Математические структуры и моделирование
2014. № 4(32). С. 53–58

УДК 519.237.07

ВЕРОЯТНОСТНЫЙ МЕТОД ФОРМИРОВАНИЯ СИМПТОМОКОМПЛЕКСОВ

В.В. Гольтяпин
старший научный сотрудник, к.ф.-м.н., доцент, e-mail: goltyapin@mail.ru

Омский филиал Учреждения Российской академии наук Института математики им. С.Л. Соболева Сибирского отделения РАН, г. Омск

Аннотация. Разработан метод и построен вычислительный алгоритм, позволяющие формировать диагностические симтокомплексы с помощью вероятностного метода распознавания. В рамках теории латентного анализа сформулированы утверждение, лемма и теорема, позволяющие находить апостериорные вероятности на базе альтернативных показателей с использованием ортогональной факторной структуры.

Ключевые слова: симптомокомплекс, факторная модель, латентная модель, корреляционный анализ, маргинальное распределение, маргинал.

1. Введение

Известно, что многомерный факторный анализ применяется преимущественно для обработки количественных показателей. В основе любого факторного исследования лежит корреляционная матрица, полученная на базе исходных количественных показателей, имеющих нормальное распределение или хотя бы близкое к нему [1, 2]. Но, к сожалению, не всегда переменные могут быть измерены количественно. Особо часто встречаются переменные, которые обладают альтернативной вариацией. Нижеследующий математический аппарат позволяет использовать модели факторного анализа для обработки альтернативных данных и формировать зависимые и независимые симптомокомплексы с помощью латентного анализа.

2. Постановка задачи

Обозначим количество объектов исследования как n — объем выборки, а количество измеряемых параметров как m — размерность выборки. Тогда исходные альтернативные данные для факторного исследования представляются в виде таблицы, столбцы которой — объекты исследования, а строки — значения измеряемых параметров у конкретного объекта. Далее полученная таблица записывается в виде матрицы Y размерности $m \times n$. Элемент этой матрицы обозначим как y_{ij}, где индекс $i{=}1, \ldots, m$ относится к параметрам, а индекс $j{=}1, \ldots, n$ – к объектам, α_i — число единиц в i-ой строке матрицы Y.

Каждый столбец этой матрицы можно рассматривать в виде вектора-объекта исследования и обозначать $\vec{y}_j = (y_{1j}, \ldots, y_{mj})$. Стандартное отклонение и выборочное математическое ожидание i-го показателя обозначим соответственно как $s_i = \sqrt{\frac{1}{n-1}\left(\alpha_i - \frac{\alpha_i^2}{n}\right)}$ и $\bar{y}_i = \frac{\alpha_i}{n}$.

Для альтернативных данных модель факторного анализа предлагается в следующем виде:

$$Z = AF, \tag{1}$$

где Z — матрица стандартизованных данных, полученных по формуле

$$z_{ij} = \frac{y_{ij} - \bar{y}_t}{s_i} = \frac{y_{ij} - \frac{\alpha_i}{n}}{\sqrt{\frac{1}{n-1}\left(\alpha_i - \frac{\alpha_i^2}{n}\right)}} \tag{2}$$

размерности $m \times n$, A — матрица факторного отображения размерности $m \times r$, F — матрица факторных значений размерности $r \times n$, r — количество выделяемых факторов. Ниже следующая теорема позволяет нам строить факторные модели на альтернативных показателях.

Теорема 1. Фундаментальная теорема Терстоуна справедлива для альтернативных данных [1].

Полагая отсутствие корреляционной зависимости между факторами, матрицу весовых нагрузок факторов A можно находить различными ортогональными методами факторного анализа (метод главных факторов с варимакс вращением, метод минимальных остатков, метод максимума правдоподобия и т.д.) [1, 2].

Для определения количества выделяемых факторов предлагается выбрать одну или несколько процедур: определение, основанное на предварительной информации; определение, основанное на собственных значениях факторов (критерий «каменистой осыпи»); определение на основе процента объяснённой дисперсии; метод расщепления и критерии значимости.

