МНОГОМЕРНОЕ МОДЕЛИРОВАНИЕ ПО ПАССИВНЫМ ВЫБОРКАМ МАЛОГО ОБЪЕМА МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ С ПРЕДВАРИТЕЛЬНОЙ ОРТОГОНАЛИЗАЦИЕЙ ФАКТОРОВ (МНКО)

Юлия Столяренко
Государственный университет им. Т.Г.Шевченко
mlcc@mail.ru

Abstract. *When building an adequate mathematical models for small samples, you must either use a method that doesn't depend on the amount of data you want, or use a technique that allows data from passive little get virtual (the equivalent of) sample of small volume.*

Ключевые слова: *моделирование по пассивным данным, выборки малого объема.*

I. Введение

При моделировании по пассивным данным в случае выборок малого объема для получения адекватных математических моделей можно воспользоваться двумя методами: модифицированным методом случайного баланса (ММСБ) и методом наименьших квадратов с предварительной ортогонализацией факторов (МНКО) [1]. Применение обоих методов возможно, однако, необходимо предварительно произвести преобразования над исходными данными методом точечных распределений (МТР) [2]. Применение МТР в итоге позволит получить многомерную виртуальную выборку достаточного объема для того чтобы, методом корреляционных плеяд отобрать некоррелированные факторы и затем отыскать математическую модель. Вся цепочка методов была объединены в одно целое и была названа «Многомерный метод точечных распределений» (ММТР). Однако при использовании этого метода большую роль играет изначальная корреляция в многомерной выборке малого объема, так как после применения ММТР возрастает внутренняя корреляция факторов, что затрудняет их отбор для дальнейшего моделирования.

II. Получение адекватной математической модели с помощью МНКО

Рассмотрим пример получения адекватной математической модели по многомерным пассивным данным малого объема с помощью МНКО. В отличие от ММСБ если выполняется условие сверхнасыщенности (количество столбцов превышает количество строк данных) плана нет необходимости выполнять отсев некоррелированных факторов перед началом построения модели. Также МНКО не столь чувствителен к объему обрабатываемой выборки, поэтому производить дополнительные преобразования над данными в этом случае нет необходимости. Пусть имеется многомерная выборка малого объема (таблица 1).

Применив МНКО для получения математической модели без какой-либо предварительной обработки данных, получим:

$$\hat{Y} = -55{,}31 - 0{,}132 X_1 + 0{,}737 X_2. \quad (1)$$

Полученная модель адекватна. В силу адекватности все полученные модели имеют право на существование, однако, точность описания ими выходной величины различна. В принципе ее можно оценить по величине коридора существования модели, однако в большом количестве случаев по вышеупомянутым причинам сама величина коридора является оценкой, иногда довольно грубой. Поэтому предлагается о качестве модели судить по количеству информации, которое она может дать, то есть по информационной ёмкости.

Таблица 1 – Исходные данные

	Y	X 1	X 2	X 3	X 4	X 5	X 6	X 7	X 8	X 9	X10	X11	X12	X13	X14	X15
1	28,06	53,19	122,99	130,35	229,7	214,83	40,67	204,64	280,5	247,01	173,48	117,25	16,06	438,44	344,09	40,73
2	25,79	48,44	119,29	131,41	220,1	206,23	42,12	198,65	271,87	248,57	156,65	121,66	16,06	437,69	336,88	44,3
3	28,08	49,4	121,09	124,19	233,2	222,92	42,21	202,11	266,45	256,1	201,97	135,21	16,43	456,77	343,95	43,86
4	27,36	52,16	121,23	128,28	231,0	211,43	42,19	203,17	273,43	248,42	185,57	129,29	15,91	447,51	342,22	48,7
5	24,33	48,15	117,07	136,29	219,9	201,87	39,12	193,99	255,51	247,7	164,42	113,77	16,09	425,83	332,02	44,72
6	26,57	50,86	120,28	126,49	231,1	212,66	40,23	200,78	265,78	246,45	160,2	130,81	15,53	461,8	341,14	42,23
7	28,63	50,84	122,4	133,8	237,4	219,41	44,1	201,29	278,09	258,01	187,11	119,54	16,63	445,12	347,8	46,24
8	30,31	51,82	126,08	138,17	247,3	220,47	41,69	211,65	294,56	253,84	190,6	135,26	16,46	456,96	353,13	45,77
9	28,45	48,81	122,19	123,89	246,9	216,29	44,24	206,33	262,72	250,9	177,27	133,47	16,43	460,14	345,38	42,41
10	27,82	52,84	122,09	123,87	208,8	215,88	40,53	201,13	282,81	249,8	181,67	129,41	15,91	446,65	346,14	41,01

