ПРИМЕНЕНИЕ МЕТОДА МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ ДЛЯ ПОИСКА СКРЫТЫХ НЕЛИНЕЙНЫХ ЗАВИСИМОСТЕЙ

APPLICATION OF THE MAXIMUM LIKELIHOOD ESTIMATION FOR THE SEARCHING OF HIDDEN NONLINEAR DEPENDENCIES

O. Buchnev

Summary. The problem of nonlinearity in data analysis is relevant for most statistical methods. Data analysis is useful, among other things, for reducing the dimension or for searching for hidden dependencies in data. For searching of hidden dependencies, factor analysis is used. The most powerful of factor analysis methods is the maximum likelihood estimation, and it is successfully used for searching of hidden linear relationships. But when the dependencies are nonlinear, the maximum likelihood estimation does not always give good results. In this paper a modification of the maximum likelihood estimation is proposed. In the proposed modification of the MLE, the pair correlation matrix is proposed to be replaced by the matrix of correlation indexes obtained by using polynomial regression. A method for converting the resulting matrix of correlation indexes into a symmetric matrix is proposed. The results of the modified MLE operation on both model and real data are presented. The proposed modification of the MLE can supplement the set of methods available to the researcher. The results of the modified MLE can provide additional information about the subject area in the analysis of data, including the search and study of hidden dependencies.

Keywords: factor analysis, maximum likelihood estimation, polynomial regression, correlation index, nonlinear dependence.

Бучнев Олег Сергеевич
К.т.н., Иркутский национальный исследовательский технический университет
buchnevo81@mail.ru

Аннотация. Проблема нелинейности при анализе данных актуальна для большинства статистических методов. Анализ данных проводится, в том числе, с целью сокращения размерности или поиска скрытых зависимостей в данных. Для поиска скрытых зависимостей применяют факторный анализ. Методы факторного анализа, и наиболее мощный из них — метод максимального правдоподобия, успешно используется для поиска скрытых линейных зависимостей. В случае, когда зависимость переменных нелинейная, метод максимального правдоподобия не всегда дает хорошие результаты. В статье предложена модификация метода максимального правдоподобия. В предложенной модификации ММП матрицу парных корреляций предложено заменить матрицей индексов корреляции, полученных с применением полиномиальной регрессии. Предложен способ преобразования получаемой матрицы индексов корреляции в симметричную матрицу. Приведены результаты работы модифицированного ММП как на модельных, так и на реальных данных. Предложенная модификация ММП может служить дополнением к набору имеющихся у исследователя методов. Результаты работы модифицированного ММП могут давать дополнительную информацию о предметной области при анализе данных, в том числе, при поиске и исследовании скрытых зависимостей.

Ключевые слова: факторный анализ, метод максимального правдоподобия, полиномиальная регрессия, индекс корреляции, нелинейная зависимость.

Введение

В настоящее время анализ данных крайне востребован, и интерес к нему продолжает нарастать. Это может быть обусловлено распространением технологий, требовательных к техническому и программному обеспечению, таких как машинное обучение, Big Data, нейронные сети. При обработке больших массивов данных для сокращения размерности и поиска скрытых зависимостей популярны так же методы многомерного статистического анализа, такие, как факторный анализ и метод главных компонент. Эти методы широко применяются для обработки и анализа данных при проведении научных исследований в различных областях — не только в сфере технических наук, но и, например, в социологии, психологии, медицине, экономике и других науках. В немалой степени этому способствует и распространение современных языков программирования с множеством библиотек, которые содержат функции анализа и статистической обработки данных.

В результате факторного анализа получают матрицу нагрузок на общие факторы L, диагональную матрицу дисперсий специфических факторов E и факторы $f_i, i = \overline{1,k}$. Для получения матрицы нагрузок на общие факторы L по ряду причин, рекомендуют метод максимального правдоподобия (ММП) [1]. Однако применение этого метода предполагает, во-первых, нормальный закон распределения исходных признаков [2], а во-вторых, в связи с тем, что сам факторный анализ позволяет найти скрытые факторы используя ковариационную (или корреляционную) матрицу исходных признаков, зависимость признаков предполагается линейной (линейная зависимость признаков является основным

предположением в факторном анализе [3]). Во многих практических приложениях зависимость между признаками может оказаться нелинейной, в этом случае традиционные методы факторного анализа оказываются малоэффективными.

За последнее время предложены подходы к решению проблемы нелинейной зависимости признаков. В работе [4] предложены две модификации ММП, использующие в качестве мер связи признаков ранговые коэффициенты корреляции Спирмена и коэффициенты Крамера. В работе [5] предложен алгоритм, который позволяет строить уточняющую модель и исследовать влияние нелинейной составляющей в факторном влиянии.

