УДК:621.383:535

Василевская Т.П.

Vasilevskaya T.P.

ПРОЦЕС ОБРАБОТКИ ДАННЫХ В ФАКТОРНОМ АНАЛИЗЕ
PROCESS OF DATA PROCESSING IN THE FACTORE ANALYSIS

Аннотация. Проводится сравнительный анализ современных методов оценки общности в моделях факторного анализа, связанных с методом наименьших квадратов, на предмет сокращения затрат на вычислительный процесс.

Summary. It is the comparative analysis of modern methods of an estimation of a generality in models of the factor analysis connected to a method of the least squares will be carried out for reduction of expenses on computing process.

Факторный анализ (ФА) – это методы выявления гипотетических (ненаблюдаемых) факторов, призванных объяснить корреляционную матрицу наблюдаемых признаков. При этом предполагается, что наблюдаемые переменные являются линейной комбинацией факторов.

В большинстве приложений ФА речь идет о выявлении в наблюдаемых признаках $x_1, ..., x_p$ некоторой латентной (скрытой) переменной f называемой *фактором.* Гипотеза о наличии этого фактора основана на предположении о существовании чего-то общего в наблюдаемых признаках. В случае существования только одного фактора суть ФА состоит в объяснении корреляции между наблюдаемыми признаками с помощью корреляции этих признаков с фактором $r(x_i, f)$, $i = 1, ..., p$. В общем случае может быть несколько факторов $f_1, ..., f_k << p$. Корреляция между наблюдаемыми признаками и факторами обозначают $r(x_i, f_j) = a_{ij}$, $i = 1, ..., p, j = 1, ..., k$. Величины a_{ij} называются факторными нагрузками *и они образуют* матрицу факторных нагрузок $A = [a_{ij}]$, $i = 1, \ldots, p, j = 1, \ldots k$.

Во многих приложениях ФА основная цель состоит в объяснении корреляционной матрицы признаков R ее матрицей факторных нагрузок A. Матрицу A находят численными методами, как правило, определяя собственные числа и векторы матрицы R при условии выполнения $k << p$. В матричной записи факторная модель имеет вид

$$X = AF + e, \tag{1}$$

где A - $[p \times k]$ – матрица нагрузок; F - k – вектор факторов; e – вектор ошибок.

Ковариационня матрица

$$K = M[XX'] = V[AFF'A'] + M[LL'] = AA' + L^2, \tag{2}$$

где $L^2 = [l_{ii}^2]$ – диагональная матрица порядка p, содержащая дисперсии ошибок.

Основной моделью факторного анализа является уравнение (2). Основное условие: L^2 – диагональная; $K_x - L^2$ – неотрицательно определенная матрица. Дополнительным

условием единственности решения является диагональность матрицы $A(L^2)^{-1} A$. Имеется множество методов решения уравнения (2). Наиболее ранним методом факторного анализа является *метод главных факторов,* в котором методика анализа главных компонент используется применительно к редуцированной корреляционной матрице R^+ *с* общностями на главной диагонали. Для оценки общностей обычно пользуются коэффициентом множественной корреляции между соответствующей переменной и совокупностью остальных переменных. Факторный анализ проводится исходя из характеристического уравнения, как и в анализе главных компонент

$$|| R^+ - \lambda I | = 0. \tag{3}$$

Этот метод широко распространен, но постепенно уступает методу наименьших квадратов.

Из всех решений выбираем только то, для которого AA – диагональная и все элементы упорядочены по убыванию и различны. Нужно найти такие оценки матриц параметров A, L^2, которые минимизируют сумму квадратов разностей между двумя соответствующими друг другу, т.е. стоящими на одинаковых местах, элементами – дисперсиями и ковариациями – ковариационной матрицы K_x и оценки ковариационной матрицы $\hat{K}_x$. Иначе говоря, требуется минимизировать по элементам матриц A и L^2 функцию

$$u = \frac{1}{2} tr[\hat{K}_x - K_x]^2 = \min \tag{4}$$

или

$$U(A, L^2) = \frac{1}{2} tr[\hat{K}_x - AA' - L^2]^2 = \min \tag{5}$$

Если количество факторов k равно числу переменных p, то вычисленные и наблюдаемые корреляционные матрицы совпадут. При использовании (МНК) считают, что $k < p$.

