ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Е.Р. Горяинова, Ю.А. Шалимова

СНИЖЕНИЕ РАЗМЕРНОСТИ МНОГОМЕРНЫХ ПОКАЗАТЕЛЕЙ СМЕШАННОЙ СТРУКТУРЫ

Препринт WP7/2014/08
Серия WP7
«Математические методы
анализа решений в экономике,
бизнесе и политике

Москва
2014

УДК 519.2
ББК 65в6
Г71

Редакторы серии WP7
«Математические методы анализа решений
в экономике, бизнесе и политике»
Ф.Т. Алескеров, В.В. Подиновский, Б.Г. Миркин

Горяинова, Е. Р., Шалимова, Ю. А.
Г71 Снижение размерности многомерных показателей смешанной структуры : препринт WP7/2014/08 [Текст] / Е. Р. Горяинова, Ю. А. Шалимова ; Нац. исслед. ун-т «Высшая школа экономики». – М. : Изд. дом Высшей школы экономики, 2014. – (Серия WP7 «Математические методы анализа решений в экономике, бизнесе и политике»). – 40 с. – 10 экз.

В работе рассмотрена задача снижения размерности многомерных векторов, компоненты которых могут измеряться в различных шкалах и иметь зависимости нелинейного характера. Для выявления общих факторов в такой системе показателей предложены модификации метода максимального правдоподобия (ММП), использующие в качестве мер связи признаков ранговые коэффициенты корреляции Спирмена и коэффициенты Крамера. С помощью обширного численного эксперимента проведён сравнительный анализ эффективности традиционного ММП и двух его модификаций. Установлено, что из трёх рассмотренных методов только адаптированный метод, использующий коэффициенты Крамера, способен верно объединить в общий фактор показатели, связанные немонотонным типом зависимости. Кроме того, коэффициенты Крамера, в отличие от коэффициентов корреляции, определены и для признаков, измеренных в номинальной шкале. Для демонстрации работоспособности указанных методов на реальных данных представлено решение задачи снижения размерности показателей динамики относительного прироста потребительских цен для группы продовольственных товаров.

УДК 519.2
ББК 65в6

Горяинова Е.Р. – Департамент математики факультета экономики НИУ ВШЭ, Москва, Россия.
Шалимова Ю.А. – студентка магистратуры факультета экономики НИУ ВШЭ, Москва, Россия.

**Препринты Национального исследовательского университета
«Высшая школа экономики» размещаются по адресу: http://www.hse.ru/org/hse/wp**

1. Введение

При изучении объектов, каждый из который характеризуется большим количеством признаков, часто возникает необходимость описать эти объекты значительно меньшим числом показателей, сохранив при этом как можно больше важной информации об объектах. Такие новые латентные (ненаблюдаемые явно) показатели принято называть общими факторами. Введение общих факторов преследует в основном следующие цели: 1) образование наиболее информативных интегративных характеристик (таких, например, как уровень жизни, успешность компании, состояние здоровья, размер одежды и т.д.); 2) визуализация изучаемых объектов с помощью проектирования значений их признаков на специальным образом выбранное двумерное или трехмерное факторное пространство; 3) сокращение объема хранимой информации. Задача факторного анализа состоит в нахождении небольшого числа некоррелированных латентных факторов, которые без существенной потери информации могут описывать наблюдаемые многомерные показатели. Выделение общих факторов в количественной шкале проводится с помощью специального представления корреляционной матрицы исходных показателей. Однако на практике нередко возникают задачи, в которых измеряемые показатели являются зависимыми, но некоррелированными. Например, при проведении исследований зависимости между возрастом и заработной платой работника часто оказывается, что выборочный коэффициент корреляции между этими показателями незначимо отличается от нуля. Однако с помощью критерия хи-квадрат можно установить, что зависимость между этими показателями имеется. Этот факт объясняется тем, что зависимость между возрастом и заработной платой нелинейная. А коэффициент корреляции не является измерителем силы связи количественных показателей, если зависимость между ними имеет нелинейный характер.

Есть и ещё одна причина, которая заставляет использовать иные меры связи признаков. Она обусловлена тем, что многие показатели в социологических и психологических исследованиях измеряются в но-

минальной шкале, и коэффициент корреляции для таких величин не определён. Например, в [Горяинова, Слепнёва, 2012] выявлена зависимость между участием респондентов в благотворительной деятельности и такими социально-демографическими характеристиками, как пол, возраст, уровень образования, регион проживания и тип населённого пункта проживания. Для того чтобы описать социально-демографический «портрет» респондента, участвующего в благотворительной деятельности, неким интегративным показателем, требуется использовать меры, которые характеризуют степень зависимости указанных номинальных и ординальных признаков.

Таким образом, если компоненты многомерного вектора показателей имеют зависимости нелинейного характера или измерены в различных шкалах (то есть имеют смешанную структуру), то процедура снижения размерности такого вектора требует корректировки. Мы предлагаем адаптировать известный в факторном анализе метод максимального правдоподобия (ММП). Модификация ММП заключается в том, что в качестве оценки неизвестной корреляционной матрицы мы будем использовать 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 2 описана модель факторного анализа, традиционный ММП, используемый в факторном анализе, а также методы определения числа общих факторов и способы вращения факторного пространства. В разделе 3 описаны адаптированные ММП и процедура компьютерного моделирования случайных векторов с линейно и нелинейно зависимыми компонентами. В разделе 4 проведено выделение общих факторов для смоделированных векторов с помощью традиционного и адаптированных методов, введён критерий эффективности различных методов сжатия и про ведён сравнительный анализ эффективности рассмотренных методов. В разделе 5 представлен пример с реальными данными, описывающий применение рассмотренных методов в задаче снижения размерности показателей изменения относительного прироста потребительских цен

в 2008–2014 гг. для группы из одиннадцати продовольственных товаров.

2. Традиционный метод максимального правдоподобия в факторном анализе

Пусть $X = (X_1, \dots, X_r)$ — r-мерный вектор наблюдаемых показателей у каждого из n объектов. Требуется найти такие некоррелированные показатели $f_1, \dots, f_k$, $k < r$, которые объясняют максимально возможную долю изменчивости наблюдаемых показателей $X_1, \dots, X_r$.

С этой целью сначала проведём центрирование и нормирование координат вектора X, используя построенные по выборкам объёма n оценки математического ожидания и дисперсии координат X_i, $i = 1, .., r$, вектора X. Определим $\vec{x} = (x_1, \dots, x_r)$, где

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

Представим теперь вектор $\vec{x} = (x_1, \dots, x_r)$ в виде

$$\begin{cases} x_1 = l_{11}f_1 + \cdots + l_{1k}f_k + \varepsilon_1, \\ \vdots \\ x_r = l_{r1}f_1 + \cdots + l_{rk}f_k + \varepsilon_r, \end{cases} \qquad (1)$$

где $f_1, \dots, f_k$ — центрированно-нормированные некоррелированные общие факторы, $\varepsilon_1, \dots, \varepsilon_r$ — центрированные частные (или специфические) факторы, такие что $\mathrm{D}\varepsilon_i = \nu_i$, $\rho(\varepsilon_i, \varepsilon_j) = 0$, $\rho(\varepsilon_i, f_m) = 0$, $i, j = 1, \dots, r, m = 1, \dots, k$.

