Лекция 11. Анализ соответствий

9.2. Анализ соответствий

Анализ соответствий[1] можно рассматривать как специальный метод исследования многомерных данных типа ТСП со многими входами. Целью анализа соответствий является представление многомерных нечисловых данных в координатном пространстве латентных переменных малой размерности в надежде получить хорошо интерпретируемую конфигурацию исследуемых объектов (признаков)-точек. Таким образом, методы анализа соответствий по своей сути похожи на методы факторного анализа (см. п. 5.2). Основной проблемой является переход от данных типа ТСП к матрицам типа «объект-объект» при исследовании пространства объектов-точек, или «признак-признак» при исследовании признаков-точек в сжатом координатном пространстве.

Основные понятия анализа соответствий

Пусть дана двухвходовая ТСП **N** размерности ($r \times c$).

Группа	*Категория*				*Всего по строкам*
	1	2	…	c	
1	n_{11}	n_{12}	…	n_{1c}	$n_{1\cdot}$
2	n_{21}	n_{22}	…	n_{2c}	$n_{2\cdot}$
⋮	⋮	⋮	⋱	⋮	⋮
r	n_{r1}	n_{r2}	…	n_{rc}	$n_{r\cdot}$
Всего по столбцам	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot c}$	$n_{\cdot\cdot} = n$

Целью анализа является представление исходных данных в пространствах меньшей размерности, которые бы сохраняли всю или почти всю информацию о различиях между строками или столбцами. Для такого представления можно использовать теорему Экарта-Юнга с различными преобразованиями исходной матрицы данных **N**. При применении метода анализа соответствий оперируют матрицей **Z**, которая является специальной стандартизованной формой матрицы **N**. Прежде чем получить матрицу **Z**, введем понятия *масса*, *инерция* и *профили строк и столбцов*, которые используются в анализе соответствий.

Масса. Сначала вычислим *относительные* частоты таблицы **N**, поделив каждый элемент на общее число наблюдений n. Тогда получим *нормированную* матрицу $\mathbf{P} = \mathbf{N}/n = \{n_{ij}/n\} = \{p_{ij}\}$, $i = 1, \ldots, r$, $j = 1, \ldots, c$, сумма всех элементов которой равна 1, и она показывает как распределена

[1] *Методы анализа соответствий разрабатывались многими авторами, и они известны под названиями: оптимальное шкалирование, оптимальная оцифровка, квантификационный анализ, взаимное усреднение, дуальное шкалирование.*

единичная *масса* по ячейкам. Масса каждой строки и столбца определяется по формулам

$$\mathbf{w}_r = \mathbf{P}\mathbf{1}_r, \quad \mathbf{w}_c = \mathbf{P}'\mathbf{1}_c, \tag{9.4}$$

где $\mathbf{1}_r$, $\mathbf{1}_c$ – векторы размерности r и c с единичными элементами.

Профили строк. Для получения профилей строк элементы p_{ij} нужно поделить на w_{r_i}: $p_{ij}^{(r)} = p_{ij} / w_{r_i}$.

Тогда получим

Группа	*Категория*				*Всего по строкам*
	1	2	⋯	c	
1	$p_{11}^{(r)}$	$p_{12}^{(r)}$	⋯	$p_{1c}^{(r)}$	1
2	$p_{21}^{(r)}$	$p_{22}^{(r)}$	⋯	$p_{2c}^{(r)}$	1
⋮	⋮	⋮	⋱	⋮	⋮
r	$p_{r1}^{(r)}$	$p_{r2}^{(r)}$	⋯	$p_{rc}^{(r)}$	1
Масса	w_{c_1}	w_{c_2}	⋯	w_{c_c}	

Масса каждого столбца w_{c_j} вычисляется по формуле (9.4). Сумма элементов строки равна 1 и каждый элемент $p_{ij}^{(r)}$ в матрице профилей строк интерпретируется как условная вероятность того, что элемент i-й строки принадлежит столбцу j.

