УДК 004.93'1

М. А. Зильберглейт, профессор; Невдах М. М., аспирант; Шпаковский Ю. Ф., ассистент

ИССЛЕДОВАНИЕ ФОРМАЛЬНЫХ ХАРАКТЕРИСТИК УЧЕБНОГО ТЕКСТА ПО ЭКОНОМИЧЕСКОЙ ТЕОРИИ МЕТОДАМИ ФАКТОРНОГО АНАЛИЗА

In the article methods of the factorial analysis (principal axis method, centroid method, principal components method) lead ordering 49 the information characteristics of the text influencing mastering of a teaching material. Processing of the received results has allowed to allocate eight conditional groups of close parameters of the text. The received calculations will be used for construction of a deciding rule of splitting.

Введение. В проведенном ранее исследовании с помощью метода кластерного анализа была проведена группировка 49-ти параметров текста, влияющих, по нашему мнению, на усвоение учебного материала. При этом на основании использования различных мер сходства (расстояние Евклида, квадрат расстояния Евклида, косинус угла, коэффициент корреляции, неравенство Чебышева, расстояние Минковского, манхэттенское расстояние) и ряда алгоритмов кластеризации исследуемые характеристики текста были разбиты на девять кластеров.

Для подтверждения и проверки согласованности результатов следует использовать и другие методы многомерной статистики. Таким образом, *цель данной работы* — систематизация параметров текста методами факторного анализа.

Основная часть. В качестве экспериментального материала в данной работе использовалось 16 отрывков из учебных изданий по экономической теории для высшей школы объемом 1800—2000 печатных знаков [1−4]. В качестве переменных было выбрано 49 признаков текста: 1) длина текста в абзацах; 2) длина текста в словах; 3) длина текста в буквах; 4) средняя длина абзаца в фразах; 5) средняя длина абзаца в словах; 6) средняя длина абзаца в буквах; 7) средняя длина абзаца в печатных знаках; 8) средняя длина предложения во фразах; 9) средняя длина предложения в словах; 10) средняя длина предложения в слогах; 11) средняя длина предложения в буквах; 12) средняя длина предложения в печатных знаках; 13) средняя длина самостоятельного предложения во фразах; 14) средняя длина самостоятельного предложения в словах; 15) средняя длина самостоятельного предложения в слогах; 16) средняя длина самостоятельного предложения в буквах; 17) средняя длина самостоятельного предложения в печатных знаках; 18) средняя длина фразы в словах; 19) средняя длина фразы в слогах; 20) средняя длина фразы в буквах; 21) средняя длина фразы в печатных знаках; 22) средняя длина слов в слогах; 23) средняя длина слов в буквах; 24) средняя длина слов в печатных знаках; 25) средняя длина слов по Деверу; 26) процент слов длиной в 5 букв и больше; 27) процент слов длиной в 6 букв и больше; 28) процент слов длиной в 7 букв и больше; 29) процент слов длиной в 8 букв и больше; 30) процент слов длиной в 9 букв и больше; 31) процент слов длиной в 10 букв и больше; 32) процент слов длиной в 11 букв и больше; 33) процент слов длиной в 12 букв и больше; 34) процент слов длиной в 13 букв и больше; 35) процент слов в 3 слога и больше; 36) процент слов в 4 слога и больше; 37) процент слов в 5 слогов и больше; 38) процент слов в 6 слогов и больше; 39) процент неповторяющихся слов; 40) средняя частота повторения слова; 41) процент неповторяющихся существительных; 42) процент повторяющихся существительных; 43) процент конкретных существительных; 44) процент абстрактных существительных; 45) процент прилагательных; 46) процент глаголов; 47) процент сложных предложений; 48) процент простых предложений; 49) процент придаточных предложений среди фраз.

Использование большого количества параметров текста является неэффективным по ряду причин [5, с. 516]:

а) сильная взаимосвязанность признаков, что приводит к дублированию информации;

б) неинформативность признаков, мало меняющихся при переходе от одного объекта к другому (малая «вариабельность» признаков);

в) возможность агрегирования (простого или «взвешенного» суммирования) по некоторым признакам.

В связи с этим представляется целесообразным перейти от p исходных показателей анализируемого материала к существенно меньшему числу наиболее информативных переменных (p') с помощью методов факторного анализа.

