8-бөлім
БИОСТАТИСТИКА

Раздел 8
БИОСТАТИСТИКА

Section 8
BIOSTATISTICS

Елеуов А.А., Тунгатаров Н.Н.

Об одном применении собственных значений матриц при обработке статистических данных

В работе вариационным методом вычисляются собственные значения матриц при обработке статистических данных. Приведенный метод главных компонент может применяться в различных задачах, где возникают симметрические матрицы. Например, когда исходной информацией об объектах служат экспертные данные о различиях между ними, выраженных числами. Разработанные алгоритмы были применены в случае самосопряженных и положительно определенных матриц, затем обобщены на случай матриц произвольного типа. Отдельно необходимо отметить, что разработанный приближенный метод был применен к процедуре приведения матриц к треугольному виду. Данный метод был применен к решению практической задачи экономики и получены удовлетворительные результаты. Следующим применением данного метода служат для задач биологии и биотехнологии. Работа может быть аттестована по физико-математическим наукам.

Ключевые слова: метод, собственные значения матрицы, симметрические матрицы, самосопряженная матрица, определенной матрица, собственный вектор, матрица, корреляционный анализ.

Eleuov A.A., Tungatarov N.N.

An application of the eigenvalues of the matrices in the processing of statistical data

In this paper a variational method to calculate the eigenvalues of the matrices in the processing of statistical data. The above method of principal components can be used in various applications where there are symmetric matrices. For example, when the initial information about the objects are expert data about the differences between them expressed by numbers. The developed algorithms have been applied in the case of self-adjoint and positive definite matrices, and then extended to the case of matrices of arbitrary type. Separately, it should be noted that an approximate method was applied to the process of reducing a matrix to triangular form. This method was applied to solve practical problems of the economy and gives satisfactory results. Following the application of the method used for problems of biology and biotechnology.

Key words: method, eigenvalues, symmetric matrices, self-adjoint matrix defined matrix eigenvector matrix, correlation analysis.

Елеуов А.А., Түнғатаров Н.Н.

Статистикалық мәліметтерді өңдеу кезінде матрицалардың меншікті мәндерін бір қолданылуы туралы

Бұл жұмыста статистикалық мәліметтерді өңдеуде матрицаның меншікті мәндері вариациялық әдіспен шығарылады. Келтірілген негізгі компонент әдісі симметриялық матрицалар кездесетін әр түрлі есептерде қолданады. Мысалы,берілген ақпаратты сандық түрде берілген әр түрлі экспертті мәліметтердің арасындағы обекттерге қызмет етеді. Өңделген алгоритмдер түйіндес және оң анықталған матрицаларға, ожан кейін жалпы еркін түрдегі матрицаларға қолданылады. Жеке түрде ескертетін жайт, өңделген жуықтау әдісі матрицаны үшбұрыштық түрге келтіру процедурасына қолданылғаню Осы әдіс экономикалық есептерді шешуге қолданылып және қанағаттанарлық нәтижелер алыңды. Осы әдіс енді биологиялық және биотехнологиялық есептерге қолданылады.

Түйін сөздер: әдіс, матрицаның меншікті мәні, симметриялық матрица, түйіндес матрица, анықталған матрица, меншікті вектор, матрица, корреляциялық талдау.

УДК 519.6

*Елеуов А.А., Тунгатаров Н.Н.

Казахский национальный университет имени аль-Фараби,
Республика Казахстан, г. Алматы
*E-mail: Eleuov@mail.ru

ОБ ОДНОМ ПРИМЕНЕНИИ СОБСТВЕННЫХ ЗНАЧЕНИЙ МАТРИЦ ПРИ ОБРАБОТКЕ СТАТИСТИЧЕСКИХ ДАННЫХ

Введение

В статье обсуждается одно полезное наблюдение, которое имеет наглядный смысл и полезно при обработке статистических данных. Материал изложен без лишних математических премудростей и доступен экономистам, социологам и специалистам в других областях, использующих статистические методы.

При статистическом анализе таблицы данных, состоящей из нескольких признаков, необходимо иметь в виду эффект существенной многомерности, из-за которого к верным выводам можно прийти лишь при одновременном учете всей совокупности взаимосвязанных признаков. К примеру, попытка различить два типа потребительского поведения семей сначала по одному признаку (расходы на питание), потом по другому (расходы на промышленные товары и услуги) не дала результата, в то время как одновременный учет обоих признаков позволил обнаружить значимое различие между анализируемыми совокупностями семей.