Функционал (5) можно представить как функцию $U(A, L^2)$ и минимизировать эту функцию.

Имеется несколько численных алгоритмов реализации оценивания по методу наименьших квадратов. Среди них наиболее широко известны следующие алгоритмы: *итерационный метод главных осей, метод Ньютона-Рафсона, метод Хармана, ориентированный на минимизацию остатка.* Здесь мы рассмотрим подробно только первый метод.

А. *Итерационный метод главных осей*

1. По матрице наблюдений "объект-признак" находим оценки вектора средних $\bar{x} = [\bar{x}_1, ..., \bar{x}_p]$ и ковариационной матрицы $\hat{K}_x = [s_{ij}]$, где $\bar{x}_i = \frac{1}{n} \sum_{k=1}^{n} x_{ki}$, (6)

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \tag{7}$$

Информация, содержащаяся в $\hat{K}_x$, может быть также представлена корреляционной матрицей $R_x = [r_{ij}]$ и вектором стандартных отклонений $s_1, s_2, \ldots, s_p$, где $r_{ij} = s_{ij} / (s_i s_j)$.

2. Выбирается число факторов k.

3. По $\hat{K}_x$ находим оценку $l_j^{(1)} = \sqrt{\frac{1}{(s_{ij})^{-1}}}$, (8)

где s_{ij} – i-й диагональный элемент обратной матрицы $\hat{K}^{-1}$; k – число факторов, p – число исходных переменных.

4. Заменяем у $\hat{K}_x$ диагональные элементы s_{ij} на $(1 - l_{ii}^2) = h_{ii}^2$, т.е. *общностями* и получаем *редуцированную корреляционную матрицу* $K_x^2 = \hat{K}_x - L_{(1)}^2$.

5. Находим условный минимум $U(A, L_{(1)}^2)$ по A при заданном $L_{(1)}^2$,

$$\frac{\partial u}{\partial A} = (\hat{K}_x - AA' - L_{(1)}^2) A = 0 \tag{9}$$

Отсюда $\hat{K}_x A = A(A'A) + L_{(1)}^2 A$;

$\hat{K}_x A - L_{(1)}^2 A = A\lambda$;

$(\hat{K}_x - L_{(1)}^2) A = A\lambda$;

$\Sigma = \hat{K}_x - L_{(1)}^2$;

$\Sigma A = A\lambda$;

$(\Sigma - I\lambda) A = 0$.

Отсюда $A_{(1)}$ – есть собственный вектор, матрицы $\Sigma = \hat{K}_x - L_{(1)}^2$. Пусть собственные числа $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ и собственные векторы $\varphi_1, \varphi_2, \ldots, \varphi_p$. Условный экстремум $U(A_{(1)}, L_{(1)}^2)$ равен

$$U(L^2_{(1)}) = \frac{1}{2}\sum_{m=k+1}^{p}\lambda^2_m \quad . \tag{10}$$

6. Минимизируем полученную функцию $U(L^2_{(1)})$ по L^2, т.е. находим

$$\frac{\partial u}{\partial L^2} = diag(\hat{K}_x - AA' - L^2) = 0 \quad . \tag{11}$$

Заменив в (11) матрицу A на $A_{(1)}$ получим второе приближение

$$L^2_{(2)}L^2_{(2)} = diag(\hat{K}_x - A_{(1)}A'_{(1)}) \quad . \tag{12}$$

7. Находим $A_{(2)}$ с помощью главных компонент матрицы $\Sigma = \hat{K}_x - L^2_{(2)}$ и т.д. В общем случае

$$L^2_{(i)} = diag(\hat{K}_x - A_{(i-1)}A'_{(i-1)} \quad . \tag{13}$$

Если элементы матрицы $L^2_{(i)}$ удовлетворяют условию $\max(l^2_{(i)} - l^2_{(i-1)}) \le \varepsilon$, например, ε = 0,005, то говорят о сходимости оценок характерностей и принимают матрицы $A_{(i\text{-}1)}$ и $L^2_{(i)} = L^2$ за конечные характеристики.

