ВЫБОР МЕТОДА ОЦЕНКИ ЗНАЧЕНИЙ ОБЩИХ ФАКТОРОВ ДЛЯ ОТДЕЛЬНЫХ ОБЪЕКТОВ В ФАКТОРНОМ АНАЛИЗЕ И ИХ СРАВНЕНИЕ ПРИ НУЛЕВЫХ НАГРУЗКАХ НА НЕКОТОРЫЕ СПЕЦИФИЧЕСКИЕ ФАКТОРЫ

УДК 519.23

Виктор Борисович Турундаевский,
к.э.н., доцент, проф. каф. Прикладной математики Московского государственного университета экономики, статистики и информатики (МЭСИ)
Тел.: 8 (495)-442-70-98
Эл. почта: VBTurundaevskiy@mesi.ru

Ирина Владленовна Орлова,
к.э.н., профессор, проф. каф. Системного анализа и моделирования экономических процессов Финансового университета при Правительстве РФ (Финуниверситет)
Тел.: 8 (499)-277-21-44
Эл. почта: IVOrlova@gmail.com

В статье рассматриваются методы оценки значений общих факторов для отдельных наблюдений, делается выбор между тремя наиболее применяемыми на практике методами оценки значений факторов: регрессионным методом, методом Бартлетта и методом «идеальных параметров» Хармана. Однако этими методами 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 рассматриваемые методы оценки значений факторов. Предлагаемый метод пригоден к использованию и в случае коллинеарности исходных признаков, что расширяет возможности применения факторного анализа. Подробнее других рассмотрен метод «идеальных параметров» Хармана. Доказаны экстремальные свойства метода.

Ключевые слова: факторный анализ, общие факторы, специфические факторы, матрица нагрузок на факторы, вырожденное распределение, методы оценки значений общих факторов, метод Бартлетта, метод «идеальных параметров» Хармана.

Viktor B. Turundaevsky,
PhD, Associate Pro-fessor, Dept. Applied Mathematics, Moscow State University of Economics, Statistics and Informatics (MESI)
Tel.: (495) 442-60-98
E-mail: vik_turund@mail.ru

Irina V. Orlova,
PhD, Professor, Dept. System analysis and modeling of economic processes of the Financial University under the Government of the Russian Federation
Tel.: (499) 277-21-44
E-mail: IVOrlova@fa.ru

SELECTION A METHOD FOR ESTIMATING THE VALUES OF COMMON FACTORS FOR INDIVIDUAL OBJECTS IN THE FACTORIAL ANALYSIS AND THEIR COMPARISON WITH ZERO LOADS ON SOME SPECIFIC FACTORS

The article considers methods of estimating the values of the common factors for individual observations, make a choice between the three most used methods in practice evaluation of factor values: the regression method, Bartlett method and the method of "ideal settings" Harman. However, these methods cannot be used if the variance of some specific factors equal to zero. A method is proposed for solving 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 apply the methods of evaluation of factor values. The proposed method is suitable for use in the case of collinearity initial signs that extends the application of factor analysis. More other method considered "ideal settings" Harman. We prove the extreme properties of the method.

Keywords: factor analysis, General factors, specific factors, the matrix of loadings on the factors, degenerate distribution, methods of evaluation values of common factors, the Bartlett method, the method of "ideal settings" Harman.

1. Введение

Пусть x_1, x_2, …, x_p – p наблюдаемых признаков, – результаты i-го наблюдения признаков, i = 1, 2, …, n, $X = (x_{ij})$ – матрица наблюдений (исходных данных).

Факторный анализ внешне связан с анализом главных компонент. В обоих случаях рассматриваются зависимости между $\boldsymbol{m}$ признаками, образующими вектор $\boldsymbol{X}$, на основе анализа ковариационной или корреляционной матрицы и расщепления её на составляющие части. Если в компонентном анализе каждый признак X_j есть линейная комбинация первых k главных компонент плюс (m-k) последних компонент, то в факторном анализе предполагается, что X_j является линейной комбинацией k линейно независимых факторов, так называемых «общих факторов» $f_1, f_2, \ldots, f_k$, плюс «специфический» для данного признака фактор u_i, некоррелированный ни с общими факторами, ни с другими специфическими факторами. Однако, если в компонентном анализе значения любых главных компонент для i-го наблюдения признаков являются линейными комбинациями значений признаков и, следовательно, могут быть вычислены на основе матрицы наблюдений признаков X, то в факторном анализе значения общих факторов являются скрытыми, непосредственно не вычисляемыми значениями. Их можно лишь оценить, приняв то или иное дополнительное допущение относительно их значений. Поэтому существует ряд методов для получения оценок общих факторов.

