СНИЖЕНИЕ РАЗМЕРНОСТИ МНОГОМЕРНЫХ ПОКАЗАТЕЛЕЙ С НЕЛИНЕЙНО ЗАВИСИМЫМИ КОМПОНЕНТАМИ

Е.Р. ГОРЯИНОВА
кандидат физико-математических наук,
доцент департамента математики, факультет экономических наук,
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: el-goryainova@mail.ru

Ю.А. ШАЛИМОВА
студентка магистратуры, факультет экономических наук,
Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20
E-mail: july.shalimova@yandex.ru

При решении задачи сжатия многомерного вектора показателей используют методы факторного анализа, одним из которых является метод максимального правдоподобия (ММП). В системе коррелированных количественных показателей он позволяет выявить некоррелированные общие факторы, которые без существенной потери информации могут представлять исходные показатели. Нахождение общих факторов проводится с помощью специального представления корреляционной матрицы наблюдаемых признаков. Однако коэффициент корреляции не определен для признаков, представленных в номинальной шкале, а для признаков, имеющих нелинейный характер зависимости, не может служить измерителем силы связи. Для таких ситуаций традиционные методы факторного анализа оказываются малоэффективными.

В статье предложены две модификации ММП, использующие в качестве мер связи признаков ранговые коэффициенты корреляции Спирмена и коэффициенты Крамера. Для сравнения качества сжатия традиционного и двух адаптированных ММП проведен численный эксперимент. С помощью метода Монте-Карло смоделированы 12-мерные векторы, состоящие из четырех независимых трехмерных подвекторов, координаты которых имеют зависимости линейного и нелинейного типа. Установлено, что из трех рассмотренных методов только адаптированный метод, использующий коэффициенты Крамера, способен верно объединить в общий фактор показатели, связанные немонотонным типом зависимости. С другой стороны, в тех случаях, когда зависимость между признаками носит монотонный характер, этот метод менее эффективен, чем два других. Для демонстрации работоспособности указанных методов на реальных данных представлено решение задачи снижения размерности динамики относительного прироста потребительских цен в 2008-2014 годах для группы продовольственных товаров.

Ключевые слова: факторный анализ, общие факторы, метод максимального правдоподобия, корреляционная матрица, матрица нагрузок, коэффициент ранговой корреляции Спирмена, коэффициент Крамера.

Цитирование: Горяинова Е.Р., Шалимова Ю.А. Снижение размерности многомерных показателей с нелинейно зависимыми компонентами // Бизнес-информатика. 2015. № 3 (33). С. 24–33.

Введение

При изучении сложных объектов исследователи пытаются описать их большим числом показателей. Как правило, это приводит к тому, что среди собранных данных имеются группы показателей, которые характеризуют одно и то же свойство объекта и поэтому являются зависимыми, а также малоинформативные показатели, которые не несут в себе существенной информации об объектах. Статистический анализ таких массивов становится затруднительным и может приводить к неверным результатам. Поэтому возникает необходимость описать наблюдаемые показатели меньшим числом интегративных показателей, сохранив при этом как можно больше важной информации об объектах.

Основная идея факторного анализа состоит в том, что структура связей между анализируемыми признаками может быть объяснена тем, что эти признаки зависят от меньшего числа других непосредственно неизмеряемых показателей, называемых общими факторами. Классическая модель факторного анализа, описанная в работе [1], предполагает, что каждая наблюдаемая переменная представляется в виде линейной комбинации некоррелированных общих факторов и одного частного фактора, оказывающего влияние только на данную переменную. Основная задача факторного анализа состоит в оценивании матрицы нагрузок, элементами которой являются корреляции между исходными показателями и общими факторами, оценивании дисперсий частных факторов и интерпретации общих факторов. Решение этой задачи позволяет в рамках факторной модели удовлетворительно воспроизводить корреляции между наблюдаемыми показателями.

Наиболее распространенными методами решения этой задачи являются метод главных факторов [2, 3], метод минимальных остатков [4] и метод максимального правдоподобия (ММП) [5]. Так, согласно методу главных факторов, требуется провести оценивание дисперсий частных факторов, а затем применить процедуры компонентного анализа [6] к редуцированной корреляционной матрице, из элементов главной диагонали которой вычтены найденные оценки дисперсий. Принцип оценивания матрицы нагрузок методом минимальных остатков основан на минимизации суммы квадратов разностей между выборочными корреляциями и корреляциями, воспроизводимыми факторной моделью с фиксированным числом факторов. В ММП предполагается, что общие и частные факторы имеют гауссовское распределение, а оценками нагрузок являются те значения, при которых достигается максимум функции правдоподобия элементов выборочной корреляционной матрицы при фиксированном числе общих факторов. Оценивание числа общих факторов в двух последних методах проводится с помощью последовательного применения хи-квадрат тестов. Заметим, что методы факторного анализа используют в качестве мер связи коэффициенты корреляции исходных показателей. Однако на практике нередко возникают задачи, в которых показатели являются зависимыми, но некоррелированными. Например, в работе [7] установлена квадратичная зависимость между вероятностью дефолта и размером активов банка. Кроме того, многие показатели в социологических и психологических исследованиях измеряются в номинальной шкале, и коэффициент корреляции для этих величин не определен. Таким образом, если компоненты вектора показателей имеют зависимости нелинейного характера или измерены в различных шкалах, то процедура снижения размерности такого вектора требует корректировки.