Следует отметить, что результаты факторного анализа являются, как правило, приближенными, и зависят, кроме прочего, от применяемого способа оценки матрицы нагрузок, общностей, вращения факторов. Кроме того, не всегда можно дать однозначную интерпретацию получаемым факторам. По этой причине, в практических приложениях, для решения задачи выделения факторов, в зависимости от условий решаемой задачи и предметной области, результаты могут зависеть от применяемых для анализа методов и техник. В данной работе предложена модификация ММП, основанная на использовании матрицы индексов корреляции, что делает возможным применение ММП для выявления факторов, порождающих нелинейные зависимости признаков.

Постановка задачи

Факторный анализ представляет собой совокупность методов, объединенных предположением о том, что изменчивость в значениях наблюдаемых признаков обусловлена наличием небольшого числа (меньшего, чем количество признаков) скрытых причин, общих для всех признаков. Эти причины называются общими факторами. Оставшаяся доля изменчивости каждого признака объясняется присутствием «частного фактора», который влияет только на этот признак, и ни на какой другой [4].

Конечная цель исследования, проводимого с привлечением методов факторного анализа, как правило, состоит в выявлении и интерпретации скрытых общих факторов. При этом исследователь преследует две противоречивые цели: необходимо минимизировать количество скрытых факторов, и в то же время минимизировать степень зависимости признаков от своих специфических остаточных случайных компонент [6]. Как и в любой модельной схеме, эта цель может быть достигнута лишь приближенно.

Таким образом, основной задачей факторного анализа является экономное описание экспериментальных данных. Он «объясняет» корреляционную матрицу системы случайных величин $x_1, \dots, x_r$ наличием небольшого числа общих гипотетических переменных (факторов), от которых зависят $x_1, \dots, x_r$. В общем виде модель факторного анализа выглядит следующим образом:

$$\vec{x} = L\vec{f} + \vec{e}, \quad (1)$$

где $L = \|l_{ij}\|$ — прямоугольная матрица нагрузок размера $r \times k$; $\vec{f} = (f_1, \dots, f_k)$ — общие факторы; $\vec{e} = (e_1, \dots, e_r)$ — частные факторы [7].

Обозначим через C матрицу вторых моментов случайной величины $\vec{x} = (x_1, \dots, x_r)$. Тогда из выражения (1) следует:

$$C = LL^T + V,$$

где V — диагональная матрица размера $r \times r$ с диагональными элементами равными остаточным дисперсиям: $De_i = v_i, i = \overline{1,r}$, остальные элементы этой матрицы равны нулю.

Пусть $\vec{x}^{(1)}, \dots, \vec{x}^{(n)}, n \geq r$,— независимые наблюдения над случайной величиной $\vec{x}$. Для этой системы случайных величин может быть построена выборочная ковариационная матрица:

$$A = \|a_{ij}\|_1^r = \frac{1}{n}\sum_{t=1}^{n} \vec{x}^{(t)}\vec{x}^{(t)'}$$

(выражение справедливо для случая, когда вектор средних значений равен нулю). Для получения оценок параметров l_{ij} и v_i можно использовать информацию, содержащуюся в A. Согласно методу максимального правдоподобия, оценки l_{ij} и v_i определяют из условия, чтобы совместная плотность элементов выборочной ковариационной матрицы, вычисленная в наблюденной точке $A = \|a_{ij}\|_1^r$, т.е. функция правдоподобия, имела наибольшее значение.

Если случайный вектор $\vec{x} = (x_1, \dots, x_r)$ распределен по нормальному закону $\mathrm{N}(0, C)$, то совместное распределение элементов матрицы $A = \|a_{ij}\|$, называемое распределением Уишарта, имеет плотность распределения вероятностей:

$$W(r, n, C) = \gamma(r, n)|C|^{-\frac{n}{2}}|A|^{\frac{(n-r-1)}{2}} e^{-\frac{n}{2}\sum_{i,j} a_{ij}c^{ij}} \quad (2)$$

где $\|c\|^{ij} = C^{-1}$ и $\gamma(r, n)$ — множитель, зависящий только от r и n.