Модель факторного анализа имеет вид:

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

Будем считать x_i центрированными, а факторы – ортогональными:

$$M(x_i) = 0,\ M(f_j) = 0,\ M(u_i) = 0,\ \sigma^2(f_j) = 1,\ \sigma^2(e_i) = v_i,\ \mathrm{cov}(f_i, f_j) = 0.$$

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 связанная с исходными наблюдаемыми переменными. Ковариационная матрица $\sum$ исходных признаков x_i, в соответствии с моделью факторного анализа (1), может быть представлена в виде

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

где $L = (l_{ij})$ – матрица нагрузок на общие факторы, $i = 1, 2, \ldots, p$, j = 1, 2, …, m, V – диагональная матрица дисперсий специфических факторов v_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.

Для оценки матрицы нагрузок L наиболее целесообразно применять метод максимального правдоподобия (детальнее этот вопрос рассмотрен в статье [1], однако он не может применяться в некоторых ситуациях, например, когда дисперсии специфических факторов равны нулю. В этой ситуации в работе [1] предлагается добавить в процесс оценивания преобразование исходных данных, с тем, чтобы к преобразованным данным можно было применить метод максимального правдоподобия. В этой работе не рассматривается вопрос оценки матрицы нагрузок L и матрицы дисперсий V. Эти вопросы рассмотрены в работе [1]. Здесь рассмотрены вопросы оценки значений общих факторов для отдельных наблюдений. Стандартные оценки в случае, когда некоторые v_j равны нулю или выборочная ковариационная матрица S вырождена, не могут быть применены. Данная работа посвящена вопросу получения оценок значений общих факторов в этой ситуации.

2. Рассматриваемая в работе ситуация, когда некоторые оценки дисперсий $\hat{V}_j = 0$, обусловливает специфику в решении задачи оценки значений общих факторов для отдельных наблюдений. Вычисление таких оценок необходимо во многих практических задачах.

Для того, чтобы дисперсии всех специфических факторов сделать отличными от нуля, прибавим к обеим частям модели (2) некоррелированный с $\overline{f}$ и $\overline{e}$ вектор $\overline{u}$ [1]. Тогда модель (2) примет вид

$$\overline{z} = L\overline{f} + \overline{g}, \qquad (3)$$

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

Матрицы нагрузок на общие факторы $\boldsymbol{L}$ в моделях (2) и (3) совпадают.

Выберем диагональную матрицу Δ дисперсий вектора $\overline{u}$ 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 (3) стали отличными от нуля. Следовательно, для оценки матрицы нагрузок $\boldsymbol{L}$ и диагональной матрицы дисперсий новых специфических факторов $\boldsymbol{V_0}$ модели (3) применим метод максимального правдоподобия [2].

В соответствии с основной моделью факторного анализа (2), наблюдаемый вектор $\overline{x}$ принадлежит p-мерному подпространству $(m + p)$-мерного пространства общих и специфических факторов. Поэтому общие и специфические факторы нельзя непосредственно выразить через $\overline{x}$. В качестве значений общих факторов выбираются «наилучшие» в некотором смысле линейные комбинации исходных переменных

$$\hat{f}_k = \beta_{k1}x_1 + \beta_{k2}x_2 + \ldots + \beta_{kp}x_p \qquad (4)$$

При этом оценки факторных значений $\hat{f}_k$ уже не будут некоррелированными между собой, если мы не выберем специальный базис, даваемый каноническим факторным анализом [3]. Оценки факторных значений $\hat{f}_k$ коррелированы не только между собой, но имеют также ненулевую корреляцию с другими факторами $\hat{f}_q (q \neq k)$.

При оценке значений общих факторов следует вернуться от модели (3) к основной модели факторного анализа (2). В противном случае наложенный на вектор $\overline{x}$ «шум» $\overline{u}$ может повлиять на оценки факторных значений.

Обзор методов оценки факторных значений содержится в работе [3]. Наиболее естественными являются 3 подхода: регрессионный метод, метод Бартлетта и метод «идеальных параметров» Хармана [4], [5]. Рассмотрим их подробнее.

a) Регрессионный метод.

Если в качестве «наилучшего» приближения общих факторов выбирается такая линейная комбинация исходных переменных x_j, которая минимизирует по множеству наблюдений квадрат разности между f_k и $\hat{f}_k$, то мы приходим к регрессионному методу оценки значений факторов [2]

$$\hat{f} = \hat{L}'S^{-1}\overline{x}$$

или, заменяя S на $\hat{L}\hat{L}' + \hat{V}$,

$$\hat{f} = (I + \hat{L}'\hat{V}^{-1}\hat{L})^{-1}\hat{L}'\hat{V}^{-1}\overline{x},$$

где $\hat{f}$ – вектор-столбец оценок факторов $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m$. $\hat{f}$ является оценкой метода наименьших квадратов вектора f.

