Топологический метод устойчивого оценивания коэффициентов многофакторного уравнения регрессии в условиях мультиколлинеарности факторов

Станислав РАДЧЕНКО

УДК 515.126.2:519.237.5

Впервые излагается метод устойчивого ($1 \le cond \le 10$) оценивания коэффициентов статистических моделей на основе топологического отображения (гомеоморфизм) прообраза планирования эксперимента в образ математического моделирования. Приведена схема формирования прообраза планирования эксперимента. Для произвольной области факторного пространства возможно устойчивое оценивание коэффициентов модели в условиях значительной ($0{,}6 < |r(X_{io}, X_{jo})| < 1$) мультиколлинеарности факторов.

A topological method of robust estimation of multifactor regression equation coefficients in *cond*itions of multicollinearity / Radchenko S.G. // Mathematical Machines and Systems. – 2001. – № 1, 2, p 114...121. UDK 515.126.2:519.237.5

A Method of robust ($1 \le cond \le 10$) estimation of statistical models coefficients for the first time is stated on the basis of topological transformation (homeomorphism) prototype of design of experiment in an image of mathematical modeling. The circuit of formation of a design of experiment prototype is given. For any area factoral space it is possible robust estimation of coefficients in *cond*itions large ($0{,}6 < |r(X_{io}, X_{jo})| < 1$) multicollinearity.

Введение

Эффективность использования статистических моделей для прогноза, управления, изучения механизма происходящих явлений зависит от правильности представления причинных, структурных и количественных связей между факторами и моделируемым критерием качества. Для большинства реальных (не стандартных) форм факторного пространства методы выделения устойчивых структур и коэффициентов (кроме метода регуляризации) не известны. Рассмотрению путей создания метода выделения устойчивых структур и коэффициентов многофакторного уравнения регрессии в условиях мультиколлинеарности факторов посвящена данная статья.

Устойчивое оценивание структуры и коэффициентов модели в стандартных областях факторного пространства

Статистические модели нашли значительное применение в решениях разнообразных научных и прикладных задач. Их получение основано на аппроксимации исходных данных. Исходные данные получают путем проведения экспериментов, использования методов статистических испытаний, экспертных оценок и др.

В большинстве случаев для построения статистических моделей используют многофакторные уравнения регрессии. Устойчивое оценивание коэффициентов этих уравнений возможно при использовании полного факторного эксперимента, многофакторных регулярных планов, $ЛП_\tau$ равномерно распределенных последовательностей[1] и системы ортогональных полиномов Чебышева [1, с. 102...103; 127...133].

Планирование эксперимента проводится, как правило, в стандартных областях факторного пространства — на кубе, сфере, симплексе (рис. 1). В этих случаях достигается максимально возможная устойчивость коэффициентов многофакторного уравнения регрессии: число обусловленности $cond = 1$ или в большинстве случаев $cond < 10$ [1, с. 128; 182; 187; 191; 208...209].

Причины мультиколлинеарности факторов в множественном регрессионном анализе

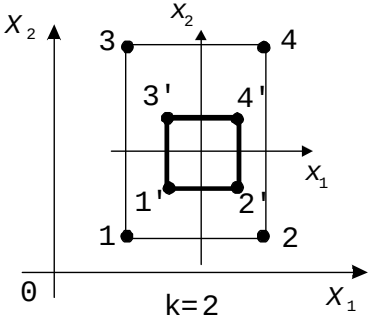
Так как планирование эксперимента в указанных стандартных областях факторного пространства не всегда возможно, а в определенных условиях для получения моделей могут использоваться результаты наблюдений, то возникает взаимная сопряженность, иначе мультиколлинеарность, факторов.