Можно показать, что детерминированные величины l_{ij} , $i = 1, \dots, r, j = 1, \dots, k$ являются коэффициентами корреляции между признаками x_i и факторами f_j, то есть $l_{ij} = \rho(x_i, f_j)$. По этой причине величины l_{ij} называют нагрузками i-го показателя на j-й фактор. Соотношение (1) можно также представить в матричном виде

$$\vec{x} = L\vec{f} + \vec{\varepsilon}, \qquad (2)$$

где в L — детерминированная матрица нагрузок размера $r \times k$, $\vec{f} = (f_1, \dots, f_k)$ — случайный вектор общих факторов, $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_r)$ — случайный вектор частных факторов. Таким образом, для получения представления (2) требуется найти матрицу нагрузок L и дисперсии $\mathrm{D}\varepsilon_i = v_i, i = 1, \dots, r$, называемые остаточными дисперсиями. Заметим, что представление вектора $\vec{x}$ в форме (2) не единственно. Действительно, если H — ортогональная матрица размера $k \times k$, то $\vec{x}$ можно также представить в виде

$$\vec{x} = (LH^T)(H\vec{f}) + \vec{\varepsilon} = L_1\vec{f}^{(1)} + \vec{\varepsilon},$$

где вектор $\vec{f}^{(1)} = H\vec{f}$ — вектор общих факторов, а $L_1 = LH^T$ — матрица нагрузок.

Умножение на ортогональную матрицу соответствует повороту осей координатной системы. Таким образом, общие факторы и нагрузки в модели (2) определяются с точностью до поворота. Поэтому после нахождения матрицы нагрузок рекомендуется решить задачу вращения, то есть найти такую систему координат, в которой матрица нагрузок позволила бы дать качественную интерпретацию общих факторов. Еще одной проблемой при решении задачи факторного анализа является выбор числа общих факторов k.

Известно несколько способов как для нахождения матрицы нагрузок, так и для определения числа общих факторов. Перейдём последовательно к изложению используемых методов. Так, для определения матрицы нагрузок и ковариационной матрицы частных факторов наиболее известны следующие четыре метода:

1) метод главных факторов (см., например, [Калинина, Соловьёв, 2003; Дронов, 2003]);

2) метод максимального правдоподобия, представленный в [Лоули, Максвелл,1967];

3) метод наименьших квадратов [Harman, Jones, 1966];

4) альфа-факторный метод Кайзера [Kaiser, Caffrey, 1965; Kaiser, Derflinger, 1990].

В данной работе будет использоваться метод максимального правдоподобия (ММП). Поэтому опишем кратко этот метод и укажем использованные нами модификации метода максимального правдоподобия. Предположим, что вектор общих факторов $\vec{f} \sim N(0, I)$, I — единичная матрица размера $k \times k$, а $\vec{\varepsilon} \sim N(0, V)$, где V — диагональная матрица размера $r \times r$ с диагональными элементами $v_i = \mathrm{D}\varepsilon_i$, $i = 1, \dots, r$. Обозначим через C ковариационную матрицу вектора $\vec{x}$. Тогда из (2) следует, что матрица C удовлетворяет соотношению

$$C = LL^T + V. \tag{3}$$

Как правило, ковариационная матрица наблюдаемого вектора неизвестна, и ее нужно оценивать. Рассмотрим здесь в качестве оценок элементов ковариационной матрицы выборочные ковариации, построенные по результатам n наблюдений за вектором $\vec{x} = (x_1, \dots, x_r)$. Обозначим через A матрицу выборочных ковариаций с элементами

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

Согласно методу максимального правдоподобия нужно выписать совместную плотность элементов выборочной ковариационной матрицы, прологарифмировать её и найти те значения l_{ij} и v_i, $i = 1, \dots, r$, $j = 1, \dots, k$, при которых достигается максимальное значение логарифмической функции правдоподобия. Логарифмическая функция правдоподобия в данном случае имеет вид

$$\ln L(C, A) = -\frac{n-1}{2} (\mathrm{tr}(AC^{-1}) + \ln|C|). \tag{4}$$

Для того чтобы найти максимум функции (4), нужно продифференцировать эту функцию по всем l_{ij} и v_i, $i = 1, \dots, r$, $j = 1, \dots, k$, и прирав-

нять частные производные к нулю. Аккуратные вычисления частных производных приведены в [Ивченко, Медведев, 2010]. Проведя эти действия, получим следующую систему уравнений для нахождения ММП-оценок:

$$\begin{cases} L^T C^{-1} - L^T C^{-1} A C^{-1} = 0, \\ \mathrm{diag}(C^{-1} - C^{-1} A C^{-1}) = 0. \end{cases}$$

Поскольку матрица L определяется с точностью до умножения на произвольную ортогональную матрицу, то с помощью введения дополнительного условия можно существенно упростить полученную систему уравнений. А именно, если потребовать, чтобы матрица $J = L^T V^{-1} L$ была диагональной с диагональными элементами, расположенными в порядке убывания, то можно показать (см. [Ивченко, Медведев, 2010, с. 552–553]), что строки матрицы L^T являются собственными векторами матрицы $V^{-1}(A - V)$. Таким образом, решение указанной задачи сводится к нахождению собственных векторов матрицы $V^{-1}(A - V)$. Отыскать соответствующие собственные векторы можно с помощью следующей итерационной процедуры, описанной в [Ивченко, Медведев, 2010, с. 554] .

Пусть имеются некоторые начальные приближения $L_{(1)}$ и $V_{(1)}$ матриц L и V. Обозначим i-е строки матриц L^T и $L_{(1)}^T$ как $\vec{l}_i^T$ и $\vec{l}_{i(1)}^T$ соответственно. На вход алгоритму подаются выборочная ковариационная матрица A и матрицы $L_{(1)}$ и $V_{(1)}$.

На первом шаге первой итерации вычисляются:

• вектор-строка $\vec{w}_1^T = \vec{l}_{1(1)}^T V_{(1)}^{-1}$;

• вектор-строка $\vec{u}_1^T = \vec{w}_1^T A - \vec{l}_{1(1)}^T$;

• число $h_1 = \vec{u}_1^T \vec{w}_1$.

Вектор $\vec{l}_{1(2)}^T = \frac{1}{\sqrt{h_1}} \vec{u}_1^T$ принимается в качестве второго приближения для $\vec{l}_1^T$.

Если число общих факторов $k > 1$, то переходим ко второму шагу первой итерации. На втором шаге вычисляются:

• вектор-строка $\vec{w}_2^T = \vec{l}_{2(1)}^T V_{(1)}^{-1}$;

• число $j_{21} = \vec{w}_2^T \vec{l}_{1(2)}$;

• вектор-строка $\vec{u}_2^T = \vec{w}_2^T A - \vec{l}_{2(1)}^T - j_{21}\vec{l}_{1(2)}^T$;

• число $h_2 = \vec{u}_2^T \vec{w}_2$.

Вектор $\vec{l}_{2(2)}^T = \frac{1}{\sqrt{h_2}}\vec{u}_2^T$ принимается в качестве второго приближения для $\vec{l}_2^T$.

Если имеется третий фактор, то делаем третий шаг первой итерации, т.е. вычисляются:

• вектор-строку $\vec{w}_3^T = \vec{l}_{3(1)}^T V_{(1)}^{-1}$;

• два числа $j_{31} = \vec{w}_3^T \vec{l}_{1(2)}$ и $j_{32} = \vec{w}_3^T \vec{l}_{2(2)}$;

• вектор-строку $\vec{u}_3^T = \vec{w}_3^T A - \vec{l}_{3(1)}^T - j_{31}\vec{l}_{1(2)}^T - j_{32}\vec{l}_{2(2)}^T$;

• число $h_3 = \vec{u}_3^T \vec{w}_3$.

