов. Поэтому слагаемое, входящее в (10) с отрицательным знаком, является константой ибо $\sum_{j=1}^r b_{ij}^2 = h_i^2$ в случае ортогонального решения. Общей мерой сложности может служить сумма q_i всех переменных:

$$q = \sum_{i=1}^n q_i = \sum_{i=1}^n \frac{\sum_{j=1}^r (b_{ij}^4) - (\sum_{j=1}^r b_{ij}^2)^2}{r^2} \quad (11)$$

Использование критерия квартимакс основано на вращении осей таким образом, чтобы результирующие нагрузки максимизировали q . При этом максимизация q эквивалентна максимизации следующего выражения:

$$Q = \sum_{i=1}^n \sum_{j=1}^r b_{ij} \quad (12)$$

Так как слагаемое со знаком минус в (11) является константой, то критерий называется кватримакс. Метод кватримакс имеет тенденцию к выделению генерального фактора.

Метод варимакс использует несколько другой критерий, в котором добиваются упрощения описания столбцов факторной матрицы. Вместо дисперсии квадратов нагрузок переменной рассматривается дисперсия квадратов нагрузок фактора. Индекс сложности v_j фактора j равен:

$$v_j = \frac{n \sum_{i=1}^n (b_{ij}^4) - (\sum_{i=1}^n b_{ij}^2)^2}{n^2} \quad (13)$$

При этом $\sum_{i=1}^n b_{ij}^2$ не является константой. Общая мера простоты задается критерием варимакс:

$$V = \sum_{i=1}^n v_i = \frac{\sum_{j=1}^r n \sum_{i=1}^n (b_{ij}^4) - \sum_{j=1}^r (\sum_{i=1}^n b_{ij}^2)^2}{n^2} \quad (14)$$

Часто факторные нагрузки нормируют, чтобы избавиться от нежелательного влияния на результаты вращения переменных с большой общностью. Варимакс дает лучшее разделение факторов чем кватримакс. Учитывая что критерий кватримакс основан на упрощении строк, а критерий варимакс - на упрощении столбцов предлагается обобщенный критерий, который имеет вид:

$$\alpha Q + \beta V = Maximum \quad (15)$$

Где α и β - веса.

Этот критерий можно записать в следующем виде:

$$\sum_{j=1}^r \sum_{i=1}^n b_{ij}^4 - \gamma \sum_{j=1}^r (\sum_{i=1}^n b_{ij}^2)^2 / n = Maximum \quad (16)$$

Где $\gamma = \beta / (\alpha + \beta)$

Если $\gamma = 0$, то имеем критерий кватримакс, а если $\gamma = 1$ то -варимакс. При $\gamma = r/2$ и $\gamma = 0.5$ получаются особые критерии, эквимакс и бикватримакс соответственно.

Методы косоугольного вращения.

Косоугольное вращение является более общим, чем ортогональное, так как здесь нет ограничений, связанных с некоррелированностью факторов. Существует два подхода к косоугольному вращению: использование вторичных осей; использование первичной матрицы факторного отображения.

Методы, основанные на введении вторичных осей.

Здесь вводится критерий, называемый кватримин:

$$N = \sum_{i=1}^n \sum_{j < k=1}^r a_i a_{ik} \quad (17)$$

Где a_i a_{ik} - проекции i -го параметра на j -ю и k -ю вторичные оси. Величина N будет нулевой, если все параметры имеют нагрузку только на один фактор. Цель вращения-нахождение таких факторных нагрузок, которые минимизируют N . Вводится критерий коваримин-минимизация ковариации проекций на вторичные оси:

$$C = \sum_{j < k=1}^r \left(n \sum_{i=1}^n a_i^2 a_{ik}^2 - \sum_{i=1}^n a_{ij}^2 \sum_{i=1}^n a_{ik}^2 \right) \quad (18)$$

Модификация этого критерия основана на нормировании- замене a_{ij}^2 на a_{ij}^2/h_i^2 . Применительно к одним и тем же данным критерий коваримин дает меньше косоугольных факторов, чем кватримин. Объединение этих двух критериев приводит к обобщенному критерию облимин:

$$B = \sum_{j < k=1}^r \left(n \sum_{i=1}^n a_i^2 a_{ik}^2 - \gamma \sum_{i=1}^n a_{ij}^2 \sum_{i=1}^n a_{ik}^2 \right) \quad (19)$$

Где $\gamma = \beta/(\alpha + \beta)$; α , β - веса соответственно N, C .