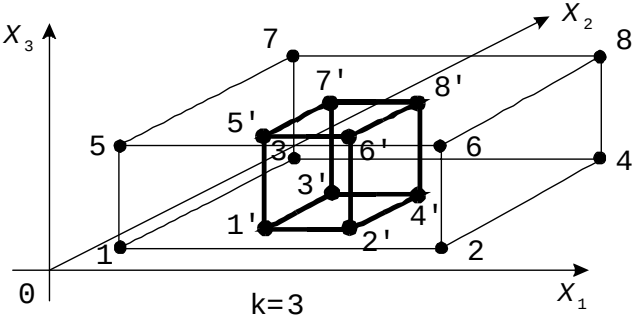
Другой причиной мультиколлинеарности факторов является гипотеза о законе корреляции параметров однородного ряда технических объектов, предложенная д.т.н. проф. А.И. Половинкиным [2, с. 338]. Она имеет

[1] Последовательность точек $P_0, P_1, \ldots, P_i, \ldots$ куба K^n называется $ЛП_\tau$ последовательностью, если любой ее двоичный участок, содержащий не менее, чем $2^{\tau+1}$, точек, представляет собой $П_\tau$-сетку.

следующую формулировку: «Однородный ряд технических объектов S_1, S_2, ..., S_k, имеющих одинаковые функцию и техническое решение, описываемое набором параметров x, y_1, ..., y_n, и отличающихся значениями главного параметра x_j, связан между собой отношениями $y_i = a_i x_j + b_i$, ($i = 1$, ..., n; $j = 1$, ..., k)».

Значительно ранее проф. А.И. Сидоров также отмечал, что в машинах все зависимости между размерами приближенно оказываются линейными зависимостями, и все размеры выражаются приближенно как функции первой степени от главного размера [3, с. 386].

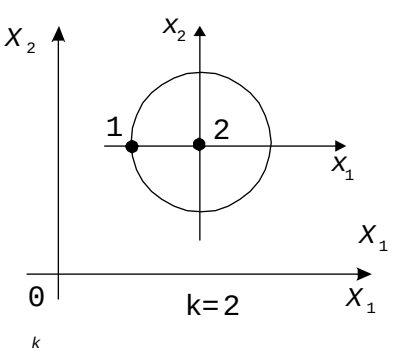


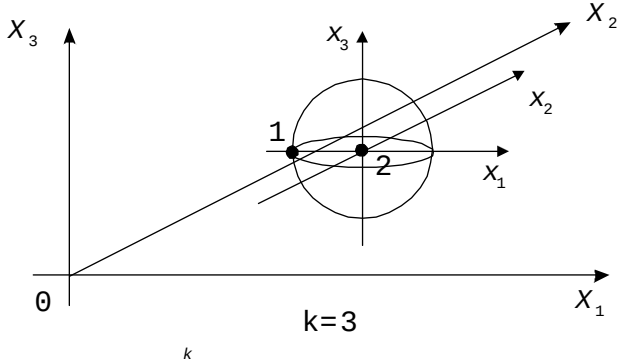


$$X_{i\min} \le X_i \le X_{i\max};$$
$$-1 \le x_i \le 1.$$

$$X_{i0} = (X_{i\min} + X_{i\max})/2,$$
$$\Delta X_i = (X_{i\max} - X_{i\min})/2.$$
$$x_i = (X_i - X_{i0})/\Delta X_i.$$

Рис. 1а. Стандартные области планирования факторного пространства: на кубе





$$\left(\sum_{i=1}^{k} X_i^2\right)^{1/2} \le \rho; \quad X_{i\min} \le X_i \le X_{i\max}.$$