Если число признаков – достаточно большое число, то разбиение множества исследуемых объектов на компактные группы (так называемые кластеры) может оказаться непростой задачей. В этом состоит задача классификации или кластер – анализ. После того, как объекты разбиты на однородные группы (классы), возникает задача изучения взаимосвязей признаков внутри отдельного класса. Если однородная группа образует «облако» эллиптического типа, то применяют методы корреляционного анализа. Когда объекты располагаются в окрестности некоторой кривой (поверхности и так далее) надо применять приемы регрессионного анализа.

Теория собственных векторов матриц и их применение в корреляционном анализе

Предположим, что каждый из n объектов описывается k признаками (рост, вес, длина черепа, длина и ширина верхней челюсти и так далее), и представим данные для отдельного класса объектов в форме таблицы $X = \left\| x_{il} \right\|_{n \times k}$.

Вычислим для каждого признака среднее значение $\overline{x}_l = \frac{1}{n}\sum_{i=1}^{n} x_{il}$ и центрируем данные: $y_{il} = x_{il} - \overline{x}_l$. Тогда $\overline{y}_l = 0$, $l = 1,\dots,k$. Обозначим через $S = \| c_{il} \|_{k\times k}$ выборочную ковариационную матрицу признаков: $S = \frac{1}{n} Y^T Y$, то есть c_{il} – выборочная ковариация i-го и l-го столбцов матрицы *Y*. Из того, что матрица ковариаций $S = \| c_{il} \|_{k\times k}$ является неотрицательно определенной матрицей, иначе говоря, самосопряженной матрицей следует ее приводимость к диагональному виду. Следовательно, существует ортогональная матрица *U*, приводящая *S* к главным осям: $U^T S U = \Lambda$. Здесь Λ – диагональная матрица с неотрицательными элементами $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$ на главной диагонали, которые являются корнями уравнения det $(S - \lambda E) = 0$. Они называются собственными значениями матрицы *S*. Предположим, что все $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$ положительны и различны. Для экспериментальных данных это условие выполняется практически всегда. Заметим также, что столбцы $u_1, u_2, \dots u_k$ матрицы *U* представляют главные оси и определяются однозначно с точностью до выбора направления оси. Они образуют ортонормированный базис в R^n, обладающий важными свойствами:

1. Проекции объектов на первую главную ось u_1 имеют наибольшую выборочную дисперсию среди проекций на всевозможные направления в пространстве R^n, причем этот максимум равен λ_1.
2. Проекции объектов на вторую главную ось u_2 имеют наибольшую выборочную дисперсию среди проекций на всевозможные направления в пространстве R^n, которые ортогональны вектору u_1. Причем этот максимум равен λ_2.
3. Сумма выборочных дисперсий исходных признаков $tr\, S = \sum_{l=1}^{k} c_{ll}$ в силу подобия матриц *S* и Λ равна $tr\,\Lambda = \lambda_1 + \dots + \lambda_k$, то есть сумме выборочных дисперсий проекций объектов на главные оси. Эта величина может рассматриваться как мера общего разброса объектов относительно их центра масс. Представляет интерес относительная доля разброса, приходящаяся на l первых главных осей,

$$\gamma_l = \frac{\lambda_1 + \dots + \lambda_l}{\lambda_1 + \dots + \lambda_k}$$

Если эта величина при некотором l достаточно близка к 1, то возможно уменьшение размерности пространства признаков за счет перехода от k исходных признаков к l новым признакам. На практике нередко удается ограничиться двумя или тремя компонентами без существенной потери информации.

Пример применения собственных векторов матриц в корреляционном анализе

В таблице указаны размеры челюстей и зубов тридцати собак (номера 1 – 30), двенадцати волков (номера 31 – 42) и ископаемого черепа неизвестного животного (номер 43), найденного в четверичном слое (по данным Де Бониса [1]). На рисунке показаны измеряемые характеристики: 1 – длина черепа, 2 – длина верхней челюсти, 3 – ширина верхней челюсти; следующие измерения относятся к зубам: 4 – длина верхнего карнивора, 5 – длина первого верхнего моляра, 6 – ширина первого верхнего моляра. Требуется узнать, к какому из классов (собак или волков) следует отнести неизвестное животное.

Здесь мы займемся более скромной задачей: найдем и интерпретируем главные компоненты для данного примера.

Алгоритм определения главных осей.

1. В каждом столбце исходной таблицы 1 находим среднее значение.
2. Из столбцов вычитаем найденные значения, соответствующие средним. Результат обозначим через таблицу 1а.
3. Затем составим новую таблицу 1б из квадратов элементов таблицы 1а. Результат обозначим через таблицу 1б.
4. В каждом столбце новой таблицы 1б находим среднее значение.
5. Столбцы таблицы 1а поделим на корни квадратные из соответствующих средних значений шага 4. Результат оформим в виде таблицы 2.
6. Таблица 2 представляет собой продолговатую матрицу (строк 43, столбцов 6). Умножим ее на ее транспонирование так, чтобы получилась матрица размерности 6 на 6.
7. Результат шага 6 поделим на 43. Смотрите таблицу 3.

Таблица 1

	1	2	3	4	5	6
1	129	64	95	17,5	11,2	13,8
2	154	74	76	20	14,2	16,5
3	170	87	71	17,9	12,3	15,9
4	188	94	73	19,5	13,3	14,8
5	161	81	55	17,1	12,1	13
6	164	90	58	17,5	12,7	14,7
7	203	109	65	20,7	14	16,8
8	178	97	57	17,3	12,8	14,3
9	212	114	65	20,5	14,3	15,5
10	221	123	62	21,2	15,2	17
11	183	97	52	19,3	12,9	13,5
12	212	112	65	19,7	14,2	16
13	220	117	70	19,8	14,3	15,6
14	216	113	72	20,5	14,4	17,7
15	216	112	75	19,6	14	16,4
16	205	110	68	20,8	14,1	16,4
17	228	122	78	22,5	14,2	17,8
18	218	112	65	20,3	13,9	17
19	190	93	78	19,7	132	14
20	212	111	73	20,5	13,7	16,6
21	201	105	70	19,8	14,3	15,9
22	196	106	67	18,5	12,6	14,2
23	158	71	71	16,7	12,5	13,3
24	255	126	86	21,4	15	18
25	234	113	83	21,3	14,8	17
26	205	105	70	19	12,4	14,9
27	186	97	62	19	13,2	14,2
28	241	119	87	21	14,7	18,3
29	220	111	88	22,5	15,4	18
30	242	120	85	19,9	15,3	17,6
31	199	105	73	23,4	15	19,1
32	227	117	77	25	15,3	18,6
33	228	122	82	24,7	15	18,5
34	232	123	83	25,3	16,8	15,5
35	231	121	78	23,5	16,5	19,6
36	215	118	74	25,7	15,7	19
37	184	100	69	23,3	15,8	19,7
38	175	94	73	22,2	14,8	17
39	239	124	77	25	16,8	27
40	203	109	70	23,3	15	18,7
41	226	118	72	26	16	19,4
42	226	119	77	26,5	16,8	19,3
43	210	103	72	20,5	14	16,7
сред. ариф. знач.	204,9535	106,4651	72,53488	21,05581	17,05814	16,8093