Этот общий критерий при $\gamma = 0$ переходит к кватримин, при $\gamma = 0.5$ - бикватримин, а при $\gamma = 1$ - в коваримин. Чаще всего в этом критерии облимин используются нормированные проекции.

Другой критерий, тесно связанный с принципами облимина называется критерием бинормамин. В нем заложена идея объективного выбора значения γ в соотношении (19). По сравнению с критерием бикватримин,

где $\gamma = 0.5$, бинормамин дает лучшие результаты для особо простых и особо сложных данных.

Прямой метод облимин.

Метод основан на упрощении матрицы нагрузок первичных факторов (без использования вторичных осей). Минимизируемая имеет следующий вид:

$$D = \sum_{j < k=1}^r [\sum_{i=1}^n b_i^2 b_{ik}^2 - d(\sum_{i=1}^n b_{ij}^2 \sum_{i=1}^n b_{ik}^2)/n] \quad (20)$$

где b_{ij} -элементы матрицы нагрузок первичных факторов. Величина d определяется аналогично γ при этом параметр α регулирует степень косоугольности получаемого решения. В случае однофакторной модели следует положить $\alpha = 0$. Существует много других методов косоугольного вращения, например методы ортоблик и метод максплейн.

Вращение с использованием целевой матрицы.

Это подход, основанный на априорной информации о факторной структуре:

1. Можно задать значения нагрузок для каждой переменной, а затем проводить вращения с целью обеспечения минимального отличия полученной факторной матрицы от заданной матрицы.

2. В качестве целевой матрицы можно использовать некоторые функции от ортогонального решения. Такой метод известен как промакс- метод косоугольного вращения. В рамках данного подхода существуют несколько алгоритмов.

3. Можно задавать целевую матрицу, состоящую из нулей и единиц. Этот подход соответствует априорной информации когда известно только от, что некоторые нагрузки должны быть велики, а другие малы. Можно задавать целевую матрицу и в более общем виде если априорная информация более полная.

Следует отметить, что при определении числа факторов может использоваться несколько критериев:

1. критерий значимости, связанный с методами максимального правдоподобия и наименьших квадратов;
2. критерии, основанные на правилах формулируемых в терминах собственных чисел;
3. критерий, основанный на величине долей дисперсии факторов;
4. критерий отсеивания;

5. критерий интерпретируемости и инвариантности.

В данной работе эти критерии не рассматриваются, желающие могут познакомиться с ними в соответствующей литературе.

Существует целый раздел факторного анализа, связанный с оценкой степени эмпирического подтверждения и адекватности факторной модели. Это направление основано на проверке определенных статистических гипотез, оно получило название конфирматорный факторный анализ. Данный раздел факторной теории нами также здесь не рассматривается.

Факторное шкалирование.

Результаты факторного анализа позволяют перейти к оценке факторных шкал, что позволяет:

- представить изучаемые объекты в терминах значений факторов, а не измеряемых переменных;
- появляется возможность использовать один или два фактора для дальнейшего в том числе визуального анализа объектов (задача сокращения описания).

Рассмотрим некоторые важные аспекты факторного шкалирования. Даже если факторная модель безошибочна для генеральной совокупности, она не будет абсолютно точно воспроизводить корреляции в выборочных данных. Поэтому, необходимо ввести критерий близости оценок и истинных значений факторов. Существует три таких критерия.

Регрессионный анализ.

Для однофакторной модели первый критерий сводится к нахождению оценки (F^*) значения фактора (F), доставляющей максимум коэффициента корреляции между F^* и F , что сводится к минимизации суммы квадратов отклонений $\sum (F - F^*)^2$ (применение регрессионного анализа). Так как факторный анализ дает значения факторных нагрузок, которые представляют собой корреляции между факторами и наблюдаемыми переменными (предикторами). При этом корреляции между предикторами являются наблюдаемыми корреляциями. Таким образом, оценка фактора может определяться из решения системы нормальных уравнений:

$$F^* = \mathbf{X}(\mathbf{B}^T \mathbf{R}^{-1}) \quad (21)$$

Где $\mathbf{B}$ - матрица факторных нагрузок; $\mathbf{X}$ - вектор наблюдаемых переменных; $\mathbf{R}$ корреляционная матрица наблюдаемых переменных. Про-

извлечение в скобках в выражении (21) задает весовые коэффициенты перехода к оценке фактора F^* от значений наблюдаемых переменных. Надежность оценки определяется из сравнения с абсолютно точной факторной моделью для генеральной совокупности.