Профили столбцов. В этом случае элементы p_{ij} нужно поделить на w_{c_j}: $p_{ij}^{(c)} = p_{ij} / w_{c_j}$. Масса каждой строки w_{r_i} вычисляется по формуле (9.4). Сумма элементов столбца равна 1 и каждый элемент $p_{ij}^{(c)}$ в матрице профилей столбцов интерпретируется как условная вероятность того, что элемент *j-го* столбца принадлежит i-й строке. Тогда получим

Группа	*Категория*				*Масса*
	1	2	⋯	c	
1	$p_{11}^{(c)}$	$p_{12}^{(c)}$	⋯	$p_{1c}^{(c)}$	w_{r_1}
2	$p_{21}^{(c)}$	$p_{22}^{(c)}$	⋯	$p_{2c}^{(c)}$	w_{r_2}
⋮	⋮	⋮	⋱	⋮	⋮
r	$p_{r1}^{(c)}$	$p_{r2}^{(c)}$	⋯	$p_{rc}^{(c)}$	w_{r_r}
Всего по столбцам	1	1	⋯	1	

Расстояние точки (объекта или признака) до центра масс определяется по формулам:

$$d_{r_i} = \left[\sum_j \frac{1}{w_{c_j}} \left(\frac{P_{ij}}{w_{r_i}} - w_{c_j} \right)^2 \right]^{1/2}, \quad d_{c_j} = \left[\sum_i \frac{1}{w_{r_i}} \left(\frac{P_{ij}}{w_{c_j}} - w_{r_i} \right)^2 \right]^{1/2}.$$

Инерция. Термин инерция является аналогом понятия «момента инерции» в прикладной математике, которая определяется как интеграл произведения элементов массы на квадрат расстояния до центра масс. Тогда инерция каждой строки и столбца определятся по формулам

$$Inr_i = w_{r_i} \cdot d_{r_i}^2, \; Inc_j = w_{c_j} \cdot d_{c_j}^2.$$

Произведя преобразования, можно показать, что инерции определяются как значения $\mathbf{X}^2$ статистики Пирсона для соответствующих строк и столбцов исходной ТСП, деленное на объем выборки n. При этом суммы инерций исходной и преобразованной систем по всем строкам или столбцам должны совпадать.

Относительная инерция. Относительная инерция каждой строки или столбца вычисляется по формуле

$$Inotr_i = \frac{w_{r_i} \cdot d_{r_i}^2}{\sum_i w_{r_i} \cdot d_{r_i}^2}, \; Inotc_j = \frac{w_{c_j} \cdot d_{c_j}^2}{\sum_j w_{c_j} \cdot d_{c_j}^2}.$$

Если вычисления проведены правильно, то сумма относительной инерции по всем строкам или столбцам должна равняться 1.

Порядок описания данных мультипликативной моделью

Для описания исходных данных мультипликативной моделью используем теорему Экарта-Юнга. Так как теорема верна при различных преобразованиях исходной ТСП, то будем применять такое преобразование, которое использует введенное выше понятие инерции. Порядок этих преобразований имеют вид:

1. Вычисляются суммарные значения $\mathbf{w}_r$, $\mathbf{w}_c$ по формулам (9.4).
2. Векторы масс $\mathbf{w}_r$ и $\mathbf{w}_c$ преобразуются в диагональные матрицы

$$\mathbf{D}_r = diag(\mathbf{w}_r^{-1/2}), \; \mathbf{D}_c = diag(\mathbf{w}_c^{-1/2}).$$

2. Тогда нормированная матрица $\mathbf{P}$ будет равна

$$\mathbf{Y} = \mathbf{D}_r \mathbf{P} \mathbf{D}_c.$$

3. Умножая $\mathbf{Y}$ справа на $\mathbf{D}_c^{-1}\mathbf{1}$ и слева на $\mathbf{1}'\mathbf{D}_r^{-1}$, получаем

$$\mathbf{Y}\mathbf{D}_c^{-1}\mathbf{1} = \mathbf{D}_r\mathbf{P}\mathbf{1} = \mathbf{D}_r^{-1}\mathbf{1};$$

$$\mathbf{1}'\mathbf{D}_r^{-1}\mathbf{Y} = \mathbf{1}'\mathbf{P}\mathbf{D}_c = \mathbf{1}'\mathbf{D}_c^{-1}.$$

