М.И. Калашникова

ТЕОРЕТИКО- ЭКСПЕРИМЕНТАЛЬНЫЕ РЕГРЕССИИ (ТЭР)

В значительном числе случаев дорогостоящие регрессионные эксперименты предваряются соответствующими теоретическими исследованиями. Таких экспериментов, как правило, бывает немного. В результате возникает задача совмещения результатов двух типов исследований.

Теоретики такие задачи трактуют так: имеем функцию Z, полученную расчетами, ее надо уточнить с помощью нескольких экспериментальных точек. Для экспериментаторов задача звучит так: по малому числу экспериментальных точек построить регрессию, опираясь на функцию Z.

Данная работа посвящена одному из возможных способов решения этой задачи. Решение ищется в виде линейной регрессии со специальным базисом, индивидуальным для каждой регрессии. Базис определяется численным путем по теоретическим данным, коэффициенты, как обычно, получают по экспериментальным данным.

В дальнейшем удобно такие линейные регрессии называть сокращенно *ТЭР* (теоретико-экспериментальные регрессии).

Первый вариант способа построения *ТЭР* был опубликован автором в работе [1], в следующей работе [2] была показана возможность выполнения на основе *ТЭР* процедуры *D*-оптимального планирования регрессионных экспериментов [3]. В обобщающей работе [4] (в соавторстве с А.Л. Поляченко и Т.А. Шапошниковой) дана сводка результатов из [1,2] и приведены конкретные численные примеры *D*-оптимальных экспериментов для типичной задачи скважинной ядерной геофизики.

В данной работе предлагаются два новых варианта построения *ТЭР*. Первый дает возможность более полно, чем в [4], использовать теоретические данные, второй вариант (со смешанным базисом) позволяет управлять процессом построения *ТЭР*. В варианте со смешанным базисом используются функции как специальные, рассмотренные автором в [4], так и классические, например, многочлены. Такие "смеси" делают процесс построения *ТЭР* более гибким и создают оправданные предпосылки для постановки задачи, в определенном смысле обратной задаче оптимального планирования экспериментов. Это задача вычисления "оптимальных" базисов, использование которых приводит к тому, что исходный план экспериментов становится оптимальным.

Следует обратить внимание на то, что, как и в [4], в данной работе мы имеем дело с готовым теоретическим и экспериментальным материалом, изменять который мы не в силах.

Автор благодарен проф. Т.А. Гермогеновой, проф. М.Б. Малютову и

К.А. Кукушкину, общение с которыми способствовало постановке задачи о смешанных базисах.

Введем обозначения.

$$x = \left(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right) - \tag{1}$$

набор непрерывных независимых переменных.

$$-\infty < a_i \le x^{(i)} \le b_i < \infty . \tag{2}$$

Ω – n-мерный параллелепипед в пространстве переменных (1), для которого выполняются неравенства (2).

$$P_i = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_{k_i}^{(i)}\right);\ i = 1,2,\ldots,n \tag{3}$$

– дискретные точки по переменным (1).

$P \in \Omega$ – полная сетка узлов переменных(1), соответствующая точкам (3).

Предположим, что с целью выявления зависимости *u(x)* проведены теоретические и экспериментальные исследования. При этом получены q экспериментальных точек

$$\overline{u}(x_j), \quad j = 1,2,\ldots,q \tag{4}$$

и теоретическая функция *Z(x)*. Считаем, что функция *Z(x)* дискретна и определена на полной сетке переменных (1), иначе говоря, *Z(x)* - n-мерная матрица (таблица) с числом элементов

$$k = \prod_{i=1}^{n} k_i .$$

Наша задача – построение *ТЭР* по эксперименту (4), опираясь на *Z(x)*.

Конкретным примером таких задач являются, например, задачи построения интерпретационных зависимостей для нейтронных методов скважинной ядерной геофизики. Для этих задач необходимое число учитываемых переменных (1) $n \ge 4$, число точек $q \le 30$.

Из сочетания этих значений n, q видна необходимость совместного учета результатов теории и эксперимента. Для указанных задач теоретическая функция *Z(x)* определяется по решению системы дифференциальных уравнений переноса излучения в горных породах (диффузионное приближение). Экспериментальные данные для этой задачи – измерения, полученные реальным скважинным прибором на натурных моделях горных пород. Крайне ограниченный обьем экспериментальных данных обьясняется трудностями строительства парка натурных моделей: далеко не все модели с необходимыми значениями переменных (1) удается реализовать, сроки жизни моделей ограничены, каждая модель уникальна.

Простейший вариант построения *ТЭР* состоит в том, что мы ищем регрессию в виде [4]

$$u(c,x)=\sum_{s=1}^{m}c_sZ_s(x);\qquad x\in P;c=(c_1,...,c_m)\in R^m, \tag{5}$$

где базисные функции $Z_s(x)$, $s=1,2,...,m$ являются дискретными функциями с разделяющимися переменными

$$Z_s(x)=f_s(x^{(1)})\times g_s(x^{(2)})\times...\times h_s(x^{(n)});\ x\in P;s=1,2,...,m\ , \tag{6}$$

являющимися в свою очередь решениями следующей задачи минимизации

$$H_0(x)=Z(x)$$

$$\min_{\tilde f_s\times\tilde g_s\times...\times\tilde h_s}\left(H_{s-1}(x)-\tilde f_s(x^{(1)})\times\tilde g_s(x^{(2)})\times...\times\tilde h_s(x^{(n)})\right)^2; \tag{7}$$

$$H_s(x)=Z(x)-\sum_{t=1}^{s}(f_t(x^{(1)})\times g_t(x^{(2)})\times...\times h_t(x^{(n)})); \tag{8}$$

$$s=1,2,...,m.$$

Здесь и далее $\|\cdot\|$ – значок среднеквадратичной нормы.

В работе [5] показано, что существует по крайней мере одно решение задачи (6)-(8), причем метод последовательного выделения компонентов $Z_s(x)$ приводит к представлению $Z(x)$ в виде убывающего ряда

$$Z(x)=\sum_{s=1}^{\infty}(f_s(x^{(1)})\times g_s(x^{(2)})\times\cdots\times h_s(x^{(n)})). \tag{9}$$

Первые m слагаемых разложения (9) и составляют базис *ТЭР* для модели (5).

Алгоритм расчета базиса (6) универсален для всех дискретных (табличных) $Z(x)$ и всех n [5,6,7]. Непрерывный его аналог впервые был, повидимому, описан Шмидтом еще в 1907г. при разложении ядра интегрального уравнения [8]. Для частного случая $n=2$ разложение (9) полностью совпадает с разложением двумерной матрицы на сумму матриц единичного ранга [9]. Время расчета базиса на компьютере, естественно, зависит от конкретной задачи. Например, в задачах, в которых автор принимал участие, расчет 5-6 базисных функций (с 6-ю значащими цифрами) для $n=4$, $k_1=6$, $k_2=k_3=4$, $k_4=3$ занимает несколько секунд на PC.

Конкретные численные примеры, дают основание предполагать, что для положительных $Z(x)$ последовательность функций f_s, $s=1,2,...,m$ обладает некоторыми свойствами собственных векторов якобиевых матриц:

– функция f_s испытывает $s-1$ перемену знака,

– нули двух последовательных функций f_s, f_{s+1} перемежаются.

Нетрудно заметить, что в *ТЭР* с моделью (5) используется только часть слагаемых разложения (9). Значительно большей "отдачи" от теоретических данных можно ожидать, если брать модель

$$u(c,x) = \sum_{s=1}^{m-1} c_s Z_s(x) + c_m(Z(x) - \sum_{s=1}^{m-1} Z_s(x));\ x \in P; \qquad (10)$$
$$c = (c_1, c_2, \ldots, c_m) \in R^m.$$

Естественным доводом в пользу модели (10) по сравнению с моделью (5) является выполнение равенства

$$u(c,x)=Z(x)$$

для $c_s=1$, $s=1,2,\ldots,m$. Очевидно, что это случай, когда все экспериментальные точки (4) лежат на *Z(x)*. Очевидно также, что модели (5) и (10) совпадают, если в разложении (9) число членов ряда конечно и равно m.

Замечания:

(i) для этой работы способы оценки коэффициентов c_s, $s=1,2,\ldots,m$ и числа m членов регрессии (5) не существенны, поэтому здесь не рассматриваются;

(ii) принимая во внимание, что для выбранного базиса (6) вопрос интерполяции в (5) сводится к интерполяции функций от одной переменной, повсюду в этой работе рассматриваются только дискретные функции;

(iii) в практических вычислениях стоит проверить необходимость использования в (10) свободного члена [7].

В процессе построения модели *ТЭР* возникает естественный вопрос: хватит ли слагаемых в разложении (9), для того чтобы можно было работать с имеющимся экспериментом (4). Ответ очевидный: не всегда. Например, для $n=1$ мы вообще имеем единственную базисную функцию $Z_1=Z$ и все выше приведенные конструкции являются тривиальными. Для $n=2$ мы имеем [5]

$$Z(x) = Z(x^{(1)}, x^{(2)}) = \sum_{s=1}^{r} Z_s(x^{(1)}, x^{(2)}) = \sum_{s=1}^{r} (f_s(x^{(1)}) \times g_s(x^{(2)})). \quad (11)$$

Здесь r – ранг матрицы Z, иначе говоря, в разложении (9) заведомо конечное число членов. Для $n > 2$ мы в (9) можем иметь как конечное, так и бесконечное число членов. Одним словом, заранее не ясно, достаточно ли членов в (9) мы имеем.

Нам представляется, что в случае необходимости разумным выходом из положения является погружение исходной задачи в задачу большей размерности, т.е. с большим числом n, необходимо также, чтобы для этой задачи большей размерности соответствующее разложение (9) имело

бы больше членов, чем в исходной задаче (это всегда можно сделать).

Рассмотрим один из подходов к этой проблеме - использование смешанного базиса. Рассмотрение целесообразно провести на простеньком примере.

Пусть $n=1$ и $l_1(x^{(1)})$, $l_2(x^{(1)})$ две дискретные функции на P_1, равные

$$l_1(x^{(1)}) = 1;\ l_2(x^{(1)}) = x^{(1)}. \tag{12}$$

Предположим, что функции

$$Z(x^{(1)}),\quad l_1(x^{(1)}),\quad l_2(x^{(1)}) \tag{13}$$

линейно независимы (очевидно, это не обременительное условие). Если совокупность (13) взять в качестве базисной для искомой регрессии, то получим

$$u(c, x^{(1)}) = c_1 Z(x^{(1)}) + c_2 + c_3 x^{(1)}. \tag{14}$$

Это легкий, но не лучший путь использования базиса. Действительно, для $c_3=0$ мы имеем часто встречающийся случай, который является обычным масштабированием и сдвигом. Однако уже в случае всех ненулевых коэффициентов в (14) регрессия представляется чем-то искусственным и "некрасивым". В связи с этим в данной работе предлагается поступить следующим образом:

1. Рассмотрим совокупность (13), как двумерную функцию $Z(x^{(1)},x^{(2)})$ от искусственной переменной $x^{(2)}$. Пусть

$$Z(x^{(1)},1) = Z(x^{(1)});\quad Z(x^{(1)},2) = l_1(x^{(1)});\quad Z(x^{(1)},3) = l_2(x^{(1)}). \tag{15}$$

2. Для $Z(x^{(1)},x^{(2)})$ найдем функции

$$Z_s(x^{(1)},x^{(2)}); s = 1,2,3, \tag{16}$$

как решение задачи (6)-(8)

3. Используем функции (16) при $x^{(2)}=1$ как базисные для регрессии

$$u(c, x^{(1)}) = \sum_{s=1}^{3} c_s Z_s(x^{(1)},1);\ x^{(1)} \in P_1;\ c = (c_1, c_2, c_3) \in R^3.$$

Из выражения (11) непосредственно следует, что

$$\sum_{s=1}^{3} Z_s(x^{(1)},1) = Z(x^{(1)}) \tag{18}$$

т.е. при $c_1=c_2=c_3=1$ в выражении (17), как и в общем случае (10), мы получаем

$$u(c, x^{(1)}) = Z(x^{(1)}).$$

Итак, для выбранной совокупности функций (13) мы получаем разложение (18) теоретической функции $Z(x^{(1)})$ на сумму слагаемых, при этом

стоящие в (17) коэффициенты дают представление о "вкладе" этих слагаемых в искомую регрессию.

Очевидно, что при других функциях, например, l_1 - константа, а l_2 - экспонента, мы получим разложение (18) с другими функциями (16). Это обстоятельство позволяет надеяться, что используя разные функции $l_1(x^{(l)})$, $l_2(x^{(l)})$, и получая разные разложения (18) и, соответственно, разные базисы (16), мы сможем научиться получать базисные функции (16), удовлетворяющие разным дополнительным требованиям. Одной из интереснейших задач такого рода является задача вычисления "оптимального" базиса, иначе говоря, базиса, для которого исходный план эксперимента является оптимальным.

Принцип использования смешанных базисов для общего случая $n>1$ тот же, что и для $n=1$. Здесь этот случай не рассматривается.

Помимо своего основного назначения – синтезирования дискретных данных разной природы – модель *ТЭР* и ее составные части могут быть использована в других задачах. Нам представляется, что наиболее интересны такие.

1. Расчет оптимальных планов эксперимента. Возможность использования модели (10) для этой задачи объясняется линейностью модели (10) по коэффициентам; разумность использования модели (10) объясняется возможностью объективного учета теоретических данных, накопленных к началу планирования экспериментальных работ. Некоторые численные примеры *D*-оптимальных планов для задач скважинной ядерной геофизики даны в работе [4].

2. Интерполяция дискретных функций, заданных на полной сетке переменных. Основание очевидно - использование в правой части разложения (9) функций от одной переменной. Однако, как и во всякой задаче интерполяции, здесь требуется опыт вычислителя.

3. Задачи, связанные с задачами распознавания образов. Это задачи компонентного анализа, фильтрации шумов и сглаживания табличных данных. В литературе для $n=2$ примеры такого рода задач даны в [10,11,12].

Важно отметить, что в случае отсутствия теоретической матрицы *Z*, некоторые способы построения регрессий с обычными базисными функциями можно свести к рассмотренному методу *ТЭР*. Предположим, что методом последовательного наращивания числа m, мы ищем регрессию

$$u(c,x) = \sum_{s=1}^{m} c_s u_s(x)$$

с обычными базисными функциями $u_s(x)$, $s=1,2,...,m$. Очевидно, что промежуточный результат при переходе от m_0 к m_0+1 можно рассматривать как "теоретическую" функцию, т.е. брать

$$Z(x) = \sum_{s=1}^{m_0} c_s u_s(x) \ .$$

Литература

1. Калашникова М.И. Алгоритм сопоставления массивов геофизической информации большой размерности. - ВНИИЯГГ. М., деп. в ВИНИТИ, 1982, N 1569-82 деп., 7с.
2. Калашникова М.И. Применение методов планирования эксперимента при ядерно-геофизических исследованиях на натурных моделях. - ВНИИЯГГ. М., деп. в ВИНИТИ, 1985, N 1894-85 деп., 14с.
3. Федоров В.В. Теория оптимального эксперимента (планирование регрессионных экспериментов). М., "Наука", 1977, 312 с.
4. M.I.Kalashnikova, A.L.Polyachenko, T.A.Shaposhnikova. Mathematical- Experimental Charts, Geophys. J., 1990. V 8(4), pp 484-492.
5. Поспелов В.В. О приближении функций нескольких переменных суммами произведений функций одного переменного. Препринт ИПМ АН СССР, N 32, М., 1972, 75с.
6. Даугавет В.А. Один практический прием приближения функций многих переменных. Сб. "Методы вычислений", N 6, изд. ЛГУ, 1970, с.3-8.
7. Шура-Бура М.Р. Аппроксимация функций многих переменных функциями, каждая из которых зависит от одного переменного. Сб."Вычислительная математика", N 2, М., 1957, с.3-19.
8. Schmidt E. Zur Theorie der linearen und nichtlinearen Integralgleichunden. Math. Ann., LXIII, 1907, p.433-476
9. G.Forsythe, M.Malcolm, C.Moler. Computer Methods for Mathematical Computations, N.Y. 1977.
10. Мещерская А.В., Руховец Л.В., Юдин М.И. и др. Естественные составляющие метеорологических полей. Л., Гидрометеоиздат, 1970, 199с.
11. Баглай Р.Д., Смирнов К.К. К обработке двумерных сигналов на ЭВМ. ЖВМ и МФ. т.15, N1, 1975, с. 241-248
12. Обухов А.М. О статистических ортогональных разложениях эмпирических функций. Изв. АН СССР, серия "Геофизическая", вып.3, 1960, с.432-439.