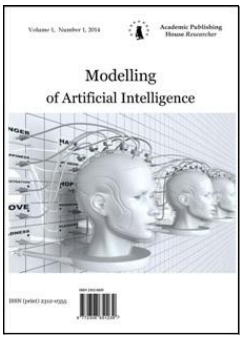
Published in the Russian Federation
Modeling of Artificial Intelligence
Has been issued since 2014.
ISSN: 2312-0355
Vol. 6, Is. 2, pp. 52-58, 2015

DOI: 10.13187/mai.2015.6.52
www.ejournal11.com



UDC 519.237.07

Algorithm of Fordiasimpt's Method on an Independent Symptom

V.V. Goltyapin

Omsk Branch of Sobolev Institute of Mathematics,
Siberian Branch of the Russian Academy of Science, Russian Federation
644043, Omsk, Pevtsova Str., 13
E-mail: goltyapin@mail.ru

Abstract

In this article we proved the theorem, allowing us to find the posterior probabilities on the basis of alternative data and orthogonal factor structure. The numerical algorithm of the probabilistic method pattern recognition is building for formation of the diagnostic scale.

Keywords: symptom, latency analysis, factor model, correlation analysis, the marginal distribution, marginal.

Введение

Основная цель работы выявить преимущества алгоритма метода ФОРДИАСИМПТ в случае независимых симптомокомплексов.

Первая задача ФОРДИАСИМПТ — сформировать набор симптомокомплексов опираясь на ортогональную факторную структуру и на уровень значимость φ коэффициента по χ^2 критерию. Вторая задача ФОРДИАСИМПТ — для каждого симптомокомплекса найти диагностическую шкалу на базе простейшей латентно-структурной модели.

В силу обоснованости использование факторного анализа для альтернативных показателей, считаем известным матрицу ортогонального факторного отображения [1][2]. С полным изложение теоретических основ алгоритма ФОРДИАСИМПТ, относящихся к поиску факторной структуры, можно ознакомится в следующей работе [3].

Особое внимание в работе уделяется непосредственно математическому аппарату используемого в построение латентной модели на базе альтернативных данных.

Материалы и методы

Обозначим количество объектов исследования за n — объем выборки, а количество измеряемых параметров за m — размерность выборки. Тогда исходные альтернативные данные представляются в виде таблицы, столбцы которой — объекты исследования, а строки — значения измеряемых параметров у конкретного объекта.

Далее введем следующие обозначения:

p_i — отношение количества объектов к n, у которых i-ый показатель равен 1;

p_{ij} — отношение количества объектов к n, у которых i-ый и j-ый показатели равны 1;

$p_{i\bar{j}}$ — отношение количества объектов к n, у которых i-ый показатель равен 1, j-ый показатель равны 0;

$p_{\bar{i}\bar{j}}$ — отношение количества объектов к n, у которых i-ый и j-ый показатели равны 0;

p_{ijk} — отношение количества объектов к n, у которых i-ый, j-ый и k-ый показатели равны 1;

$p_{i\bar{j}k}$ — отношение количества объектов к n, у которых i-ый и k-ый показатели равны 1, а j-ый показатель равен 0;

$p_{\bar{i}jk}$ — отношение количества объектов к n, у которых i-ый и j-ый показатели равны 0, а k-ый показатель равен 1;

$\tilde{\phi}(x_t)$ — частота, соответствующая относительному объему t-го класса;

$\tilde{f}_i(x_t)$ — вероятность значения 1 по i-му показателию у объекта находящегося в t-ом классе;

$\tilde{f}_{ik}(x_t)$ — вероятность значения 1 по i-му и k-му показателям у объекта находящегося в t-ом классе;

$\tilde{f}_{ijk}(x_t)$ — вероятность значения 1 по i-му, j-му и k-му показателям у объекта находящегося в t-ом классе;

Однозначное разделение объектов по трем показателям на два латентных класса позволяет сформировать разрешимую систему уравнений с дискретными переменными:

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

Определение. Отношения p_i, p_{ij}, p_{ijk} называются маргиналами.

Определение. Латентно-структурная модель называется простейшей, если для её построения используются три альтернативных показателя.

Определение. Диагностической шкалой называется набор апостериорных вероятностей полученных с помощью простейшей латентно-структрной модели и формулы Баейса, позволяющей отнести объект исследования к одному из двух сформированных классов.

Определение. Симптомокомплекс — тройка альтернативных показателей, используемых для построения диагностической шкалы в методе ФОРДИАСИМПТ.

Определение. Два симптомокомплекса считаются зависимыми, если они содержат один и более общих параметров, в противном случае считается, что эти два симтомокомплекса независимы.

