УДК 519.23

ВЫБОР МЕТОДА ОЦЕНКИ МАТРИЦЫ НАГРУЗОК В ФАКТОРНОМ АНАЛИЗЕ И АЛГОРИТМ ОЦЕНКИ ПРИ НУЛЕВЫХ НАГРУЗКАХ НА ЧАСТЬ СПЕЦИФИЧЕСКИХ ФАКТОРОВ

[1]Орлова И.В., [2]Турундаевский В.Б.

[1]Финансовый университет при Правительстве Российской Федерации, Москва, e-mail: IVOrlova@gmail.com;

[2]Московский государственный университет экономики, статистики и информатики, Москва, e-mail: vik_turund@mail.ru

В статье делается выбор между двумя наиболее применяемыми на практике методами оценки матрицы нагрузок: методом главных осей и методом максимального правдоподобия Лоули. В силу ряда причин рекомендуется использовать метод максимального правдоподобия. Однако методом нельзя пользоваться, если дисперсии некоторых специфических факторов равны нулю. Предлагается метод решения задачи в этом случае. Смысл метода состоит в добавлении к исходным переменным искусственно сгенерированных специфических факторов, с тем, чтобы к преобразованным данным можно было применить метод максимального правдоподобия. Предлагаемый метод пригоден к использованию и в случае коллинеарности исходных признаков, что расширяет возможности применения факторного анализа. Статья содержит 6 подразделов: 1. Введение. 2. Выбор метода оценки матрицы нагрузок на общие факторы. 3. Получение оценок матриц L и V. 4. Оценка числа общих факторов. 5. Случай вырожденного распределения.

Ключевые слова: наблюдаемые признаки, ковариационная матрица, корреляционная матрица, выборочная ковариационная матрица, факторный анализ, общие факторы, специфические факторы, матрица нагрузок на факторы, вырожденное распределение, метод максимального правдоподобия, распределение Уишарта

THE CHOICE OF ASSESSMENT METHOD OF THE MATRIX OF LOADINGS IN FACTOR ANALYSIS AND ALGORITHM EVALUATION IN THE ABSENCE OF SOME SPECIFIC FACTORS

[1]Orlova I.V., [2]Turundaevskiy V.B.

[1]Financial University under the Government of the Russian Federation, Moscow, e-mail: IVOrlova@gmail.com;

[2]Moscow state University of Economics, Statistics and Informatics, Moscow, e-mail: vik_turund@mail.ru

This paper makes a choice between the two most used methods in practice evaluation of the matrix of loadings: principal axis and maximum likelihood method Lawley. Due to a number of reasons, it is recommended to use the maximum likelihood method. However, the method cannot be used if the variance of some specific factors equal to zero. We propose a method of the solution of the problem in this case. The meaning of the method consists in adding to the original artificially generated variables specific factors, in order to transformed data it was possible to replace the maximum likelihood method. The proposed method is suitable for use and in the case of collinearity source characteristics that enhances the use of factor analysis. The article contains 6 subsections: 1. Introduction. 2. The choice of assessment method matrix of loadings on the common facto-ry. 3. Estimation of the matrices L and V. 4. The estimated number of common factors. 5. Case of a degenerate distribution.

Keywords: observable traits, covariance matrix, correlation matrix, the sample covariance matrix, factor analysis, General factors, specific factors, the matrix of loadings on the factors, degenerate distribution, maximum likelihood method, the distribution Wishart

1. Пусть x_1, x_2, …, x_p – p наблюдаемых признаков, $\overline{x_i}' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ – результаты i-го наблюдения признаков, $i = 1, 2, \ldots, n$, $\mathbf{X} = (x_{ij})$ – матрица наблюдений (исходных данных). В факторном анализе предполагается, что x_j является линейной комбинацией m линейно независимых факторов, так называемых «общих факторов» $f_1, f_2, \ldots, f_m$, плюс «специфический» для данного признака фактор e_i, некоррелированный ни с общими факторами, ни с другими специфическими факторами,

$$x_i = l_{i1}f_1 + l_{i2}f_2 + \ldots + l_{im}f_m + e_i = \sum_{j=1}^{m} l_{ij}f_j + e_i. \quad (1)$$

Будем считать x_i центрированными, а факторы – ортогональными:

$$M(x_i) = 0; \quad M(f_j) = 0;$$

$$M(e_i) = 0; \quad \sigma^2(f_j) = 1;$$

$$\sigma^2(e_i) = v_i; \quad \mathrm{cov}(f_i, f_j) = 0.$$

Общие факторы f_j являются «причиной» корреляций между признаками x_i. Эти факторы представляют собой непосредственно не измеряемые, скрытые (латентные) переменные, в той или иной мере связанные с исходными наблюдаемыми переменными. Ковариационная матрица $\sum$ исходных