Опуская члены, не зависящие от C, из (2) получаем, что искомые оценки для l_{ij} и v_i доставляют минимум функции

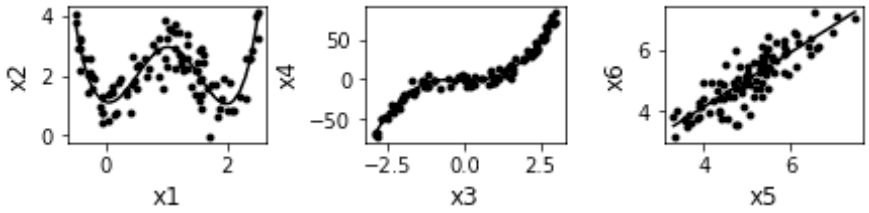

Рис. 1. Корреляционные поля сгенерированных наборов данных

$$L = \begin{vmatrix} -0{,}001 & 0{,}108 & 0{,}067 \\ -0{,}005 & -0{,}037 & 0{,}085 \\ \mathbf{0,972} & -0{,}076 & -0{,}145 \\ \mathbf{0,978} & -0{,}049 & 0{,}131 \\ -0{,}125 & \mathbf{-0,906} & -0{,}018 \\ -0{,}115 & \mathbf{-0,979} & 0{,}007 \end{vmatrix} V$$

$$= \begin{vmatrix} 0{,}98 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}99 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}03 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}02 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}03 \end{vmatrix}$$

Рис. 2

$$L = ln|C| + \sum_{i,j} a_{ij} c^{ij} \quad (3)$$

Чтобы получить соответствующие уравнения, следует продифференцировать (3) по l_{ij} и v_i, и приравнять частные производные нулю. В результате будут получены уравнения максимального правдоподобия для нахождения оценок l_{ij} и v_i [7]:

$$L^T C^{-1} - L^T C^{-1} A C^{-1} = 0,$$

$$diag(C^{-1} - C^{-1} A C^{-1}) = 0,$$

где $diag(M)$ обозначает матрицу, у которой на главной диагонали стоят диагональные элементы матрицы M, а все остальные элементы — нули.

К преимуществам метода максимального правдоподобия можно отнести хорошее приближение корреляционной матрицы, даже если вектор наблюдаемых переменных $\vec{x} = (x_1, \ldots, x_r)$ не имеет многомерное нормальное распределение, при этом метод максимального правдоподобия имеет под собой строгое математическое обоснование, и оценки максимального правдоподобия обладают такими свойствами, как состоятельность, асимптотическая несмещённость и асимптотическая эффективность [1].

Не смотря на то, что в практических приложениях факторного анализа зависимости переменных, как правило, линейны, может возникнуть задача выявления фактора, порождающего нелинейную зависимость исходных признаков. В этом случае коэффициент корреляции при измерении силы связи перестает быть информативным, и имеющаяся, даже очевидная связь, не находит отражения в корреляционной матрице и в матрице нагрузок.

Для экспериментальной проверки выдвигаемых гипотез на языке Python 3 написана программа, с помощью которой можно получать наборы данных как с линейными, так и с нелинейными (описываемыми полиномом, порядок которого может быть задан пользователем) зависимостями между признаками. Для стандартного факторного анализа получаемых наборов данных использована функция FactorAnalysis бесплатной библиотеки машинного обучения Scikit-learn.

С помощью написанной программы получен набор данных из ста наблюдений, каждое наблюдение имеет шесть признаков. При этом первый и второй признаки связаны нелинейной зависимостью, которая может быть аппроксимирована выражением: $\bar{x}_2 = 2x_1^4 - 8x_1^3 + 8x_1^2 + 1$. Зависимость третьего и четвертого признаков можно аппроксимировать: $\bar{x}_4 = 3x_3^3 + x_3$. Пятый и шестой признак связаны линейной зависимостью с коэффициентом корреляции равным 0,8. Корреляционные поля соответствующих зависимостей показаны на рисунке 1.

Нагрузочная матрица и остаточные дисперсии сгенерированного набора данных, полученные с помощью функции FactorAnalysis, для случая, когда число факторов равно трем (рис. 2).

Как видно из нагрузочной матрицы, удалось выявить лишь два фактора, которые обусловили зависимость между переменными x_3, x_4 и x_5, x_6. По значениям элементов остаточной матрицы видно, что переменные x_1, x_2 не могут быть описаны этими двумя факторами. Этот вывод подтверждает значение критерия χ^2=5 при критическом значении $\chi^2_{(0,05;4)}$=9,5.

Таким образом, результат эксперимента указывает на то, что при наличии очевидной зависимости между исходными переменными, в случае, если она имеет нелинейный характер, методом максимального правдоподобия выделить порождающий ее фактор не удается. Поэтому поставлена задача модифицировать метод максимального правдоподобия для определения факторных нагрузок и матрицы остаточных дисперсий для случая, когда зависимость между переменными нелинейная.