В целях дальнейшего изложения теоретического аппарата введем следующие обозначения:

$$A_{ij} = \begin{pmatrix} p_{ij}\, p_{\bar{i}\bar{j}} \\ p_{\bar{i}\bar{j}}\, p_{\bar{i}\bar{j}} \end{pmatrix},\ A_{ij|k} = \begin{pmatrix} p_{ijk}\, p_{\bar{i}\bar{j}k} \\ p_{\bar{i}jk}\, p_{\bar{i}\bar{j}k} \end{pmatrix},\ A_{ij|\bar{k}} = \begin{pmatrix} p_{ij\bar{k}}\, p_{i\,\overline{jk}} \\ p_{\bar{i}j\bar{k}}\, p_{\overline{ijk}} \end{pmatrix}.$$

Тогда определители вышеуказанных матриц равны $|A_{ij}| = p_{ij}\, p_{\bar{i}\bar{j}} - p_{\bar{i}\bar{j}}\, p_{\bar{i}\bar{j}}$, $|A_{ij|k}| = p_{ijk}\, p_{\bar{i}\bar{j}k} - p_{\bar{i}jk}\, p_{\bar{i}jk}$, $|A_{ij|\bar{k}}| = p_{ij\bar{k}}\, p_{\bar{i}\bar{j}\bar{k}} - p_{\bar{i}j\bar{k}}\, p_{i\,\overline{jk}}$ и называются произведением i-го и j-го показателей при условии (или без такового), что k-ый показатель равен 0 или 1.

Теорема. Наличие всех маргиналов в простейшей латентно-стуктурной модели позволяет свести поиск всех неизвестных вероятностей к решению трех квадратных уравнений:

$$x^2 - \left(1 - \frac{|A_{ij|\bar{k}}|}{|A_{ij}|} + \frac{|A_{ij|k}|}{|A_{ij}|}\right)x + \frac{|A_{ij|k}|}{|A_{ij}|} = 0,\ y^2 - \left(1 - \frac{|A_{jk|\bar{i}}|}{|A_{jk}|} + \frac{|A_{jk|i}|}{|A_{jk}|}\right)y + \frac{|A_{jk|i}|}{|A_{jk}|} = 0,$$

$$z^2 - \left(1 - \frac{|A_{ik|\bar{j}}|}{|A_{ik}|} + \frac{|A_{ik|j}|}{|A_{ik}|}\right)z + \frac{|A_{ik|j}|}{|A_{ik}|} = 0.$$

где под x, y и z подразумеваются искомые вероятности.

Далее предполагается совместное использование латентной модели и ортогональной факторной структуры для построения алгоритма метода ФОРДИАСИМПТ вероятностного метода распознавания на базе альтернативных показателей. Для упрощения в целях дальнейшего изложения введем функцию

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

где l — номер класса и может принимать значение 1 или 2, k — номер номер симптомокомплекса, $\tilde{f}_{ik}(x_l)$ — вероятность положительного ответа респондента из l-ого класса на i-ый вопрос, выбранный исследователям как параметр составляющий симптомокомплекс на основании анализа факторной структуры.

Тогда вероятность принадлежности первому классу можно определить посредством формулы Баейса с использованием введенной функции

$$P(1|\, y_{a_k j}, y_{b_k j}, y_{c_k j}) = \frac{\gamma_{1k}(y_{a_k j})\gamma_{1k}(y_{b_k j})\gamma_{1k}(y_{c_k j})\tilde{\phi}_k(x_1)}{\sum_{i=1}^{2}\gamma_{ik}(y_{a_k j})\gamma_{ik}(y_{b_k j})\gamma_{ik}(y_{c_k j})\tilde{\phi}_k(x_i)}, \qquad (2)$$

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

Метод ФОРДИАСИМПТ можно использовать для распознаванию двух образов в пространстве двоичных признаков при совпадении количества выделенных измеряемых факторов и полученных независимых симптомокомплексов. В силу ортогональности выделяемых факторов, получаем, что группы параметров, наполняющие тот или иной фактор, очень слабо коррелируют между собой. Тогда можно использовать условия независимости частных апостериорных вероятностей для получения общей формулы апостериорной вероятности:

$$P(1|y_{a_1}j, y_{b_{1j}}, y_{c_{1j}}, \ldots, y_{a_r}j, y_{b_r}j, y_{c_r}j) = \sum_{i=1}^{r} P(1|y_{a_i}j, y_{b_i}j, y_{c_i}j). \quad (3)$$