признаков x_i, в соответствии с моделью факторного анализа (1), может быть представлена в виде

$$\Sigma = L \cdot L' + V, \quad (2)$$

где $L = (l_{ij})$ – матрица нагрузок на общие факторы, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, m$; V – диагональная матрица дисперсий специфических факторов e_i. Диагональные элементы матрицы $\Sigma^+ = L \cdot L'$ представляют собой дисперсии признаков, объясняемые m общими факторами. Эти элементы называются общностями, а сама матрица $\sum^+$ – редуцированной корреляционной матрицей.

Выбор метода оценки матрицы нагрузок на общие факторы

Оценки матрицы нагрузок L обычно получают одним из двух способов: методом главных осей или методом максимального правдоподобия [1, 2, 4]. В методе главных осей в качестве оценок матрицы нагрузок выбирают первые m собственных векторов редуцированной корреляционной матрицы $\sum^+$, соответствующие наибольшим собственным значениям матрицы $\sum^+$. В методе максимального правдоподобия оценка матрицы нагрузок получается путём максимизации функции правдоподобия, считая, что вектор наблюдаемых признаков $\overline{x}$ имеет многомерное нормальное распределение. Хотя оба метода направлены на максимальное приближение внедиагональных элементов корреляционной матрицы, тем не менее методы дают несколько различные результаты. При этом, как показали результаты численного моделирования, метод максимального правдоподобия приближает корреляционную матрицу немного лучше метода главных осей, даже если вектор наблюдаемых переменных $\overline{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ не имеет многомерное нормальное распределение. При этом метод максимального правдоподобия имеет под собой строгое математическое обоснование и оценки максимального правдоподобия обладают рядом хороших свойств, как-то: состоятельность, асимптотическая эффективность и асимптотическая несмещённость. Поэтому выбор метода максимального правдоподобия является предпочтительным. Однако метод максимального правдоподобия для оценки матрицы нагрузок не может применяться в некоторых ситуациях, например когда дисперсии специфических факторов равны нулю. В этой ситуации нами предлагается добавить в процесс оценивания преобразование исходных данных, с тем чтобы к преобразованным данным можно было применить метод максимального правдоподобия. Рассмотрим подробнее метод решения задач факторного анализа в этих ситуациях.

При практическом использовании факторного анализа часто возникают следующие ситуации:

1) некоторые специфические факторы отсутствуют в факторной модели;

2) выборочная ковариационная матрица исходных переменных не является положительно определенной.

Пусть $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_{n+1}$ – выборка из p-мерной нормальной совокупности $N(0, \sum)$.

Допустим, что вектор $\overline{x}$ генеральной совокупности удовлетворяет модели факторного анализа (1)

$$\overline{x} = L\overline{f} + \overline{e}, \quad (3)$$

где $\overline{x}' = (x_1, x_2, \ldots, x_p)$ – вектор наблюдаемых переменных, $\overline{f}' = (f_1, f_2, \ldots, f_m)$; $\overline{e}' = (e_1, e_2, \ldots, e_p)$ – векторы общих и специфических факторов соответственно, не наблюдаемые непосредственно; $L = (l_{jk})$ – матрица нагрузок $\overline{x}$ на общие факторы.

В модели (3) предполагается, что $(m + p)$-мерный вектор факторов $\overline{F} = \begin{pmatrix} \overline{f} \\ \overline{e} \end{pmatrix}$ имеет нулевое математическое ожидание и ковариационная матрица $\overline{F}$ имеет вид

$$M(\overline{F}\overline{F}') = \begin{pmatrix} I & 0 \\ 0 & V \end{pmatrix},$$

где $I = M(\overline{f} \cdot \overline{f}')$ – единичная матрица; $V = M(\overline{e} \cdot \overline{e}')$ – диагональная матрица.

Из модели (3) вытекает, что ковариационная матрица вектора $\overline{x}$ равна

$$\Sigma = LL' + V. \quad (4)$$

В приложениях факторного анализа возникают следующие задачи [1]:

1. Получение оценок матриц L и V.

2. Оценка числа общих факторов m.

3. Решение первых двух задач в условиях сильной коррелированности переменных, когда некоторые из них являются линейными комбинациями остальных.

