Математические структуры и моделирование 2019. № 1(49). С. 65–72

УДК 519.237.7
DOI: 10.25513/2222-8772.2019.1.65-72

ФАКТОРНЫЙ АНАЛИЗ ДЛЯ ВОССТАНОВЛЕНИЯ ПРОБЕЛОВ ДАННЫХ АРТЕРИАЛЬНОЙ ГИПЕРТЕНЗИИ

В.А. Шовин
научный сотрудник, e-mail: v.shovin@mail.ru

Институт математики им. С.Л. Соболева Сибирского отделения РАН
(Омский филиал), Омск, Россия

Аннотация. Разработан алгоритм заполнения пробелов данных на базе восстановления вектора показателей объектов из факторной структуры данных, вычисляемой с помощью метода штрафных функций. Пробелы в данных и соответствующие им уравнения факторной модели для отдельных объектов не учитывались в критерии оптимизации невязок уравнений факторной модели, что позволяет достоверно оценить значения пробелов данных. Проведён численный эксперимент, подтверждающий работоспособность алгоритма, и создана программа с интерфейсом, позволяющая пользователю загружать новые данные.

Ключевые слова: факторный анализ, метод штрафных функций, артериальная гипертензия.

Введение

Факторный анализ позволяет вычислить факторную структуру данных, определив тем самым связь между латентными переменными (факторами) и исходными переменными. В классическом факторном анализе связь между факторами и исходными переменными является линейной. В факторном анализе решаются несколько задач. Это поиск матрицы факторной структуры, определяющей нагрузки переменных на факторы, т. е. соответствует коэффициентам корреляции между исходными переменными и факторами. А также определение значений факторов (новых «латентных» переменных) у объектов. Зная значение факторов и матрицу факторной структуры, можно восстановить значения исходных переменных, тем самым очистив исходные данные от шума.

Существует несколько подходов к вычислению матрицы факторной структуры. Одним из продвинутых подходов к определению всех неизвестных параметров модели факторного анализа является метод оптимизации с проверкой на дополнительные условия или ограничения, а именно метод штрафов. В данном методе для оценки параметров и значений латентных переменных модели, задаваемой линейными уравнениями, может быть использован критерий минимальных невязок как сумма невязок модели вычисленных для всей выборки

различных объектов. Дополнительно на параметры и значения латентных переменных могут быть заданы ограничительные условия. Для решения задачи минимизации невязок модели предлагается использовать методы нелинейной оптимизации с условиями: метод конфигураций. Метод штрафных функций позволяет учитывать ограничения, накладываемые на значения параметров и латентных переменных модели.

С помощью метода штрафов можно определить неизвестные параметры факторной модели и из факторной структуры данных восстановить исходные данные, тем самым восстановив пробелы в данных. Чтобы задача оптимизации невязок факторной модели была определена, в данном алгоритме восстановления пробелов данных предлагается предварительно исключить из критерия оптимизации отдельные уравнения объектов, соответствующие пробелам данных.

1. Факторный анализ как частный случай структурных уравнений

В теории структурных уравнений используются следующие типы матриц.

Матрица $Z \underset{m\times n}{\leftrightarrow} z_{ij}$ — матрица значений измеряемых переменных у исследуемых объектов или состояний объекта размерности $m \times n$, где m — число измеряемых параметров, n — число объектов или состояний объекта (объём выборки).

Матрица $P \underset{g\times n}{\leftrightarrow} p_{ij}$ — матрица значений латентных переменных объектов размерности $g \times n$, где g — число латентных параметров.

Матрица $A \underset{k\times s}{\leftrightarrow} a_{ij}$ — матрица параметров структурных уравнений размерности $k \times s$, где k — число структурных уравнений, s — число параметров в структурных уравнениях.

Система структурных уравнений задаётся в виде:

$$\begin{cases} f_1\left(a_{11}, a_{12}, \ldots, a_{1s}; p_{1t}, p_{2t}, \ldots, p_{gt}; z_{1t}, z_{2t}, \ldots, z_{mt}\right) + \varepsilon_{1t} = 0, \\ f_2\left(a_{21}, a_{22}, \ldots, a_{2s}; p_{1t}, p_{2t}, \ldots, p_{gt}; z_{1t}, z_{2t}, \ldots, z_{mt}\right) + \varepsilon_{2t} = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ f_k\left(a_{k1}, a_{k2}, \ldots, a_{ks}; p_{1t}, p_{2t}, \ldots, p_{gt}; z_{1t}, z_{2t}, \ldots, z_{mt}\right) + \varepsilon_{kt} = 0, \end{cases}$$

где $f_1, f_2, \ldots, f_k$ — в общем случае нелинейные функции своих переменных, $\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{kt}$ — невязки модели для t-го объекта или состояния объекта.