Сравнив полученные выражения с формулой (9.3), видим, что $\mathbf{D}_r^{-1}\mathbf{1}$ и $\mathbf{1}'\mathbf{D}_c^{-1}$, есть пара сингулярных векторов, соответствующих *единичному* сингулярному значению. Так как матрица $\mathbf{Y}$ положительная и состоит из элементов, меньших единицы, то из теоремы Фробиниуса-Перрона сле-

дует, что единичное сингулярное значение является максимальным. Элементы собственных векторов, соответствующие этому сингулярному значению, равны суммам строк и столбцов и не могут быть использованы непосредственно для вывода координатного описания мультипликативной модели. Исключив влияние эффектов строк и столбцов, получаем следующее сингулярное разложение

$$\mathbf{Z} = \mathbf{D}_r\mathbf{P}\mathbf{D}_c - \mathbf{D}_r^{-1}\mathbf{1}\mathbf{1}'\mathbf{D}_c^{-1} = \sum_{i=2}\sigma_i\mathbf{u}_i\mathbf{v}_i' = \mathbf{U}\Sigma\mathbf{V}'. \qquad (9.5)$$

Правая часть выражения (9.5) есть сингулярное разложение матрицы **Z**, элементы которой равны

$$z_{ij} = \frac{n_{ij}}{\sqrt{n_{i\bullet}n_{\bullet j}}} - \frac{\sqrt{n_{i\bullet}n_{\bullet j}}}{n} = \frac{n_{ij} - n_{i\bullet}n_{\bullet j}/n}{\sqrt{n_{i\bullet}n_{\bullet j}}}. \qquad (9.6)$$

Выражение (9.6) представляет собой квадратный корень из элемента χ^2-статистики Пирсона, деленное на n, которая применяется для проверки гипотезы независимости групп-строк категорий-столбцов (см. (3.5)).

Таким образом, анализ соответствий можно рассматривать как декомпозицию χ^2-статистики для ТСП, которая обеспечивает два вида шкалирования: для групп-строк и для категорий-столбцов

$$\mathbf{F} = \mathbf{D}_r\mathbf{U}\Sigma, \qquad (9.7)$$

$$\mathbf{G} = \mathbf{D}_c\mathbf{V}\Sigma. \qquad (9.8)$$

Выбор размерности координатного пространства. Выбор размерности пространства можно осуществить отдельно как для групп-строк, так и для категорий-столбцов. Размер базиса для групп-строк или категорий-столбцов в евклидовом пространстве выбирается, как и в факторном анализе, по доле вклада собственных значений или доли инерции (следовательно, и величины хи-квадрат) строк (столбцов) в общую инерцию в зависимости от числа координат пространства.

Оценка качества решения. Качество решения определяется точностью представления расстояний между точками в пространстве более низкой размерности. Если используются максимальная размерность $(\min(r,c)-1)$, то все расстояния воспроизводятся без ошибок.

Сначала рассмотрим задачу для разложения по строкам. Найдем значения относительной инерции при выборе координаты $F^{(k)}$ к общей инерции исходной системы

$$InotF_i^{(k)} = \frac{w_{r_i}\cdot(F_i^{(k)})^2}{\sum_i w_{r_i}\cdot(F_i^{(k)})^2}, \quad k = 1,\ldots,m;\ i = 1,\ldots,r, \qquad (9.9)$$

где $m < r$ – размерность выбранного координатного пространства. По величине $InotF_i^{(k)}$ можно судить, какая координата вносит больший вклад в относительную инерцию системы.

Оценку качества решения по координате $F_i^{(k)}$ можно определить по отношению величин инерций полученной и исходной координат

$$\gamma(F_i^{(\kappa)}) = \frac{w_{r_i} \cdot (F_i^{(k)})^2}{w_{r_i} \cdot d_{r_i}^2} = \frac{(F_i^{(k)})^2}{d_{r_i}^2}.$$

Величина $\gamma(F_i^{(k)})$ интерпретируется как корреляция i-го объекта с координатной осью k. Качество решения для координатного пространства размерности m определяется по формуле

$$\gamma(F_i^{(1)} \ldots F_i^{(k)}) = \sum_{k=1}^{m} \gamma F_i^{(k)}, \quad .$$

Если величина $\gamma(F_i^{(1)} \ldots F_i^{(k)})$ для i-й строки мала (например, меньше 0,1), то выбранный размер m координатного пространства мал и недостаточно хорошо представляет данную строку.

Качество решения для разложения по точкам столбцам по каждой отдельно взятой координате G_j можно определить по вышеприведенным формулам, заменив F на G, а индексы i на j, r на c.

Тогда получим

$$InotG_j^{(k)} = \frac{w_{c_j} \cdot (G_j^{(k)})^2}{\sum_j w_{c_j} \cdot (G_j^{(k)})^2}, \quad k = 1,\ldots,m; \; i = 1,\ldots,c,$$

где $m < c$ – размерность выбранного координатного пространства

$$\gamma\left(G_j^{(\kappa)}\right) = \left(G_j^{(k)}\right)^2 / d_{c_j}^2, \; \gamma_j\left(G_j^{(1)} \ldots G_j^{(k)}\right) = \sum_{k=1}^{m} \gamma G_j^{(k)}.$$

Углы точек, исходящие от центра масс. Углы при разложении по строкам (формула (9.7)) и столбцам (формула (9.8)) относительно абсциссы (строки) и ординаты (столбца) определяются по формулам

$$\alpha F_i^{(k)} = \arccos(\sqrt{\gamma(F_i^{(\kappa)})}, \quad \alpha G_j^{(k)} = \arccos(\sqrt{\gamma(G_j^{(\kappa)})}.$$

Статистическая значимость анализа соответствий. Анализ соответствий является разведочным методом, и он разработан на методологии построения моделей с точки зрения их соответствия данным, а не наоборот. Отсюда следует, что не существует статистических гипотез, которые могут быть применены для проверки результатов этого анализа.

Пример 9.1[1]. Рассмотрим построение структурной модели дифференциальной диагностики заболевания артериальной гипертонией по исходным данным, полученным на основе клинических карт 88 больных, находящихся на стационарном лечении (табл. 9.1).

---

[1] *Данные получены канд. мед. наук Н. А. Маршалкиной., клиника СГМУ. г. Саратов.*

Таблица 9.1. Частота признака при различных диагнозах

Признаки	*Значения признака*	*Д1*	*Д2*	*Д3*	*Д4*	*Всего по строке*	*Признаки*	*Значения признака*	*Д1*	*Д2*	*Д3*	*Д4*	*Всего по строке*
Пол	Муж.	15	2	17	4	38	Причины для госпитализации	1	3	1	26	1	31
	Жен.	15	7	26	2	50		2	16	5	12	4	37
Увеличение левого желудочка	Нет	14	3	24	3	44		3	10	2	5	0	17
	Есть	16	6	19	3	44		4	1	1	0	1	3
Диабет	Нет	23	9	19	4	75	Учащение пароксизмов мерцательной аритмии	Нет	12	3	23	2	40
	Есть	7	0	4	2	13		Есть	18	6	20	4	48
Гипертония	Нет	3	2	7	1	13	Функциональный класс недостаточности кровообращения	1	7	2	11	2	22
	Есть	27	7	36	6	75		2	21	6	31	4	62
Пароксизмальная наджелудочная тахикардия	Нет	11	2	21	2	36		3	2	1	1	0	4
	Есть	19	17	22	4	52	Иммуноглобулины класса M (хламидия пневмонии)	0	30	7	39	6	82
Гипоталамические нарушения	Нет	25	7	34	4	70		1	0	10	0	1	1
	Есть	5	2	9	2	18		2	0	1	4	0	5
Инфаркт миокарда	Нет	25	2	26	6	59	Иммуноглобулины класса G (хламидия пневмонии)	0	12	0	12	1	25
	Есть	5	7	17	0	29		1	18	9	31	5	63
Хроническая обструктивная болезнь легких	Нет	23	5	35	5	68	Иммуноглобулины класса M (цитомегаловирус)	0	29	9	43	6	87
	Есть	7	4	8	1	20		1	1	0	0	0	1
Патология щитовидной железы	Нет	28	9	38	5	80	Иммуноглобулины класса G (цитомегаловирус)	0	5	2	11	2	20
	Есть	2	0	5	1	8		1	25	7	32	4	68
С-реактивный белок	Нет	26	7	34	6	73	**Всего по столбцу**		**510**	**153**	**731**	**102**	**1496**

Причины госпитализации: 1 – развитие нестабильной стенокардии, 2 – прогрессирование недостаточности кровообращения, 3 – учащение пароксизмов мерцания, 4 – «плановая» госпитализация. В соответствии с диагнозами больные разделены на 4 группы: D1 – ишемическая кардиопатия; D2 – перенесенный инфаркт; D3 – нестабильная стенокардия; D4 – стабильная стенокардия. Экспертами отобраны 18 наиболее значимых признаков. Большая часть признаков выражены дихотомическими данными, остальная часть – полихотомическими. Все признаки преобразованы к квазиквантативным данным, что увеличило размерность признаков до 39.

Для построения структурной модели диагнозов применим процедуру анализа соответствий. После сингулярного разложения матрицы данных получаем сингулярныечисла σ_i, определяем доли собственных значений $\lambda := \sigma^2$ в процентах от общей суммы $\Delta\lambda_j = (\lambda_j/\Sigma_j\lambda_j)$ 100 % и накопленный процент. Находим χ^2- статистику с ч.с.с. $(r-1)(c-1)$ $\chi^2 := \Sigma_i\Sigma_j(Z_{i,j})n$ и инерции по диагнозам.

Собственные числа и их доли приведены в табл. 9.2. При отображении конфигурации в двухмерном пространстве сохраняется 86,62 % информации об исходных данных относительно диагнозов.

Таблица 9.2. Собственные числа конфигурации

Число координат	λ_j	*Пропорции* λ_j	
		Доли, %	*Накопленный, %*
1	0,030734	46,98	46,98
2	0,025928	39,63	86,62
3	0,008755	13,38	100
Всего	0,065417		

Полная сводка результатов анализа соответствий при выборе двухмерной конфигурации приведена в табл. 9.3.

Таблица 9.3. Сводка результатов сингулярного разложения по столбцам

Диагноз	*Масса*	*Инерци* Inc_j	*Дистанция* d_{c_j}	*Координаты*		*Качество в пространстве координат*		
				G1	*G2*	*γG1*	*γG1*	*γG1 G2*
1	0,316	0,243	0,047	0,199	–0,044	0,852	0,042	0,894
2	0,233	0,368	0,236	–0,239	–0,422	0,242	0,756	0,998
3	0,321	0,220	0,029	–0,123	0,120	0,512	0,488	1,000
4	0,130	0,168	0,061	0,241	–0,007	0,361	0,000	0,362

Из таблицы видно, что основную инерцию составляет диагноз D2, диагнозы D1 и D3 несут приблизительно одинаковую инерцию, а диагноз D4 слабо связан с гипотетическим центром тяжести конфигурации.

Из показателей качества отображения в двухмерное пространство видно, что диагнозы D2 и D3 полностью описываются выбранным пространством, а качество описания диагнозов D1 и D4 составляет соответственно 89 и 36 %.

Найденная конфигурация точек структурной модели изображена на рис. 9.1, из которого видно, что диагнозы D1 и D4 расположены близко, что позволяет отнести их в одну группу, т. е. можно говорить о большом сходстве заболеваний ишемической кардиопатией и стабильной стенокардией.

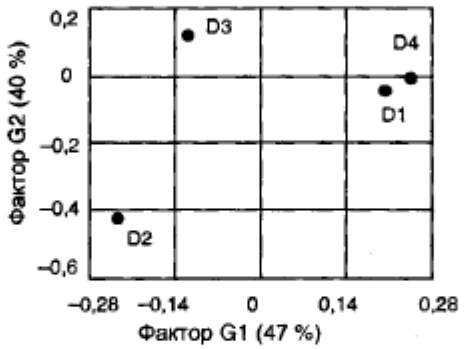

Рис. 9.1. Конфигурация диагнозов заболевания

Полученная конфигурация интерпретирована врачами-экспертами следующим образом: при этих диагнозах в организме больных протекают схожие процессы и, следовательно, должны быть близкими методы лечения и применяемых лекарственных средств.

При этом развитие гипертонической болезни протекает от ишемической кардиопатии (D1), через перенесенный инфаркт (D2), она переходит к нестабильной стенокардии (D3), которая затем стабилизируется и становится стабильной стенокардией (D4). При этом стабильная стенокардия аналогична ишемической кардиопатии, но на новом качественном уровне.