Таблица 2

	1	2	3	4	5	6
1	-2,81171	-2,86441	2,491943	-1,3857	-0,32938	-1,23658
2	-1,88624	-2,18987	0,384368	-0,41145	-0,1607	-0,1271
3	-1,29394	-1,31298	-0,17026	-1,22982	-0,26753	-0,37365
4	-0,6276	-0,84081	0,051593	-0,6063	-0,21131	-0,82566
5	-1,62711	-1,7177	-1,94506	-1,54158	-0,27878	-1,56532
6	-1,51605	-1,11062	-1,61228	-1,3857	-0,24504	-0,86675
7	-0,07232	0,170986	-0,83581	-0,13866	-0,17195	-0,00382
8	-0,99779	-0,63845	-1,72321	-1,46364	-0,23942	-1,03112
9	0,260853	0,508252	-0,83581	-0,2166	-0,15508	-0,53802
10	0,594022	1,11533	-1,16858	0,056189	-0,10448	0,078361
11	-0,81269	-0,63845	-2,27783	-0,68424	-0,2338	-1,35986
12	0,260853	0,373345	-0,83581	-0,52836	-0,1607	-0,33256
13	0,557004	0,710611	-0,28118	-0,48939	-0,15508	-0,49693
14	0,408929	0,440799	-0,05933	-0,2166	-0,14946	0,366005
15	0,408929	0,373345	0,273443	-0,56733	-0,17195	-0,16819
16	0,001722	0,238439	-0,50303	-0,09969	-0,16633	-0,16819
17	0,853154	1,047877	0,606218	0,562797	-0,1607	0,407097
18	0,482966	0,373345	-0,83581	-0,29454	-0,17757	0,078361
19	-0,55356	-0,90826	0,606218	-0,52836	6,462765	-1,1544
20	0,260853	0,305892	0,051593	-0,2166	-0,18882	-0,08601
21	-0,14635	-0,09883	-0,28118	-0,48939	-0,15508	-0,37365
22	-0,33145	-0,03137	-0,61396	-0,996	-0,25067	-1,07221
23	-1,73816	-2,39223	-0,17026	-1,69746	-0,25629	-1,44204
24	1,852661	1,31769	1,493618	0,134129	-0,11572	0,489281
25	1,075267	0,440799	1,160843	0,095159	-0,12697	0,078361
26	0,001722	-0,09883	-0,28118	-0,80115	-0,26191	-0,78457
27	-0,70164	-0,63845	-1,16858	-0,80115	-0,21693	-1,07221
28	1,334398	0,845518	1,604543	-0,02175	-0,13259	0,612557
29	0,557004	0,305892	1,715468	0,562797	-0,09323	0,489281
30	1,371417	0,912971	1,382693	-0,45042	-0,09885	0,324913
31	-0,22039	-0,09883	0,051593	0,913526	-0,11572	0,941293
32	0,816135	0,710611	0,495293	1,537044	-0,09885	0,735833
33	0,853154	1,047877	1,049918	1,420135	-0,11572	0,694741
34	1,001229	1,11533	1,160843	1,653954	-0,01451	-0,53802
35	0,96421	0,980424	0,606218	0,952496	-0,03138	1,146753
36	0,37191	0,778064	0,162518	1,809833	-0,07636	0,900201
37	-0,77567	-0,43609	-0,39211	0,874556	-0,07074	1,187845
38	-1,10884	-0,84081	0,051593	0,445888	-0,12697	0,078361
39	1,260361	1,182783	0,495293	1,537044	-0,01451	4,187559
40	-0,07232	0,170986	-0,28118	0,874556	-0,11572	0,776925
41	0,779116	0,778064	-0,05933	1,926743	-0,0595	1,064569
42	0,779116	0,845518	0,495293	2,121592	-0,01451	1,023477
43	0,186816	-0,23373	-0,05933	-0,2166	-0,17195	-0,04491

Таблица3

1	0,958741	0,348183	0,612949	-0,032121	0,587251
0,958741	1	0,200333	0,661002	-0,085869	0,594653
0,348183	0,200333	1	0,369962	0,120454	0,354777
0,612949	0,661002	0,369962	1	-0,015032	0,762643
-0,03212	-0,085869	0,120454	-0,015032	1	-0,120108
0,587251	0,594653	0,354777	0,762643	-0,120108	1

Таблицы 2 и 3 вычислены в популярной программе для обработки электронных таблиц Microsoft Excel. Собственные векторы и собственные значения матрицы, приведенной в таблице 7, вычислены с использованием вариационных методов. В диссертационной работе [2] нами предложены различные алгоритмы вычисления собственных значений и собственных векторов матриц на основе вариационного метода. В работе [3] эти методы применялись для некоторых задач экономики. В данной работе предлагается применение указанных алгоритмов к некоторым задачам статистических данных.

$$\vec{c}_1=\begin{bmatrix}0.43\\0.43\\0.23\\0.44\\0.46\\0.42\end{bmatrix},\ \vec{c}_2=\begin{bmatrix}0.23\\0.38\\-0.89\\-0.07\\0.02\\-0.10\end{bmatrix},\ \vec{c}_3=\begin{bmatrix}0.53\\0.39\\0.38\\-0.40\\-0.27\\-0.44\end{bmatrix},\ \vec{c}_4=\begin{bmatrix}0.11\\0.01\\-0.02\\-0.52\\-0.31\\0.79\end{bmatrix},\ \vec{c}_5=\begin{bmatrix}0.05\\-0.20\\0.00\\-0.58\\0.78\\-0.09\end{bmatrix},\ \vec{c}_6=\begin{bmatrix}0.68\\-0.69\\-0.13\\0.18\\-0.09\\-0.01\end{bmatrix}$$

$$\lambda_1=4.100,\ \lambda_2=0.883,\ \lambda_3=0.639,\ \lambda_4=0.259,\ \lambda_5=0.097,\ \lambda_6=0.022$$

След матрицы равен 6, при этом

– первое собственное значение составляет 68.3% от следа,

– сумма первых двух собственных значений составляет 83.0%,

– сумма первых трех собственных значений составляет 93.7%.