Объектом исследования данной работы являются методы (в частности, ММП) снижения размерности в модели факторного анализа, а предметом исследования – адаптация методов сжатия для векторов с нелинейной структурой зависимости компонент. Предлагаемая нами модификация заключается в том, что в качестве оценки неизвестной корреляционной матрицы будут использоваться матрицы коэффициентов ранговой корреляции Спирмена и матрицы коэффициентов Крамера. С помощью компьютерного моделирования будет показано, что адаптированный ММП является более эффективным для решения задачи снижения размерности многомерного вектора с нелинейно зависимыми компонентами.

Данная работа имеет следующую структуру. В разделе 1 представлена модель факторного анализа и традиционный ММП, используемый в факторном анализе. В разделе 2 описаны адаптированные ММП и процедура компьютерного моделирования случайных векторов с линейно и нелинейно зависимыми компонентами. В разделе 3 проведен сравнительный анализ качества сжатия смоделированных векторов. В разделе 4 с помощью рассмотренных методов решена задача снижения размерности показателей изменения относительного

прироста потребительских цен в 2008-2014 году для группы продовольственных товаров.

1. Задача факторного анализа

Пусть $X = (X_1,\ldots,X_r)^T$ – r-мерный вектор наблюдаемых показателей у каждого из n объектов. Обозначим вектор стандартизированных наблюдений через $\vec{x} = (x_1,\ldots,x_r)^T$, где

$$x_i = \frac{X_i - \overline{X}_{i\cdot}}{s_i},\ \overline{X}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} X_{ij},\ s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2.$$

Согласно канонической модели факторного анализа вектор $\vec{x}$ представляется в виде

$$\vec{x} = L\vec{f} + \vec{\varepsilon}, \qquad (1)$$

где L – детерминированная матрица $r \times k$, $k < r$, $\vec{f} = (f_1,\ldots,f_k)^T$ – случайный вектор центрированно-нормированных некоррелированных общих факторов, $\vec{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_r)^T$ – случайный вектор центрированных частных факторов, таких что, коэффициенты корреляции $\rho(\varepsilon_i,\varepsilon_j) = 0$; $\rho(\varepsilon_i,f_m) = 0$; $i, j = 1, ..., r$; $m = 1, ..., k$.

Из формулы (1) следует, что ковариационная матрица C вектора $\vec{x}$ удовлетворяет соотношению

$$C = LL^T + V, \qquad (2)$$

где V – диагональная матрица размера $r \times r$ с диагональными элементами $D\varepsilon_i = v_i$, а элементы l_{ij}, $i = 1, ..., r$; $j = 1, ..., k$ матрицы L являются коэффициентами корреляции между признаками x_i и факторами f_j, то есть $l_{ij} = \rho(x_i, f_j)$. По этой причине L называют матрицей нагрузок.

Предположим дополнительно, что вектор общих факторов $\vec{f} \sim N(0, I)$, I – единичная матрица размера $k \times k$, а $\vec{\varepsilon} \sim N(0, V)$.

Основная задача факторного анализа состоит в оценивании матрицы нагрузок L и дисперсий v_i, $i = 1, ..., r$. Выше было сказано о том, что разработано несколько методов решения этой задачи. Поскольку в данной работе моделируются гауссовские показатели, для решения задачи будет использован оптимальный в этой ситуации ММП, дающий асимптотически эффективные оценки указанных параметров [3].

Традиционно в качестве оценок элементов матрицы C используются выборочные ковариации, построенные по результатам n наблюдений вектора $\vec{x} = (x_1,\ldots,x_r)^T$. Обозначим через A матрицу выборочных ковариаций с элементами

$$a_{ij} = \frac{1}{n}\sum\nolimits_{m=1}^{n} x_{im}x_{jm},\ i,j = 1,..., r.$$

Следуя ММП, для оценивания l_{ij} и v_i, $i = 1, ..., r$; $j = 1, ..., k$ нужно выписать совместную плотность элементов матрицы A, прологарифмировать ее и найти те значения l_{ij} и v_i, при которых достигается максимальное значение логарифмической функции правдоподобия. Как показано в работе [8], решение этой задачи сводится к нахождению собственных векторов матрицы $V^{-1}(A - V)$, найти которые можно с помощью итерационной процедуры. Соответствующий итерационный алгоритм был реализован нами в среде Matlab и подробно описан в работе [9]. Отметим, что поскольку $\vec{x}$ – стандартизированный, то ковариационная матрица C является корреляционной, а A – выборочной корреляционной матрицей. Вообще говоря, ММП позволяет выбирать в качестве матрицы C как ковариационную матрицу, так и корреляционную.

Заметим, что L и f в формуле (1) определяются с точностью до вращения, цель которого в получении качественной интерпретации факторов. Наиболее распространенными методами вращения являются варимакс и квартимакс [10].

Еще одной проблемой при решении задачи факторного анализа является выбор числа общих факторов k. Существует несколько способов решения этой задачи, как теоретически обоснованных, так и эмпирических. Если в факторном анализе применяется ММП, то определение числа общих факторов основывается на проверке статистической гипотезы о том, что число общих факторов равно заданной величине k. Тестовая статистика отношения правдоподобия при сделанных предположениях имеет распределение хи-квадрат.