Алгоритм метода ФОРДИАСИМПТ

1. Из матрицы Y путем элементарного преобразования получаем матрицу Z размерности $m \times n$ и вычисляем корреляционную матрицу R размерности $m \times m$.
2. С целью исключения незначимых показателей, вычисляем вероятностные значения уровней зависимости по формуле $\chi^2 = n \cdot \varphi$ при единичной степени свободы.
3. Определяем наименьшее количество выделяемых факторов (критерий Гуттмана, критерий «каменной осыпи» или другой адекватный критерий)[1].
4. Находим общности любым из известных методов (лучше взять метод минимальных остатков) [1],[2].
5. 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. Полученную на предыдущем шаге матрицу весовых нагрузок подвергаем ортогональномоу вращению в соответствии с варимакс критерием [2][6][7].
7. Осуществляем анализ ортогональной факторной структуры, полученной после вращения и формируем зависимые и независимые симптомокомплексы.
8. Для каждого симптомокомплекса формируем диагностическую шкалу вычисляя маргиналы и решая систему уравнений (1), используя результаты теоремы.
9. По формуле (2) вычисляем частные апостериорные вероятности для все объектов исследования.
10. По формуле (3) вычисляем вероятность принадлежности объекта исследования к классу А или к классу В.

Основные этапы работы алгоритма представлены ниже таблицами.

Таблица 1

Исходные объекты исследования с частотой встречаемости в каждом классе

Класс А						Частота	Класс В						Частота
0	1	1	1	1	1	0,015625	0	0	0	1	1	1	0,372093
0	0	0	1	0	0	0,078125	1	0	0	0	1	1	0,034884
1	1	1	1	1	1	0,046875	1	0	0	1	1	1	0,058140
1	0	1	1	0	1	0,046875	0	1	0	0	1	0	0,011628
0	0	0	0	0	0	0,171875	0	0	0	0	1	0	0,046512
1	0	0	0	0	1	0,015625	0	0	0	0	1	1	0,197674
0	0	0	0	0	1	0,093750	0	0	1	1	0	1	0,011628
1	0	0	1	0	0	0,015625	1	0	0	1	1	0	0,023256
0	1	0	0	0	1	0,031250	0	0	1	1	1	0	0,034884
1	1	1	0	1	1	0,093750	0	0	0	1	0	1	0,058140
1	0	1	0	0	1	0,015625	0	0	0	1	1	0	0,034884
1	0	1	0	0	0	0,031250	1	0	0	1	0	1	0,011628
1	1	0	0	1	1	0,015625	0	0	1	1	1	1	0,034884
1	0	0	0	0	0	0,078125	0	1	0	1	1	1	0,011628
1	0	1	1	1	1	0,062500	0	0	1	0	1	1	0,034884
1	1	1	0	0	0	0,015625	0	1	0	1	0	1	0,011628
1	1	0	0	0	1	0,015625	1	0	0	0	1	0	0,011628
1	1	0	1	1	1	0,015625							
1	1	1	0	1	0	0,015625							
0	1	1	0	1	1	0,015625							

0	0	1	0	1	0	0,015625							
0	1	1	0	0	1	0,015625							
0	1	1	0	0	0	0,031250							
0	0	1	1	0	0	0,015625							
1	1	1	1	1	0	0,015625							
1	0	1	0	1	1	0,015625							

Таблица 2

Матрица коэффициентов корреляции

1	0,293	0,362	-0,061	-0,022	-0,016
0,293	1	0,418	-0,173	0,0189	0,0586
0,362	0,418	1	-0,016	0,0509	0
-0,061	-0,173	-0,016	1	0,23	0,207
-0,022	0,0189	0,0509	0,23	1	0,39
-0,016	0,0586	0	0,207	0,39	1

Таблица 3

Матрица значимости φ коэффициентов по χ^2 распределению

1,000	0,999	0,999	0,545	0,207	0,154
0,999	1,000	0,999	0,167	0,183	0,527
0,999	0,999	1,000	0,156	0,466	0,000
0,545	0,167	0,156	1,000	0,995	0,988
0,207	0,183	0,466	0,995	1,000	0,999
0,154	0,527	0,000	0,988	0,999	1,000

Таблица 4

Матрица ортогонального факторного отображения

Фактор №1	Фактор №2
0,699	0,0441
0,769	0,012
0,785	-0,061
-0,198	**-0,597**
0,0517	**-0,784**
0,0505	**-0,765**