$$\left(\sum_{i=1}^{k} x_i^2\right)^{1/2} \le 1; \quad -1 \le x_i \le 1.$$

Рис. 1б. Стандартные области планирования факторного пространства на сфере

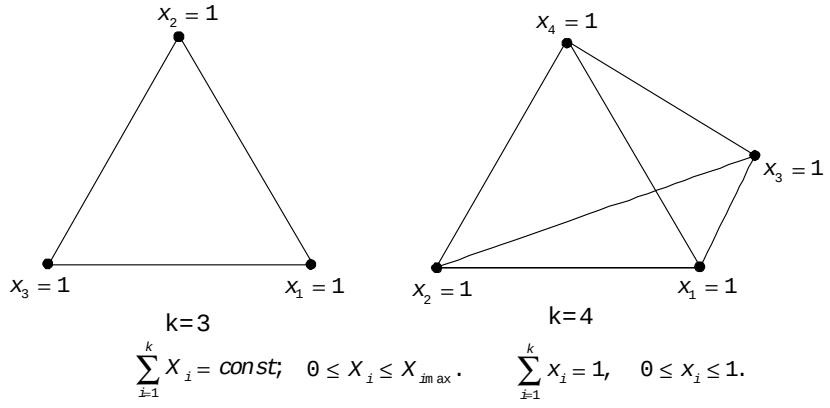


Рис. 1в. Стандартные области планирования факторного пространства на симплексе

Немецкие исследователи в области обработки металлов резанием приводят примеры фактических областей факторного пространства рабочих режимов технических и технологических систем во многих важных исследовательских задачах, существенно отличающихся от стандартных [4, с. 46].

На проблему мультиколлинеарности факторов в множественном регрессионном анализе обращают внимание многие специалисты. Мультиколлинеарность является причиной понижения точности оценки коэффициентов уравнения регрессии, искажения оценок их дисперсий и ковариаций между ними, искажения дисперсии остатков. С возрастанием закоррелированности факторов вычисленные коэффициенты теряют прикладной смысл при попытке их интерпретации. Среднеквадратичные ошибки коэффициентов существенно возрастают, а сами коэффициенты становятся очень чувствительны к выборочным наблюдениям. Незначительное изменение исходных данных «может привести к очень сильным сдвигам в значениях оценок» [5, с. 215].

Д.ф.-м.н. Е.З. Демиденко, специалист в области линейной и нелинейной регрессии, делает вывод, что «мультиколлинеарность — одно из основных препятствий эффективного применения аппарата регрессионного анализа» [6, с. 186].

Глубинный смысл последствий мультиколлинеарности факторов связан с неполной вырожденностью факторного пространства, в котором невоз-

можны независимые оценки эффектов и, следовательно, наиболее точное их определение. С увеличением закоррелированности факторов, что проистекает из прикладных системных причин, задача становится некорректно поставленной, а ее решение с приемлемыми критериями качества получаемых моделей — невозможным.

Из приведенного следует, что проблема устойчивого ($1 \le cond < 10$) оценивания коэффициентов в условиях мультиколлинеарности факторов является одной из основных проблем решения некорректно поставленных задач.

Анализ мультиколлинеарности факторов показывает, что она обусловлена метрическими свойствами пространства (размер, угол, форма, длина) для множества точек X факторного пространства, которые представляют в совокупности значения исходных данных, т.е. значения факторов.

Топологический метод устойчивого оценивания структуры и коэффициентов модели

Произвольное метрическое пространство может также рассматриваться как топологическое пространство. Для плохо обусловленного факторного пространства можно найти топологически эквивалентное (или гомеоморфное) хорошо обусловленное факторное пространство, в котором решение поставленной задачи будет хорошо обусловленным.

Будем различать множество точек прообраза планирования эксперимента $X_{\text{пр}}$ и множество точек образа математического моделирования $X_{\text{о}}$. Можно специальными методами сконструировать множество точек $X_{\text{пр}}$ и получить план эксперимента с наилучшими возможными статистическими свойствами. Множество точек $X_{\text{о}}$ задается комплексом условий в предметной области и, как правило, не может быть изменено.

Предлагается использовать метод топологического отображения (гомеоморфизм) хорошо обусловленного факторного пространства — прообраза планирования эксперимента (множество точек $X_{\text{пр}}$) в плохо обусловленное — образ математического моделирования (множество точек $X_{\text{о}}$).

Множества $X_{\text{пр}} \subset R^n$ и $X_{\text{о}} \subset R^n$ называются топологически эквивалентными (гомеоморфными), если существует такая взаимно однозначная функция $f: X_{\text{пр}} \to X_{\text{о}}$, что как сама функция f, так и обратная функция $f^{-1}: X_{\text{о}} \to X_{\text{пр}}$ непрерывны. Такая функция f называется топологическим отображением, или гомеоморфизмом.