В качестве второго приближения для $\vec{l}_3^T$ выбирается $\vec{l}_{3(2)}^T = \frac{1}{\sqrt{h_3}}\vec{u}_3^T$, и так далее. Если число общих факторов равно k, то необходимо сделать k таких шагов на каждой итерации. В конце первой итерации мы получим второе приближение для матрицы L, обозначим его $L_{(2)}$. Второе приближение матрицы V, обозначаемое $V_{(2)}$, мы получаем, подставляя элементы $L_{(2)}$ в равенство $v_i = a_{ii} - \sum_{j=1}^{k} l_{ij}^2$.

Матрицы $L_{(2)}$ и $V_{(2)}$ принимаются за начальные для второй итерации алгоритма. На ней мы аналогичным способом получаем третьи приближения матриц L и V, обозначаемые $L_{(3)}$ и $V_{(3)}$. И так далее. Процедура повторяется до тех пор, пока

$$\left\|V_{(m-1)} - V_{(m)}\right\| < \delta, \left\|L_{(m-1)} - L_{(m)}\right\| < \delta,$$

где δ — заранее выбранная константа, а m — номер итерации. Условия сходимости описанного алгоритма не установлены. Однако известно, что на практике обычно сходимость есть, причём довольно быстрая [Ивченко, Медведев, 2010, с. 554]. Этот итерационный алгоритм был реализован нами в среде Matlab и оказался сходящимся.

Перейдем теперь к проблеме выбора числа общих факторов. 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 H_0 о том, что число общих факторов равно заданной величине k. Для того чтобы выписать тестовую статистику отношения правдоподобия T для проверки гипотезы H_0, потребуем, чтобы $(r-k)^2 > r+k$ и введём ещё одну оценку ковариационной матрицы C вида $\hat{C} = \hat{L}\hat{L}^T + \hat{V}$, где матрицы $\hat{V}$ и $\hat{L}$ построены указанным выше способом в предположении, что число факторов равно k.

Тогда, учитывая сделанные предположения о гауссовости общих и частных факторов, статистика отношения правдоподобия

$$T = n\left(\ln\frac{|\hat{C}|}{|A|} + \mathrm{tr}(A\hat{C}^{-1}) - r\right) \tag{5}$$

имеет асимптотически распределение хи-квадрат с числом степеней свободы $s = \frac{1}{2}\left((r-k)^2 - (r+k)\right)$. Таким образом, если реализация статистики T оказывается меньше $(1-\alpha)$-квантили распределения хи-квадрат с s степенями свободы, то на уровне значимости α принимается гипотеза H_0. Если реализация статистики T больше указанной квантили, то число общих факторов k следует увеличить на единицу и повторить процедуру нахождения матриц L и V для числа факторов равного $k+1$.

Помимо указанного теоретически обоснованного способа, отметим ещё несколько хорошо зарекомендовавших себя эмпирических способов:

1) выбрать число общих факторов k равным числу собственных значений матрицы A, принимающих значения больше единицы [Иберла, 1980];

2) выбрать такое число общих факторов k, чтобы $\frac{1}{r}\sum_{j=1}^{k}\lambda_j \approx 0.7 \div 0.8$, где $\lambda_1, \dots, \lambda_r$ — упорядоченные по возрастанию собственные числа матрицы A [Калинина, Соловьёв, 2003];

3) проранжировать факторы по величине доли объясняемой суммарной дисперсии признаков, а затем исключить из рассмотрения те факторы, которые не объясняют долю дисперсии, превышающую некоторый заранее заданный порог [Ким и др., 1989].

Оговорим в этом разделе ещё два технических момента реализованной итерационной процедуры.

Первое. Поскольку мы рассматриваем центрировано-нормированный вектор признаков $\vec{x}$, то ковариационная матрица C является корреляционной, а матрица A — выборочной корреляционной матрицей. Отметим, что ММП позволяет выбирать в качестве матрицы C как ковариационную матрицу, так и корреляционную.

Второе. В итерационной процедуре можно использовать различные способы задания начального приближения $L_{(1)}$ и $V_{(1)}$ матриц L и V. Так, в [Иберла, 1980] в качестве первого приближения диагональных элементов $v_i^{(1)}$, $i = 1, \dots, r$ диагональной матрицы $V_{(1)}$ предлагается выбирать значения $v_i^{(1)} = 1 - \max_j |a_{ij}|$, где a_{ij} — элемент матрицы A; в [Ким и др., 1989] предлагается выбрать $v_i^{(1)} = \frac{1}{b_{ii}}$, где b_{ii} — элемент (i, i) матрицы A^{-1}. После выбора первого приближения для матрицы V первое приближение $L_{(1)}$ для матрицы L определяется из условия

$$L_{(1)} L_{(1)}^T = A - V_{(1)}.$$

В данной работе в качестве первого приближения матрицы L была выбрана матрица размера $r \times k$, состоящая из случайных величин, равномерно распределенных на отрезке $[0,1]$, в качестве первого приближения матрицы V — диагональная матрица размера $r \times r$, все диагональные элементы которой равны 0.001.

Наконец, после вычисления матриц L и V производится вращение факторного пространства. Наиболее распространенными методами вращения являются варимакс и квартимакс.

Методом варимакс находится такая матрица L с элементами l_{ij}, чтобы значение выражения

$\sum_{j=1}^{k} \frac{1}{r} \sum_{i=1}^{r} \left(l_{ij}^2 - \frac{1}{r} \sum_{h=1}^{r} l_{hj}^2 \right)^2$ было максимально. Столбцы полученной матрицы будут максимально сильно отличаться друг от друга, что означает разделение факторов за счет максимального уменьшения числа признаков, связанных с каждым фактором.

Квартимакс находит матрицу L из условия максимума значения выражения $\sum_{i=1}^{r} \frac{1}{k} \sum_{j=1}^{k} \left(l_{ij}^2 - \frac{1}{k} \sum_{h=1}^{k} l_{ih}^2 \right)^2$. В этом случае строки полученной матрицы будут различаться максимально сильно, что означает, что с каждым признаком связано минимальное возможное число факторов [Калинина, Соловьёв, 2003].

3. Модификации ММП

Как было 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 или коэффициент Крамера.

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

$$\rho_{YZ} = \frac{\sum_{m=1}^{n} (R_m - \overline{R})(S_m - \overline{S})}{\sqrt{\sum_{m=1}^{n} (R_m - \overline{R})^2 \sum_{m=1}^{n} (S_m - \overline{S})^2}},$$

в которой R_m — ранг элемента Y_m в выборке $Y_1, \dots, Y_n$, а S_m — ранг элемента Z_m в выборке $Z_1, \dots, Z_n$, $\overline{R} = \frac{1}{n} \sum_{m=1}^{n} R_m$, $\overline{S} = \frac{1}{n} \sum_{m=1}^{n} S_m$ — средние арифметические рангов.