Критерий наименьших квадратов.

В однофакторной модели каждая переменная считается взвешанной суммой общих и характерных факторов:

$$X_j = b_j F + d_j u_j \quad (22)$$

Предположим, что вместо F взята оценка F^* . Поскольку критерий наименьших квадратов определяется оценкой, минимизирующей сумму квадратов:

$$\sum_i \sum_j (X_{ij} - b_j F^*)^2 \quad (23)$$

то получаем следующую оценку:

$$F^* = \mathbf{X}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B} \quad (24)$$

Отличие (24) от (21) состоит в том что здесь используются воспроизведенные в модели корреляции $\mathbf{B}\mathbf{B}^T$ вместо R .

Таким образом регрессионный анализ и критерий наименьших квадратов приводит к одни и тем же оценкам, когда выборочные корреляции совпадают с корреляциями для генеральной совокупности.

Критерий Бартлетта.

Для данного подхода включается в рассмотрение выборочная изменчивость. Если характерную долю дисперсии отнести на счет ошибок наблюдения, то лучше уменьшить вес тех переменных, которые имеют большие дисперсии ошибок. Водится следующий критерий:

$$\sum_i \sum_j (X_{ij} - b_j F^*)^2 / d_j^2 \quad (25)$$

В результате параметры с меньшими общностями получают и меньший вес. Поэтому, оценка факторной шкалы, получаемая с помощью критерия Бартлетта имеет вид:

$$F^* = \mathbf{X}\mathbf{U}^{-2}\mathbf{B}(\mathbf{B}^T\mathbf{U}^{-2}\mathbf{B})^{-1} \quad (26)$$

Где U^{-2} -диагональная матрица характеристик. Наличие U^{-2} может рассматриваться как результат взвешивания.

Несколько общих факторов.

Три рассматриваемых критерия можно обобщить для много факторного случая как для ортогонального, так и для косоугольного решения. Все результаты, для оценок фактора, справедливы и для нескольких факторов. Но при этом возникают следующие вопросы:

- будут ли факторные шкалы ортогональны друг другу, если скрытые факторы ортогональны;
- будет ли каждая шкала коррелировать только с соответствующим ей фактором (факторная шкала называется монохроматичной, если ее частные коэффициенты корреляции с другими факторами нулевые)

В общем случае всем этим требованиям не удовлетворяет ни одна из оценок. Однако в частном случае перечисленные требования выполняются когда:

- факторная модель точно соответствует экспериментальным данным и отсутствует выборочная изменчивость и ошибки измерений;
- каждая переменная имеет нагрузку только на один фактор.

При этом задача сводится к однофакторной модели без ошибок и все три оценки, рассмотренные выше, эквивалентны. Но такая идеализированная ситуация практически неосуществима. Но для некоторых факторных шкал есть дополнительные условия ортогональности и монохроматичности. Так например, если первоначальные факторы (до вращения) были выделены с использованием критерия максимального правдоподобия, то регрессионная оценка и оценка Бартлетта для факторных шкал будут ортогональны и монохроматичны. Правда после вращения для регрессионной оценки уже ни одно из этих условий не выполняется, а для оценки Бартлетта остается справедливым только условие монохроматичности (ортогональность отсутствует).

Связи с этим появился четвертый критерий (Андерсена-Рубина), который является модификацией критерия Бартлетта. Минимизируется взвешанная сумма квадратов, используемая в критерии Бартлетта, при условии, что получаемые шкалы ортогональны друг другу. Следовательно, независимо от вращения критерий дает некоррелированные шкалы. Но монохроматичность при этом критерии не обеспечивается.

Выбор метода шкалирования.

При выборе метода необходимо проанализировать свойства получаемых шкал. Если рассматривать корреляции между скрытыми факто-

рами и их шкалами, то регрессионный метод предпочтительнее метода Бартлетта, а метод Бартлетта предпочтительнее метода наименьших квадратов.