Метод решения

Для случая, когда зависимость между переменными нелинейная, вместо коэффициента корреляции Пирсона предлагается использовать индекс корреляции:

$$\eta_{xy} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}}, \quad (4)$$

где $\hat{y}_i$ — значения функции нелинейной регрессии, которую можно получить, применив МНК [8]. При этом для индекса корреляции выполняется: $0 \le \eta_{xy} \le 1$. Из индексов корреляции строится матрица Σ. Поскольку $\eta_{xy} \ne \eta_{yx}$, при формировании для симметричности матрицы Σ из двух индексов корреляции будем выбирать наибольший:

$$\eta_{ij} = \max\left\{\eta_{x_i x_j}; \eta_{x_j x_i}\right\}, \text{при } i \ne j, i, j = \overline{1,r}. \quad (5)$$

Для того, чтобы вектор наблюдений допускал интерпретацию в рамках модели факторного анализа (1), исходная матрица ковариаций (или корреляций) должна удовлетворять определенным требованиям [9, 10]. Кроме этого, существуют требования к соотношению размерности исходного пространства r и числа общих факторов k. Одним из общих требований является возможность представления исходной матрицы ковариаций в виде суммы диагональной матрицы с положительными элементами и матрицы ранга k с положительными собственными значениями. Следует отметить, что вопросы разрешимости задачи факторного анализа еще до конца не решены, поэтому покажем лишь, что свойства получаемой матрицы индексов корреляции Σ совпадают со свойствами традиционно применяемой в факторном анализе ковариационной (корреляционной) матрицы. Во-первых, в силу (5), получаемая матрица симметрична. Во вторых, неотрицательная определенность вытекает из критерия Сильвестра: $\eta_{x_1 x_1} = 1 > 0$, $1 - \eta^2_{x_1 x_2} \ge 0$ в силу свойства индекса корреляции.

После построения матрицы Σ для сгенерированных ранее данных, следует определить общности и выбрать начальное приближение в ММП. Для этого применен метод главных факторов. В результате выполнения итераций ММП получены нагрузочная матрица и матрица остаточных дисперсий:

$$L = \begin{vmatrix} 0{,}276 & 0{,}152 & \mathbf{0{,}949} \\ 0{,}248 & 0{,}103 & \mathbf{0{,}743} \\ \mathbf{0{,}995} & -0{,}078 & 0{,}009 \\ \mathbf{0{,}992} & -0{,}129 & -0{,}008 \\ 0{,}345 & \mathbf{0{,}938} & -0{,}017 \\ 0{,}401 & \mathbf{0{,}812} & -0{,}033 \end{vmatrix}$$

$$V = \begin{vmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}377 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}004 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}178 \end{vmatrix}$$

По результатам факторного анализа видно, что модифицированный ММП выделил первый фактор, порождающий нелинейную зависимость между x_1 и x_2. Значения факторных нагрузок достаточно велики (больше 0,6), нагрузки специфических факторов малы, поэтому можно говорить о том, что полученные результаты отражают реальные зависимости, которые были заложены при генерировании тестового набора данных.

Экспериментальная часть и результаты

Для демонстрации результатов работы модифицированного ММП на реальных данных с сайта www.kaggle.com выбран набор данных «Red wine quality». В наборе данных 1600 наблюдений и 11 признаков: титрируемые кислоты, летучие кислоты, лимонная кислота, сахар, соль, свободный диоксид серы, общее содержание диоксида серы, плотность (зависит от содержания алкоголя и сахара), кислотность, винно-сульфатная добавка, этанол.

Выберем число факторов, равное четырем. Факторные нагрузки, полученные методом максимального правдоподобия, приведены в таблице 1.

Таблица 1. Факторные нагрузки, полученные методом максимального правдоподобия.

ТК	**0,67**	-0,17	-0,03	**0,66**
ЛК	0,02	0,08	0,05	-0,4
Лим. к.	0,37	0,01	0,13	**0,68**
Сахар	0,36	0,17	-0,04	-0,1
Соль	0,21	0,08	0,97	0
СДС	-0,02	**0,67**	-0,05	-0,05
ОСДС	0,08	**0,99**	-0,05	0
Плотн.	**0,98**	0	-0,01	0
Ph	-0,34	-0,1	-0,19	**-0,7**
ВСД	0,15	0,05	0,34	0,2
Этанол	-0,5	-0,17	-0,1	0,32

$$V = |0{,}10{,}780{,}380{,}8300{,}55000{,}410{,}820{,}61|$$

Таблица 2. Факторные нагрузки, полученные модифицированным методом максимального правдоподобия.