Снижение размерности набора переменных в методах факторного анализа базируется в основном на взаимной коррелированности исходных признаков [5, с. 547]. В связи с этим первый этап исследования заключался в вычислении корреляционной матрицы, фрагмент которой представлен в табл. 1.

При изучении экспериментальных данных было установлено, что первые три фактора объясняют около 64% разброса дисперсии (табл. 2.).

Таблица 1

Корреляционная матрица исходных признаков

№ п/п	1	2	3	4	5	6	7	8	9	10	…	49
1	1,00	-0,40	-0,66	-0,75	-0,85	-0,86	-0,87	0,32	-0,34	-0,33	…	-0,29
2	-0,40	1,00	0,56	0,56	0,53	0,41	0,41	0,05	0,19	-0,11	…	0,56
3	-0,66	0,56	1,00	0,66	0,67	0,70	0,71	-0,25	-0,02	0,04	…	0,12
4	-0,75	0,56	0,66	1,00	0,89	0,87	0,87	0,01	0,03	-0,02	…	0,10
5	-0,85	0,53	0,67	0,89	1,00	0,99	0,98	-0,25	0,32	0,27	…	0,35
6	-0,86	0,41	0,70	0,87	0,99	1,00	1,00	-0,29	0,28	0,30	…	0,26
7	-0,87	0,41	0,71	0,87	0,98	1,00	1,00	-0,29	0,28	0,29	…	0,25
8	0,32	0,05	-0,25	0,01	-0,25	-0,29	-0,29	1,00	0,16	0,07	…	-0,26
9	-0,34	0,19	-0,02	0,03	0,32	0,28	0,28	0,16	1,00	0,86	…	0,54
10	-0,33	-0,11	0,04	-0,02	0,27	0,30	0,29	0,07	0,86	1,00	…	0,23
…	…	…	…	…	…	…	…	…	…	…	…	…
49	-0,29	0,56	0,12	0,10	0,35	0,26	0,25	-0,26	0,54	0,23	…	1,00

Таблица 2

Объясненная дисперсия исследуемых параметров текста

Метод	Фактор	Исходные собственные значения		
		Собственные значения	Процент дисперсии	Кумулятивный процент
Метод главных факторов	1	15,87079	32,38937	32,38937
	2	9,16900	18,71225	51,10163
	3	6,53391	13,33452	64,43615
	4	4,35898	8,89588	73,33203
	5	3,20449	6,53978	79,87181
Центро-идный метод	1	15,21835	31,05785	31,05785
	2	9,26161	18,90124	49,95909
	3	6,84211	13,96349	63,92258
	4	4,16999	8,51018	72,43276
	5	3,42540	6,99061	79,42337
Метод главных компонент	1	16,05031	32,75574	32,75574
	2	9,31615	19,01254	51,76828
	3	6,74153	13,75823	65,52651
	4	4,52624	9,23722	74,76373
	5	3,42584	6,99151	81,75525

Так как факторный анализ является методом сокращения числа переменных, то возникает вопрос, какие из факторов следует оставить для дальнейшей обработки. Исследователи рекомендуют руководствоваться здравым смыслом и оставлять только те факторы, которые имеют понятную или логическую интерпретацию. Однако установить заранее назначение каждого фактора не всегда представляется возможным, поэтому для начала были использованы формальные критерии: критерий Кайзера [6] и критерий «каменистой осыпи» Р. Кетелла [7].

На основании первого критерия, предложенного Кайзером в 1960 году, для дальнейшего анализа необходимо сохранить те факторы, чьи собственные значения превышают единицу. В данном случае следует оставить восемь факторов для всех методов факторного анализа. Критерий «каменистой осыпи» является графическим методом. Для выделения факторов используется график их собственных значений (рис. 1).

По утверждению Р. Кэттела следует найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Анализ графиков для всех методов показал, что целесообразно оставить от 4 до 6 факторов.