В. Метод Ньютона-Рафсона

Будем минимизировать функционал (5) одним алгоритмом. Представим этот функционал как функцию $U(A, L)$ от A и L

$$U(A,L) = \frac{1}{2}tr[\hat{K}_x - AA' - L^2]^2 \quad . \tag{14}$$

Минимизация $U(A, L)$, как и в методе главных осей, проводится в два этапа. Сначала находится условный экстремум по A при заданном L. Полученная функция $U(L^2_{(1)})$ минимизируется численно по L, с помощью метода Ньютона-Рафсона. Для минимизации функции $U(L)$ по этому методу нужны производные функции $U(L)$ по L первого и второго порядков. Первые производные функции $U(L)$ равны $\frac{\partial u}{\partial l_i} = -2l_i\sum_{m=k+1}^{p}\lambda_m\varphi^2_{im}$. Если все собственные числа $\lambda_{k+1},...,\lambda_p$ матрицы близки к нулю, то вторые производные определяются по формуле

$$\frac{\partial^2 u}{\partial l_i \partial l_j} \approx 4l_i l_j\left(\sum_{m=k+1}^{p}\varphi_{im}\varphi_{im}\right)^2 \quad . \tag{15}$$

Основной алгоритм минимизации. Пусть L обозначает вектор-столбец с элементами $l_1, l_2, \ldots, l_p$ и пусть h и H обозначают соответственно вектор-столбец и матрицу производных, $\frac{\partial u}{\partial L}, \frac{\partial^2 u}{\partial L \partial L'}$. Обозначим $L^{(s)}$ значение L на s-й итерации, а $h^{(s)}$ и $H^{(s)}$ – соответствующие вектор первых производных и матрица вторых производных. ” Тогда итерационная процедура Ньютона-Рафсона запишется в виде

$$\begin{aligned} H^{(s)}\delta^{(s)} &= h^{(s)}, \\ L^{(s+1)} &= L^{(s)} - \delta^{(s)}, \end{aligned} \tag{16}$$

где $^{(s)}$ – вектор-столбец поправок, определенный в (16). Эта процедура проста для применений: основные вычисления на каждой итерации состоят в нахождении собственных чисел и собственных векторов и решении симметричной системы (16). По определению общность h^2_{ii} переменной i равна сумме квадратов нагрузок общих факторов этой переменной

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2 \quad (17)$$

Общности являются диагональными элементами редуцированной корреляционной матрицы. В случае использования корреляционной матрицы общности могут принимать значения от 0 до 1. Имеется 12 различных способов предварительной оценки общности. Наиболее часто используются три метода, которые кратко описаны ниже.

1. *Способ наибольшей корреляции.* В этом случае общность переменной приравнивается наибольшему коэффициенту корреляции данной переменной с остальными и на главной диагонали редуцированной корреляционной матрицы записывается этот коэффициент без учета знака. Этот способ оценки общностей хорош при большом числе переменных, порядка 20.

2. *Способ квадрата коэффициента множественной корреляции.* Для общности справедливо неравенство

$$h_i^2 \geq R_{i\cdot 12\ldots)i(\ldots p}^2, i = 1,\ldots, p \quad (18)$$

где $R_{i\cdot 12\ldots)i(\ldots p}^2$ – квадрат множественной корреляции (КМК), который известен из регрессионного анализа как *коэффициент детерминации.* Значение КМК является мерой дисперсии переменной, общей с другими переменными исследуемого множества, в то время как общность является мерой дисперсии *i*-й переменной, обусловленной общими для нескольких переменных факторами. Значения КМК для каждой переменной удобно вычислять с помощью обратной корреляционной матрицы по формуле

$$R_{i\cdot 12\ldots)i(\ldots p} = 1 - \frac{1}{r_{ii}^{-1}},$$

где r_{ii}^{-1} – диагональный элемент обратной матрицы R_x^{-1}.