Обозначим через S выборочную ковариационную матрицу вектора $\overline{x}$.

Допустим, что ковариационная матрица $\sum$ не вырождена. Тогда матрица S имеет распределение Уишарта, и оценка максимального правдоподобия матриц L и V получается из условия максимизации функции Z_0 [4, 5]:

$$Z_0 = -\frac{1}{2} n \left\{ \ln \left| \hat{L}\hat{L}' + \hat{V} \right| + tr \left[S \left(\hat{L}\hat{L}' + \hat{V} \right)^{-1} \right] \right\}. \quad (5)$$

В литературе описаны два основных подхода к решению задачи максимизации функции Z_0 [2, 3, 4]. В обоих известных

методах, основанных на них, предполагается, что все оценки дисперсий специфических факторов $\hat{V}_j$ отличны от нуля. Однако в практических исследованиях встречаются такие матрицы S, для которых некоторые оценки $\hat{V}_j$ близки к нулю. Это может явиться следствием одной из следующих причин:

а) в модели (3) некоторые дисперсии специфических факторов v_j равны нулю, т.е. размерность вектора факторов $\overline{F}$ меньше $m + p$. В данной ситуации при любом объеме выборки некоторые оценки $\hat{V}_j$ могут быть близки к нулю;

б) в модели (3) некоторые v_j близки к нулю; в этом случае, если объем выборки был бы достаточно велик, все оценки дисперсий $\hat{V}_j$ были бы больше нуля.

В практических расчетах матрица S иногда не является положительно определенной. Это может быть вызвано тем, что

а) в модели факторного анализа (3) больше, чем m, специфических факторов имеют нулевую дисперсию ($\overline{x}$ имеет вырожденное распределение);

б) определитель $|\sum|$ близок к нулю, и вследствие недостаточно большого объема выборки или ошибок округления матрица S может оказаться не положительно определённой.

Если S не является положительно определенной, то плотность распределения Уишарта равна нулю и мы не можем воспользоваться для оценки матриц нагрузок L и V функцией максимального правдоподобия [4].

Поскольку проверка гипотез о числе общих факторов производится после того, как определены оценки $\hat{L}$ и $\hat{V}$ [4], то в рассматриваемых ситуациях мы не сможем проверить эти гипотезы. Если максимум функции правдоподобия (5) ищется методом Лоули [4], то число общих факторов m и начальные приближения оценок $\hat{L}$ и $\hat{V}$ часто находят центроидным методом. Следует отметить, что в рассматриваемых ситуациях мы не сможем оценить число общих факторов также и в центроидном методе.

Итак, в ряде случаев нельзя использовать разработанный аппарат оценок максимального правдоподобия матриц $\hat{L}$ и $\hat{V}$ и, соответственно, проверить гипотезы о числе общих факторов.

Для решения этих задач можно предложить искусственно увеличивать дисперсии специфических факторов.

Получение оценок матриц L и V

Пусть $\overline{u} \in N(0, \Delta)$ – случайная величина с диагональной ковариационной матрицей, не зависящая от $\overline{x}$. Обозначим через $\hat{\Delta}$ выборочную ковариационную матрицу случайного вектора $\overline{u}$ и через $\hat{\Delta}_{xu}$ – матрицу выборочных коэффициентов ковариации векторов $\overline{x}$ и $\overline{u}$,

$$\hat{\Delta}_{xu} = \frac{1}{n+1}\sum_{i=1}^{n+1} \overline{x}_i \overline{u}_i',$$

где $\overline{x}_i$, $\overline{u}_i$ – векторы значений $\overline{x}$ и $\overline{u}$ в i-м наблюдении, $n + 1$ – объем выборки.

Для того, чтобы дисперсии всех специфических факторов сделать отличными от нуля, прибавим к обеим частям модели (2) вектор $\overline{u}$. Тогда модель (2) примет вид

$$\overline{z} = L\overline{f} + \overline{g}, \qquad (6)$$

где $\overline{z} = \overline{x} + \overline{u}$; $\overline{g} = \overline{e} + \overline{u}$.

Матрицы нагрузок на общие факторы L в моделях (3) и (6) совпадают.

Вектор $\overline{z}$ имеет многомерное нормальное распределение $N(0, \Sigma_0)$, где $\Sigma_0 = \Sigma + \Delta$.

Выберем диагональную матрицу Δ таким образом, чтобы S_0 – выборочная ковариационная матрица вектора $\overline{z} = \overline{x} + \overline{u}$ стала положительно определенной и оценки дисперсий всех специфических факторов модели (6) стали отличными от нуля.