Таблица 5

Основные показатели симптомокомплекса №1

Маргиналы	Значения частот и априорных вероятностей	Варианты ответов			Апостериорная вероятность
$p_1 = 0,266$	$\tilde{\phi}_1 = 0,770$	1	1	1	0,007
$p_2 = 0,166$	$\tilde{\phi}_2 = 0,230$	0	1	1	0,086
$p_3 = 0,3$	$\tilde{f}_1(x_1) = 0,048$	0	0	1	0,705
$p_{12} = 0,113$	$\tilde{f}_2(x_1) = 0,174$	0	0	0	0,992
$p_{13} = 0,153$	$\tilde{f}_3(x_1) = 0,096$	1	0	0	0,914
$p_{23} = 0,1$	$\tilde{f}_1(x_2) = 0,849$	1	1	0	0,294
$p_{123} = 0,08$	$\tilde{f}_2(x_2) = 0,564$	1	0	1	0,162
	$\tilde{f}_3(x_2) = 0,723$	0	1	0	0,837

Таблица 6

Основные показатели симптомокомплекса №2

Маргиналы	Значения частот и априорных вероятностей	Варианты ответов			Апостериорная вероятность
$p_1 = 0,66$	$\tilde{\phi}_1 = 0,746$	1	1	1	0,995
$p_2 = 0,7$	$\tilde{\phi}_2 = 0,254$	0	1	1	0,972
$p_3 = 0,51$	$\tilde{f}_1(x_1) = 0,842$	0	0	1	0,486
$p_{12} = 0,546$	$\tilde{f}_2(x_1) = 0,858$	0	0	0	0,046
$p_{13} = 0,393$	$\tilde{f}_3(x_1) = 0,615$	1	0	0	0,220
$p_{23} = 0,406$	$\tilde{f}_1(x_2) = 0,125$	1	1	0	0,913
$p_{123} = 0,333$	$\tilde{f}_2(x_2) = 0,235$	1	0	1	0,857
	$\tilde{f}_3(x_2) = 0,214$	0	1	0	0,643

Выводы

В рамках теории латентного анализа сформулирована теорема о сведении решения системы уравнений простешей латентно-структурной модели к решению трех квадратных уравнений. Предложен вычислительный алгоритм ФОРДИАСИМПТ позволяющий строить диагностические симптомокомплексы на базе альтернативных данных, ортогональной факторной структуры, простешей латентно-структурной модели и формулы Баейса. Показана целесообразность применение алгоритма ФОРДИАСИМПТ в распознновании объектов иссследования в случае независимых симптомокомплесов при адекватной статистической информации.

Примечания:

1. Иберла К. Факторный анализ. М.: Статистика. 1989.
2. Харман Г. Современный факторный анализ. М.: Статистика. 1972.

3. Гольтяпин В.В. Реализация вычислительного алгоритма метода ФОРДИАСИМПТ на примере альтернативных показателей артериальной гипертензии. // Современные наукоемкие технологии. 2014. №11. С. 50-55.
4. Осипов Г.В. Методы измерения в социологии. М.: Наука. 2003.
5. Lazarsfeld P.F. The logical and mathematical foundation of latent structure analysis. In: Measurement and Prediction. N. Y. 1950.
6. Kaiser H.F. The varimax criterion for analytic rotation in factor analysis. // Psyhometrika №23. С. 187-200. 1958.
7. D. Saunders. The rationale for an “oblimax” method of transformation in factor analysis. // Psyhometrika №26. С. 317-324. 1961.

References:

1. Iberla K. Faktornyi analiz. M.: Statistika. 1989.
2. Kharman G. Sovremennyi faktornyi analiz. M.: Statistika. 1972.
3. Gol'tyapin V.V. Realizatsiya vychislitel'nogo algoritma metoda FORDIASIMPT na primere al'ternativnykh pokazatelei arterial'noi gipertenzii. // Sovremennye naukoemkie tekhnologii. 2014. №11. S. 50-55.
4. Osipov G.V. Metody izmereniya v sotsiologii. M.: Nauka. 2003.
5. Lazarsfeld P.F. The logical and mathematical foundation of latent structure analysis. In: Measurement and Prediction. N. Y. 1950.
6. Kaiser H.F. The varimax criterion for analytic rotation in factor analysis. // Psyhometrika №23. S. 187-200. 1958.
7. D. Saunders. The rationale for an “oblimax” method of transformation in factor analysis. // Psyhometrika №26. S. 317-324. 1961.

УДК 519.237.07

Алгоритм метода фордиасимпт на независимых симптомокомплексах

В.В. Гольтяпин

Омский филиал Института математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Российская Федерация
644077, г.Омск, ул.Певцова, 13
E-mail: goltyapin@mail.ru

Анотация. В данной статье в рамках теории латентного анализа сформулирована теорема, позволяющая находить апостериорные вероятности на базе альтернативных показателей с использованием ортогональной факторной структуры. Построен вычислительный алгоритм, позволяющий строить диагностические симптомокомплексы на базе вероятностного метода распознавания образов.

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