Представим исследуемый объект контроля в виде двух систем: системы факторов X и системы выходных показателей качества Y. В случае, когда не имеется математического описания взаимодействия этих систем, энтропии их равны $H(X)$ и $H(Y)$, а энтропия объединенной системы будет максимальна и равна $H(X,Y) = H(X) + H(Y)$. После получения сведений о характере взаимодействия обеих систем X и Y в виде математической модели $\hat{Y} = f(X)$, «остаточная» энтропия и есть информация

$$I_{Y \to X} = H(X) + H(Y) - H(X, \hat{Y}). \qquad (2)$$

Согласно известной теореме, энтропия объединённой системы (в данном случае математической модели) равна энтропии одной из её составных частей плюс условная энтропия второй части относительно первой, то есть

$$H(X, \hat{Y}) = H(X) + H(\hat{Y} / X), \qquad (3)$$

где $H(\hat{Y} / X)$ – условная энтропия модели системы $\hat{Y}$ относительно X.

Подставляя (2) в (3), получим выражение для полной информации о системе Y, содержащейся в системе X, с помощью модели системы $\hat{Y}$

$$I_{Y \to X} = H(Y) - H(\hat{Y} / X). \qquad (4)$$

Это означает, что количество информации, получаемое за счёт знания характеристик взаимодействия (математических моделей) систем Y и X равно разности двух энтропий: энтропии системы, состояние которой описывается случайной величиной Y с определенным рядом распределения (его можно представить в виде гистограммы опытных данных $Y_1, Y_2, ...Y_j, ...Y_n$; Y_j – величина центра j-го разряда гистограммы, n – число разрядов), и условной энтропии модели системы $\hat{Y}$ при условии, что система X находится в состоянии Z_{ki}, то есть каждый k-й эффект $(k = \overline{1,m})$, включенный в модель, находится в i-м состоянии $(i = \overline{1, l_i})$. При этом значение Z_{ki} есть величина центра i-го разряда гистограммы эффекта Z_k.

Выражение для полной информации с учетом (4) и (6) примет вид

$$I_{Y \to Z} = -\sum_{j=1}^{n} p(Y_j) \log_2 p(Y_j) + \sum_{k=1}^{m} \sum_{i=1}^{l} \sum_{j=1}^{n} q_k \, p\left(\hat{Y}_k / Z_{ki}\right) p\left(\hat{Y}_{kj} / Z_{ki}\right) \log_2 p\left(\hat{Y}_{kj} / Z_{ki}\right). \qquad (5)$$

Полученное выражение можно использовать в качестве критерия для оценки информационной емкости математической модели.

Для практического использования выражения (5) необходимо установить порядок определения входящих в него вероятностей. Это можно сделать на основе теоремы Бернулли, которая позволяет заменять вероятности событий их частостями. После необходимых преобразований получим следующую формулу для практического применения

$$I_{Y\to Z} = -\sum_{j=1}^{n} \frac{N_j}{N}\log_2\frac{N_j}{N} + \sum_{k=1}^{m}\sum_{i=1}^{l}\sum_{j=1}^{n} \frac{t_k}{\sum_{k=1}^{m} t_k}\cdot\frac{N_{ki}}{N}\cdot\frac{N_{kij}}{N_{ki}}\log_2\frac{N_{kij}}{N_{ki}}. \qquad (6)$$

Информационная емкость для полученной модели (1) равна 70%.

Перед проведением работ по получению математической модели во всех случаях рекомендуется сократить первоначальный список факторов до возможного минимума, так как с ростом числа факторов трудоемкость моделирования растет как степенная функция. Отсев факторов можно производить по двум критериям: факторы незначимые, то есть не влияющие на целевую функцию и внесенные в первоначальный список факторов ошибочно, и факторы коррелированные, то есть имеющие сильную внутреннюю связь.

Одним из способов понижения размерности факторного пространства из-за сокращения сильно коррелированных факторов являются корреляционные плеяды, основанные на анализе корреляционной матрицы.

Корреляционная матрица представляет собой симметричную квадратную матрицу размером *M*M*, где *M* – число исследуемых факторов, главная диагональ которой заполнена единицами (или нулями для удобства дальнейшего анализа), а недиагональные элементы представляют собой меру тесноты связи между парой факторов (коэффициент корреляции, корреляционное отношение, модифицированный индекс Фехнера и т.д.).

Непосредственный анализ корреляционной матрицы представляет значительную трудность, так как корреляционные связи между факторами образуют деревья, цепи, циклы и другие фигуры графов. Для выделения главных зависимостей следует прибегнуть к одному из методов анализа таких матриц, простейшим из которых является метод корреляционных плеяд.