3. Вероятностный метод формирования диагностических симптомокомплексов

Суть латентного анализа состоит в обработке теста или анкеты, состоящей из r вопросов, которые относятся к изучаемой скрытой характеристике. Выделенные вопросы называют явными переменными, а скрытую характеристику — латентной переменной. В теории тестов скрытая характеристика интерпретируется как одномерный латентный континуум (непрерывная латентная переменная) [4, 5]. Переходя непосредственно к построению латентной модели на базе альтернативных данных, введём следующие обозначения: p_i — отношение числа лиц, положительно ответивших на i-ый вопрос к общему числу респондентов; p_{ij} — отношение числа лиц, положительно ответивших на i-ый

и j-ый вопросы к общему числу респондентов; $p_{i\bar{j}}$ — отношение числа лиц, положительно ответивших на i-ый и отрицательно на j-ый вопросы, к общему числу респондентов; $p_{\bar{i}\bar{j}}$ — отношение числа лиц, отрицательно ответивших на i-ый и j-ый вопросы к общему числу респондентов; p_{ijk} — отношение числа лиц, положительно ответивших на i-ый, j-ый и k-ый вопросы, к общему числу респондентов; $p_{i\bar{j}k}$ — отношение числа лиц, положительно ответивших на i-ый и k-ый вопросы и отрицательно на j-ый, к общему числу респондентов; $p_{\bar{i}\bar{j}k}$ — отношение числа лиц, отрицательно ответивших на i-ый и j-ый вопросы при положительном ответе на k-ый вопрос, к общему числу респондентов; $\tilde{\varphi}(x_i)$ — частота, соответствующая относительному объему i-го класса; $\tilde{f}_j(x_i)$ — вероятность положительного ответа респондента на j-ый вопрос, находясь в i-ом классе; $\tilde{f}_{jk}(x_i)$ — вероятность положительного ответа респондента на j-ый и k-ый вопросы, находясь в i-ом классе; $\tilde{f}_{123}(x_i)$ — вероятность положительного ответа респондента на первый, второй и третий вопросы, находясь в i-ом классе.

Используя эти данные при построении латентной модели на базе альтернативных при наличии трех вопросов и двух латентных классов, получаем дискретные классы респондентов и разрешимую систему уравнений с дискретными переменными:

$$\begin{cases} \tilde{\phi}(x_1) + \tilde{\phi}(x_2) = 1 \\ p_1 = \tilde{f}_1(x_1)\tilde{\phi}(x_1) + \tilde{f}_1(x_2)\tilde{\phi}(x_2) \\ p_2 = \tilde{f}_2(x_1)\tilde{\phi}(x_1) + \tilde{f}_2(x_2)\tilde{\phi}(x_2) \\ p_3 = \tilde{f}_3(x_1)\tilde{\phi}(x_1) + \tilde{f}_3(x_2)\tilde{\phi}(x_2) \\ p_{12} = \tilde{f}_{12}(x_1)\tilde{\phi}(x_1) + \tilde{f}_{12}(x_2)\tilde{\phi}(x_2) \\ p_{13} = \tilde{f}_{13}(x_1)\tilde{\phi}(x_1) + \tilde{f}_1(x_2)\tilde{\phi}(x_2) \\ p_{23} = \tilde{f}_{23}(x_1)\tilde{\phi}(x_1) + \tilde{f}_{23}(x_2)\tilde{\phi}(x_2) \\ p_{123} = \tilde{f}_{123}(x_1)\tilde{\phi}(x_1) + \tilde{f}_{123}(x_2)\tilde{\phi}(x_2). \end{cases} \tag{3}$$

Для решения системы уравнения (3) предлагается воспользоваться результатами и формулами из ниже следующего утверждения, леммы и теоремы.

Утверждение. При известности вероятностей латентной модели с двумя классами и тремя вопросами нахождение частоты $\tilde{\varphi}(x_i)$ сводится к каноническому уравнению прямой с точкой (p_1, p_2, p_3) и направляющим вектором

$$\vec{n} = (f_1(x_i) - f_1(x_j), f_2(x_i) - f_2(x_j), f_3(x_i) - f_3(x_j)),$$

где $i \neq j$.

Лемма. Отношение условного произведения трёх вопросов к произведению двух вопросов в латентной модели равно произведению вероятностей положительного ответа респондента на условный вопрос.

Теорема 2. Наличие всех маргиналов для латентной модели с двумя классами и тремя вопросами позволяет свести поиск всех неизвестных вероятностей к решению трёх квадратных уравнений.

Далее предполагается совместное использование латентной модели и ортогональной факторной структуры для построения алгоритма метода. Первая задача метода — сформировать набор симптомокомплексов, опираясь на ортогональную факторную структуру с учётом уровня значимости φ коэффициента по χ^2 критерию. Вторая задача метода — для каждого симптомокомплекса найти диагностическую шкалу на базе простейшей латентно-структурной модели.

Для упрощения в целях дальнейшего изложения введём функцию

$$\gamma_{lk}(y_{i,j}) = \begin{cases} \tilde{f}_{ik}(x_l), & \text{если } y_{ij} = 1 \\ 1 - \tilde{f}_{ik}(x_l), & \text{если } y_{ij} = 0, \end{cases} \tag{4}$$

где l — номер класса и может принимать значение 1 или 2, k — номер симптомокомплекса, $\tilde{f}_{ik}(x_l)$ — вероятность положительного ответа респондента из l-ого класса на i-ый вопрос, выбранный как параметр, составляющий симптомокомплекс. Условием вхождения параметров в зависимый или независимый симтокомплекс является значение весовых нагрузок соответствующего фактора на уровне не ниже $0,5$.