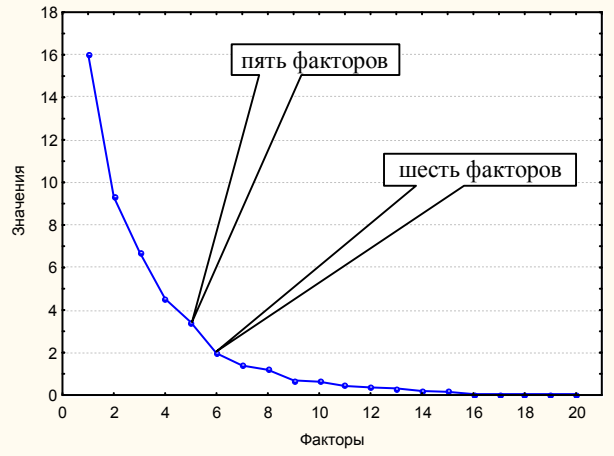

Рис. 1. График собственных значений для метода главных компонент

Следует отметить, что первый критерий, как правило, сохраняет слишком много факторов, в то время как второй — слишком мало, поэтому решение об оптимальном количестве факторов можно принять только после их вращения и интерпретации.

Целью вращения факторов является получение простой структуры, которой соответствует большое значение нагрузки каждой переменной только по одному фактору и малое по всем остальным факторам. Нагрузка (значение лежит в пределах от –1 до 1) отражает связь между переменной и фактором. В работе использовались ортогональные методы вращения: варимакс, квартимакс и эквимакс. В результате были получены матрицы нагрузок для переменных. Фрагмент представлен в табл. 3.

Изучение результатов с использованием всех методов факторного анализа и методов вращения позволило выявить, как признаки распределились между четырьмя факторами (табл. 4).

Как видно из таблицы, факторы по всем методам вращения практически идентичны. Сравнение данных, полученных ранее с помощью кластерного анализа показало, что результаты не совпадают.

Для более ясного представления о распределении переменных использовались диаграммы рассеяния. Для трех факторов диаграммы изображены в трехмерном пространстве (рис. 2).

Результаты, полученные методом главных факторов, центроидным методом и методом главных компонент, позволяют выделить восемь условных групп близких параметров текста.

Первая группа. Признаки 1, 4, 8, 13, 22—25, 40, 42—44 и 46 — длина текста в абзацах, средняя длина абзаца во фразах, средняя длина

Таблица 3

Факторные веса при анализе 49-ти информационных характеристик текста с использованием метода главных факторов и вращением варимакс

Параметры текста	Фактор 1	Фактор 2	Фактор 3	Фактор 4
1. Длина текста в абзацах	-0,177	-0,281	-0,145	-0,789
2. Длина текста в словах	-0,711	-0,013	0,212	0,550
3. Длина текста в буквах	0,013	0,131	-0,066	0,676
4. Средняя длина абзаца в фразах	-0,053	-0,203	-0,067	0,931
5. Средняя длина абзаца в словах	0,027	0,162	0,152	0,912
6. Средняя длина абзаца в буквах	0,170	0,186	0,092	0,905
7. Средняя длина абзаца в печатных знаках	0,163	0,194	0,075	0,903
8. Средняя длина предложения в фразах	-0,404	-0,183	-0,582	-0,038
9. Средняя длина предложения в словах	-0,231	0,750	0,110	0,204
10. Средняя длина предложения в слогах	0,083	0,800	-0,111	0,156
…	…	…	…	…
49. Процент придаточных предложений среди фраз	-0,442	0,385	0,590	0,250

Таблица 4

Распределение характеристик текста с использованием различных методов факторного анализа и методов вращения

Метод вращения	Метод факторного анализа											
	метод главных факторов				центроидный метод				метод главных компонент			
	фактор 1	фактор 2	фактор 3	фактор 4	фактор 1	фактор 2	фактор 3	фактор 4	фактор 1	фактор 2	фактор 3	фактор 4
варимакс	2, 22–25, 27–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7
квартимакс	2, 22–25, 27–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7
эквимакс	2, 22–25, 27–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–37	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7	2, 22–38	9–12, 14, 16–18, 20, 21	13, 15, 19, 42–44	1, 4–7