Отметим, что критерий для проверки гипотезы о независимости двух случайных величин Y и Z, основанный на коэффициенте Спирмена, является состоятельным против альтернатив о том, что связь между Y и Z описывается монотонной функцией [Горяинова, Панков, Платонов, 2012, с.117–121]. Сам коэффициент Спирмена ρ_{YZ} может служить оценкой степени монотонной зависимости между величинами Y и Z. 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 $(Y_1, Z_1), \dots, (Y_n, Z_n)$ двумерного вектора (Y, Z). Для этого построим таблицу сопряжённости признаков Y и Z следующим образом. Разобьём область V_Y возможных значений величины Y на l непересекающихся интервалов $\Delta_{Y,i}$, $i = 1, \dots, l$, так, что $\bigcup_{i=1}^{l} \Delta_{Y,i} = V_Y$, а область V_Z возможных значений величины Z на s непересекающихся интервалов $\Delta_{Z,j}$, $j = 1, \dots, s$, так, что $\bigcup_{j=1}^{s} \Delta_{Z,j} = V_Z$. Пусть n_{ij} — число пар выборки $(Y_1, Z_1), \dots, (Y_n, Z_n)$, попавших в прямоугольник $\Delta_{Y,i} \times \Delta_{Z,j}$, $i = 1, \dots, l$, $j = 1, \dots, s$. Обозначим через $n_{i.} = \sum_{j=1}^{s} n_{ij}$ количество наблюдений выборки Y, попавших в интервал $\Delta_{Y,i}, i = 1, \dots, l$, а через $n_{.j} = \sum_{i=1}^{l} n_{ij}$ количество наблюдений выборки Z, попавших в интервал $\Delta_{Z,j}, j = 1, \dots, s$. Тогда коэффициент Крамера определяется как

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

где

$\hat{\chi}^2_{YZ} = n \sum_{i=1}^{l} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{n_{i.} n_{.j}}$ — статистика критерия хи-квадрат, предназначенного для проверки гипотезы о независимости случайных величин Y и Z.

Известно, что критерий хи-квадрат является состоятельным против любых альтернатив о зависимости Y и Z. В [Крамер, 1975] показано, что коэффициент Крамера, принимающий значения в интервале [0,1], может служить мерой, характеризующей силу связи между признаками Y и Z.

Особо отметим случай, когда признаки Y и Z измеряются в номинальной шкале. В этой ситуации таблица сопряжённости строится естественным образом. А именно, количество l интервалов Δ_Y совпадает с количеством градаций (возможных категорий) признака Y, а число s интервалов Δ_Z — с количеством градаций признака Z, соответственно n_{ij} — число тех наблюдений, у которых признак Y имеет i-ю категорию, а признак Z — j-ю категорию. Обозначим через K матрицу с элементами k_{ij}, где $k_{ij} = k_{x_i x_j}$ — коэффициент Крамера показателей x_i и x_j. Если признаки измерены в номинальной шкале, то для построения матрицы K не требуется проводить описанную в предыдущем разделе стандартизацию признаков.

Рассмотрим теперь следующие две модификации ММП, описанного в предыдущем разделе. Далее будем называть «вторым» адаптированным ММП метод, в котором матрица выборочных коэффициентов корреляции A заменяется матрицей коэффициентов Спирмена P, и, соответственно, «третьим» адаптированным ММП — метод, 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 вторым методом, а при наличии нелинейных немонотонных связей — третьим методом. Это предположение проверяется на тестовых данных с помощью обширного численного эксперимента.

А именно, 12-мерные векторы $\vec{x} = (x_1, \dots, x_{12})$ 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 «зашумлённой» функциональной зависимостью линейного типа, признаки третьей группы связаны «зашумлённой» функциональ-

ной зависимостью нелинейного монотонного типа, признаки четвёртой группы — «зашумлённой» функциональной зависимостью немонотонного типа. Опишем кратко моделирование этих групп. Моделирование первой группы проводилось следующим образом. С помощью встроенного в Matlab датчика генерировалась стандартная нормальная случайная величина x_1, затем стандартная нормальная величина x_2, имеющая со случайной величиной x_1 коэффициент корреляции равный 0,7, затем стандартная нормальная величина x_3, имеющая коэффициент корреляции равный 0,7 с величиной x_2. Принцип моделирования коррелированных величин основан на использовании следующего утверждения.

Утверждение. Пусть случайные величины Y и W независимы и имеют конечные дисперсии, а случайная величина $Z = \alpha W + Y$. Тогда коэффициент корреляции $\rho_{ZW} = \rho$ случайных величин Z и W связан с константой α соотношением

$$\alpha = \sqrt{\frac{\rho^2}{1-\rho^2}} \cdot \sqrt{\frac{DY}{DW}} \operatorname{sign}(\rho). \tag{6}$$

Доказательство. Найдём дисперсию $DZ = \alpha^2 DW + DY$ и ковариацию

$$\operatorname{cov}(W, Z) = \alpha DW + \operatorname{cov}(W, Y) = \alpha DW.$$

Тогда

$$\rho_{ZW} = \frac{\operatorname{cov}(W, Z)}{\sqrt{DW \cdot DZ}} = \frac{\alpha DW}{\sqrt{DW(\alpha^2 DW + DY)}}.$$

Выражая из последнего соотношения α, получим (6).

Таким образом, случайные величины Z и W будут иметь заданный коэффициент корреляции ρ, если Z представляется $Z = \alpha W + Y$, где α имеет вид (6), а величины Y и W независимы.

Принцип генерации второй, третьей и четвёртой групп следующий. Пусть случайные величины a_1, a_2, a_3 имеют усеченное стандартное нормальное распределение, а величины $e_1, \ldots, e_9$ — нормальное рас-

пределение $N(0, 0.1)$. Тогда значения признаков $x_4, \dots, x_{12}$ вычисляются по следующим формулам:

$$x_4 = a_1 + e_1, \quad x_5 = f(a_1) + e_2, \quad x_6 = f(f(a_1)) + e_3,$$
$$x_7 = a_2 + e_4, \quad x_8 = g(a_2) + e_5, \quad x_9 = g(g(a_2)) + e_6,$$
$$x_{10} = a_3 + e_7, \quad x_{11} = h(a_3) + e_8, \quad x_{12} = h(h(a_3)) + e_9,$$

где функция $f(\cdot)$ — линейная функция, $g(\cdot)$ — нелинейная монотонная функция, $h(\cdot)$ — нелинейная функция. Реализации значений пар признаков для каждой из четырёх групп объёма 10000 представлены на рис. 1.

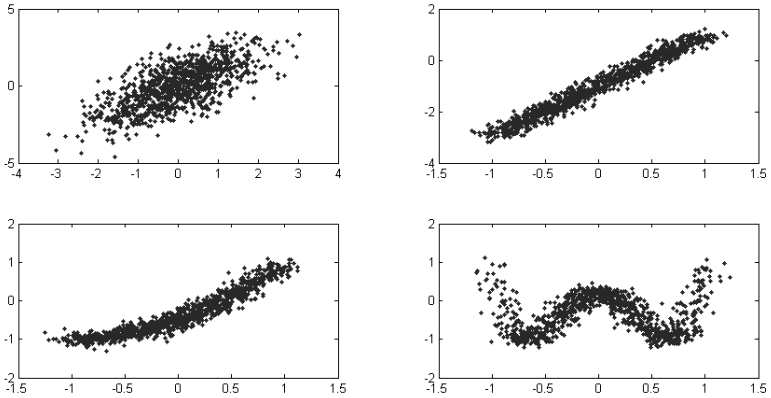

Рис. 1. Реализации признаков в группах демонстрационных данных

Помимо модификации ММП, обусловленной введением иных мер связи признаков, нам потребовалось также применить отличный от описанного в предыдущем разделе способ определения числа общих факторов. Это связано с тем, что метод определения общих факторов, основанный на критерии с тестовой статистикой отношения правдоподобия (5), показал неудовлетворительные результаты на моделированных данных. Неработоспособность этого метода объясняется следующим образом. Тестовая статистика (5) имеет распределение хи-квадрат в том случае, когда наблюдения имеют гауссовское распределение,

а компоненты $x_7, \dots, x_{12}$ сгенерированного вектора $\vec{x}$ являются нелинейными преобразованиями гауссовских случайных величин и, следовательно, не являются гауссовскими. Поэтому для определения числа общих факторов нами был реализован эмпирический метод, основанный на величине приращения среднеквадратического отклонения, объясняемого общими факторами. Опишем алгоритм этого метода.