Метод заключается в том, что в корреляционной матрице находится недиагональный элемент с максимальной по модулю величиной $|r_{ij}|=max$. Из матрицы вычеркиваются столбцы с номерами *i* и *j*, а из строк с номерами *i* и *j* выбирается следующий максимальный по модулю элемент, например $|r_{il}|$. Столбец с номером *l* вычеркивается, а из строк с номерами *i, j* и *l* выбирается следующий максимальный по модулю элемент, и так далее до исчерпания данных.

Результат такой работы удобно представить на рисунке в виде графа, вершинами которого являются факторы, ребрами – максимальные связи, причем длины ребер обратно пропорционально величине соответствующих коэффициентов корреляции. Выбрав некоторое пороговое значение коэффициента корреляции, например $|r_{пор}|=0{,}5$, можно отделить по этому признаку плеяды друг от друга.

Внутри каждой плеяды связь между факторами признается тесной, а между плеядами – слабой. Это означает, что если от каждой плеяды выбрать по одному представителю, то новое общее количество факторов, сокращенное до количества плеяд, будет нести об исследуемом объекте практически ту же информацию, что и раньше. При этом факторы новой таблицы данных будут слабо коррелированными между собой, что является одним из главных условий перехода к математическому моделированию.

Задача выбора одного фактора из плеяды – неформальная задача и решать ее надо всеми возможными методами с учетом мнения специалистов (например, технологов исследуемого процесса) лучше всего экспертными методами. Это значит, что в обязательном порядке надо сопоставлять корреляционные плеяды, полученные на основе анализа корреляционных матриц, составленных не только из коэффициентов корреляции, но и корреляционных отношений, и МИФ. Этим самым уменьшается ошибка от неучета нелинейного характера связи между факторами, а также влияние хоть и принадлежащих к данной двумерной совокупности, но нетипичных пар данных. Также одним из подобных методов является метод весовых коэффициентов

важности, который обладает меньшей неопределенностью и более удобен для эксперта с психологической точки зрения.

Для реализации метода весовых коэффициентов важности необходимо соблюдение определенных правил:

1. Опрос экспертов производится только письменно и только в виде специально разработанной анкеты.
2. Анкета должна состоять из пунктов (объектов), в которых сформулированы некоторые утверждения (не вопросы).
3. Пункты анкеты должны быть сформулированы таким образом, чтобы на них каждый эксперт мог ответить однозначно.
4. Отбор экспертов производится исследователем по возможности из разнородных групп.
5. Опрос экспертов должен производится индивидуально.
6. Обработка анкет должна вестись объективными методами. Должны быть некоторые контрольные критерии проверки.
7. После обработки анкет должно быть достаточно убедительное представление результатов.

Кроме того, следует также использовать вспомогательный прием – построение корреляционных ядер по всем трем мерам тесноты связи.

После выделения по одному представителю от каждой плеяды можно из таблицы исходных данных построить таблицу некоррелированных (фактически слабо коррелированных) данных, информационная емкость которой практически не изменяется, а размерность факторного пространства сокращается в несколько раз. Однако в силу того, что плеяды и ядра учитывают не все связи, а только максимальные, следует полученную таблицу предполагаемых некоррелированных данных заново проверить на наличие парных корреляционных зависимостей, вновь составить корреляционную матрицу и проанализировать её с помощью новых плеяд и ядер. В случае обнаружения достаточно сильной корреляционной зависимости ее следует уничтожить (уменьшить до приемлемого уровня) путем замены прежнего фактора на другого представителя из соответствующей первоначальной плеяды. Работа должна быть продолжена до тех пор, пока очередная проверка не подтвердит создание таблицы действительно некоррелированных (фактически слабо коррелированных) данных, которая на самом деле и является исходной для нахождения математических моделей [1].

После построения корреляционных плеяд были отобраны самые слабо коррелированные факторы, а именно, *X3, X6, X14, X15* для дальнейшего построения модели с помощью МНКО.

В этом случае была получена также модель адекватная модель

$$\hat{Y} = -68{,}15 - 0{,}008X_1 + 0{,}27X_2. \qquad (7)$$

Далее была определена информационная емкость полученной модели (7), равная 90%.

III. Заключение

На основании всего изложенного и после рассмотрения математических моделей полученных в обоих случаях, можно сделать вывод о том, что возможно применение МНКО для построения адекватных моделей по пассивным данным малого объема без предварительной обработки.

IV. Библиография:

1. Долгов Ю.А. Статистическое моделирование: Учебник для вузов. – 2-е изд., доп.- Тирасполь: Полиграфист, 2011. – 352 с.

2. Столяренко Ю.А. Контроль кристаллов интегральных схем на основе статистического моделирования методом точечных распределений. – Дисс. на соиск. уч. степ. канд. техн. наук по спец. 05.27.01. – М. ГУП «СПУРТ»,2006. – 192 с.