С точки зрения требования монохроматичности оценка шкал по критерию Бартлетта является наилучшей, а если брать свойство ортогональности, то предпочтительнее критерий Андерсена-Рубина. Так как заранее не известно, ортогональны ли скрытые факторы, выбирать следует либо регрессионный анализ, либо метод Бартлетта. Но все это имеет лишь академический интерес и для оценивания шкал хорош любой способ. Выбор метода факторного шкалирования зависит от специфики решаемой задачи. Если факторное шкалирование используется совместно скакими то новыми внешними переменными, то некоторые методы хуже приспособлены для такой задачи (в частности регрессионный анализ). Если задача состоит только в использовании факторных шкал как новых предикторов для внешних выходных переменных, то регрессионный анализ является наилучшим. Но всегда важна эвристическая качественная оценка получаемых результатов.

Рассмотрев основные аспекты достаточно глубокой теории факторного анализа, приведем пример его использования для решения задачи сокращения размерности средствами *python* и *R*.

Пример использования методов факторного анализа для сокращения числа предикторов средствами *python* и *R*.

В качестве исходных данных используем информацию об основных социально-экономических показателях российских регионов за 2018 год. Исходная матрица предикторов имеет размерность 81x16 в качестве выходной переменной используется номинальная переменная тип экономики региона. В реализации на Python переменной "тип экономики": А-аграрный; D-добывающий, Р- промышленный; а при реализации на R добавляется значение: Т-торговый.

Код на *Python*

Здесь из исходной таблицы выбираются три столбца Р10,Р11,Р14 , а выходная переменная Р18 принимает три значения А,Д,Р

```
““python
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
import scipy.stats as st
```

```

import matplotlib.pyplot as plt
from pandas import Series, DataFrame
import pandas as pd
#from sklearn import decomposition
from sklearn.decomposition import FactorAnalysis

result =pd.read_table('dan04.txt',sep='\s+')
print(result)

y=result['P18']

X =DataFrame(result, columns=['P10','P11','P14'])
y=np.array(y)
X=np.array (X)

SUMX=X[:,0] +X[:,1] + X[:,2]
X[:,0] =X[:,0] /SUMX
X[:,1] =X[:,1] /SUMX
X[:,2] =X[:,2] /SUMX

#описательные статистики

print('Описательные статистики предикторов')

for num in range(3) :
    print('Номер переменной: %.0f' %num)
    print('число элентов массива =',len(X[:,num]))
    print('среднее значение =',np.mean(X[:,num]))
    print('стандартное отклонение =',np.std(X[:,num]))
    print('мин макс =',np.min(X[:,num]), np.max(X[:,num]))

#построение исходного графика 3 D

fig = plt.figure(1)
ax = fig.add_subplot(111, projection='3d')
xs = X[:,0]
ys = X[:,1]
zs =X[:,2]

```

```

ax.scatter(xs, ys, zs, c=y, cmap=plt.cm.nipy_spectral,
edgecolor='k')
ax.set_title("Исходное пространство предикторов")
ax.set_xlabel('P10')
ax.set_ylabel('P11')
ax.set_zlabel('P14')

# Преобразование данных , уменьшающее размерность до 2

transformer = FactorAnalysis(n_components=2,
rotation='varimax', random_state=0)
X_transformed = transformer.fit_transform(X)
DimX=X_transformed.shape
print(DimX)
y = np.choose(y, [1, 2, 0]).astype(np.float)
fig = plt.figure(2)
plt.clf()
plt.cla()
plt.scatter(X_transformed[:, 0], X_transformed[:, 1],
c=y, cmap=plt.cm.nipy_spectral, edgecolor='k')
plt.title("Снижение размерности методом FactorAnalysis")
plt.xlabel("F1")
plt.ylabel("F2")
plt.show()

'''

```

В модуле python "FactorAnalysis" для определения матрицы нагрузок используется оценка максимального правдоподобия с последующем вращением "варимакс". Для факторного шкалирования применяется подход, основанный на SVD.