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

$$\mu_j = \sqrt{\sum_{i=1}^{r} l_{ij}^2}, j = 1, \dots, r. \quad (7)$$

Каждый из коэффициентов μ_j является суммой квадратов элементов j-го столбца матрицы нагрузок и показывает количество суммарной дисперсии признаков, которое объясняется добавлением j-го фактора к уже имеющимся $j-1$ факторам $f_1, \dots, f_{j-1}$. Заметим, что поскольку столбцы матрицы нагрузок являются собственными векторами, соответствующими собственным числам, упорядоченным в порядке убывания, то и значения $\mu_j, j = 1, \dots, r$ будут упорядочены в порядке убывания. Из представления (3) следует, что

$$\mathrm{D}x_i = \sum_{j=1}^{k} l_{ij}^2 + v_i, i = 1, \dots, r.$$

Заметим, что мы рассматриваем нормированные значения признаков. Поэтому, если значение μ_j для некоторого столбца j оказывается больше единицы, то это означает, что фактором f_j можно объяснить вариацию по крайней мере двух признаков. Такой фактор включается в число общих факторов. Если коэффициент μ_j оказался значительно меньше единицы, то фактор f_j объясняет вариацию, обусловленную помехами (шумом), и его следует отбросить. Если значение μ_j близко к единице, то, скорее всего, фактор f_j объясняет изменчивость только

одного признака, и решение о включении его в число общих факторов остаётся на усмотрение исследователя. Таким способом определяется число общих факторов. На втором шаге запускается алгоритм ММП с выбранным числом факторов. Этот алгоритм показал адекватные результаты на моделированных данных.

4. Сравнительный анализ традиционного и адаптированных ММП в факторном анализе

Для того чтобы проводить сравнение различных методов, предназначенных для решения задачи сжатия многомерного вектора, требуется ввести количественный показатель, характеризующий качество выделения общих факторов. В задаче с реальными данными с априорно неизвестной структурой связей такой показатель вряд ли можно определить. Однако моделированные данные имеют определённую заданную структуру зависимостей, и, следовательно, матрица нагрузок сгенерированного вектора также должна иметь известную структуру. Так, если сгенерированный вектор $\vec{x}$ состоит из m подвекторов, и при этом компоненты, принадлежащие одному подвектору, зависимы между собой и независимы с компонентами других подвекторов вектора $\vec{x}$, то идеальная матрица нагрузок для вектора $\vec{x}$ будет иметь m столбцов, состоящих из нулей и единиц. Каждый столбец будет соответствовать общему фактору, объединяющему компоненты одного подвектора. Единичные значения каждого столбца будут располагаться в строках, соответствующих зависимым компонентам данного подвектора. Остальные элементы столбца будут нулевыми, так как компоненты вектора $\vec{x}$, принадлежащие разным подвекторам, независимы. Назовём такую «идеальную» матрицу эталонной. Будем считать, что из двух рассматриваемых методов более эффективным является метод, определяющий такую нагрузочную матрицу, столбцы которой наиболее близки (в смысле среднеквадратического отклонения) к столбцам эталонной матрицы.

Так, для демонстрационных данных $\vec{x} = (x_1, \dots, x_{12})$, структура которых описана в предыдущем разделе, эталонная матрица должна иметь вид

$$\begin{matrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{matrix}$$

Заметим, что столбцы этой матрицы могут быть расположены и в другом порядке. Однако без потери общности можно зафиксировать этот порядок. Первый столбец соответствует первой группе признаков, второй — второй группе и т.д. Получив с помощью одного из методов реальную нагрузочную матрицу, сравним ее столбцы с эталонными столбцами. Для этого сначала необходимо понять, какому из эталонных столбцов соответствует каждый реальный столбец. С этой целью для каждого столбца реальной нагрузочной матрицы определим среднеквадратическую ошибку отклонения этого столбца от первого столбца эталонной матрицы как

$$\gamma(et, l) = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(et_i - l_i)^2},$$

где r — количество признаков, et_i — i-й элемент данного эталонного столбца, l_i — i-й элемент текущего реального столбца. Среди всех столбцов реальной нагрузочной матрицы выберем тот, который будет иметь минимальное значение γ. Этот столбец соответствует той же группе, что и первый столбец эталонной матрицы. Значение $1 - \gamma_1$, где γ_1 — полученное минимальное значение, будем называть коэффициентом эффективности метода для этой группы. Исключаем выбранный

столбец из дальнейшего рассмотрения и вычисляем коэффициенты γ для второго столбца эталонной матрицы и оставшихся столбцов реальной матрицы и т.д. Процедура повторяется k раз, где k — число общих факторов. Таким образом, с помощью коэффициентов $\gamma_1, \dots, \gamma_k$ мы определяем эффективность метода выделения общих факторов при известной структуре зависимостей компонент многомерного вектора.

Перейдём к представлению результатов работы трёх рассмотренных методов для вектора $\vec{x} = (x_1, \dots, x_{12})$, структура которого описана в разделе 2. Сначала будем применять все три метода, полагая, что число общих факторов нам известно и равно 4. Матрица нагрузок, полученная традиционным ММП (метод 1), имеет вид

-0.0048	-0.0061	0.7121	-0.0028
-0.0028	0.0123	0.9879	-0.0006
-0.0112	0.0154	0.6913	0.0108
0.0281	0.9826	-0.0024	-0.0005
0.0290	0.9950	-0.0004	-0.0005
0.0290	0.9986	-0.0008	0.0001
0.9367	-0.0117	-0.0026	0.0063
0.9996	-0.0119	0.0001	-0.0000
0.9207	-0.0138	0.0016	0.0015
0.0017	-0.0046	-0.0068	-0.8116
-0.0012	-0.0046	-0.0113	0.0011
0.0030	-0.0060	0.0117	-0.0092

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). В качестве четвёртого фактора метод выделяет лишь признак x_{10} и не выделяет в четвёртую группу зависимые с x_{10} признаки x_{11} и x_{12}. Коэффициенты эффективности для групп 1–4 у этого метода соответственно равны: 0,8780; 0,9896; 0,9672; 0,5902.

Применение к этой матрице нагрузок методов вращения «варимакс» и «квартимакс» не внесло в нее никаких существенных изменений. Практически не изменились после вращения и коэффициенты эффективности.

Матрица нагрузок, полученная адаптированным ММП (метод 2), имеет вид

0.0099	-0.6891	0.0163	-0.0132
-0.0009	-0.9916	0.0244	0.0001
0.0071	-0.6871	0.0153	0.0001
0.9850	0.0014	0.0013	-0.0023
0.9961	-0.0007	0.0006	0.0013
0.9990	0.0000	-0.0005	-0.0002
-0.0004	-0.0305	-0.9777	-0.0031
-0.0058	-0.0325	-0.9781	0.0047
0.0018	-0.0215	-0.8860	-0.0085
0.0032	0.0036	-0.0054	0.0026
-0.0185	-0.0127	-0.0003	-0.0395
0.0010	-0.0091	0.0042	0.6735

И этот метод верно выделяет три группы зависимых признаков — признаки с линейным типом зависимости (имеют высокие нагрузки на первый фактор), сильно коррелированные признаки (имеют высокие нагрузки на второй фактор) и признаки с монотонным нелинейным ти-

пом зависимости (имеют высокие нагрузки на третий фактор). Как и традиционный ММП, этот метод не выявляет четвёртую группу признаков, связанных немонотонным типом зависимости. Высокую нагрузку на четвёртый фактор имеет здесь лишь признак x_{12}. Коэффициенты эффективности для групп в порядке с первой по четвертую: 0,8718; 0,9919; 0,9645; 0,5893. Методы вращения снова не внесли никаких существенных изменений.