Выборочная ковариационная матрица S_0 вектора $\overline{z}$ будет иметь распределение Уишарта $w(\Sigma_0, n)$. Так как S_0 положительно определена, плотность распределения Уишарта в точке S_0 отлична от нуля [4]. Следовательно, для оценки матриц L и V_0 модели (6) применим метод максимального правдоподобия. Функцию максимального правдоподобия получим, заменив в (5) $\hat{V}$ на $\hat{V}_0$, S на S_0 и $\hat{L}$ на $\hat{L}_0$. В силу выбора Δ оценки дисперсий специфических факторов положительны. Поэтому для максимизации функции правдоподобия можно воспользоваться любым из двух описанных в литературе подходов. При этом мы получим состоятельные, асимптотически несмещенные и эффективные оценки матриц L и V_0. Оценка матрицы нагрузок V на специфические факторы в модели (3) определяется из соотношения

$$\hat{V} = \operatorname{diag}\left(S - \hat{L}_0 \hat{L}_0'\right).$$

В практических задачах часто приходится рассматривать в качестве исходной выборочную корреляционную, а не ковариационную матрицу переменных. В этом случае полученную оценку матрицы нагрузок на общие факторы вектора $\overline{z}$ необходимо преобразовать, чтобы получить оценку матрицы нагрузок на вектор $\overline{x}$.

Пусть $\hat{L}_1$ – оценка матрицы нагрузок на общие факторы нормированного вектора $\overline{z}$,

$\hat{R}_z$ и $\hat{R}_x$ – оценки корреляционных матриц нормированных векторов $\overline{z}$ и $\overline{x}$ соответственно.

Очевидно,

$$\overline{z}_i = \hat{D}^{-1/2}(\overline{x}_i + \overline{u}_i), \quad i = 1,2,..., n+1, \tag{7}$$

где $\hat{D} = I + \mathrm{diag}\left(\hat{\Delta} + 2\hat{\Delta}_{xu}\right)$.

Из (7) получаем

$$\hat{R}_z = \hat{D}^{-1/2}\left(\hat{R}_x + \hat{\Delta} + \hat{\Delta}_{xu} + \hat{\Delta}'_{xu}\right)\hat{D}^{-1/2}.$$

Считая $\hat{D}$ не зависящей от выборки, легко показать, что

$$\hat{L}_0 = \hat{D}^{1/2}\hat{L}_1,$$

где $\hat{L}_0$ – оценка матрицы L модели (6).

Оценка $\hat{V}$ вычисляется по формуле

$$\hat{V} = I - \mathrm{diag}\left(\hat{L}_0\hat{L}'_0\right).$$

Оценка числа общих факторов

Для проверки гипотез о числе общих факторов используется статистика [4]

$$Z_1 = n\left\{\ln\frac{\left|\hat{L}_0\hat{L}'_0 + \hat{V}_0\right|}{|S_0|} + tr\left[S_0\left(\hat{L}_0\hat{L}'_0 + \hat{V}_0\right)^{-1}\right] - p\right\}.$$

Так как матрицы S_0 и $\hat{L}_0\hat{L}'_0 + \hat{V}$ не вырождены, то можно теперь для модели (6) проверить гипотезы о числе общих факторов.

Числа общих факторов в моделях (6) и (3) равны между собой.

Очевидно, при фиксированной матрице Δ, $P_m \to 1$ по вероятности при $n \to \infty$, где P_m – вероятность принять гипотезу H_0 о числе общих факторов в модели (6), равном m.

Рассмотрим зависимость статистики Z_1 от выбора матрицы Δ при фиксированном объеме выборки.

Выборочная ковариационная матрица вектора $\overline{z} = \overline{x} + \overline{u}$ равна

$$S_0 = S + \hat{\Delta}_{xu} + \hat{\Delta}'_{xu} + \hat{\Delta}. \tag{8}$$

Оценка максимального правдоподобия матрицы V_0 связана с S_0 и $\hat{L}_0$ соотношением [4]

$$\hat{V}_0 = \mathrm{diag}\left(S_0 - \hat{L}_0\hat{L}'_0\right). \tag{9}$$

Учитывая (8) и (9), нетрудно получить, что при $\hat{\Delta}_{11} \to \infty, \hat{\Delta}_{22} \to \infty, ..., \hat{\Delta}_{pp} \to \infty$.

$$\frac{\left|\hat{L}_0\hat{L}'_0 + \hat{V}_0\right|}{\left|\hat{\Delta}\right|} \to 1; \quad \frac{|S_0|}{\left|\hat{\Delta}\right|} \to 1; \tag{10}$$

$$tr\left[S_0\left(\hat{L}_0\hat{L}'_0 + \hat{V}_0\right)^{-1}\right] \to p \tag{11}$$

по вероятности.