Исходное пространство предикторов размерности 3D имеет следующий вид:

Трансформированное пространство размерности 2D общих факторов F_1, F_2 имеет следующий вид:

Код на R

Здесь из исходной таблицы выбираются четыре столбца (предиктора)

Исходное пространство предикторов

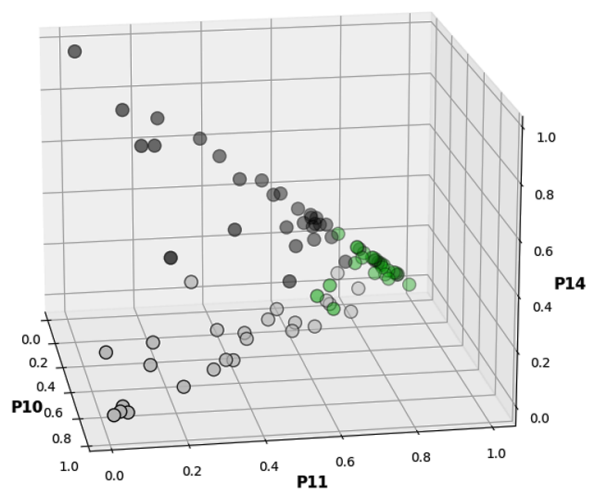


Рис. 2: ДАННЫЕ В ТРЕХМЕРНОМ ПРОСТРАНСТВЕ

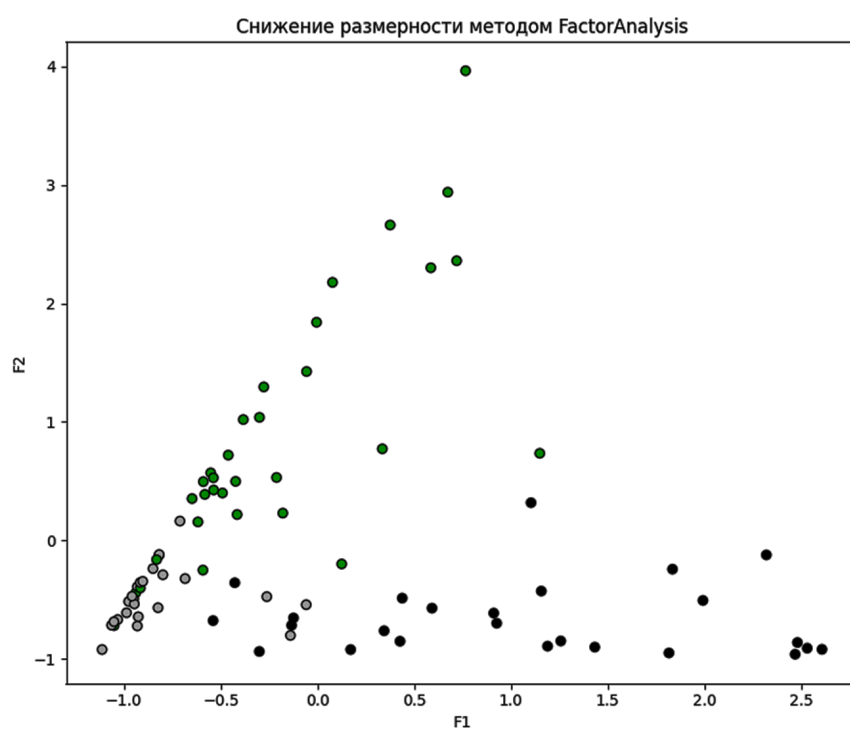


Рис. 3: ДАННЫЕ В ТРАНСФОРМИРОВАННОМ ПРОСТРАНСТВЕ

P10,P11,P14,P16 , а выходная переменная P18 принимает четыре значения A,D,P,T

```
““r
library(gmodels)
library(lattice)
library(vegan)
library(grDevices)
library(reshape2)
library(ggplot2)
library(ggthemes)
library(psych)

load("xRegion.RData")

#ФОРМИРОВАНИЕ data.frame ОСНОВНЫЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ
# ПОКАЗАТЕЛИ РЕГИОНОВ РФ
xreg <-data.frame(P1=x$Pok_001,P2=x$Pok_002,
P3=x$Pok_003,P4=x$Pok_004,
P5=x$Pok_005,P6=x$Pok_006,P7=x$Pok_007,
P8=x$Pok_008,P9=x$Pok_009,P10=x$Pok_010,
P11=x$Pok_011,P12=x$Pok_012,P13=x$Pok_013,P
14=x$Pok_014,P15=x$Pok_015,P16=x$Pok_016,
P17=x$Pok_017,P18=x$Pok_018)
#xreg
#Формирование фактора тип экономики региона
xreg_f <- as.factor(x$Pok_018)

#выбор исследуемого подмножества показателей регионов
w=c(10,11,14,16)
reg.ekon <-xreg[,w]
Y <-reg.ekon
X=Y
X <- as.matrix(X)

XX=cor(X)
paf2=fa(X,nfactors=2,rotate="varimax",SMC=T,symmetric=T, fm="pa")
```

```

paf2$e.values # собственные значения исходной матрицы
paf2$communality # Оценки общности для каждого элемента.
paf2$rotation # какая ротация была запрошена?
paf2$complexity # Индекс сложности Хоффмана для каждого элемента.
paf2$fit # показывает Насколько хорошо факторная модель
#воспроизводит корреляционную матрицу
paf2$R2 #Множественный квадрат R между факторами
# и оценками факторных баллов
paf2$weights #веса, чтобы найти оценки факторов
paf2$valid #Оценка достоверности факторов
print(paf2,sort=TRUE)
f3l <- paf2$loadings
f3l
x11()
fa.diagram(f3l,main="input from a matrix")

P1=vector(mode = "numeric", 81)