Множества $X_{\text{пр}} \subset R^n$ и $X_{\text{о}} \subset R^n$ должны соответствовать топологическим свойствам компактности и связности.

Множество $X \subset R^n$ называется ограниченным, если оно содержится в некотором достаточно большом шаре: существуют такие точка x_0 и число $r > 0$, что $X \subset N(x_0, r)$. Множество будет компактным, если оно будет ограниченным.

Пространство, не допускающее никакого разбиения, называется связным. Разбиением пространства X называется пара A, B непустых множеств X, таких, что $A \cup B = X$, $A \cap B = \varnothing$ и A и B открыты в X.

Произвольное факторное пространство X будем задавать формализованно, путем задания ограничивающих его точек $1_{\text{о}}, 2_{\text{о}}, ..., N_{\text{о}}$ (рис. 2), где о — индекс образа математического моделирования. Аппроксимируя определенные группы точек зависимостями $X_{j\text{о}} = f_{j\text{о}}(X_{1\text{о}}, ..., X_{k\text{о}})$, получаем ограничительные линии в виде отрезков (или поверхности, если $k \geq 3$) выделенного факторного пространства.

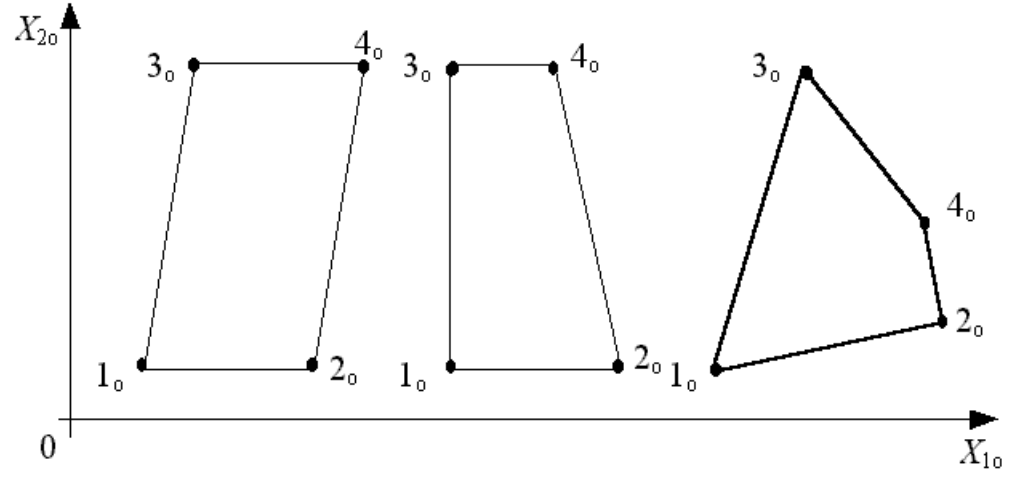


Рис. 2. Возможные нестандартные области факторного пространства

Отображение прообраза планирования эксперимента в образ математического моделирования осуществляется по множествам заданных точек и условиям отображения их в виде эквивалентных наименований точек: $1_{\text{пр}} \to 1_{\text{о}}$; $2_{\text{пр}} \to 2_{\text{о}}$; ...; $N_{\text{пр}} \to N_{\text{о}}$. Указанные точки и их эквивалентность задаются специалистом, решающим задачу, исходя из анализа условий решения поставленной задачи в предметной (содержательной) области.

Алгоритм RASTA 4

Отображение области прообраза планирования эксперимента в область образа математического моделирования при ограничениях факторного пространства линейчатыми поверхностями проводится по алгоритму RASTA 4.

Исходные данные включают следующую информацию. Прямоугольная таблица (матрица) координат точек образа математического моделирования при условии ограничения факторного пространства (размерности $k = 2, \ldots, 5$) линейчатыми поверхностями с количеством точек ограничения в общем случае $N = 2^k$.

Рассматриваются выпуклые области образа математического моделирования. Обозначение точек

$$2^k \,//\, N$$

прообраза планирования эксперимента в кодированных значениях, т.е. «–1» и «+1» , должно соответствовать принятому.