Регрессионный метод приводит к смещенным оценкам значений факторов [3]. Оценка $\hat{f}$ приводит также к смещенным оценкам коэффициентов уравнения регрессии зависимой переменной y по общим факторам.

Оценкой $\hat{f}$ нельзя пользоваться в случае, если дисперсии каких-то специфических факторов равны нулю.

Исходные переменные $x_1, x_2, \ldots, x_p$ должны удовлетворять основным требованиям регрессионного анализа, и, в частности, степень корреляции переменных не должна быть высокой. Последнее ограничение является существенным, ибо для его выполнения необходимо выводить из модели факторного анализа часть переменных. Это ограничение на модель факторного анализа лишает нас одного из преимуществ этого метода в отборе переменных для модели. Поэтому в практических исследованиях возможности регрессионного метода оценки факторов ограниченны.

b) Метод Бартлетта.

Если для получения оценок факторных значений минимизируется по множеству наблюдений сумма квадратов нормированных специфических факторов $\dfrac{\sum_{j=1}^{p}\left(x_j - \sum_{k=1}^{m}\hat{l}_{jk}\hat{f}_k\right)^2}{\hat{V}_j}$, то приходим к методу минимизации остатков, предложенному Бартлеттом [2].

$$\hat{f} = (\hat{L}'\hat{V}^{-1}\hat{L})^{-1}\hat{L}'\hat{V}^{-1}\overline{x} \qquad (5)$$

Если какой-то из специфических факторов имеет оценку дисперсии $\hat{V}_j = 0$, то оценка (5) теряет смысл.

c) Метод «идеальных параметров» Хармана.

Оценка «идеальных параметров» Хармана почти не используется при оценке значений факторов и в уравнениях регрессии на общих факторах, хотя обладает рядом экстремальных свойств.

К этой оценке приходим, если станем минимизировать по множеству наблюдений сумму квадратов специфических факторов

$$\sum_{j=1}^{p}\left(x_j - \sum_{k=1}^{m}\hat{l}_{jk}\hat{f}_k\right)^2 \qquad (6)$$

Оценка имеет вид [4]:

$$\tilde{f} = (\hat{L}'\hat{L})^{-1}\hat{L}'\overline{x}. \quad (7)$$

Нетрудно видеть, что оценка (7) минимизирует (6) не только по всей совокупности наблюдений, но и для каждого наблюдения в отдельности.

Для ковариационной матрицы оценки $\tilde{f}$ получаем выражение

$$\tilde{R} = I + (\hat{L}'\hat{L})^{-1}\hat{L}'V\hat{L}(\hat{L}'\hat{L})^{-1}. \quad (8)$$

Как известно [6], $\hat{L} = V_0^{1/2}\Omega(\theta - I)^{1/2}$, откуда

$$(\hat{L}'L)^{-1} = (\theta - I)^{-1/2}(\Omega'\hat{V}_0\Omega)^{-1}(\theta - I)^{-1/2}, \quad (9)$$

где θ – диагональная матрица m первых собственных чисел матрицы $\hat{V}_0^{-1/2}S_0\hat{V}_0^{-1/2}$, Ω – $(p \times m)$ – матрица, столбцы которой являются соответствующими собственными векторами $\hat{V}_0^{-1/2}S_0\hat{V}_0^{-1/2}$.

Из (9) вытекает, что обратная матрица $(\hat{L}'L)^{-1}$ существует и, следовательно, оценкой факторных значений (7) можно пользоваться и в том случае, если некоторые $\hat{V}_j = 0$.

Для построения уравнения регрессии зависимой переменной y по общим факторам необходимо иметь оценки коэффициентов корреляции между y и факторами $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m$.

Обозначим через $\hat{R}_{\tilde{f}y}$ вектор оценок коэффициентов корреляции общих факторов с y, основанных на оценке «идеальных параметров» Хармана, через $\hat{r}_{\tilde{f}y}, \hat{r}_{\overline{x}y}, \hat{r}_{\hat{f}y}, \hat{r}_{\overline{\gamma}y}$ – векторы выборочных коэффициентов ковариации y с $\tilde{f}$, $\overline{x}$, $\hat{f}$, $\overline{\gamma}$ соответственно.