ТК	**0,93**	0,37	-0,02	0
ЛК	0,22	0,22	0,02	0,36
Лим. к.	**0,6**	0,36	-0,02	0,36
Сахар	0,4	-0,26	0,31	-0,05
Соль	0,4	-0,3	0,07	**0,68**
СДС	0,22	0,13	**0,97**	0
ОСДС	0,22	0,04	**0,7**	0,12
Плотн.	**0,91**	-0,41	0	0
Ph	**0,6**	0,45	-0,04	0,33
ВСД	0,23	0,05	0,05	**0,68**
Этанол	0,54	-0,33	0,04	0,2

$$V = |00{,}80{,}380{,}680{,}2800{,}45000{,}330{,}480{,}56|$$

Анализируя факторные нагрузки, можно сделать вывод о том, что свободный диоксид серы и общее содержание диоксида серы коррелируют друг с другом. Также в зависимости находятся показатели величины титрируемых кислот и плотности, соль образует отдельный фактор, четвертый фактор связывает показатели титрируемых кислот, лимонной кислоты и кислотности. Летучие кислоты, этанол, сахар и содержание винно-сульфатной добавки не зависят от выделенных факторов. Полученные результаты не противоречат результатам, полученным в [11].

Применение модифицированного ММП к имеющемуся набору данных позволяет получить нагрузки на факторы, приведенные в таблице 2.

Как видно из результатов, этанол, сахар и летучие кислоты, как и в предыдущем случае, не зависят от выделенных факторов. Соль и винно-сульфатные добавки выделились в отдельный фактор, что логически может быть обосновано. Титрируемые кислоты, лимонная кислота, Ph и плотность так же объединились в один фактор. Таким образом, можно говорить о том, что модифицированный ММП дает несколько иные результаты, чем стандартный. Результат работы на экспериментальном наборе данных модифицированного ММП, в первом приближении, выглядит более обоснованным.

Заключение

В статье рассмотрен ММП для решения задачи факторного анализа. На экспериментальных данных показано, что традиционный ММП не всегда может выявить нелинейные зависимости между показателями. Предложено модифицировать традиционный ММП, заменив матрицу парных корреляций индексами корреляции, полученными с использованием регрессионных полиномов. Предложен способ преобразования матрицы индексов корреляции в симметричную. На реальных данных проведен эксперимент, показавший, что предложенная модификация способна улучшить результаты факторного анализа. Предложенная модификация ММП обогащает арсенал имеющихся у исследователя методов, и может оказаться полезной, например, для уточнения некоторых результатов или когда стандартные методы исследования не позволяют получить приемлемый результат.

ЛИТЕРАТУРА

1. Орлова И.В., Турундаевский В. Б. Выбор методы оценки матрицы нагрузок в факторном анализе и алгоритм оценки при нулевых нагрузках на часть специфических факторов // Фундаментальные исследования № 6, 2015. С. 161–165.
2. Харман Г. Современный факторный анализ // М.: Статистика, 1972–488 с.
3. Окунь Я. Факторный анализ // М.: Статистика, 1974–200 с.
4. Горяинова Е.Р., Шалимова Ю. А. Снижение размерности многомерных показателей с нелинейно зависимыми компонентами // Бизнес-информатика. 2015. № 3 (33). С. 24–33.
5. Шовин В. А. Нелинейные структурные уравнения и квадратичный факторный анализ // Математические структуры и моделирование 2018. № 2(46). С. 51–61.
6. Айвазян С.А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности: Справ. изд. // М.: Финансы и статистика, 1989–607 с.: ил.
7. Ивченко Г.И., Медведев Ю. И., Введение в математическую статистику: Учебник. М.: Издательство ЛКИ, 2010–600 с.
8. Ферстер Э. Ренц Б. Методы корреляционного и регрессионного анализа // М.: Финансы и статистика, 1983–303 с.
9. Anderson T.W., Rubin H. Statistical inference in factor analysis. // Proc. 3 Berkeley Symp. Math. Statist. And Probab.— Univ. Calif. Press, 1956, 5.— P. 11–50.
10. Лоули Д., Максвелл А. Факторный анализ как статистический метод // М.: Мир, 1967–145 с.
11. Умарова Н.Н., Давлетшина Ф. И., Вильданова А. И., Евгеньев М. И. Многомерный анализ качества вин // Вестник технологического университета. 2016. Т. 19, № 13 с. 145–148.

Журнал «Современная наука: актуальные проблемы теории и практики»

Г. Иркутск