Расположение факторов $X_{i\text{о}}$, их нумерация, а также нумерация значений координат точек $X_{i\text{о}}$ должны соответствовать принятым.

Шаг 1. Вычисляются координаты центра области образа математического моделирования:

$$X_{i\text{оц}} = \sum_{u=1}^{N} X_{i\text{о}} / N.$$

Строится матрица натуральных значений факторов прообраза планирования эксперимента по данным образа математического моделирования.

Натуральные значения уровней факторов, соответствующих значениям «–1», суммируются и делятся на число слагаемых, т.е. определяется среднее значение для нижнего уровня «–1» прообраза планирования эксперимента

$$X_{i\text{пр}}{}^{(-1)} = \left(\sum_{u=1}^{N^{(-1)}} X_{iu} \right) \Big/ N^{(-1)},$$

где u – текущее значение $X_{i\text{пр}}$ на нижнем уровне.

Аналогично определяются натуральные значения уровней факторов для верхнего уровня «+1» прообраза планирования эксперимента

$$X_{i\text{пр}}{}^{(+1)} = \left(\sum_{u=1}^{N^{(+1)}} X_{iu} \right) \Big/ N^{(+1)},$$

где u – текущее значение $X_{i\text{пр}}$ на верхнем уровне.

Шаг 2. Вычисляются значения коэффициентов парных корреляций r_{ij} $(1 \le i < j \le k)$ для факторов $X_{i\text{о}}$ и $X_{i\text{пр}}$ и представляют их в таблицах.

Шаг 3. По таблице натуральных значений прообраза планирования эксперимента строится таблица в кодированных значениях «–1» и «+1» факторов $x_{i\text{пр}}$.

Шаг 4. Используя координаты точек образа математического моделирования и полученную на шаге 3 таблицу в кодированных значениях факторов x_i, вычисляют систему математических моделей, связывающих прообраз планирования эксперимента с образом математического моделирования.

$$\begin{cases} X_{1\text{о}} = f_{1\text{пр}}(x_{1\text{пр}},\ ...,\ x_{k\text{пр}}) \\ ..... \\ X_{k\text{о}} = f_{k\text{пр}}(x_{1\text{пр}},\ ...,\ x_{k\text{пр}}). \end{cases} \qquad (1)$$

Математические модели получают по схеме полного факторного эксперимента типа $2^k \,//\, N$. Вычисляются все коэффициенты математических моделей.

Шаг 5. В целях возможного упрощения полученных на шаге 4 математических моделей (1) можно отбросить достаточно малые по абсолютной величине $|b_i'| \le b_з$ $(1 \le i' \le 2^k$; $b_з$ — заданное критическое значение коэффициента) коэффициенты моделей, пересчитать по моделям с отброшенными коэффициентами значения $X_{i\text{о}}$, полученные результаты принять как заданные и включить их в таблицу координат точек образа математического моделирования вместо первоначально принятых. Перейти на шаг 1.

Шаг 6. Выбрать число уровней по каждому фактору $X_{i\text{пр}}$, план эксперимента — полный факторный эксперимент, многофакторные регулярные планы, планы на основе ЛП_τ равномерно распределенных последовательностей — для получения математической модели (моделей) области математического моделирования. По матрице выбранного плана эксперимента и матрице кодированных значений факторов $x_{i\text{пр}}$ рассчитывается матрица значений уровней факторов $X_{i\text{о}}$.

Шаг 7. По матрице плана эксперимента, полученной на шаге 6, провести эксперимент и получить математическую модель (модели):

$$\hat{y}_j = f_{j\text{пр}}(X_{1\text{пр}}, ..., X_{k\text{пр}}), \qquad (2)$$

используя значения факторов прообраза планирования эксперимента, выбранного на шаге 6.

Шаг 8. Задаемся значениями факторов $X_{1\text{о}}$, ..., $X_{k\text{о}}$ и, используя систему уравнений (1), численными методами находим значения $X_{1\text{пр}}$, ..., $X_{k\text{пр}}$ для области прообраза планирования эксперимента.

