УДК 681.5.015

ВЫБОР КОМПОЗИЦИИ МНОГОМЕРНЫХ ЗАВИСИМОСТЕЙ В ЗАДАЧЕ НЕЛИНЕЙНОЙ РЕГРЕССИИ

Д. К. Тюмиков

Государственное образовательное учреждение высшего профессионального образования «Самарский государственный университет путей сообщения»,
443066, г. Самара, 1-й Безымянный пер., 18
E-mail: dktyumikov@mail.ru

Рассматривается подход к выбору композиции многомерной нелинейной зависимости исходя из парных функций. Он основан на анализе элементов разложения множественного дисперсионного отношения (МДО) на парные дисперсионные отношения (ПДО), дисперсионно-корреляционные отношения эффектов взаимосвязей (ДКОЭВС) и дисперсионные отношения эффектов взаимодействия (ДОЭВД). Показано, что при наличии в разложении только ПДО рекомендуется аддитивная композиция парных нелинейных функций. В разложении с преобладанием ДОЭВД предлагается мультипликативная комбинация. В смешанном наборе дисперсионных отношений предпочтительны смешанные комбинации. ДКОЭВС участвуют в выборе доминантных переменных. Идентичность моделей определяется МДО. Приводится пример, иллюстрирующий различные комбинации многомерных зависимостей, и анализ дисперсионных отношений.

Ключевые слова: нелинейная регрессия, многомерные модели, дисперсионные отношения, аддитивная и мультипликативная композиция.

Введение. При обработке экспериментальных данных часто возникает необходимость нахождения многомерной нелинейной зависимости при априорной её неопределённости. Вид парных нелинейных зависимостей можно с заданной точностью определить по двумерным графикам. Для многомерных зависимостей такой подход практически не реализуем.

Известные методы выбора композиции многомерной зависимости имеют свои недостатки, главный из которых заключается в сведении выбора в самостоятельную сложную задачу. Так, метод перебора достаточно трудоёмок. Метод, основанный на использовании многомерных ортогональных рядов, порождает дополнительные вычислительные трудности, связанные с условием ортогонализации [1]. При этом зависимость получается сложной, как и с применением многомерных рядов Тейлора, что затрудняет интерпретацию экспериментальных результатов. Кроме того, появляются дополнительные трудности в оценивании параметров (коэффициентов) зависимости. Метод группового учёта аргументов, автоматизирующий поиск многомерной зависимости на базе внешних и внутренних критериев, также выдаёт громоздкую модель [2].

Цель данной работы — получить зависимость минимальной сложности и с приемлемой погрешностью. Полагается, что выбор доминантных переменных был произведён [3]. Для этого предлагается подход, основанный на анализе разложения множественного дисперсионного отношения (МДО) [3, 4]. Далее будет показано, что различные сочетания элементов разложения соответствуют различным структурам многомерной зависимости.

Постановка задачи. По данным $\{x, y\}$, $x \in X^n$, $y \subset Y'$, характеризующим состояние объекта исследования, на определённых элементах множества функций $\Phi = \{f_1(x_i), \ldots, f_n(x_i)\}$, $i = \overline{1,n}$, с помощью бинарных операций $\Omega = \{+, -, *, /, **\}$ и множества коэффициентов $C = \{c_1, \ldots, c_k\}$ необходимо сконструировать многомерную зависимость

$$\hat{y} = M[y \mid x] + \Xi = f(x_1, \ldots, x_n, c_1, \ldots, c_k) + \Xi,$$

где $M[\Xi] = 0$, $M[\Xi_i \Xi_j] = 0$ при $j \neq i$, $j, i = 1, \dots, N$ (N — число экспериментов), $D[\Xi] = \sigma_\varepsilon^2$ (значения помехи центрированы, не коррелированы, а их измерения равноточны).

Тогда в статическом варианте в общем виде модель объекта исследования будет описываться системой

$$\Sigma_M = (X, Y, \Omega, \Phi, C, F).$$

Здесь F — критерий параметрического оценивания (например, МНК).