2. Адаптация ММП для нелинейно зависимых показателей

Как показано в предыдущем разделе, модель факторного анализа предполагает, что значения признаков линейно зависят от общих факторов, а в качестве меры связи самих признаков используются коэффициенты корреляции. Если же признаки связаны нелинейной зависимостью или измеряются в номинальной шкале, то коэффициент корреляции теряет свою информативность как измеритель силы связи. Поэтому в качестве мер связи таких признаков надо использовать другие коэффициенты, например, коэффициент ранговой корреляции Спирмена [11] или коэффициент Крамера [12].

Коэффициентом ранговой корреляции Спирмена ρ_{YZ} случайных величин Y и Z, построенным по наблюдениям (Y_1, Z_1), ..., (Y_n, Z_n), называется статистика

$$\rho_{YZ} = \frac{\sum_{m=1}^{n}\left(R_m - \frac{n+1}{2}\right)\left(S_m - \frac{n+1}{2}\right)}{\sqrt{\sum_{m=1}^{n}\left(R_m - \frac{n+1}{2}\right)^2 \sum_{m=1}^{n}\left(S_m - \frac{n+1}{2}\right)^2}},$$

в которой R_m – ранг элемента Y_m в выборке Y_1, ..., Y_n, а S_m – ранг элемента Z_m в выборке Z_1, ..., Z_n.

Отметим, что ρ_{YZ} может служить оценкой степени монотонной зависимости между величинами Y и Z [13]. Обозначим через P матрицу с элементами ρ_{ij}, $1 \le i, j \le r$, где $\rho_{ij} = \rho_{x_i x_j}$ – ранговый коэффициент корреляции Спирмена показателей x_i и x_j.

Дадим определение коэффициента Крамера для наблюдений (Y_1, Z_1), ..., (Y_n, Z_n) двумерного вектора (Y, Z). Для этого разобьем область V_Y возможных значений величины Y на l непересекающихся интервалов $\Delta_{Y,i}$, $i = 1, ..., l$, так, что $\bigcup_{i=1}^{l}\Delta_{Y,i} = V_Y$, а область V_Z возможных значений величины Z на s непересекающихся интервалов $\Delta_{Z,j}$, $j = 1, ..., s$, так, что $\bigcup_{j=1}^{s}\Delta_{Z,j} = V_Z$. Пусть n_{ij} – число пар выборки (Y_1, Z_1), ..., (Y_n, Z_n), попавших в прямоугольник $\Delta_{Y,i} \times \Delta_{Z,j}$, $i = 1, ..., l, j = 1, ..., s$.

Обозначим $n_{i.} = \sum_{j=1}^{s} n_{ij}$, а $n_{.j} = \sum_{i=1}^{l} n_{ij}$.

Тогда коэффициент Крамера определяется как

$$k_{YZ} = \sqrt{\frac{\widehat{\chi}^2_{YZ}}{n \cdot min\{(l-1),(s-1)\}}}, \text{ где}$$

$$\widehat{\chi}^2_{YZ} = n\sum_{i=1}^{l}\sum_{j=1}^{s}\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{n_{i.}n_{.j}} -$$

статистика критерия хи-квадрат. В работе [12] показано, что коэффициент Крамера, принимающий значения в интервале [0,1], может служить мерой, характеризующей силу связи между признаками Y и Z. Обозначим через K матрицу с элементами k_{ij}, $1 \le i, j \le r$, где $k_{ij} = k_{x_i x_j}$ — коэффициент Крамера показателей x_i и x_j.

Рассмотрим следующие две модификации ММП. Назовем «модификацией 1» адаптированный ММП, в котором матрица выборочных коэффициентов корреляции A заменяется матрицей коэффициентов Спирмена P, и, соответственно, «модификацией 2» – адаптированный ММП, в котором матрица A заменена матрицей коэффициентов Крамера K. Наше предположение состоит в том, что при наличии монотонных, но нелинейных зависимостей между компонентами вектора $\vec{x}$ задачу выделения общих факторов эффективнее решать, используя модификацию 1, а при наличии нелинейных немонотонных связей – модификацию 2. Это предположение проверяется на тестовых данных с помощью обширного численного эксперимента.

В рамках эксперимента 12-мерные векторы $\vec{x} = (x_1, \ldots, x_{12})^T$ были сгенерированы таким образом, чтобы компоненты вектора образовывали 4 независимые группы по 3 признака в каждой группе. При этом признаки первой группы сильно коррелированны между собой, признаки второй группы связаны «зашумленной» функциональной зависимостью линейного типа, признаки третьей группы связаны «зашумленной» функциональной зависимостью нелинейного монотонного типа, а признаки четвертой группы – «зашумленной» функциональной зависимостью немонотонного типа.

Принцип моделирования коррелированных величин базируется на использовании следующего свойства, доказанного в работе [9]. Если случайные величины Y и W независимы и имеют конечные дисперсии, а величина $Z = \alpha W + Y$, то коэффициент корреляции $\rho_{ZW} = \rho$ величин Z и W связан с константой α соотношением

$$\alpha = \sqrt{\frac{\rho^2}{1-\rho^2}} \cdot \sqrt{\frac{DY}{DW}} sign(\rho). \qquad (3)$$

Теперь с помощью встроенного в Matlab датчика генерируется стандартная нормальная случайная величина $x_1 \sim N(0; 1)$; затем, используя соотношение (3), генерируется $x_2 \sim N(0; 1)$, такая что $\rho_{x_1 x_2} = 0,7$; затем $x_3 \sim N(0; 1)$, такая что $\rho_{x_3 x_2} = 0,7$.

Принцип генерации второй, третьей и четвертой групп следующий. Пусть случайные величины α_1, α_2, α_3 имеют усеченное стандартное нормальное распределение, а величины ε_1, ..., $\varepsilon_9 \sim N(0; 1)$. Тогда значения признаков x_4, ..., x_{12} вычисляются по следующим формулам:

$$x_4 = \alpha_1 + \varepsilon_1, \quad x_5 = f(\alpha_1) + \varepsilon_2, \quad x_6 = f(f(\alpha_1)) + \varepsilon_3,$$
$$x_7 = \alpha_2 + \varepsilon_4, \quad x_8 = g(\alpha_2) + \varepsilon_5, \quad x_9 = g(g(\alpha_2)) + \varepsilon_6,$$
$$x_{10} = \alpha_3 + \varepsilon_7, \quad x_{11} = h(\alpha_3) + \varepsilon_8, \quad x_{12} = h(h(\alpha_3)) + \varepsilon_9,$$

где функция $f(\cdot)$ — линейная функция, $g(\cdot)$ — нелинейная монотонная функция, $h(\cdot)$ — нелинейная функция. Реализации значений пар признаков для каждой из четырех групп объема 10 000 представлены на *рис. 1.*

Помимо указанных модификаций ММП потребовалось применить другой способ определения числа общих факторов, так как критерий, основанный на статистике отношения правдоподобия,

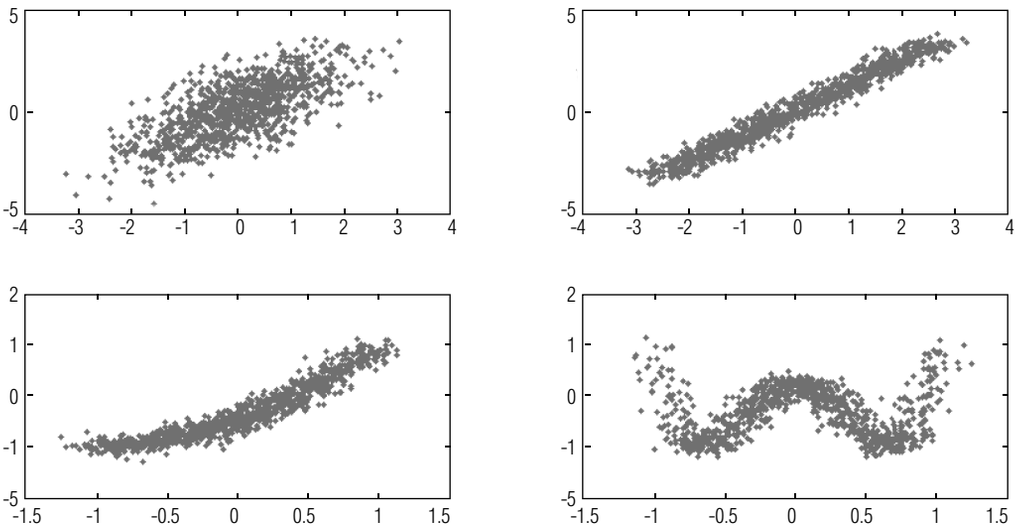

Рис. 1. Реализации признаков в группах демонстрационных данных

оказался неработоспособным на моделированных данных. Этот факт объясняется тем, что тестовая статистика имеет распределение хи-квадрат в случае гауссовских наблюдений, а компоненты x_7, ..., x_{12} сгенерированного вектора $\vec{x}$ являются нелинейными преобразованиями гауссовских случайных величин и, следовательно, не являются гауссовскими. Поэтому для определения числа общих факторов нами был реализован следующий эмпирический метод.

На первом шаге применяется ММП с числом общих факторов равным числу признаков. Затем для полученной матрицы нагрузок L вычисляются коэффициенты

$$\mu_j = \sqrt{\sum_{i=1}^{r} l_{ij}^2}, j = 1,\ldots,r. \qquad (4)$$

Каждый из коэффициентов μ_j показывает количество суммарного среднеквадратического отклонения признаков, которое объясняется добавлением j-го фактора к уже имеющимся $j-1$ факторам $f_1,\ldots,f_{j-1}$. В случае нормированных признаков положим число общих факторов равным k, если $\mu_k \geq 1$, а $\mu_{k+1} < 1$. На втором шаге запускается алгоритм ММП с выбранным числом факторов. Обоснование такого способа выбора приведено в работе [9].

3. Сравнительный анализ традиционного и адаптированных ММП

Перейдем к представлению результатов сжатия вектора $\vec{x} = (x_1,\ldots,x_{12})^T$, структура которого описана в разделе 2. Последовательно применим к демонстрационным данным все три метода с максимальным числом общих факторов равным 12. Значения μ_1, ..., μ_{12}, вычисленные по формуле (4), для традиционного ММП и двух его модификаций представлены на *рис. 2, 3* и *4* соответственно.

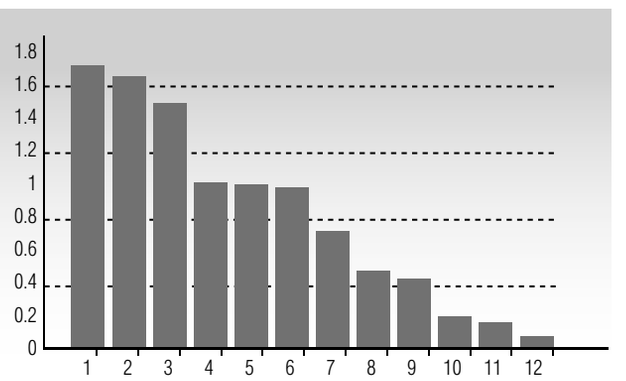

Рис. 2. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для традиционного метода

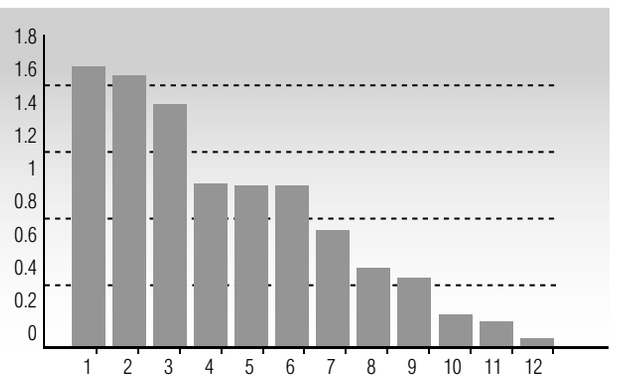

Рис. 3. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для модификации 1

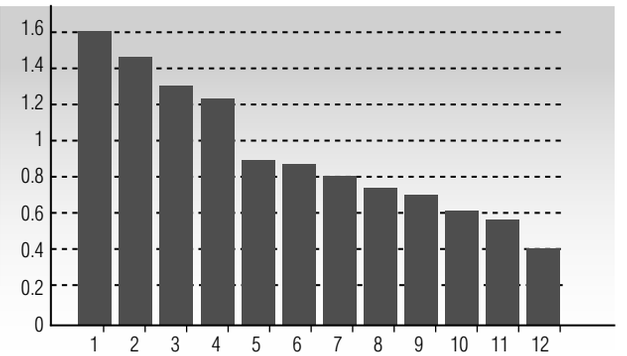

Рис. 4. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для модификации 2

Согласно *рис. 2* и *3*, значение $\mu_j > 1$ получено при $j = 1, 2, 3$ и $\mu_j \approx 1$ при $j = 4, 5, 6$. Поэтому для традиционного ММП и модификации 1 считаем число общих факторов $k = 6$. Матрица нагрузок традиционного ММП имеет следующий вид:

-0.0022	-0.0042	**0.7144**	0.0057	-0.0010	0.0077
-0.0047	0.0077	**0.9845**	-0.0001	0.0009	-0.0001
0.0003	-0.0046	**0.7079**	-0 0045	0 0180	-0.0007
0.0033	**0.9842**	0.0009	0.0000	-0.0018	0.0012
0.0020	**0.9957**	-0.0001	0.0011	0.0004	-0.0004
0.0028	**0.9989**	-0.0006	-0.0003	0.0001	-0.0000
-0.9379	0.0017	-0.0040	0.0039	-0.0042	-0.0046
-0.9994	0.0018	-0.0001	-0.0001	-0.0000	0.0000
-0.9217	0.0068	-0.0047	-0.0029	0.0013	0.0052
0.0016	0.0099	-0.0023	-0.0160	-0.0232	**0.8265**
-0.0136	-0.0026	0.0041	**0.8483**	-0.1786	0.0078
0.0130	0.0097	0.0158	-0.1914	**-0.8366**	-0.0208

Видно, что этот метод правильно определяет общие факторы, соответствующие группе признаков с монотонным нелинейным типом зависимости (высокие нагрузки этих признаков на первый фактор выделены в столбце 1), группе с линейным типом зависимости (высокие нагрузки этих признаков на второй фактор выделены в столбце 2) и группе сильно коррелированных признаков (высокие нагрузки этих признаков на третий фактор выделены в столбце 3). Признаки x_{10}, x_{11}, x_{12} имеют высокие нагрузки на шестой, четвертый и пятый факторы соответственно. Таким образом, традиционный метод выделяет в отдельные факторы признаки связанные немонотонным типом зависимости.

Матрица нагрузок модификации 1 имеет следующий вид:

0.0045	-0.0070	**0.6911**	0.0080	-0.0020	0.0067
0.0099	-0.0081	**0.9782**	-0.0024	-0.0008	0.0004
0.0133	-0.0100	**0.6984**	0.0017	-0.0033	0.0118
0.9850	-0.0016	-0.0024	0.0002	-0.0013	0.0022
0.9964	-0.0015	0.0005	-0.0007	0.0002	-0.0004
0.9989	-0.0005	-0.0005	0.0002	0.0000	-0.0000
-0.0102	**-0.9792**	-0.0026	-0.0003	0.0008	-0.0014
-0.0094	**-0.9787**	-0.0047	0.0003	-0.0024	0.0005
-0.0094	**-0.8860**	-0.0084	0.0078	-0.0063	0.0061
0.0012	-0.0013	0.0164	0.0018	-0.0174	**-0.7782**
-0.0030	-0.0074	-0.0096	**-0.8918**	-0.0138	-0.0005
-0.0029	0.0211	-0.0081	0.0196	**-0.8486**	0.0112

Этот способ также верно выделяет три группы зависимых признаков – признаки с линейным типом зависимости (первый фактор), признаки с монотонным нелинейным типом зависимости (второй фактор) и сильно коррелированные признаки (третий фактор). Как и традиционный ММП, модификация 1 не выявляет четвертую группу признаков, связанных немонотонным типом зависимости.

На *рис. 4* видно, что больше единицы оказались значения μ только для четырех факторов. Матрица нагрузок для модификации 2 при $k = 4$ имеет следующий вид:

0.0343	-0.0386	0.1032	**-0.4380**
0.0338	-0.0503	0.1406	**-0.6847**
0.0300	-0.0499	0.1020	**-0.4317**
0.7777	0.0080	-0.0028	0.0003
0.9019	0.0168	-0.0108	0.0059
0.9278	0.0209	-0.0091	0.0044
0.0415	**-0.7596**	-0.0391	0.0163
0.0413	**-0.8061**	-0.0485	0.0172
0.0376	**-0.6744**	-0.0318	0.0138
0.0305	-0.0502	**0.4291**	0.0461
0.0388	-0.0682	**0.7326**	0.1241
0.0373	-0.0638	**0.5750**	0.0875

Из трех рассмотренных способов только этот способ верно выделяет все четыре группы зависимых признаков. Так, в первый фактор выделены признаки с линейным типом зависимости, во второй – признаки с нелинейным монотонным типом зависимости, в третий – признаки с немонотонным типом зависимости, а в четвертый – сильно коррелированные признаки. Однако следует заметить, что нагрузки для групп показателей с монотонными типами связи ниже, чем у двух предыдущих методов.

Отметим, что на других смоделированных данных аналогичной структуры представленный эмпирический метод определения числа факторов продемонстрировал адекватные результаты. Применение к матрицам нагрузок методов вращения не внесло существенных изменений.

При попытке задать в традиционном ММП и в модификации 1 число общих факторов $k = 4$ были получены матрицы нагрузок, у которых в четвертый фактор выделялась лишь одна из компонент четвертого подвектора.

4. Пример с реальными данными

Продемонстрируем работу трех рассмотренных методов на реальных данных. Для демонстрации эффективной работы методов сжатия многомерных признаков хотелось выбрать такие показатели, чтобы наличие зависимости между ними было в значительной степени предсказуемо из соображений здравого смысла. Мы выбрали еженедельные средние потребительские цены на некоторые продукты питания за период с января 2008 г. по апрель 2014 г. В данном случае признаками являются цены на конкретные товары, а наблюдениями – цены на товары в фиксированные моменты времени. Согласно модели факторного анализа, наблюдения за каждым признаком должны быть независимы и одинаково распределены. Но, поскольку цены на товары растут с течением времени, то в качестве реализации X_{ij} i-го признака для j-го наблюдения будем рассматривать не саму цену i-го товара в момент времени j (обозначим ее c_{ij}), а величину относительного прироста цены, т.е.

$$X_{ij} = \frac{c_{ij} - c_{i(j-1)}}{c_{i(j-1)}}.$$

В качестве признаков были выбраны относительные приросты цен на следующие товары: говядина, сосиски и сардельки, колбаса полукопченая и варено-копченая, колбаса вареная I сорта, говядина и свинина тушеная консервированная, масло сливочное, сметана, творог жирный, сыры сычужные твердые и мягкие, мука пшеничная, хлеб и булочные изделия из пшеничной муки. Еженедельные средние потребительские цены на эти продукты за указанный период взяты с сайта Федеральной службы государственной статистики (*www.gks.ru*). Понятно, что первые пять продуктов образуют «мясную» группу, следующие четыре продукта – «молочную» группу, а последние два продукта – «мучную» группу.

Применим последовательно все три способа сжатия к имеющимся данным. Для определения числа общих факторов вычислим для каждого метода коэффициенты μ_1, ..., μ_{11} по формуле (4). Для обеих модификаций значения больше единицы имели первые три коэффициента, поэтому число общих факторов $k = 3$. У традиционного ММП близким к единице оказался и μ_4, что вызывает некоторые сомнения относительно включения четвертого фактора. Мы приняли решение о включении трех факторов. Отметим, что в отличие от моделированных данных, для реальных данных потребовалось применить методы вращения нагрузочной матрицы. Это позволило существенно улучшить интерпретируемость результатов каждого из трех методов. Поэтому опустим представление матриц нагрузок, полученных до процедуры вращения.

Матрица нагрузок традиционного ММП после вращения имеет следующий вид:

0.7837	0.0763	-0.1608
0.7953	0.0406	-0.2426
0.8927	0.0537	-0.0971
0.4513	-0.0657	0.1389
0.6518	-0.0541	0.0374
-0.0711	0.0205	-0.6946
0.0975	-0.4298	-0.4831
0.2239	-0.0078	-0.8595
0.0544	0.2281	-0.7201
-0.0472	-0.9172	-0.0053
0.0166	-0.7206	0.1819

Как и ожидалось, признаки отчетливо объединились в три группы. Первый фактор объединяет продукты «мясной» группы, второй – «мучной» группы, а третий – «молочной». Однако из общей картины несколько выбиваются строки, соответ-

ствующие приросту цен на вареную колбасу (строка 4) и сметану (строка 7). Видно, что прирост цен на колбасу имеет существенно меньшую нагрузку на «мясной» фактор, чем остальные признаки из этой группы, а прирост цен на сметану имеет немалую нагрузку 0,429 и в «мучной» группе.

Матрица нагрузок модификации 1 после вращения имеет следующий вид:

0.0273	-0.6796	-0.0774
0.1401	-0.7700	-0.0513
0.0854	-0.7521	-0.0426
0.0247	-0.7064	-0.1153
0.0110	-0.7383	-0.1250
0.7496	-0.0413	-0.0142
0.7817	-0.1850	-0.0820
0.8146	-0.1559	-0.1012
0.7728	0.1204	0.1601
0.0432	-0.0625	-0.7701
-0.0180	-0.1533	-0.6487

Этот способ также позволяет явно выделить три фактора, соответствующих «молочной» (первый фактор), «мясной» (второй фактор) и «мучной» (третий фактор) группам. Но, в отличие от результатов традиционного ММП, четвертая и седьмая строки, соответствующие вареной колбасе и сметане, мало отличаются от других строк своих групп. То есть, разбиение строк на группы «похожести» оказывается более четким, чем в традиционном методе.

Матрица нагрузок модификации 2 после вращения имеет следующий вид:

0.4365	0.1265	–0.1382
0.5407	0.1425	–0.0716
0.5426	0.1227	–0.1277
0.5381	0.1130	–0.1235
0.5009	0.1265	–0.1350
0.1331	0.5229	–0.1304
0.1839	0.5642	–0.0820
0.1573	0.5590	–0.0389
0.0704	0.4679	–0.2211
0.1558	0.1338	–0.4246
0.1625	0.1216	–0.5062

Этот способ также правильно выделяет три фактора, причем картина разбиения признаков на похожие группы достаточно отчетливая. Однако все признаки имеют на «свои» факторы меньшие нагрузки, чем в двух предыдущих матрицах.

Заключение

В данной работе рассмотрена задача снижения размерности многомерного вектора показателей. При решении этой задачи применен традиционный ММП и две модификации этого метода, использующие в качестве мер связи признаков ранговые коэффициенты корреляции Спирмена (модификация 1) и коэффициенты Крамера (модификация 2). Для сравнения качества сжатия этими методами проведен численный эксперимент, в ходе которого сгенерированы 12-мерные случайные векторы, состоящие из четырех независимых подвекторов. При этом компоненты первого подвектора являлись сильно коррелированными, компоненты второго – связанными «зашумленной» функциональной зависимостью линейного типа, компоненты третьего – связанными «зашумленной» функциональной зависимостью монотонного нелинейного типа, а компоненты четвертого – немонотонной «зашумленной» функциональной зависимостью. Оказалось, что традиционный ММП достаточно хорошо выделяет в общие факторы коррелированные признаки и признаки, связанные зависимостями линейного и монотонного типа. Однако этот метод не способен выделить в единую группу признаки, связанные немонотонной зависимостью. Модификация 1 показала аналогичные результаты, и только модификация 2 правильно выделила все четыре группы связанных признаков. Это объясняется тем, что коэффициенты Крамера, использованные в модификации 2, основаны на статистике критерия хи-квадрат, который является состоятельным против любого вида альтернатив о зависимости случайных величин. Критерии же, основанные на выборочном коэффициенте корреляции, используемом в качестве меры связи признаков в традиционном методе, или на ранговом коэффициенте Спирмена, используемом в модификации 1, являются состоятельными лишь против альтернатив о линейной или монотонной зависимости признаков соответственно. Однако универсальность коэффициента Крамера имеет и негативную сторону: его применение при выявлении линейных и монотонных зависимостей менее эффективно, чем использование коэффициента корреляции.

Рассмотренные методы показали адекватные результаты в практической задаче снижения размерности вектора относительного прироста цен на продовольственные товары. Поскольку все три способа сжатия выделили одинаковые факторы, следует признать, что истинные зависимости между показателями имеют монотонный характер. Наиболее четкую структуру матрицы нагрузок показала модификация 1. Этот факт, по-видимому, говорит о том, что существенный вклад в вариацию признаков вносят частные факторы, а коэффициенты Спирмена, как более робастные оценки истинных коэффициентов корреляции, лучше улавливают наличие линейной зависимости зашумленных данных, чем выборочные коэффициенты корреляции. ■

Литература

1. Anderson T.W., Rubin H. Statistical inference in factor analysis // Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol. 5. Berkeley: University of California Press. 1956. P. 111–150.
2. Harman H. Modern factor analysis. Chicago: University of Chicago Press, 1960. 469 p.
3. Прикладная статистика: Классификации и снижение размерности / С.А. Айвазян и [др.]; под ред. С.А. Айвазяна. М.: Финансы и статистика, 1989. 607 с.
4. Harman H., Jones W. Factor analysis by minimizing residuals (minres) // Psychometrika. 1966. Vol. 31, No. 3. P. 351–369.
5. Lawley D., Maxwell A.F. Factor analysis as astatistical method. London: Butterworths, 1963. 145 с.
6. Лагутин М.Б. Наглядная математическая статистика. М.: Бином. Лаборатория знаний, 2007. 472 с.
7. Карминский А.Н., Костров А.В. Моделирование вероятности дефолта российских банков: расширенные возможности // Журнал Новой Экономической Ассоциации. 2013. № 1. С. 64–86
8. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. М.: ЛКИ, 2010. 600 с.
9. Горяинова Е.Р., Шалимова Ю.А. Снижение размерности показателей смешанной структуры / Препринт WP7/2014/08, серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике». М.: ИД ВШЭ, 2014. – 40 с.
10. Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким и [др.]. М.: Финансы и статистика, 1989. 216 с.
11. Kendall M.G. Rank correlation methods. London: Griffin, 1970. 272 p.
12. Cramer G. (1961) Matematical methods of statistics. NY: Princeton, 1961. 575 p.
13. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: ИД ВШЭ, 2012. 310 с.

REDUCING THE DIMENSIONALITY OF MULTIVARIATE INDICATORS CONTAINING NON-LINEARLY DEPENDENT COMPONENTS

Elena R. GORYAINOVA
Associate Professor, Department of Mathematics, Faculty of Economic Sciences, National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: el-goryainova@mail.ru

Julia A. SHALIMOVA
Graduate Student, Faculty of Economic Sciences, National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow, 101000, Russian Federation
E-mail: july.shalimova@yandex.ru

To solve the problem of reduction of the multidimensional vector of indicators methods of factor analysis are used. One of them is the maximum likelihood method (MLM). It allows to identify uncorrelated common factors among the set of correlated quantitative indicators. The uncorrelated common factors can represent initial indicators without significant loss of information. Common factors are detected using a special representation of the correlation matrix of the observed indicators. However, the correlation coefficient is not defined for the characteristics measured in a nominal scale. In addition, it cannot serve as a measure for the strength of the coupling indicators with nonlinear dependence. Traditional methods of factor analysis are ineffective for such situations. Two MLM modifications are proposed in the paper. They use the rank Spearman correlation coefficients and Cramer coefficients as measures of relationship between variables. 12-dimensional vectors with their coordinates dependent on each other with linear and nonlinear dependency were simulated, using the Monte Carlo method. Then a comparative analysis of the effectiveness of the traditional MLM and the two proposed modifications of the MLM was carried out for these data. It is shown that only adapted method that uses the Cramer coefficients is able to combine correctly the indicators related with nonmonotonic dependency in the common factor. On the other hand, this method has a lower efficiency than the other two methods in the cases where the dependency between variables is linear or monotonic. To demonstrate the efficiency of these methods on real data, the task of reducing the dimension of the dynamics of the relative consumer price growth in the years 2008-2014 for a group of food products has been solved.

Key words: factor analysis, common factors, the maximum likelihood method, correlation matrix, matrix of loadings, Spearman rank correlation coefficient, Cramer coefficient.

Citation: Goryainova E.R., Shalimova Ju.A. (2015) Snizhenie razmernosti mnogomernyh pokazatelei s nelineino zavisimymi komponentami [Reducing the dimensionality of multivariate indicators containing non-linearly dependent components]. *Business Informatics*, no. 3 (33), pp. 24–33 (in Russian).

References

1. Anderson T. W., Rubin H. (1956) Statistical inference in factor analysis. Proceedings of the *Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 5. Berkeley: University of California Press, pp. 111–150.
2. Harman H. (1960) *Modern factor analysis*. Chicago: University of Chicago Press.
3. Ajvazyan S.A., Buhshtaber V.M., Enukov I.S., Meshalkin L.D. (1989) *Prikladnaya statistika: Klassifikaciya i snizenie razmernosti* [Applied statistics: Classification and reducing the dimension]. Moscow: Finansy i statistika (in Russian).
4. Harman H., Jones W. (1966) Factor analysis by minimizing residuals (minres), *Psychometrika*, vol. 31, no. 3, pp. 351–369.
5. Lawley D., Maxwell A.F. (1963) *Factor analysis as astatistical method*. London: Butterworths.
6. Lagutin M.B. (2007) *Naglyadnaya matematicheskaya statistika* [Visual mathemathical statistics]. Moscow: Binom. Laboratoria znanij (in Russian).
7. Karminsky A., Kostrov A. (2013) Modelirovanie verojatnosti defolta rossijskih bankov: rasshirennye vozmozhnosti [Modeling the default probabilities of Russian banks: Extended abillities], *Journal of the New Economic Association*, no. 1 (17), pp. 64–86 (in Russian).
8. Ivchenko G.I., Medvedev Yu.I. (2010) *Vvedenie v matematicheskuju statistiku* [Introduction to mathematical statistics]. Moscow: LKI (in Russian).
9. Goryainova E., Shalimova Ju. (2014) *Snizhenie razmernosti pokazatelej smeshannoj struktury* [Reduction of dimensionality for the indicators that have a mixed structure]. Working paper WP7/2014/8. Moscow: HSE (in Russian).
10. Kim J.-O., Mueller C.U., Klecka C. (1989) *Faktornij, diskriminantnij i klasternij analiz* [Factor, discriminant and cluster analysis]. Moscow: Finansy i statistika (in Russian).
11. Kendall M.G. (1970) *Rank correlation methods*, London: Griffin.
12. Cramer G. (1961) *Matematical methods of statistics*. NY: Princeton.
13. Goryainova E.R., Pankov A.R., Platonov E.N. (2012) *Prikladnye metody analiza statisticheskih dannyh* [Applied methods of statistical data analysis]. Moscow: HSE (in Russian).