P2=vector(mode = "numeric", 81)
for(i in 1:81)
{P1[i]=X[i,]%*%paf2$weights[,1];
P2[i]=X[i,]%*%paf2$weights[,2]}

#нормирующая функция-----
norm <-function(x){
return ((x-min(x))/(max(x)-min(x)))
}

P1=norm (P1)
P2=norm (P2)

PA <-data.frame(PA1=P1,PA2=P2,Class=xreg_f)

#График регионов по классам в пространстве 2 компонент FA
x11()
g3 <-ggplot(data= PA, aes(x=PA1,y=PA2,colour = Class),colors = "Accent")
g3 <-g3 +geom_point(size=4)
g3 <-g3 +labs(title ="Регионы в пространсте двух компонент FA",

```

```

x="Фактор 1", y="Фактор 2",colour = "Классы")
#g3 <-g3 + theme_fivethirtyeight() + scale_color_fivethirtyeight()
g3 <-g3 + theme_stata() + scale_colour_stata()
g3

'''

```

Получены следующие результаты:

Собственные значения исходной матрицы

```
2.20023426 1.00324134 0.74231453 0.05420988
```

Оценки общности для каждого элемента

	P10	P11	P14	P16
	1.0.1982701	0.9736012	0.1095078	0.9579688

Какая ротация была запрошена?

```
"varimax"
```

Индекс сложности Хорфмана для каждого элемента

	P10	P11	P14	P16
	2.1.008731	1.564938	1.008690	1.978491

Насколько хорошо факторная модель воспроизводит корреляционную матрицу

```
0.7753318
```

Множественный квадрат R между факторами и оценками факторных баллов

	PA1	PA2
--	-----	-----

0.8237262 0.6305974

Веса, чтобы найти оценки факторов

	PA1	PA2
P10	0.10551671	-0.1437470
P11	1.46052478	-1.0597592
P14	-0.08346587	0.1311754
P16	-0.66092642	1.6662982

Standardized loadings (pattern matrix) based upon correlation matrix

item	PA1	PA2	h2	u2	com
P11	2 0.86	0.48	0.97	0.026	1.6
P16	4 0.73	0.66	0.96	0.042	2.0
P10	1 0.44	-0.03	0.20	0.802	1.0
P14	3 0.02	0.33	0.11	0.890	1.0

	PA1	PA2
SS loadings	1.47	0.77
Proportion Var	0.37	0.19
Cumulative Var	0.37	0.56
Proportion Explained	0.66	0.34
Cumulative Proportion	0.66	1.00

	PA1	PA2
Correlation of (regression) scores with factors	0.91	0.79
Multiple R square of scores with factors	0.82	0.63
Minimum correlation of possible factor scores	0.65	0.26

Loadings:

	PA1	PA2
P10	0.444	
P11	0.862	0.480
P14		0.330
P16	0.727	0.655

	PA1	PA2
SS loadings	1.470	0.769
Proportion Var	0.368	0.192
Cumulative Var	0.368	0.560

input from a matrix

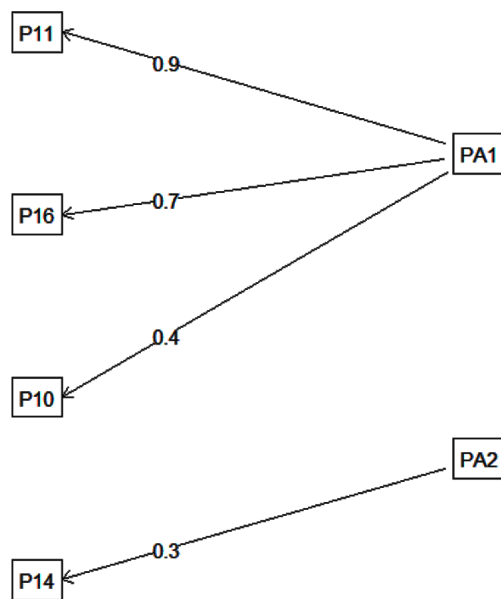


Рис. 4: ФАКТОРНАЯ ДИАГРАММА

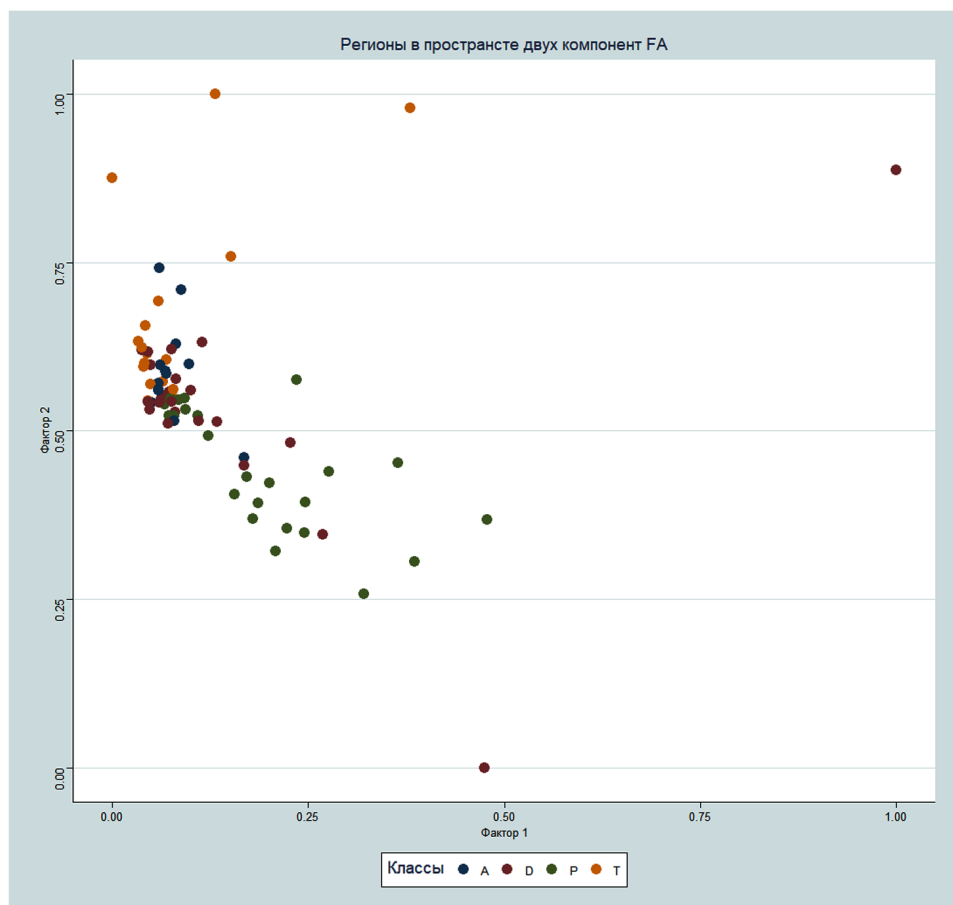


Рис. 5: ДАННЫЕ В ТРАНСФОРМИРОВАННОМ ПРОСТРАНСТВЕ