Множественное дисперсионное отношение или остаточную дисперсию можно принять как внутренний критерий, характеризующий идентичность регрессионной зависимости объекту. Дополнительным, внешним, критерием может быть критерий сложности зависимости, выраженный количеством коэффициентов многомерной зависимости.

Однако существуют трудности решения поставленной задачи в таком виде.

1. Не задан способ генерации многомерной зависимости.

2. Не определена предполагаемая точность зависимости сгенерированной структуры, а следовательно, не произведён анализ структур на оптимальность в смысле представленных выше критериев.

Далее приводятся утверждения, позволяющие обойти указанные трудности.

Разложение множественного дисперсионного отношения. Выбор композиции многомерной модели основывается на разложении МДО [3, 4]. Введём обозначения:

$$n_1 = C_n^{n-j},\ j = \overline{0, n-1}; \quad n_2 = C_{n-j}^{n-j-i},\ i = \overline{1, n-j}; \quad M[y \mid X_{v,k}] = M[y] \ \text{при}\, k = 0.$$

Дисперсия зависимой переменной разлагается на остаточную дисперсию и МДО. Разложение МДО имеет вид

$$D[y] = M\Big[(y - M[y \mid X^n])^2\Big] + \sum_{j=0}^{n-1} \sum_{v=1}^{n_1} M\Big[\Big(M[y \mid X_{v,n-j}] + \sum_{i=1}^{n-j} \sum_{v'=1}^{n_2} (-1)^i M[y \mid X_{v',n-j-i}]\Big)^2\Big] +$$

$$+ 2 \sum_{j=0}^{n-1} \sum_{v=0}^{n_1} \sum_{j'=0}^{n'-1} \sum_{v''=1}^{n_1'} M\Big[\Big(M[y \mid X_{v,n-j}] + \sum_{i=1}^{n-j} \sum_{v'=1}^{n_2} (-1)^i M[y \mid X_{v',n-j-i}]\Big) \times$$

$$\times \Big(M[y \mid X_{v'',n-j'}] + \sum_{i'=1}^{n-j'} \sum_{v'''=1}^{n_2'} (-1)^{i'} M[y \mid X_{v''',n-j'-i'}]\Big)\Big], \quad X_{v',n-j-i} \subset X_{v,n-j}. \tag{1}$$

Если поделить левую и правую части разложения на $D[y]$ и переобозначить соответствующие слагаемые, то получим для двух переменных относительные величины

$$1 = \theta_{yx_1x_2} + \eta_{y(x_1x_2)} + \eta_{yx_1} + \eta_{yx_2} + 2(\eta_{y(x_1x_2)(x_1)} + \eta_{y(x_1x_2)(x_2)} + \eta_{y(x_1)(x_2)}),$$

где $\theta_{yx_1x_2}$ — дисперсионное отношение остаточной дисперсии (ДООД); $\eta_{y(x_1x_2)}$ — дисперсионное отношение эффекта взаимодействия (ДОЭВД) двух переменных; η_{yx_1}, η_{yx_2} — парные дисперсионные отношения; $\eta_{y(x_1x_2)(x_i)}$ — дисперсионно-корреляционные отношения эффектов влияния взаимосвязей (ДКОЭВС) переменных x_1, x_2 и x_i, $i = 1, 2$.

Аддитивная многомерная композиция парных нелинейных функций.

Утверждение. Если переменные x_r, $r = 1, m$, статистически независимы и ДОЭВД отсутствуют или ими можно пренебречь, то многомерная модель $y = M[y \,|\, x] + \varepsilon = f(x, c) + \varepsilon$ представляется в виде парных регрессионных зависимостей

$$y_r = M[y \,|\, x_r] + \varepsilon_r, \quad r = \overline{1, m}, \tag{2}$$

соотношением

$$y = M[y \,|\, x] + \varepsilon = \sum_{r=1}^{m} M[y \,|\, x_r] - (m-1)M[y] + \varepsilon.$$

Доказательство. Действительно, из (1) следует, что статистическая связь переменных x_r, $r = 1, m$, с y оценивается только парными дисперсионными отношениями, а значит, имеем

$$D[y] = M[D[y \,|\, x_r]] + \varepsilon = \sum_{r=1}^{m} M[(M[y \,|\, x_r] - M[y])^2] + \varepsilon.$$