Матрица нагрузок, полученная адаптированным ММП (метод 3), имеет 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 трёх рассмотренных методов только этот метод верно выделяет все 4 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 и в четвёртый — сильно коррелированные признаки. Соответствующие коэффициенты эффективности

для групп равны: 0,7476; 0,9203; 0,8646; 0,7720. Однако следует заметить, что коэффициенты эффективности этого метода для групп показателей с монотонными (как линейными, так и нелинейными) типами связи ниже, чем у двух предыдущих методов.

Теперь будем считать, что мы не имеем априорной информации о количестве общих факторов и применим описанный в предыдущем разделе эмпирический способ определения числа общих факторов, основанный на величине приращения среднеквадратического отклонения.

Применим к демонстрационным данным традиционный ММП с максимальным числом общих факторов равным 12 и вычислим по формуле (7) значения коэффициентов $\mu_1, \dots, \mu_{12}$. На рис. 2 представлены значения приращений объясняемого среднеквадратического отклонения $\mu_1, \dots, \mu_{12}$.

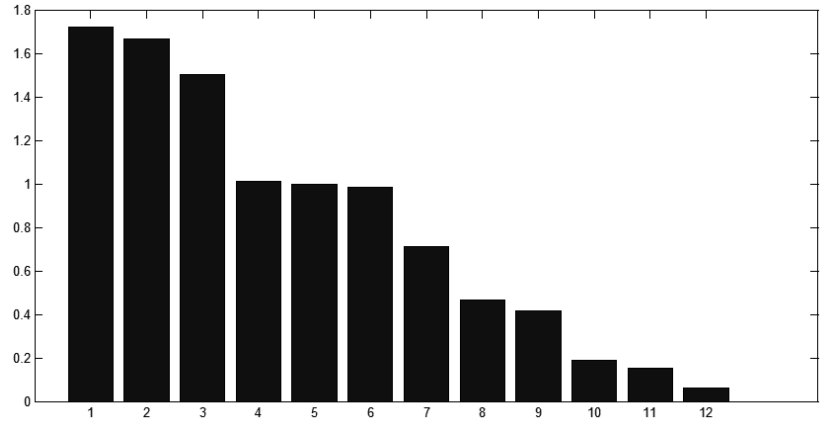

Рис. 2. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для традиционного метода

Значения μ_j больше единицы имеют первые три фактора, значения μ_j, практически равные единице, имеют следующие три фактора, оставшиеся шесть факторов имеют μ_j существенно меньше единицы.

Таким образом, число общих факторов считаем равным шести. Применяя традиционный ММП для $k = 6$, получим матрицу нагрузок

-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

Действительно, признаки x_{10}, x_{11}, x_{12} имеют высокие нагрузки на шестой, четвёртый и пятый факторы соответственно. Таким образом, традиционный метод выделяет в отдельные факторы признаки, связанные немонотонным типом зависимости.

Теперь применим к нашим данным второй метод, предварительно задав ему максимальное число факторов равное 12. Полученные значения коэффициентов $\mu_1, \dots, \mu_{12}$ представлены на рис. 3.

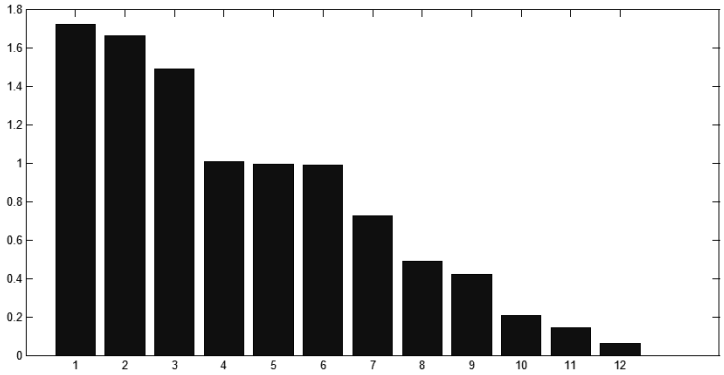

Рис. 3. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для второго метода

Этот рисунок также указывает на то, что следует выбрать число общих факторов равное шести. Матрица нагрузок второго метода при шести общих факторах имеет 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

Интерпретация полученного результата такая же, как и для традиционного метода.

Наконец, применим к моделированным данным третий метод, который обнаружил все четыре зависимые группы признаков. Значения μ_j для двенадцати факторов представлены на рис. 4.

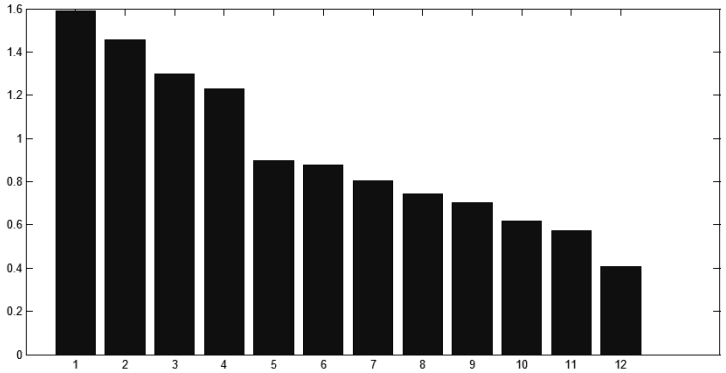

Рис. 4. Приращения объясняемого среднеквадратического отклонения для факторов с первого по двенадцатый для третьего метода

На графике видно, что больше единицы оказались значения μ_j только для четырех факторов. Таким образом, этот адаптированный метод верно определяет число общих факторов. Матрица нагрузок третьего метода при неизвестном числе факторов остаётся такой же, которая была получена при априорном предположении о том, что $k = 4$.

Важно отметить, что на других искусственно смоделированных данных (той же структуры, но отличных от демонстрационных) представленный эмпирический метод определения числа факторов продемонстрировал столь же адекватные результаты.

5. Снижение размерности показателей изменения прироста цен для группы продовольственных товаров

Продемонстрируем теперь работу трёх рассмотренных методов на реальных данных. Для демонстрации эффективной работы методов

сжатия многомерных признаков хотелось бы выбрать такие показатели, чтобы наличие зависимости между ними было в значительной степени предсказуемо из соображений здравого смысла. В качестве таких данных было решено выбрать еженедельные средние потребительские цены с января 2008 г. по апрель 2014 г. на некоторые продукты питания. 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 мы будем рассматривать не саму цену i-го товара в момент времени j (обозначим её c_{ij}), а величину относительного прироста цены, то есть $X_{ij} = \frac{c_{ij} - c_{i(j-1)}}{c_{i(j-1)}}$.

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 период времени с января 2008 г. по апрель 2014 г. были взяты с сайта Федеральной службы государственной статистики http://www.gks.ru.

Понятно, что первые пять продуктов образуют одну «мясную» группу, следующие четыре продукта образуют «молочную» группу, а последние два продукта — «мучную» группу.

Применим последовательно все три метода к имеющимся данным. Оговорим сразу, что в отличие от моделированных данных, для реальных данных потребовалось применить методы вращения варимакс и квартимакс. Вращение нагрузочной матрицы позволило существенно улучшить интерпретируемость результатов каждого из трёх рассматриваемых методов. Поэтому мы опустим представление тех матриц нагрузок, которые были получены до процедуры вращения. Отметим также, что матрицы, полученные методами варимакс и квартимакс, различались очень мало.