Обсуждение и интерпретация полученных результатов. На первые 3 компоненты приходится 93.7% полной дисперсии «облака». При этом первая компонента имеет смысл общего размера. Это следует из того, что все компоненты у $\vec{c}_1$ одного знака и примерно одинаковы по величине, то есть при проектировании на эту ось координаты нормированных признаков складываются. Вторая компонента в основном отвечает за ширину верхней челюсти (признак 3), поскольку третья координата у $\vec{c}_2$ по абсолютной величине равна 0.89 (почти 1), а вторая – 0.38. Так как знаки этих координат разные, то эти признаки отражают различие в пропорциях челюстей и отличают удлиненные формы от укороченных (гончих и колли от бульдогов и боксеров). Второй и третий признаки у волков и немецких овчарок почти одинаковы. Третья ось противопоставляет размеры челюстей размерам зубов: первые три координаты у $\vec{c}_3$ примерно равны по сумме без знака последним трем, но противоположны по знаку. Эта ось позволяет отличить животных с развитыми зубами (волки, немецкие овчарки, доберманы) от собак других пород (сенбернары, сеттеры).

Заключение

Приведенный метод главных компонент может применяться в различных задачах, где возникают симметрические матрицы. Например, когда исходной информацией об объектах служат экспертные данные о различиях между ними, выраженных числами.

Литература

1 Жамбю М. Иерархический кластер – анализ и соответствия. – М.: Наука «Финансы и статистика», 1988. – 385 с.
2 Елеуов А.А., Отелбаев М.О. Вычисление собственных чисел и собственных векторов матриц // Евразийский математический журнал ЕНУ им. Л.Н. Гумилева и МГУ им. М.В. Ломоносова, г. Астана, 2005. – № 1. – С. 57-78.
3 Елеуов А.А., Алгоритмы счета собственных чисел и собственных векторов матриц // Вестник КазНПУ им. Абая. Серия физика, математика, информатика, 2007. – №1(17). – С.23-28.
4 Елеуов А.А., Отелбаев М.О. Об одном методе нахождения всех собственных чисел матрицы // Совместный выпуск по материалам международной конференции / Вестник КазНУ. Серия математика, механика, информатика. – Алматы, Новосибирск, 2004. – С. 190-193.
5 Елеуов А.А., Отелбаев М.О. К вопросу нахождения всех собственных чисел матрицы // Материалы 10-ой международной межвузовской конференции по математике и механике. – Алматы, 2004. – С. 138-140.
6 Елеуов А.А., Отелбаев М.О. К численному методу треугольного представления матриц // Материалы II международной научно-методической конференция «Математическое моделирование информационных технологий в образовании и науке» (ММ ИТОН), посвященной 75-летию КазНПУ им Абая. – Алматы, 2003. – С. 280-283.

References

1 Zhambju M. Ierarhicheskij klaster – analiz i sootvetstvija. – M.: Nauka «Finansy i statistika», 1988. – 385 s.
2 Eleuov A.A., Otelbaev M.O. Vychislenie sobstvennyh chisel i sobstvennyh vektorov matric // Evrazijskij matematicheskij zhurnal ENU im. L.N. Gumileva i MGU im. M.V. Lomonosova, g. Astana, 2005. – № 1. – S. 57-78.
3 Eleuov A.A., Algoritmy scheta sobstvennyh chisel i sobstvennyh vektorov matric // Vestnik KazNPU im. Abaja. Serija fizika, matematika, informatika, 2007. – №1(17). – S.23-28.
4 Eleuov A.A., Otelbaev M.O. Ob odnom metode nahozhdenija vseh sobstvennyh chisel matricy // Sovmestnyj vypusk po materialam mezhdunarodnoj konferencii / Vestnik KazNU. Serija matematika, mehanika, informatika. – Almaty, Novosibirsk, 2004. – S. 190-193.
5 Eleuov A.A., Otelbaev M.O. K voprosu nahozhdenija vseh sobstvennyh chisel matricy // Materialy 10-oj mezhdunarodnoj mezhvuzovskoj konferencii po matematike i mehanike. – Almaty, 2004. – S. 138-140.
6 Eleuov A.A., Otelbaev M.O. K chislennomu metodu treugol'nogo predstavlenija matric // Materialy II mezhdunarodnoj nauchno-metodicheskoj konferencija «Matematicheskoe modelirovanie informacionnyh tehnologij v obrazovanii i nauke» (MM ITON), posvjashhennoj 75-letiju KazNPU im Abaja. – Almaty, 2003. – S. 280-283.