Из (7) получаем

$$\hat{r}_{\tilde{f}y} = (L'L)^{-1}\hat{L}'\hat{r}_{\overline{x}y}, \quad (10)$$

откуда

$$\hat{R}_{\tilde{f}y} = s_y^{-1}(\mathrm{diag}\ \tilde{R})^{-1}(L'L)^{-1}\hat{L}'\hat{r}_{\overline{x}y},$$

где s_y^2 – оценка дисперсии y.

Можно доказать, что оценка (10) минимизирует на множестве $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m$ линейных комбинаций вида (4) сумму квадратов

$$\varphi(\tilde{f}) = \sum_{j=1}^{p}\left(\hat{r}_{x_j y} - \sum_{k=1}^{m}\hat{l}_{jk}\hat{r}_{\hat{f}_k y}\right)^2,$$

где $\tilde{f}' = (\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_m)$.

Для доказательства рассмотрим соотношения

$$\overline{x}_i = \hat{L}\tilde{f}_{(i)} + \overline{\gamma}_i, \quad (11)$$

где $\tilde{f}'_{(i)} = (\tilde{f}_{i1}, \tilde{f}_{i2}, \ldots, \tilde{f}_{im})$ – вектор значений функций $\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_m$ вида (4) в i-ом наблюдении, $i = 1, 2, \ldots, n+1$, невязка $\overline{\gamma}_i$ определяется соотношением (11).

Из (11) получаем

$$\hat{r}_{\overline{x}y} = \hat{L}\hat{r}_{\tilde{f}y} + \hat{r}_{\overline{\gamma}y}. \quad (12)$$

По свойству метода наименьших квадратов из (12) вытекает, что $\min\sum_{j=1}^{p}\hat{r}^2_{\gamma_j y}$ достигается на векторе $\hat{r}_{\tilde{f}y}$, определяемом соотношением (10), и, следовательно, $\varphi(\tilde{f}) = \min_{\hat{f}}\varphi(\hat{f})$.

3. Итак, в тех случаях, когда-либо заранее известно, что некоторые дисперсии специфических факторов в основной модели факторного анализа (2) равны нулю, либо в процессе получения оценок матриц $\boldsymbol{L}$ и $\boldsymbol{V}$ некоторые оценки оказались равными нулю, либо выборочная ковариационная матрица $\boldsymbol{S}$ вырождена, метод максимального правдоподобия для оценивания матриц $\boldsymbol{L}$ и $\boldsymbol{V}$ неприменим. В этих случаях оценки матриц $\boldsymbol{L}$ и $\boldsymbol{V}$ можно получить с помощью предлагаемого в [1] метода. Для вычисления в рассматриваемых ситуациях оценок значений общих факторов для отдельных наблюдений метод Бартлетта неприменим, регрессионный метод также не всегда может быть использован. Оценки метода «идеальных параметров» Хармана можно использовать во всех рассматриваемых ситуациях. В данной работе рассмотрены некоторые свойства оценок Хармана.

Следует отметить, что при практическом использовании предлагаемого метода нет необходимости накладывать «шум» $\overline{u}$ на исходную статистическую информацию — достаточно получить выборочные ковариационные матрицы $\hat{\Delta}$ и $\hat{\Delta}_{xu}$. Отметим также, что при выборе диагональной ковариационной матрицы Δ рекомендуется брать отличными от нуля только те элементы Δ_{jj}, которым соответствуют оценки $\hat{V}_j = 0$. При этом сами элементы Δ_{jj} не должны быть слишком большими.

Литература

1. Орлова И.В., Турундаевский В.Б. Выбор метода оценки матрицы нагрузок в факторном анализе и алгоритм оценки при нулевых нагрузках на часть на часть специфических факторов // Фундаментальные исследования. – 2015. – №6 (часть 1).

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

3. Lawley D.N. Some new results in maximum likelihood factor analysis // Proceeding of Royal Society of Edinburgh – 1966–1967, v. A67, p. 256–264.

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

5. Окунь Я. Факторный анализ: Пер. с польск. – М.: Статистика, 1974.

References

1. Orlova I. V., Turundaevskiy V. B. The choice of assessment method of the matrix of loadings in factor analysis and the estimation algorithm at zero loads on the part of specific factors // Fundamental research. – 2015. – No. 6 (part 1).

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

3. Lawley D. N. Some new results in maximum likelihood factor analysis. Proceeding of Royal Society of Edinburgh, 1966–1967, v. A67, p. 256–264.

4. Harman H. Modern factor analysis. – M.: Statistics,1972.

5. Okun J. Factor analysis: Trans. with the Polska. – M.: Statistics, 1974.