Выбор КМК в качестве оценки общности в настоящее время наиболее теоретически обоснован и чаще всего рекомендуется. Обычно значения КМК получают как побочный результат при использовании метода главных компонент.

3. *Итеративный метод.* Вначале выбирается *k* факторов с помощью методов, описанных выше, затем вычисляются значения общностей, например, по КМК. После этого выделяются факторы методом главных факторов и вычисляют новые оценки общностей. Опять выполняется процедура выделения *k* факторов по корреляционной матрице с новыми общностями. Процесс продолжается до тех пор, пока диагональные элементы корреляционной матрицы не будут отличаться от итерации к итерации на заданную величину.

На первом этапе факторного анализа определяется минимальное число факторов, которые воспроизводят наблюдаемые между исходными переменными. После нахождения факторов нужна интерпретация этим факторам. Простую легко интерпретируемую структуру факторного пространства можно получить методом вращения осей.

Под простой факторной структурой понимается такая структура, в которой каждая переменная имеет не нулевую нагрузку только на один общий фактор. Если общих факторов больше или равно двум, то в простой матрице нагрузок каждая строка будет содержать только один ненулевой элемент, каждый столбец будет иметь несколько нуле и для каждой пары столбцов нулевые элементы не совпадают. Примером такой структуры может служить матрица

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

Такая идеальная структура для реальных данных недостижима, поэтому задача состоит в возможно большом приближении к такой структуре. Для получения простой структуры для реальных данных нужно вращать оси так, чтобы они проходили через скопление точек факторного пространства.

Рассмотрим факторную матрицу (табл. 1), полученную в результате применения *y* факторного анализа для шести переменных. Преобразование для точки 2, соответствующей переменной x_2, с координатами (a_1, a_2). Координаты этой точки после вращения осей против часовой стрелки на угол будут равны (рис. 1)

$$a_1' = a_1 \cos\theta - a_2 \sin\theta,$$
$$a_2' = a_1 \cos\theta + a_2 \sin\theta. \qquad (20)$$

a_1

a_2

Таблица 1 – Элементы факторной матрицы

Переменные	f_1	f_1
x_1	0,60	0,60
x_2	0,40	0,40
x_6	-0,10	0,50

Пусть угол вращения равен 30°, следовательно sin30° = 0,5000, cos30° = 0,8660. Новые координаты точек получаются при перемножении матриц:

$$\begin{pmatrix} 0,60 & 0,60 \\ 0,40 & -0,40 \\ -0,30 & 0,80 \\ -0,20 & 0,70 \\ -0,20 & 0,60 \\ -0,10 & 0,50 \end{pmatrix} \begin{pmatrix} 0,866 & -0,500 \\ 0,500 & 0,866 \end{pmatrix} = \begin{pmatrix} 0,8196 & 0,2196 \\ 0,5464 & 0,1464 \\ 0,1402 & 0,8428 \\ 0,1768 & 0,7062 \\ 0,1268 & 0,6196 \\ 0,1634 & 0,4830 \end{pmatrix}$$

$$AT = A,$$

где T матрица ортогонального вращения; A факторная система в новой системе координат. Таким образом, перевод одной системы координат в другую может быть записано в матричной форме $AT = A$, где T матрица вращения.

Для двухфакторного пространства T равна

$$T = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$ – при вращении против часовой стрелки,

$$T = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$ – при вращении по часовой стрелке.

Разработанные алгоритмы реализации описанных методов оценки общностей **в** моделях факторного анализа имеют ряд значительных преимуществ по сравнению с ранее используемыми, что связанно прежде всего, с уменьшением затрат на аппаратурные и программные средства. Это позволяет усовершенствовать процесс обработки данных и использовать его при обработке сложных сигналов.

Литература

1. *Хартман Г.* Современный факторный анализ. – Н.: Мир, 1972.

2. Панфилов И.П., Плотников И.М., Тррад И.С. Факторный анализ сжатия данных в процессе многомерной решеточной дискретизации // Наукові праці УДАЗ ім. О.С. Попова. – 2001. – № 1. – С. 3-14.