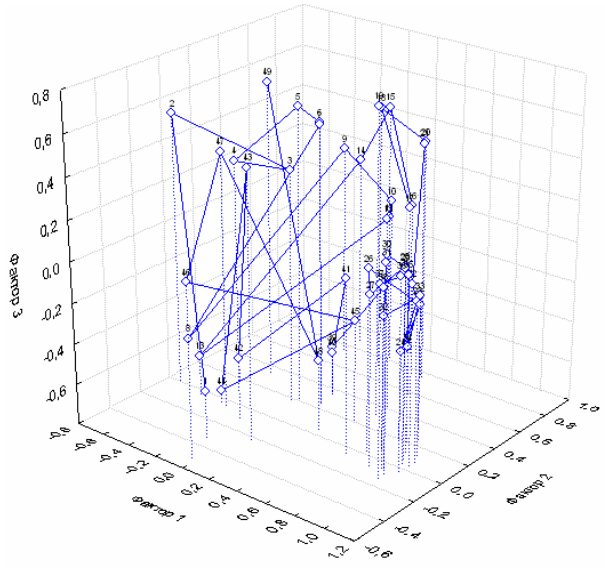

Рис. 2. Диаграмма рассеяния признаков для метода главных факторов

предложения во фразах, средняя длина самостоятельного предложения во фразах, средняя длина слов в слогах, средняя длина слов в буквах, средняя длина слов в печатных знаках, средняя длина слов по Деверу, средняя частота повторения слова, процент повторяющихся существительных, процент конкретных существительных, процент абстрактных существительных и процент глаголов.

Вторая группа. Признаки 2, 9, 14 и 18 — длина текста в словах, средняя длина предложения в словах, средняя длина самостоятельного предложения в словах и средняя длина фразы в словах.

Третья группа. Признаки 3, 39, 45 и 48 — длина текста в буквах, процент неповторяющихся слов, процент прилагательных и процент простых предложений.

Четвертая группа. Признаки 5—7 — средняя длина абзаца в словах, средняя длина абзаца в буквах и средняя длина абзаца в печатных знаках.

Пятая группа. Признаки 10—12, 16 и 17 — средняя длина предложения в слогах, средняя длина предложения в буквах, средняя длина предложения в печатных знаках, средняя длина самостоятельного предложения в буквах и средняя длина самостоятельного предложения в печатных знаках.

Шестая группа. Признаки 15, 19—21 — средняя длина самостоятельного предложения в слогах, средняя длина фразы в слогах, средняя длина фразы в буквах и средняя длина фразы в печатных знаках.

Седьмая группа. Признаки 26—38 и 41 — процент слов длиной в 5 букв и больше, процент слов длиной в 6 букв и больше, процент слов длиной в 7 букв и больше, процент слов длиной в 8 букв и больше, процент слов длиной в 9 букв и больше, процент слов длиной в 10 букв и больше, процент слов длиной в 11 букв и больше, процент слов длиной в 12 букв и больше, процент слов длиной в 13 букв и больше, процент слов в 3 слога и больше, процент слов в 4 слога и больше, процент слов в 5 слогов и больше, процент слов в 6 слогов и больше и процент неповторяющихся существительных.

Восьмая группа. Признаки 47 и 49 — процент сложных предложений и процент придаточных предложений среди фраз.

Выводы. Завершающим и наиболее сложным этапом факторного анализа является интерпретация результатов. На данный момент исследования логическое объяснение всем выделенным факторам найти не удалось. Так как в будущем полученные расчеты будут использованы для построения решающего правила разбиения, то следует надеяться, что это позволит дать соответствующую интерпретацию результатам факторного анализа.

Литература

1. Сажина, М. А. Основы экономической теории: учебное пособие для неэкономических специальностей вузов / М. А. Сажина, Г. Г. Чибриков; отв. ред. и руководитель авт. коллектива П. В. Савченко. — М.: Экономика, 1995.

2. Экономическая теория: учебник / Н. И. Базылев, А. В. Бондарь, С. П. Гурко и др.; под общ. ред. Н. И. Базылева, С. П. Гурко. — Мн.: Экоперспектива, 1997.

3. Экономическая теория: учебник для студентов вузов / Под ред. В. Д. Камаева. — 6-е изд., перераб. и доп. — М.: ВЛАДОС, 2001.

4. Экономическая теория: учебное пособие / Л. Н. Давыденко, А. И. Базылева, А. А. Дичковский и др.; под общ. ред. Л. Н. Давыденко. — Мн.: Вышэйшая школа, 2002.

5. Айвазян, С. А. Прикладная статистика и основы эконометрики: учебник для вузов / С. А. Айвазян, В. С. Мхитарян. — М.: ЮНИТИ, 1998. — 1022 с.

6. Kaiser, H. F. The application of electronic computers to factor analysis / H. F. Kaiser // Educational and Psychological Measurement. — 1960. — № 20. — P. 141–151.

7. Cattell, R. B. The scree test for the number of factors / R. B. Cattell // Multivariate Behavioral Research. — 1966. — № 1. — P. 245–276.