Шаг 9. Задаемся значениями факторов $X_{1\text{пр}}$, ..., $X_{k\text{пр}}$ и, используя систему уравнений (1), находим значения $X_{1\text{о}}, ..., X_{k\text{о}}$. По прямоугольной сетке значений факторов прообраза планирования эксперимента строим прямолинейную сетку для значений факторов в области образа математического моделирования.

Шаг 10. По математической модели (моделям) (2) получаем всевозможную информацию о значении (значениях) моделируемых критериев качества системы, процесса, объекта.

Пример отображения областей факторного пространства

Рассмотрим пример отображения областей для $k = 3$. В табл. 1 приведены рабочие матрицы областей образа математического моделирования и прообраза планирования эксперимента, полученного по приведенному алгоритму. На рис. 3 указанные области, построенные по исходным значениям, заданным в табл. 1, показаны в совмещенных натуральных системах координат.

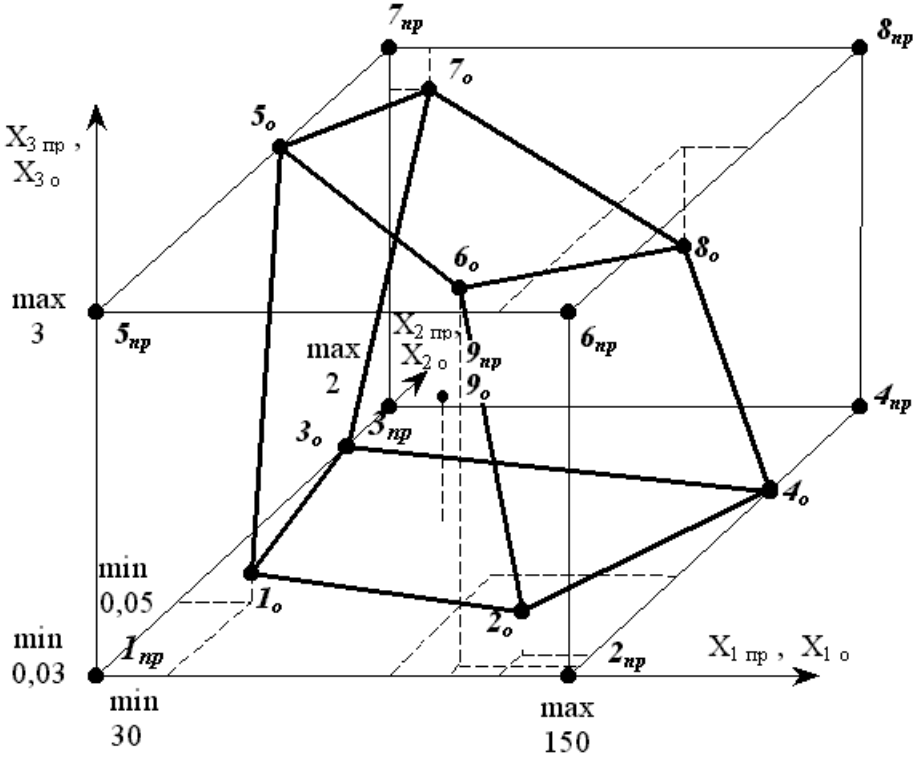


Рис. 3. Области образа и прообраза фазового пространства

Таблица 1

Рабочая матрица плана эксперимента отображения области прообраза планирования эксперимента в область образа математического моделирования

№ точек	Область прообраза планирования эксперимента			№ точек	Область образа математического моделирования		
	$X_{1\,пр}$	$X_{2\,пр}$	$X_{3\,пр}$		$X_{1о}$	$X_{2о}$	$X_{3о}$
1пр	37,5	0,5	0,124	$1_о$	50	0,6	0,206
2пр	140,0	0,5	0,124	$2_о$	140	0,1	0,23
3пр	37,5	1,447	0,124	$3_о$	30	1,6	0,03
4пр	140,0	1,447	0,124	$4_о$	150	1,1	0,03
5пр	37,5	0,5	2,89	$5_о$	30	1,25	3
6пр	140,0	0,5	2,89	6о	130	0,05	2,93
7пр	37,5	1,447	2,89	7о	40	2	2,8
8пр	140,0	1,447	2,89	8о	140	1,088	2,83
9пр	88,75	0,9735	1,507	9о	88,75	0,9735	1,507