На значения параметров и значения латентных переменных могут накладываться дополнительные условия в виде равенств и неравенств.

Оптимальными значениями параметров и латентных переменных считаются те значения, которые минимизируют абсолютные значения невязок модели и удовлетворяют всем дополнительным условиям.

В данной работе рассмотрен частный случай структурной модели – линейная факторная модель, описываемая следующими уравнениями [1–3]:

$$\begin{cases} z_{1t} = a_{11}p_{1t} + a_{12}p_{2t} + \ldots + a_{1g}p_{gt} + \varepsilon_{1t}, \\ z_{2t} = a_{21}p_{1t} + a_{22}p_{2t} + \ldots + a_{2g}p_{gt} + \varepsilon_{2t}, \\ \vdots \\ z_{mt} = a_{m1}p_{1t} + a_{m2}p_{2t} + \ldots + a_{mg}p_{gt} + \varepsilon_{mt}, \end{cases} \tag{1}$$

где матрица $A \underset{m\times g}{\leftrightarrow} a_{ij}$ называется матрицей факторной структуры размерности $m \times g$ весовых коэффициентов. Где m — число изучаемых параметров, g — число общих факторов.

На вид факторной структуры A налагаются дополнительные ограничения:

— общности переменных факторной структуры должны быть не больше 1, а также не меньше определённого порога значимости:

$$h_i = \sqrt{\sum_{k=1}^{g} a_{ik}^2} \leqslant 1, \quad h_i \geqslant p; \tag{2}$$

— критерий оптимизации задаётся в следующем виде:

$$K = \sum_{t=1}^{n}\sum_{k=1}^{m} \varepsilon_{kt}^2; \tag{3}$$

— минимизация критерия K и учёт дополнительных условий на вид факторной структуры приводит к оптимальному решению для варьируемых значений элементов факторной структуры a_{ij} и факторов p_{ij}.

Оптимизацию суммы квадратов невязок линейных уравнений факторной структуры как функций от независимых переменных матрицы факторной структуры и значений факторов с ограничениями предлагается осуществлять методом штрафных функций [4]. В качестве метода безусловной оптимизации метода штрафных функций был выбран метод конфигураций [5].

Вычислительный алгоритм

Алгоритм построения линейной факторной модели:

1. Определение числа факторов числом $g < m$.
2. Определение начальных приближений матрицы A линейной части размерности $m \times g$ и матрицы P значений факторов размерности $g \times n$ случайными числами из диапазона $[-1; 1]$.
3. Минимизация критерия (3) суммы квадратов невязок структурных уравнений (1) как функций от независимых переменных матриц A факторной структуры и значений факторов P с ограничениями (2) методом штрафных функций и методом конфигураций.

2. Алгоритм восстановления пробелов данных

На базе метода штрафов возможно вычислить факторную структуру данных и восстановить пробелы данных:

1. Выявить пробелы данных.
2. Исключить из критерия невязок уравнений факторной модели уравнения объектов, соответствующие пробелам данных.
3. Выполнить оценку неизвестных параметров линейной факторной модели по методу штрафов как задачи оптимизации невязок модели.
4. После процедуры оценки матрицы факторной структуры и значений факторов у объектов по методу штрафов заменить значения переменных, соответствующие пробелам данных восстановленными значениями, минимизирующих невязки факторной модели.

3. Отбраковка грубых ошибок (проверка на однородность выборки)

Таблица экспериментальных данных может содержать грубые ошибки. Грубые ошибки могут быть следствием нарушения основных условий измерения, неправильного чтения показаний измерительного прибора, просчёта, неверной записи при внесении результата измерения в таблицу. Внешним признаком результата, содержащего грубую ошибку, является его резкое отличие по величине от результатов остальных измерений.

Для отбраковки грубых ошибок предлагается использовать следующий алгоритм:

1. Необходимо проверить, является ли выборка симметричной или нет.
2. Строится вариационный ряд: $x'_1 \dots x'_n \quad \left(x'_1 \leqslant x'_2 \leqslant \dots \leqslant x'_{n-1} \leqslant x'_n\right)$, где x'_i — элементы вариационного ряда, полученного из элементов x_i проверяемой выборки. Анализируются крайние элементы вариационного ряда.
3. Делается предположение, что элемент x'_n померен с грубой ошибкой.
4. Берётся для исследования выборка $x'_1 \dots x'_{n-1}$.
   По выборке $x'_1 \dots x'_{n-1}$ строится интервал $(\overline{x} - 3S_x, \overline{x} + 3S_x)$, если выборка $x'_1 \dots x'_{n-1}$ симметрична, или строится интервал $(\overline{x} - 5S_x, \overline{x} + 5S_x)$, если выборка $x'_1 \dots x'_{n-1}$ несимметрична, где $\overline{x}$ — выборочное математическое ожидание величин $x_1 \dots x_n$, S_x — выборочное стандартное отклонение величин $x_1 \dots x_n$.
5. Если $x'_n \in$ интервалу, то грубой ошибки нет.
6. Аналогично проверяется x'_1.
7. Если x'_n померен с грубой ошибкой, то он отбрасывается.
   Элемент x'_1 рассматривается относительно выборки $x'_2 \dots x'_{n-1}$.
8. И т. д.