Вероятность принадлежности первому классу вычисляется посредством формулы Баейса с использованием введённой функции

$$p(1|y_{a_kj}, y_{b_kj}, y_{c_kj}) = \frac{\gamma_{1k}(y_{a_kj})\gamma_{1k}(y_{b_kj})\gamma_{1k}(y_{c_kj})}{\sum\limits_{i=1} 2\gamma_{ik}(y_{a_kj})\gamma_{ik}(y_{b_kj})\gamma_{ik}(y_{c_kj})}, \tag{5}$$

где a_k, b_k, c_k — номера трёх параметров k-го симптомокомплекса.

Алгоритм метода:

1. Из матрицы Y путём элементарного преобразования получаем матрицу Z размерности $m \times n$.
2. Вычисляем корреляционную матрицу R.
3. С целью исключения незначимых показателей вычисляем вероятностные значения уровней зависимости по формуле $\chi^2 = n \cdot \varphi$ при единичной степени свободы.
4. Определяем наименьшее количество выделяемых факторов (критерий Гуттмана, критерий «каменной осыпи» или другой адекватный критерий) [1].
5. Находим общности любым из известных методов (лучше взять метод минимальных остатков) [1, 2, 6].
6. Вычисляем первичную ортогональную матрицу весовых нагрузок факторов A размерности $m \times r$ (метод главных факторов, метод минимальных остатков или любой другой адекватный метод) [1, 2, 6].
7. Полученную на предыдущем шаге матрицу весовых нагрузок подвергаем ортогональномоу вращению в соответствии с варимакс критерием [2, 6].

8. Осуществляем анализ ортогональной факторной структуры, полученной после вращения, и формируем зависимые и независимые симптомокомплексы.
9. Для каждого симптомокомплекса формируем диагностическую шкалу, вычисляя маргиналы и решая систему уравнений (3), используя результаты теоремы 2.
10. По формуле (4) вычисляем частные апостериорные вероятности для всех объектов исследования.

4. Заключение

Дано математическое обоснование возможности применения альтернативных данных в факторном исследовании. Для латенто-структурной модели сформулированы и доказаны: Утверждение, позволяющее находить относительный объем соответствующего класса через каноническое уравнение прямой; Лемма о соотношении условного произведения трёх вопросов к произведению двух вопросов; Теорема о сведении решения системы уравнений латентно-структурной модели к решению трёх квадратных уравнений.

На базе полученных теоретических выкладок построен вычислительный алгоритм, позволяющий строить диагностические симтокомплексы на базе альтернативных данных, оптимальной ортогональной факторной структуры, простейшей латентно-структурной модели и формулы Баейса.

Литература

1. Иберла К. Факторный анализ. М.: Статистика, 1980.
2. Харман Г. Современный факторный анализ. М.: Статистика, 1972.
3. Кендалл М.Дж., Стюарт А.Т. Статистические выводы и связи. М.: Наука 1973.
4. Осипов Г.В. Методы измерения в социологии. М.: Наука, 2003.
5. Lazarsfeld P.F. The logical and mathematical foundation of latent structure analysis // Measurement and Prediction. N.Y., 1950.
6. Гольтяпин В.В. Вычислительные аспекты метода минимальных остатков при разрешении варианта Хейвуда // Сибирский журнал индустриальной математики. 2005. Том VII, № 3(23). С. 145–151.

THE PROBABILISTIC METHOD OF THE SET OF SYMPTOMS FORMATION

V.V. Goltyapin
PhD(Math.), Associate Professor, Senior Reseacher, e-mail: goltyapin@mail.ru

Omsk Branch of Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Science, Omsk

Abstract. This paper presents a method for calculating the redistribution of the initial and / or ongoing resources in a wide range of practical problems of optimal control. The presented method is based on a special extension of the dynamic equations of the system that formalizes the original problem. Additional terms of the right sides of the equations describe the managed switch in currents linking the components of the system state. Within a given intensity-sharing switch allows arbitrary reallocation of module components (resources) while maintaining their current amount. The form of the function, majorizing flow rate, is determined by the type of task. In problems of the first type accommodation of the initial resources is a part of optimizable initial conditions, and the reallocation of resources to the control interval is prohibited or physically impossible. In problems of the second type initial conditions are hard coded, but it is allowed to reallocate current total resources in the control interval. In problems of the third type it is allowed to optimize the initial resource accommodation as well as the reallocation of resources in the control interval.

Keywords: accommodation, reallocation, resources, flows, switching, dynamic systems, neural networks.