Вычтем из левой и правой частей (2) $M[y]$, возведём в квадрат и применим к обеим частям операцию усреднения. Последовательно получим

$$(y - M[y]) = \sum_{r=1}^{m} M[y \,|\, x_r] - mM[y] + \varepsilon,$$

$$(y - M[y])^2 = \Big(\sum_{r=1}^{m} M[y \,|\, x_r] - mM[y] + \varepsilon\Big)^2,$$

$$\underset{yx_1 \ldots x_m}{M}[(y - M[y])^2] = \underset{yx_1 \ldots x_m}{M}\Big[\Big(\sum_{r=1}^{m} M[y \,|\, x_r] - mM[y] + \varepsilon\Big)^2\Big].$$

Воспользовавшись свойством аддитивности математического ожидания, перепишем

$$D[y] = \underset{yx_1 \ldots x_m}{M}\Big[\sum_{r=1}^{m}(M[y \,|\, x_r] - M[y])^2\Big] + \underset{yx_1 \ldots x_m}{M}[\varepsilon^2] + \underset{yx_1 \ldots x_m}{M}\Big[2\Big(\sum_{r=1}^{m}(M[y \,|\, x_r] - M[y])\Big)\varepsilon\Big].$$

Учитывая свойство аддитивности, а также взаимонезависимость переменных, для первого слагаемого имеем

$$M\Big[\sum_{r=1}^{m}(M[y \,|\, x_r] - M[y])^2\Big] = \sum_{r=1}^{m} M[(M[y \,|\, x_r] - M[y])^2].$$

Для второго слагаемого получаем

$$\underset{yx_1 \ldots x_m}{M}[\varepsilon^2] = \underset{yx_1 \ldots x_m}{M}[(y - M[y \,|\, x])^2] = M[D[y \,|\, x]].$$

Третье слагаемое равно нулю, если произвести алгебраические операции, воспользоваться взаимонезависимостью переменных и тем же свойством аддитивности. Следовательно, утверждение доказано.

Утверждение и его доказательство формализуют условия применения одного из методов построения многомерной регрессионной зависимости, основанного на анализе остатков и последовательном нахождении парных зависимостей [5]. Схематически метод заключается в следующем.

Пусть отыскивается $y = f(x, c) + \varepsilon$ по выборке $\{x, y\}$, $x \in X^n$, $y \subset Y'$. Сначала строится зависимость $\bar{y}_1 = f_1(x_1)$ и из y вычитается влияние x_1:

$$y_1^\varepsilon = y - \bar{y}_1 = y - f_1(x_1).$$

Затем строится зависимость

$$\bar{y}_1^\varepsilon = f_2(x_2)$$

и т. д. (до m включительно). Восстановленные этим методом функции $f_r(x_r)$, $r = \overline{1, m}$, составляют многомерную зависимость

$$y = \sum_{r=1}^{m} f_r(x_r) + \varepsilon.$$

Метод построения многомерной зависимости, вытекающей из утверждения, более прост (нет необходимости в пересчёте y_r^ε, $r = \overline{1, m}$) и естествен для интерпретации практических исследований. С точки зрения проблемы идентификации сравниваемые методы одинаковы. Это следует из эквивалентных преобразований. Пусть $m = 2$,

$$y_1^\varepsilon - M[y \,|\, x_1], \quad y_2^\varepsilon = y_1^\varepsilon - M[y_1^\varepsilon \,|\, x_2] = y - M[y \,|\, x_1 x_2] = \varepsilon.$$

Вместо y_1^ε при соответствующей подстановке получаем

$$\varepsilon = y - M[y \,|\, x_1] - \underset{y \,|\, x_2}{M} [y - M[y \,|\, x_1]] = y - M[y \,|\, x_1] - M[y \,|\, x_2] + M[y],$$

так как

$$\underset{y \,|\, x_2}{M} [M[y \,|\, x_1]] = M[y]$$

в силу независимости $M[y \mid x_1]$ от x_2, следовательно,

$$M[y \,|\, x_1 x_2] = M[y \,|\, x_1] + M[y \,|\, x_2] - M[y] = f(x_1 x_2),$$

что соответствует утверждению и, значит, доказывает идентичность методов.

Наконец, следует отметить, что рассматриваемый метод имеет существенный недостаток: из-за неточности восстановления парных зависимостей $M[y \,|\, x_1]$, $M[y_1^\varepsilon \,|\, x_2]$ и других через величины y_r^ε будет проявляться погрешность аппроксимаций парных зависимостей, что приведёт к увеличению погрешности многомерной зависимости.

Неаддитивные многомерные композиции. Пусть в результате анализа был определён статистически значимый эффект взаимодействия двух переменных x_1 и x_2.

Дисперсионное отношение такого эффекта имеет вид

$$\eta_{y(x_1x_2)} = \frac{M[(M[y\,|\,x_1x_2] - M[y\,|\,x_1] - M[y\,|\,x_2] + M[y])^2]}{D[y]}.$$

Наличие этой величины в разложении (1) приводит к выводу: многомерную функцию $M[y\,|\,x_1x_2]$ нельзя представить в виде суммы парных регрессионных зависимостей, так как здесь ДОЭВД будет равно нулю, что противоречит рассматриваемому случаю. Однако охватить все виды комбинаций одномерных функций для этого случая невозможно из-за многообразия вариантов. Поэтому выделим наиболее распространённые схемы:

— представление $M[y\,|\,x]$ произведением парных зависимостей:

$$y = c\prod_{r=1}^{m} f(x_r) + \varepsilon; \tag{3}$$

— представление $M[y\,|\,x]$ в виде некоторой функции от одномерных функций:

$$y = \Psi(f_1(x_1), \ldots, f_m(x_m)).$$

Общих утверждений, позволяющих однозначно дать рекомендации по синтезу многомерной модели аналогично вышеприведённому, найти не удалось. Тем не менее наличие ДОЭВД указывает на необходимость привлечения мультипликативных операций между парными функциями соответствующих ДОЭВД входных переменных. Рассмотрим схему (3) более подробно. Пусть строится зависимость

$$\bar{y} = c\prod_{r=1}^{m} f_r(x_r).$$

На первом шаге подбирается функция $\bar{y}_1 = f_1(x_1)$ по выборке $\{y, x_1, \ldots, x_n\}$ и вычисляется величина

$$y_1^{\varepsilon} = y/f_1(x_1).$$

На втором шаге строится $y_2^{\varepsilon} = f_2(x_2)$ по данным (y_1^{ε}, x_2) и вычисляется величина

$$y_m^{\varepsilon} = y/(f_1(x_1)f_2(x_2))$$

и т. д. до

$$y_m^{\varepsilon} = y\Big/\prod_{r=1}^{m} f_r(x_r) = \varepsilon_m.$$

Зададимся целью отыскать зависимость (3) в выражении через парные регрессионные зависимости $M[y\,|\,x_r] = f_r(x_r)$, сделав допущение, что $M[y\,|\,x]$ может быть представлена в виде (3). Тогда для $m = 2$ имеем

$$y_1 = M[y\,|\,x_r], \quad y_1^{\varepsilon} = y/M[y\,|\,x_1], \quad y_2 = M[y_1^{\varepsilon}\,|\,x_r], \quad y_2^{\varepsilon} = y_1^{\varepsilon}/M[y_1^{\varepsilon}\,|\,x_2] = \varepsilon_2.$$

Произведя соответствующие замены переменных, получим

$$\varepsilon_2 = \frac{y}{M[y\,|\,x_1]M\Big[\dfrac{y}{M[y\,|\,x_1]}\,\Big|\,x_2\Big]} = \frac{yM[y]}{M[y\,|\,x_1]M[y\,|\,x_2]}.$$

Следовательно, $M[y\,|\,x]$ можно записать как

$$M[y\,|\,x] = \frac{c}{(m-1)M[y]}\prod_{r=1}^{m} M[y\,|\,x_r]. \qquad (4)$$

Выражение (4) позволяет избежать внесения дополнительных погрешностей при пересчёте y_1^{ε} из-за неточности представления $f_r(x_r)$.

Естественно ожидать, что при расчёте элементов разложения (1) могут возникнуть и ПДО, и ДОЭВД. Тогда необходимо использовать поочерёдно построение аддитивных и мультипликативных комбинаций в последовательности, соответствующей ранжированному ряду дисперсионных отношений. При появлении в разложении (1) ДКОЭВС следует производить выбор доминантных переменных с учётом этих характеристик.

Описанный подход может быть полезен при аппроксимации многомерных функций рядами. ПДО и ДОЭВД будут указывать на аддитивные и мультипликативные составляющие ряда соответствующих входных переменных. Перебор при синтезе модели исключён. Для получения требуемой точности модели может потребоваться увеличение порядка учитываемых составляющих ряда.

Пример. На основе трёх парных зависимостей формируется трёхмерная композиция, которая табулируется, образуя массив экспериментальных данных. Рассчитываются дисперсионные отношения и сводятся для соответствующей композиции в табл. 1–6. В табл. 4–6 кроме вида композиции указаны зависимости между входными переменными. В диагонали представлены ПДО ($X_1 - X_1, \ldots$) и ДОЭВД ($X_1X_2 - X_1X_2, \ldots$), вне диагонали — ДКОЭВС $(X_1)(X_2), \ldots, (X_1X_2X_3)(X_1X_3)$. ПВП обозначает подмножество входных переменных. Под последними колонками таблиц указана величина МДО. Значения в таблицах симметричны относительно главной диагонали. Пусть заданы парные зависимости для трёх переменных:

$$f1(x) := (x_1)^2, \quad f2(x) := \sin(2x_2), \quad f3(x) := \mathrm{e}^{-x_3}.$$

Таблица 1

$y := f1(x_1) + f2(x_2) + f3(x_3)$

ПВП	X_1	X_2	X_3	X_1X_2	X_2X_3	X_1X_3	$X_1X_2X_3$
X_1	**0,208**	0	0	0	0	0	0
X_2	0	**0,045**	0	0	0	0	0
X_3	0	0	**0,660**	0	0	0	0
X_1X_2	0	0	0	**0**	0	0	0
X_2X_3	0	0	0	0	**0**	0	0
X_1X_3	0	0	0	0	0	**0**	0
$X_1X_2X_3$	0	0	0	0	0	0	**0**

0,9199

Таблица 2

$$y := f1(x_1)f2(x_2) + f3(x_3)$$

ПВП	X_1	X_2	X_3	X_1X_2	X_2X_3	X_1X_3	$X_1X_2X_3$
X_1	**0**	0	0	0	0	0	0
X_2	0	**0,108**	0	0	0	0	0
X_3	0	0	**0,690**	0	0	0	0
X_1X_2	0	0	0	**0,103**	0	0	0
X_2X_3	0	0	0	0	**0**	0	0
X_1X_3	0	0	0	0	0	**0**	0
$X_1X_2X_3$	0	0	0	0	0	0	**0**

0,9162

Таблица 3

$$y := f1(x_1)f2(x_2)f3(x_3)$$

ПВП	X_1	X_2	X_3	X_1X_2	X_2X_3	X_1X_3	$X_1X_2X_3$
X_1	**0,018**	0	0	0	0	0	0
X_2	0	**0,242**	0	0	0	0	0
X_3	0	0	**0,014**	0	0	0	0
X_1X_2	0	0	0	**0,237**	0	0	0
X_2X_3	0	0	0	0	**0,228**	0	0
X_1X_3	0	0	0	0	0	**0,016**	0
$X_1X_2X_3$	0	0	0	0	0	0	**0,223**

0,9780

Таблица 4

$$y := f1(x_1) + f2(x_2) + f3(x_3), \quad x_2 := 2x_1$$

ПВП	X_1	X_2	X_3	X_1X_2	X_2X_3	X_1X_3	$X_1X_2X_3$
X_1	**0,158**	0,158	0	–0,158	–0,158	0	0
X_2	0,158	**0,158**	0	0	0	0	0
X_3	0	0	**0,744**	0	0	0	0
X_1X_2	–0,158	0	0	**0,158**	0	0	0
X_2X_3	–0,158	0	0	0	**0**	0	0
X_1X_3	0	0	0	0	0	**0**	0
$X_1X_2X_3$	0	0	0	0	0	0	**0**

0,9026

Таблица 5

$$y := f1(x_1)f2(x_2) + f3(x_3), \quad x_2 := 2x_1$$

ПВП	X_1	X_2	X_3	X_1X_2	X_2X_3	X_1X_3	$X_1X_2X_3$
X_1	**0,280**	0,280	0	–0,280	–0,280	0	0
X_2	0,280	**0,280**	0	0	0	0	0
X_3	0	0	**0,636**	0	0	0	0
X_1X_2	–0,280	0	0	**0,280**	0	0	0
X_2X_3	–0,280	0	0	0	**0**	0	0
X_1X_3	0	0	0	0	0	**0**	0
$X_1X_2X_3$	0	0	0	0	0	0	**0**

0,9165

Таблица 6

$$y := f1(x_1)f2(x_2)f3(x_3), \quad x_2 := 2x_1$$

ПВП	X_1	X_2	X_3	X_1X_2	X_2X_3	X_1X_3	$X_1X_2X_3$
X_1	**0,448**	0,448	0	–0,448	–0,448	0	0
X_2	0,448	**0,448**	0	0	0	0	0
X_3	0	0	**0,129**	0	0	0	0
X_1X_2	–0,448	0	0	**0,448**	0	0	–0,407
X_2X_3	–0,448	0	0	0	**0,407**	0,407	–0,407
X_1X_3	0	0	0	0	0,407	**0,407**	0
$X_1X_2X_3$	0	0	0	–0,407	–0,407	0	**0,407**

0,9840

В табл. 1 представлены результаты расчётов для аддитивной комбинации. Наличие только ПДО подтверждает справедливость утверждения. С введением мультипликативных комбинаций появляются соответствующие ДОЭВД (см. табл. 2, 3). Зависимость между входными переменными способствует появлению не только ДКОЭВС, но и ДОЭВД, что усложняет анализ дисперсионных характеристик. При этом первоначально заданная трёхмерная композиция заменяется двумерной, что видно из соответствующих подстановок в формулах. В таком случае выбор доминантных переменных с учётом ДКОЭВС значительно упрощает выбор типа многомерной зависимости.

Заключение. Анализ дисперсионных характеристик в разложении множественного дисперсионного отношения позволил определить условия применения аддитивной многомерной зависимости на основе парных регрессионных функций. Эти условия заключаются в наличии в разложении МДО только парных дисперсионных отношений. При появлении ДОЭВД требуется применение мультипликативных или более сложных комбинаций. Предложенный подход не требует генерации вариантов многомерной зависимости. Рекомендуемый вариант определяется по индексам входных переменных в выбранных ПДО и ДОЭВД. По полученным дисперсионным характеристикам определяется МДО — достижимая величина идентичности зависимости объекту исследования. Подход может быть формализован для компьютерной обработки данных и выбора типа многомерной зависимости. Приведённый пример показал работоспособность предложенного подхода.

СПИСОК ЛИТЕРАТУРЫ

1. **Никольский С. М.** Приближение функций многих переменных и теоремы вложения. М.: Наука, 1969. 480 с.
2. **Ивахненко А. Г.** Долгосрочное прогнозирование и управление сложными системами. Киев: Технiка, 1975. 312 с.
3. **Кацюба О. А., Тюмиков Д. К.** Алгоритм определения доминантных переменных в задачах нелинейной регрессии // Автометрия. 1983. № 3. С. 18–24.
4. **Тюмиков Д. К.** Идентификация многомерных нелинейных объектов на основе дисперсионных отношений. Самара: СНЦ РАН, 2008. 162 с.
5. **Дрейпер Н., Смит Г.** Прикладной регрессионный анализ. М.: Диалектика, 2007. 912 с.

Поступила в редакцию 29 января 2010 г.