При $\Delta_{jj} \to \infty$ оценки $\hat{\Delta}_{jj} \to \infty$ по вероятности, откуда, с учетом соотношений (10) и (11), получаем, что статистика Z_1 при $\hat{\Delta}_{11} \to \infty, \hat{\Delta}_{22} \to \infty, ..., \hat{\Delta}_{pp} \to \infty$ стремится к нулю по вероятности. Следовательно, в этом случае $P_0 \to 1$ по вероятности, где P_0 – вероятность принять гипотезу H_0 о том, что число общих факторов равно нулю.

Поскольку статистика Z_1 с ростом дисперсии «шума» убывает, при решении практических задач оценка числа общих факторов вследствие недостаточного объема выборки может оказаться заниженной. Поэтому численные значения Δ_{jj} следует выбирать не слишком большими, лишь бы только новые оценки дисперсий специфических факторов в модели (6) не получились равными нулю. Поскольку дисперсии оценок зависят от объема выборки, то и выбор численных значений Δ_{jj} будет зависеть в этом случае от объема выборки.

Факторный анализ направлен на анализ структуры внедиагональных элементов ковариационных матриц. Чем меньшие значения Δ_{jj} будут выбраны, тем меньше будут отличаться внедиагональные элементы выборочных ковариационных матриц S и S_0 и, следовательно, тем меньшее влияние на оценку матрицы нагрузок L окажет наложенный на статистические данные «шум» (при фиксированном объеме выборки). Это также является аргументом в пользу выбора небольших значений Δ_{jj}.

Случай вырожденного распределения $\overline{x}$

В практических исследованиях может встретиться ситуация, когда какая-то компонента вектора $\overline{x}$, например x_1, является линейной комбинацией остальных. Рассмотрим, как и выше, вектор $\overline{z} = \overline{x} + \overline{u}$ и допустим, что $\Delta_{11} > 0$ (Δ_{11} – дисперсия u_1). Тогда, если независимые переменные $x_2, x_3, \ldots, x_p$ имеют невырожденное многомерное нормальное распределение, то и вектор $\overline{z}$ будет иметь невырожденное нормальное распределение и можно, таким образом, включить x_1 в факторную модель (6). Это дает большую свободу в отборе переменных для факторного анализа, а также может быть использовано при построении уравнения регрессии с помощью факторного анализа.

Список литературы

1. Дубров А.М., Турундаевский В.Б., Френкель А.А. О задачах факторного анализа при отсутствии части специфических факторов // Учёные записки по статистике. т. 33. Прикладной многомерный статистический анализ. – М.: Наука, 1978.

2. Окунь Я. Факторный анализ: пер. с польск. – М.: Статистика, 1974.

3. Харман Г. Современный факторный анализ. – М.: Статистика, 1972.

4. Lawley D.N., Maxwell A.E. Factor Analysis as a Statistical Method, 2nd ed. – London: Butterworths, 1971.

5. Lawley D.N. Some new results in maximum likelihood factor analysis. Proceeding of Royal Society of Edinburgh, 1966–1967, v. A67.

References

1. Dubrov A.M., Turundaevskij V.B., Frenkel A.A. O zadachah faktornogo analiza pri otsutstvii chasti specificheskih faktorov // Uchjonye zapiski po statistike. t. 33. Prikladnoj mnogomernyj statisticheskij analiz. M.: Nauka, 1978.

2. Okun Ja. Faktornyj analiz: per. s polsk. M.: Statistika, 1974.

3. Harman G. Sovremennyj faktornyj analiz. M.: Statistika, 1972.

4. Lawley D.N., Maxwell A.E. Factor Analysis as a Statistical Method, 2nd ed. London: Butterworths, 1971.

5. Lawley D.N. Some new results in maximum likelihood factor analysis. Proceeding of Royal Society of Edinburgh, 1966–1967, v. A67.

Рецензенты:

Кобелев Н.Б., д.э.н., профессор кафедры «Системный анализ и моделирование экономических процессов», Финансовый университет при Правительстве РФ, Президент НП «Ремесленная палата России», г. Москва;

Киселёва И.А., д.э.н., профессор кафедры «Прикладная математика», МЭСИ, г. Москва.