Структура математических моделей (1), получаемых на шаге 4, задается схемой полного факторного эксперимента $2^3//8$ и выражением [1, с. 80]:

$(1+x_1)(1+x_2)(1+x_3) = 1+x_1+x_2+x_3+x_1x_2+x_1x_3+x_2x_3+x_1x_2x_3$.

Общий вид системы математических моделей следующий:

$$\begin{cases} X_{1о} = b_{0пр}^{(1)} + b_1^{(1)}x_{1пр} + b_2^{(1)}x_{2пр} + b_3^{(1)}x_{3пр} + b_{12}^{(1)}x_{1пр}x_{2пр} + b_{13}^{(1)}x_{1пр}x_{3пр} + \\ \quad + b_{23}^{(1)}x_{2пр}x_{3пр} + b_{123}^{(1)}x_{1пр}x_{2пр}x_{3пр}, \\ X_{2о} = b_{0пр}^{(2)} + b_1^{(2)}x_{1пр} + b_2^{(2)}x_{2пр} + b_3^{(2)}x_{3пр} + b_{12}^{(2)}x_{1пр}x_{2пр} + b_{13}^{(2)}x_{1пр}x_{3пр} + \\ \quad + b_{23}^{(2)}x_{2пр}x_{3пр} + b_{123}^{(2)}x_{1пр}x_{2пр}x_{3пр}, \\ X_{3о} = b_{0пр}^{(3)} + b_1^{(3)}x_{1пр} + b_2^{(3)}x_{2пр} + b_3^{(3)}x_{3пр} + b_{12}^{(3)}x_{1пр}x_{2пр} + b_{13}^{(3)}x_{1пр}x_{3пр} + \\ \quad + b_{23}^{(3)}x_{2пр}x_{3пр} + b_{123}^{(3)}x_{1пр}x_{2пр}x_{3пр}, \end{cases}$$

где X_{1o}, X_{2o}, X_{3o} – натуральные значения факторов образа математического моделирования;

(1), (2), (3) – индексы коэффициентов математических моделей функций отображения $f_{1пр}$, $f_{2пр}$, $f_{3пр}$ для факторов X_{1o}, X_{2o}, X_{3o} соответственно;

$x_{1пр}, x_{2пр}, x_{3пр}$ – кодированные значения факторов для прообраза планирования эксперимента.

С использованием программного средства «Планирование, регрессия и анализ моделей», разработанного на кафедре технологии машиностроения НТУУ «КПИ», была получена система математических моделей:

$$\begin{cases} X_{1o} = 88{,}75 + 51{,}25x_{1пр} + 1{,}25x_{2пр} - 3{,}75x_{3пр} + 3{,}75x_{1пр}x_{2пр} - 1{,}25x_{1пр}x_{3пр} + \\ \quad + 3{,}75x_{2пр}x_{3пр} - 3{,}75x_{1пр}x_{2пр}x_{3пр}, \\ X_{2o} = 0{,}9735 - 0{,}389x_{1пр} + 0{,}4735x_{2пр} + 0{,}1235x_{3пр} + 0{,}036x_{1пр}x_{2пр} - \\ \quad - 0{,}139x_{1пр}x_{3пр} - 0{,}0265x_{2пр}x_{3пр} + 0{,}036x_{1пр}x_{2пр}x_{3пр}, \\ X_{3o} = 1{,}507 - 0{,}00199999x_{1пр} - 0{,}0845x_{2пр} + 1{,}383x_{3пр} + 0{,}00949999x_{1пр}x_{2пр} - \\ \quad - 0{,}008x_{1пр}x_{3пр} + 0{,}00949998x_{2пр}x_{3пр} + 0{,}0155x_{1пр}x_{2пр}x_{3пр}, \end{cases}$$

где

$x_{1пр} = 0{,}0195122(X_{1пр} - 88{,}75)$;

$x_{2пр} = 2{,}11193(X_{2пр} - 0{,}9735)$;

$x_{3пр} = 0{,}723066(X_{3пр} - 1{,}507)$.