$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ — выборочное среднее.
$S_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$ — выборочная дисперсия.
Выборка считается симметричной, если величина $\delta = |h_x - \overline{x}| \leqslant 3 \cdot \frac{S_x}{\sqrt{n}}$, где

h_x — медиана,

$$h_x = \begin{cases} 0.5 \cdot \left(x'_l + x'_{l+1}\right), & n = 2l; \\ x'_{l+1}, & n = 2l+1. \end{cases}$$

4. Численный эксперимент

В качестве исходных данных были взяты 38 биофизических показателей для 131 лица с артериальной гипертензией начальной стадии. Некоторые показатели из выборки:

1) *вес,*
2) *индекс массы тела (ИМТ),*
3) *частота дыхания (ЧД),*
4) *сегментоядерные нейтрофилы (С),*
5) *лимфоциты (Л),*
6) *конечно-систолический размер левого желудочка (КСР),*
7) *конечно-систолический объём левого желудочка (КСО),*
8) *конечно-диастолический размер левого желудочка (КДР),*
9) *конечно-диастолический объём левого желудочка (КДО),*
10) *ударный объём (УО),*
11) *минутный объём сердца (МОС),*
12) *общее периферическое сосудистое сопротивление (ОПСС),*
13) *индекс Хильдебрандта (ИХ),*
14) *фракция выброса левого желудочка (ФВ),*
15) *фракция укорочения левого желудочка (ФУ).*

Исходные данные содержали пробелы, соответствующие грубым ошибкам и выбросам из нормального распределения показателей.

С помощью алгоритма восстановления пробелов данных было выполнено предварительное исключение уравнений объектов факторной модели, соответствующих пробелам данных. Была проведена оценка неизвестных параметров линейной факторной модели. Количество факторов было выбрано по принципу «каменистой осыпи» независимого классического факторного анализа исходных данных. После вычисления факторной структуры пробелы данных были восстановлены.

Оказалось, что 65 % восстановленных пробелов, соответствующих алгоритму отбраковки грубых ошибок, оказались вне интервала минимальных и максимальных значений отдельных переменных. Тогда как только 35 % пробелов оказались в рамках таких интервалов. Данный факт можно интерпретировать как ошибочное определение грубых ошибок в виду рассмотрения независимых нормальных распределений отдельных переменных, а не многомерного нормального распределения. Скорее всего, лишь 35 % выявленных грубых ошибок оказались истинными.

Алгоритм, учитывающий многомерность распределения данных, не смог исправить 65 % грубых ошибок. То есть это были не грубые ошибки, а результат совместного однонаправленного воздействия со стороны различных факторов.

Такие выбросы из многомерного или части одномерных нормальных распределений являются естественным результатом однонаправленного воздействия различных факторов.

Оставшиеся 35 % выявленных грубых ошибок как выбросы из одномерных нормальных распределений оказались после восстановления в рамках нормального распределения, что свидетельствует о том, что это были истинные выбросы. То есть после восстановления эти выбросы оказались в рамках нормальных распределений. В то время как остальные 65 % остались вне нормальных распределений.

Всё это свидетельствует о том, что при оценке значений и проверке их на нормальное распределение или выбросы нельзя рассматривать независимо только это значение и этот показатель отдельно от других. В том числе нельзя утверждать, что такова организационная система показателей: в данном случае человек имеет плохую биологическую организацию и плохое функциональное состояние. Нарушение значений показателей из одномерного нормального распределения может быть результатом компенсационных функциональных процессов при воздействии однонаправленных негативных факторов.

Например, некоторые негативные факторы могут понемногу увеличиваться в одном скоординированном направлении, что может приводить к большим выбросам значений отдельных показателей, находящихся в функциональном взаимоотношении с данными факторами. Функциональное состояние — это система функций от значений показателей объекта, т. е. как должны изменяться показатели объекта при изменении других, например, возрастать или убывать.

Можно сделать вывод, что в данном численном эксперименте в множестве объектов лишь 35 % объектов имели нарушенное функциональное состояние или имели какую-либо ошибку в измерении показателей. Поэтому в диагностике объектов нельзя использовать один показатель, в данном случае это был показатель повышенного артериального давления. Такое повышенное артериальное давление может быть нормальным значением в виду скоординированного небольшого воздействия нескольких негативных факторов, вызывающих сильное отклонение от нормы данного показателя. Данные объекты при исключении вредных факторов должны показать улучшение значения отдельного диагностического показателя. Такая система в последствии будет продиагностирована как нормальная, поскольку изначально не имела нарушенного функционального взаимоотношения показателей. Её отдельные показатели придут в интервалы нормальных значений, если негативные факторы выйдут из скоординированного воздействия или сами потеряют свой негативный статус. Подобные функциональные системы или организмы и так были в норме или в индивидуальной норме. Они продолжают жить в рамках своих нормальных функциональных взаимоотношений. Помещённые в нормальные условия такие объекты должны продемонстрировать нормальные значения отдельных показателей. Поэтому можно рекомендовать нормальные условия существования для систем, и в случае обнаружения негативных факторов нивелировать их скоординированное негативное воздействие. Опасным представляется нормализация значений отдельных показателей без нормализации негативных факторов, по-

скольку в такой системе могут в дальнейшем проявиться различные нарушения функционального состояния, что в свою очередь может вывести всю систему из нормального функционирования, и система может начать распадаться на подсистемы, в которых ещё выполняется нормальное функциональное взаимоотношение показателей. В то же время не зависимые друг от друга подсистемы могут быть организованы вместе для выполнения определённых действий, для поддержания функционирования всего организма. Такого рода системы работают под действием внешней организующей силы либо существуют из-за работы подсистем. Чтобы снять нагрузку с организующей силы в системах, это перекладывается на функциональное взаимоотношение подсистем. В контексте медицины это означает, что в случае плохого функционального состояния необходимо поддержание всего организма и отдельных его функций с помощью постоянного лечения. В случае хорошего функционального состояния можно рекомендовать лечение по одновременному воздействию на группы показателей, соответствующих факторам заболевания. Устранить негативный статус факторов заболевания. Нарушить скоординированное воздействие таких факторов. Возможно, достаточно устранить всего один фактор риска.

5. Заключение

На базе метода штрафных функций и минимизации невязок уравнений факторной модели был разработан алгоритм восстановления пробелов данных. С помощью численного эксперимента была подтверждена работоспособность алгоритма.

Литература

1. Шовин В.А. Нелинейные структурные уравнения и квадратичный факторный анализ // Математические структуры и моделирование. 2018. № 2(46). С. 51–61.
2. Иберла К. Факторный анализ / Пер. с нем. В.М. Ивановой; Предисл. А.М. Дуброва. М. : Статистика, 1980.
3. Харман Г. Современный факторный анализ / Пер. с англ. В.Я. Лумельского; Научное редактирование и вступительная статья Э.М. Бравермана. М. : Статистика, 1972.
4. Банди Б. Методы оптимизации. Вводный курс. М. : Радио и связь, 1988. 128 с.
5. Кокуев А.Г. Оптимальное управление. Поиск экстремумов многомерных функций. Астрахань : АГТУ, 2011. 34 с.

FACTOR ANALYSIS FOR RESTORING DATA GAPS OF HYPERTENSION

V.A. Shovin
Scientist Researcher, e-mail: v.shovin@mail.ru

Institute of Mathematics S.L. Soboleva of Siberian Branch of RAS
(Omsk Branch), Omsk, Russia

Abstract. An algorithm for filling data gaps on the basis of restoring the vector of object indices from the factorial data structure, calculated with penalty method, is developed. Data gaps and the corresponding factor model equations for individual objects were not taken into account in the optimization criterion for residuals of the factor model equations. That allows to reliably estimate the values of data gaps. A numerical experiment confirming the operability of the algorithm is carried out and a program with an interface that allows the user to upload new data is created.

Keywords: factor analysis, penalty method, hypertension.

References

1. Shovin V.A. Nelineinye strukturnye uravneniya i kvadratichnyi faktornyi analiz. Matematicheskie struktury i modelirovanie, 2018, no. 2(46), pp. 51–61. (in Russian)
2. Iberla K. Faktornyi analiz. Per. s nem. V.M. Ivanovoi, Predisl. A.M. Dubrova, Moscow, Statistika Publ., 1980. (in Russian)
3. Kharman G. Sovremennyi faktornyi analiz. Per. s angl. V.Ya. Lumel'skogo, Nauchnoe redaktirovanie i vstupitel'naya stat'ya E.M. Bravermana, Moscow, Statistika Publ., 1972. (in Russian)
4. Bandi B. Metody optimizatsii. Vvodnyi kurs. Moscow, Radio i Svyaz' Publ., 1988, 128 p. (in Russian)
5. Kokuev A.G. Optimal'noe upravlenie. Poisk ekstremumov mnogomernykh funktsii. Astrakhan', AGTU Publ., 2011, 34 p. (in Russian)

Дата поступления в редакцию: 10.12.2018