Применим традиционный ММП, считая сначала число факторов равным 11, и вычислим коэффициенты $\mu_1, \dots, \mu_{11}$ по формуле (7). Значения $\mu_1, \dots, \mu_{11}$ представлены на рис. 5.

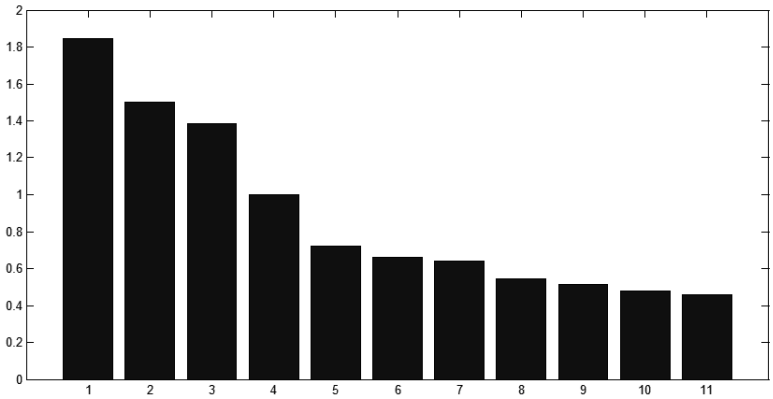

Рис. 5. Приращения объясняемого среднеквадратического отклонения факторов с первого по одиннадцатый для традиционного метода

Значения μ_j для первых трёх факторов больше единицы, и, следовательно, три фактора должны быть включены в модель. Здесь есть некоторые сомнения относительно включения четвёртого фактора, так как значение μ_4 близко к единице. Примем решение о том, что в модели будет три фактора, и, получив нагрузочную матрицу, проведём её вра-

щение методом квартимакс. Матрица нагрузок после вращения имеет вид

$$\begin{array}{ccc}
\boxed{0.7837} & 0.0763 & -0.1608 \\
\boxed{0.7953} & 0.0406 & -0.2426 \\
\boxed{0.8927} & 0.0537 & -0.0971 \\
\boxed{0.4513} & -0.0657 & 0.1389 \\
\boxed{0.6518} & -0.0541 & 0.0374 \\
-0.0711 & 0.0205 & \boxed{-0.6946} \\
0.0975 & -0.4298 & \boxed{-0.4831} \\
0.2239 & -0.0078 & \boxed{-0.8595} \\
0.0544 & 0.2281 & \boxed{-0.7201} \\
-0.0472 & \boxed{-0.9172} & -0.0053 \\
0.0166 & \boxed{-0.7206} & 0.1819
\end{array}$$

Как и ожидалось, признаки достаточно отчётливо объединились в три группы. Первый фактор объединяет продукты «мясной» группы, второй — «мучной» группы, третий — «молочной». Однако из общей картины несколько «выбиваются» строки, соответствующие приросту цен на сметану (строка 7) и на варёную колбасу (строка 4). Видно, что прирост цен на колбасу имеет существенно меньшую нагрузку на «мясной» фактор, чем остальные признаки из этой группы, а прирост цены на сметану имеет немалую нагрузку 0.429 и в «мучной» группе.

Перейдём ко второму адаптированному методу, использующему в качестве мер связи коэффициенты Спирмена. Коэффициенты $\mu_1, \dots, \mu_{11}$ представлены на рис. 6.

В данном случае, согласно используемому эмпирическому методу, следует принять решение о наличии трёх общих факторов, так как только три первых значения μ_j оказались больше единицы, а оставшиеся восемь — меньше. Матрица нагрузок, полученная вторым методом, после процедуры вращения варимакс имеет вид

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

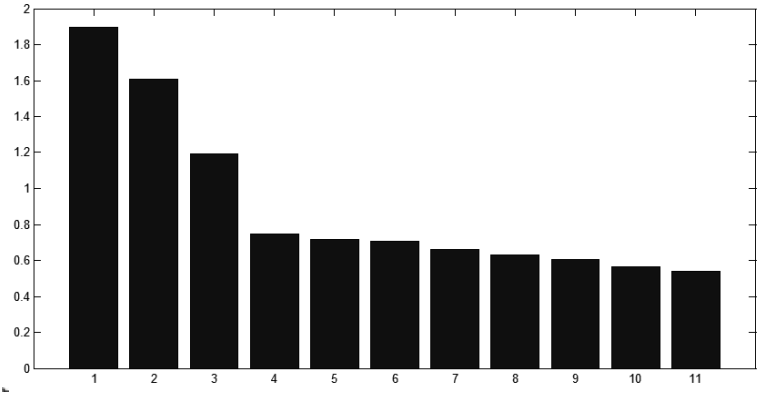

Рис. 6. Приращения объясняемого среднеквадратического отклонения для факторов с первого по одиннадцатый для второго метода

Этот метод также явно выделяет три фактора, соответствующих «молочной» (первый фактор), «мясной» (второй фактор) и «мучной» (третий фактор) группам. Но в отличие от результатов традиционного метода, 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.

Наконец, применим третий метод, использующий в качестве меры связи коэффициенты Крамера. Коэффициенты $\mu_1, \dots, \mu_{11}$ представлены на рис. 7.

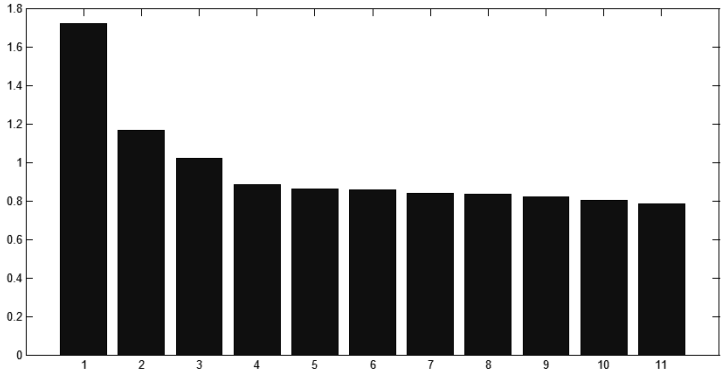

Рис. 7. Приращения объясняемого среднеквадратического отклонения для факторов с первого по одиннадцатый для третьего метода

Только три первых значения μ_j оказались больше единицы. Применим третий адаптированный метод для трёх общих факторов и проведём процедуру вращения полученной матрицы методом варимакс. Матрица нагрузок после вращения имеет вид

$$
\begin{array}{ccc}
\boxed{0.4365} & 0.1265 & -0.1382 \\
\boxed{0.5407} & 0.1425 & -0.0716 \\
\boxed{0.5426} & 0.1227 & -0.1277 \\
\boxed{0.5381} & 0.1130 & -0.1235 \\
\boxed{0.5009} & 0.1265 & -0.1350 \\
0.1331 & \boxed{0.5229} & -0.1304 \\
0.1839 & \boxed{0.5642} & -0.0820 \\
0.1573 & \boxed{0.5590} & -0.0389 \\
0.0704 & \boxed{0.4679} & -0.2211 \\
0.1558 & 0.1338 & \boxed{-0.4246} \\
0.1625 & 0.1216 & \boxed{-0.5062}
\end{array}
$$

Этот метод также правильно выделяет три фактора, и картина разбиения признаков на группы «похожести» достаточно отчётливая. Однако все признаки имеют на «свои» факторы меньшие нагрузки, чем в двух предыдущих методах.

При работе с реальными данными, в отличие от эксперимента с искусственно смоделированными данными, мы не можем утверждать, что матрица нагрузок рассматриваемых показателей имеет вполне определённый вид. Это объясняется следующими причинами: во-первых, реальные данные могут быть значительно зашумлены (то есть большой вклад в дисперсию признака вносит специфический фактор); во-вторых, изменения цен на продукты из разных групп, скорее всего, зависимы, пусть и значительно слабее, чем внутри групп; в-третьих, зависимости внутри групп носят более сложный характер, чем в случае искусственно сгенерированных данных. Допустим, что мы проигнорировали указанные причины и условно считаем, что эталонная нагрузочная матрица имеет вид

$$Et = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Тогда мы условно имеем возможность вычислить коэффициенты эффективности $1-\gamma_1$, $1-\gamma_2$, $1-\gamma_3$ для каждого метода и провести сравнение методов. Полученные коэффициенты эффективности для традиционного метода: 0,7679; 0,7652; 0,8243; для второго метода: 0,7894; 0,8573; 0,8523; для третьего метода: 0,6520; 0,6973; 0,7436. Поскольку все три метода выделили одинаковые факторы, то следует признать, что зависимость между показателями не носит немонотонный

характер. Более высокие коэффициенты эффективности имеет здесь второй метод, использующий в качестве меры связи ранговые коэффициенты Спирмена. Это говорит о том, что данные, возможно, сильно зашумлены, и коэффициенты Спирмена, как более робастные оценки истинных коэффициентов корреляции, лучше «улавливают» наличие линейной зависимости загрязнённых данных, чем выборочные коэффициенты корреляции.

6. Заключение

В данной работе была рассмотрена задача снижения размерности многомерного вектора показателей. При решении этой задачи был применён традиционный ММП и две модификации этого метода, использующие в качестве мер связи признаков ранговые коэффициенты корреляции Спирмена (метод 2) и коэффициенты Крамера (метод 3). Затем для трёх рассмотренных методов был проведён сравнительный анализ качества сжатия смоделированного многомерного вектора заданной структуры. С этой целью было введено формальное определение коэффициентов эффективности метода, а для определения числа общих факторов был программно реализован эмпирический метод, основанный на величине приращения объясняемого факторами среднеквадратического отклонения признаков.

В ходе проведения численного эксперимента были сгенерированы 12-мерные случайные векторы, состоящие из четырёх независимых подвекторов. При этом компоненты первого подвектора являлись сильно коррелированными, компоненты второго — связанными «зашумлённой» функциональной зависимостью линейного типа, компоненты третьего — связанными «зашумлённой» функциональной зависимостью монотонного нелинейного типа, компоненты четвёртого — немонотонной «зашумлённой» функциональной зависимостью. При применении традиционного ММП для снижения размерности смоделированного вектора оказалось, что этот метод достаточно хорошо выделяет в общие факторы коррелированные признаки и признаки, которые связаны зависимостями линейного и монотонного типа. Однако этот метод не способен выделить в единую группу признаки, связанные не-

монотонной зависимостью. Второй адаптированный метод показал аналогичные результаты, и коэффициенты эффективности этого метода практически не отличаются от коэффициентов традиционного метода. И только третий адаптированный метод правильно выделил все четыре группы связанных признаков. Этот факт имеет следующее объяснение. В этом методе были использованы в качестве мер связи признаков коэффициенты Крамера, основанные на статистике хи-квадрат. А критерий, основанный на этой статистике, 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 во втором методе, являются состоятельными лишь против альтернатив и линейной или монотонной зависимости признаков соответственно. Однако универсальность коэффициента Крамера имеет и негативную сторону. Его использование при выявлении линейных и монотонных зависимостей имеет более низкую эффективность, чем при использовании коэффициентов корреляции.

Все три рассмотренных метода показали адекватные результаты в реальной задаче снижения размерности вектора относительного прироста цен на продовольственные товары. Более эффективным методом в этой задаче можно условно считать второй метод. По-видимому, это обусловлено тем, что истинные зависимости между показателями имеют монотонный характер, и существенный вклад в вариацию признаков вносят частные факторы.

Литература

Горяинова Е.Р., Слепнёва Т.И. Методы бинарной классификации объектов с номинальными показателями // Журнал Новой экономической ассоциации. 2012. № 2. С. 27–49.

Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: ВШЭ, 2012.

Дронов С.В. Многомерный статистический анализ. Барнаул: Алтайский государственный университет, 2003.

Иберла К. Факторный анализ. М.: Статистика, 1980.

Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. М.: Издательство ЛКИ, 2010.

Калинина В.Н., Соловьев В.И. Введение в многомерный статистический анализ. М.: ГУУ, 2003.

Крамер Г. Математические методы статистики. М.: Мир, 1975.

Лоули Д., Максвелл А. Факторный анализ как статистический метод. М.: Мир, 1967.

Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким и др. М.: Финансы и статистика, 1989.

Harman H., Jones W. Factor analysis by minimizing residuals (minres) // Psychometrika. 1966. Vol. 31. No. 3. P. 351–369.

Kaiser H.F., Caffrey J. Alfa factor analysis // Psychometrika. 1965. Vol. 30. No. 1. P. 1–14.

Kaiser H.F., Derflinger G. Some Contrasts Between Maximum Likelihood Factor Analysis and Alpha Factor Analysis // Applied Psychological measurement. 1990. Vol. 14. No. 1. P. 24–32.

Goryainova, E. R., Shalimova J.
Reduction of dimensionality for the indicators that have a mixed structure : Working paper WP7/2014/08 [Text] / E. R. Goryainova, J. Shalimova ; National Research University Higher School of Economics. – Moscow : Higher School of Economics Publ. House, 2014. – 40 p. – 10 copies.

In many areas of science components of multidimensional vectors can be measured in different scales and be nonlinearly dependent. This paper considers the problem of reducing the dimensionality of such vectors. To identify common factors in such a system of indicators two modifications of the maximum likelihood method (MLM) are proposed. These modifications use Spearman rank correlation coefficients and Cramer coefficients as measures to connect indicators. Comparative analysis of the effectiveness of the traditional MLM and its two modifications was carried out using a numerical experiment. It was revealed that the adapted method that uses Cramer coefficients is more preferable. It turned out that only it can correctly unite in a common factor variables with non-monotonic type of communication. Furthermore, this method is also applicable to the nominal indicators. The proposed methods were tested on real data. Namely, we solve the problem of reducing the dimensionality of the dynamics of the relative rate of consumer prices for the group of foodstuffs.

Goryainova Elena – Ph. D., Department of Mathematics for Economics National Research University Higher School of Economics, Moscow, Russia

Shalimova Julia – Graduate Student Faculty of Economics National Research University Higher School of Economics, Moscow, Russia

Препринт WP7/2014/08
Серия WP7
Математические методы анализа решений
в экономике, бизнесе и политике

Горяинова Е.Р., Шалимова Ю.А.

Снижение размерности многомерных показателей смешанной структуры

Зав. редакцией оперативного выпуска *А.В. Заинченко*
Технический редактор *Ю.Н. Петрина*

Отпечатано в типографии
Национального исследовательского университета
«Высшая школа экономики» с представленного оригинал-макета
Формат 60×84 1/16. Тираж 10 экз. Уч.-изд. л. 2,5.
Усл. печ. л. 2,4. Заказ № . Изд. № 1908

Национальный исследовательский университет
«Высшая школа экономики»
125319, Москва, Кочновский проезд, 3
Типография Национального исследовательского университета
«Высшая школа экономики»