Анализ отображения областей факторного пространства

Средняя абсолютная погрешность отображения прообраза планирования эксперимента в образ математического моделирования для заданных в таблице 1 значений составил по моделям $4{,}39 \cdot 10^{-8} \ldots 8{,}02 \cdot 10^{-10}$, т. е. отображение следует считать точным для прикладных целей.

Контрольный счет различных точек образа математического моделирования факторов X_{1o}, X_{2o}, X_{3o} по полученным моделям полностью подтвердил теоретические положения и полилинейный вид отображения: прямоугольная сетка прообраза планирования эксперимента отображается в

прямолинейную сетку образа математического моделирования при условиях $X_{i\text{пр}} = var$, $X_{j\text{пр}} = const$, $1 \le i \ne j \le k$.

Область прообраза планирования эксперимента представляет многомерный параллелепипед (куб). Поэтому возможно использование планов экспериментов, указанных на шаге 6 алгоритма. Статистические свойства математических моделей (2), полученных по плану эксперимента на шаге 6, будут соответствовать наилучшим из возможных [1, с. 102–103, 127–133]. Коэффициенты полученных математических моделей будут максимально устойчивы: число обусловленности $cond = 1$ или в большинстве случаев $cond < 10$. Для области же образа математического моделирования взаимная (исходная) сопряженность факторов может быть значительной: $0{,}6 < |r(X_{i\text{o}}, X_{j\text{o}})| < 1$.

Выводы

1. Впервые разработаны теоретические основы устойчивого оценивания коэффициентов статистических моделей в условиях исходной сопряженности (мультиколлинеарности) факторов на основе топологической эквивалентности (гомеоморфности) образа математического моделирования и прообраза планирования эксперимента.

2. Приведена схема формирования прообраза планирования эксперимента по исходным линейным условиям ограничения образа математического моделирования с минимально возможной деформацией отображаемого пространства.

3. Предложенный метод позволяет устойчиво оценивать коэффициенты статистических моделей ($1 \le cond \le 10$) для сравнительно произвольных (нестандартных) областей факторного пространства в условиях исходной значительной сопряженности факторов — $0{,}6 < |r(X_{i\text{o}}, X_{j\text{o}})| < 1$.

Источники информации:

1. Радченко С.Г. Математическое моделирование технологических процессов в машиностроении. – К.: ЗАО «Укрспецмонтажпроект», 1998. – 274 с.
2. Половинкин А.И. Основы инженерного творчества: Учеб. пособие для студ. втузов. – М.: Машиностроение, 1988. – 368 с.
3. Сидоров А.И. Основные принципы проектирования и конструирования машин. – М.: Макиз, 1929. – 428 с.
4. Якобс Г.Ю., Якоб Э., Кохан Д. Оптимизация резания. Параметризация способов обработки резанием с использованием технологической оптимизации / Пер. с нем. – М.: Машиностроение, 1981. – 279 с.
5. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов / Пер. с нем. и предисл. В.М. Ивановой. – М.: Финансы и статистика, 1983. – 304 с.
6. Демиденко Е.З. Линейная и нелинейная регрессии. – М.:Финансы и статистика, 1981. – 302 с.

Контактная информация:

Лаборатория экспериментально-статистических методов исследований
http://www.n-t.org/sp/lesmi/

Об авторе:

Радченко Станислав Григорьевич
http://www.n-t.org/ac/rsg/

Впервые опубликовано:

Топологический метод устойчивого оценивания коэффициентов многофакторного уравнения регрессии в условиях мультиколлинеарности факторов / Радченко С.Г. // ISSN 1028-9763 Математичні машини і системи. – 2001. – №1, 2, с 114...121.

Дата публикации:

9 октября 2001 года

Электронная версия: