ВЫБОР ГРУППЫ СЛУЧАЙНЫХ АРГУМЕНТОВ ДЛЯ НАИЛУЧШЕГО ЛИНЕЙНОГО ПРОГНОЗА*

В. А. КАМИНСКИЙ

Московский Инженерно-строительный институт

(Поступило: 15 июня 1981 г.)

Представлено: проф. Рейманн Й., Кафедра Математики, СТР. БТУ

Summary

POINTING OUT THE GROUP OF INDEPENDENT VARIABLES FOR AN OPTIMUM LINEAR PROGNOSTIC — A method has been outlined for solving a classical problem in mathematical statistics: how to select from a set of numerical, independent random variables a group in the closest linear group-wise connection with the function of these variables, in order to predict function x_1. The solution method relies on the characteristics of so-called „best groups". The selection algorithm of the group is obtained by computing the complete covariance matrix, then submatrices are tested all along, with certain coordinate groups fixed.

Резюме

В статье рассмотрен метод решения классической задачи математической статистики о выборе из некоторого множества случайных числовых аргументов $x_2, \ldots, x_n$ группы аргументов $x_{i_1}, \ldots, x_{i_p}$, $i_k \in N = \{1, \ldots, n\}$, $p \leq n$, $x_1(x_2, \ldots, x_n)$, имеющих наиболее тесную групповую связь с функцией зависящей от этих аргументов, с целью использования этой группы аргументов для прогноза функции x_1. Метод решения базируется на вводимой характеристике «наилучшей группы». Алгоритм выбора этой группы сводится к вычислению полной ковариационной матрицы, а затем полного перебора ее подматриц в связи с фиксацией определенных групп координат.

В заметке рассмотрен метод решения классической задачи математической статистики о выборе из некоторого множества случайных числовых аргументов $x_2, \ldots, x_n$ группы аргументов $x_{i_1}, \ldots, x_{i_p}$, $i_k \in N = \{1, \ldots, n\}$, $p \leq n$, имеющих наиболее тесную линейную групповую связь с функцией $x_1(x_2, \ldots, x_n)$, зависящей от этих аргументов, с целью использования этой аргументов для прогноза функции x_1. Метод решения базируется на вводимой характеристике «наилучшей группы».

1. Пусть $x = (x_1, \ldots, x_n) \in E_n$, где E_n-эвклидово пространство, $x_1 = x_1(x_2, \ldots, x_n)$-случайная функция числовых аргументов x_j $j = 2, \ldots, n$, при этом плотность распределения случайного вектора x — $f(x) = ae^{-1/2(x-\bar{x})'A(x-\bar{x})}$ можно считать без ограничения общности нормальной (см. замечание 3), здесь a-нормировочная постоянная, A-эрмитова-вещественная, симметри-

* Статья публикуется в рамках договора о сотрудничестве МИСИ и БТУ.

ческая, положительно определённая матрица, $\bar{x}$-среднее значение вектора x а $(\,.\,)'$-операция транспонирования вектора $(\,.\,)$. Из работы [1] (см. теорему 2.3.1 стр. 29 из [1]) следует, что матрица $\mathbf{A}$ является обратной к матрице ковариаций К, т. е. $\mathbf{A} = \mathrm{K}^{-1}$, а постоянная $a = (2\pi)^{-(1/2)n}|A|^{-1/2}$. Напомним, что матрицей ковариаций вектора x называется матрица $\mathrm{K} = [\kappa_{ij}]i,j = 1, \ldots, n$, элементами которой будут $\kappa_{ij} = (M(x-\bar{x}))(x-\bar{x})$, где $M(y)$-операция нахождения среднего для случайного вектора y. Поверхностями уровня $S(\tilde{c})$ функции $f(x)$ будут поверхности

$$f(x) = 2\pi^{-1/2}\,|A|^{-1/2}e^{-1/2(x-\bar{x})'A(x-\bar{x})} = \tilde{c} \tag{1}$$

являющиеся эллипсоидами. Переходя от (1) к виду

$$(x-\bar{x})'\mathbf{A}\,(x-\bar{x}) = C, \tag{2}$$

где $C = -2\ln(\tilde{c}\cdot 2\pi^{1/2n}|A|^{1/2})$, положим для упрощения $\bar{x} = \Theta$, что соответствует уже центрированной векторной случайной величине x и возьмём число $\tilde{c}$ достаточно малым (например, таким, чтобы $\left|\int\limits_{V} f(x)dx - \int\limits_{E_n} f(x)dx\right| < \varepsilon$, где ε достаточно малое фиксированное число, V — «объём», ограниченный поверхностью $S(\tilde{c})$, а под символом $\int$ понимается «n»-кратный интеграл. Из выбора $\tilde{c}$ вытекает что вероятностная мера «объёма» V отличается от полной меры (равной 1) на величину ε. Таким образом можно считать, что генеральная совокупность значений случайного вектора x находится в «объёме» V, ограниченном поверхностью $S(\tilde{c})$.

Замечание I. Из эрмитовости матрицы $\mathbf{A}$ следует её простота (см., например, [2], стр. 76, теорема 2.9.4), т. е. равенство кратности каждого собственного числа матрицы её геометрической кратности.

Пусть $\lambda_1, \ldots, \lambda_k, \ldots, \lambda_n$-собственные числа, а $e_1, \ldots, e_k, \ldots, e_n$-собственные векторы соответственно матрицы $\mathbf{A}$, притом $\lambda_k \neq 0\ k = 1, \ldots, n$ так как матрица $\mathbf{A}$ является не вырожденной (в противном случае функция $f(x)$ не могла бы представлять из себя плотность распределения некоторого случайного вектора); тогда при $\lambda_i \neq \lambda_j$ имеем ортогональность векторов e_i и e_j. Отсюда и из замечания I вытекает существование базиса пространства E_n, состоящего из попарно ортогональных собвенных векторов матрицы $\mathbf{A}$.

Найдём связь между расположением двух эллипсоидов $S(c_1)$ и $S(c_2)$ в пространстве E_n, для чего сравним уравнения поверхностей $x'\mathbf{A}x = c_1$ и $x'\mathbf{A}x = C_2$. Второе уравнение перепишем в виде $x'\left(\frac{c_1}{c_2}\mathbf{A}\right)x = c_1\,(c_2 \neq 0)$, обозначая $\frac{c_1}{c_2}\mathbf{A} = \tilde{\mathbf{A}}$. Матрицы $\mathbf{A}$ и $\tilde{\mathbf{A}}$ подобны, т. е. $\mathbf{A} - \mathbf{A}$, отсюда следует равенство их спектров, а, значит, и совпадение наборов соответственно собственных векторов. Так как матрица $\mathbf{A}$ является простой, то она подобна диагональной

матрице $\mathbf{\Lambda} = [\lambda_{ij}]i, j = 1, \ldots, n$, где $\lambda_j = \lambda_{jj} \neq \Theta (j = 1, \ldots, n)$-собственные числа матрицы $\mathbf{A}$, а $\lambda_{jj} = 0$, если $i \neq j$ (см., например, [2] стр. 62 теорема 2.4.2). В силу отсутствия в квадратичной форме (2) линейной части в x оба эллипсоида $S(c_1)$ и $S(c_2)$ имеют один и тот же центр в точке $x = \bar{x} = \Theta$. Отсюда фактически вытекает следующая лемма.

Лемма. Эллипсоиды $S(c_1)$ и $S(c_2)$ имеют общий центр $x = \bar{x} = \Theta$ и гомотетичны относительно $\bar{x} = \Theta$ с коэффициентом $\gamma = \sqrt{\frac{c_1}{c_2}}$.

Следствие. Задача нахождения главной системы полуосей эллипсоида $S(c)$, т. е. системы осей, на которых лежат собственные векторы матрицы $\mathbf{A}$, определяющей квадратичную форму (2), не зависит от постоянной C.

2.° Будем считать для достаточно малых c эллипсоид $S(c)$ полем рассеяния случайного вектора x, плотность распределения для которого определяется в (1). Задача построения линейного регресионного уравнения является по существу экстремальной задачей, в которой для множества точек из V с мерой $f(\xi)$ из (1) ограниченного поверхностью $S(c)$ ищется наилучшая гиперплоскость L^* среди различных гиперплоскостей L пространства E_n, т. е. задачей минимизации

$$\sup_{\xi \in V} \inf_{\eta \in L} [\varrho(\xi, \eta) \cdot f(\xi)] \to \min_{L}, \tag{3}$$

где $\xi, \eta \in E_n$, а $f(\xi)$-плотность вероятной меры (1), а расстояние $\varrho(.,.)$ может пониматься в различных смыслах: а) $\varrho(\xi, \eta) = \|\xi - \eta\|_{E_n}$ для $\xi, \eta \in E_n$, что соответствует обычному эвклидовому расстоянию в пространстве E_n; в) $\varrho(\xi, \eta) = |\xi_1 - \eta_1|$ при условии совпадения остальных координат $\xi_j = \eta_j$ $j = 2, \ldots, n$ у пары точек ξ и η из E_n, т. е. соотвествует расстоянию между точками ξ и η по функционалу (в случае линейной аппроксимации эллипсоида этот функционал линеен) с) методу наименьших крадратов соответствует задача 3с:

$$\int_V \varrho^2(\xi, \eta)\, f(\xi) d\xi \to \min_{L}, \tag{3c}$$

где расстояние $\varrho(\xi, \eta) = |\xi_1 - \eta_1|$ берётся только по парам точек $\xi \in V$ $\eta \in L$, для которых $\xi_j = \eta_j$ $j = 2, \ldots, n$ (расстояние по координате x_1). Отметим, что задача 3а (или 3в, 3с) является конечномерной, так как гиперплоскость L в пространтсве E_n определяется "$n + 1$" параметром — "n" координатами нормального вектора и постоянной.

В силу симметричности функции $f(x)$ относительно $x = \Theta$ гиперплоскость L_a^* (L_b^* или, соответственно, L_c^*) является наилучшей в смысле задачи 3а (3в или, соответственно, 3с) и проходит через точку $x = \Theta$. В общем случае

гиперплоскости L_a^*, L_b^* и L_c^*, конечно, не совпадают в силу разной топологии измеряемого расстояния $\varrho(\xi, \eta)$ для случаев "3а", "3в", "3с".

Остановимся на случае расстояния, вводимого в задаче 3а, т. е. на наиболее простом и удобном случае. С силу эрмитовости матрицы **A** существует унитарное линейное преобразование T базиса в пространстве E_n $x = Ty$ такое, что $T^{-1}AT = \Lambda$, где Λ-диагональная матрица подобная матрице **A**; отсюда следует, что в новом базисе квадратичная форма (2) имеет канонический вид

$$\varphi(x) = \tilde{\varphi}(y) = y'(T^{-1}AT)y = \sum_{j=1}^{n} \lambda_j y_j^2. \tag{4}$$

Без ограничения общности положим

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq \ldots \geq \lambda_n, \tag{5}$$

обозначив соответственно набор собственных векторов $e_1, \ldots, e_k, \ldots, e_n$.

Преобразовав уравнение эллипсоида $S(c)$ в форму

$$\sum_{j=1}^{n} y_j^2 \left(\sqrt{\frac{\bar{c}}{\lambda_j}}\right)^2 = 1, \tag{6}$$

где $\bar{c} \neq 0$, легко найти решение задачи 3а в виде гиперплоскости $L_a^* = \{y \in E_n : y_1 = 0\}$ нормальной собственному вектору, соответствующему наибольшему собственному числу $\lambda_1 = \max\limits_{i \leq j \leq n} (\lambda_j)$. Возвращаясь к исходному базису пространства E_n и учитывая инвариантность решения задачи 3а относительно выбора базиса можно получить следующий критерий элемента L_v^*.

Теорема. Для того, чтобы гиперплоскость L_a^* была решением задачи 3а необходимо и достаточно, чтобы она проходила через точку $x = \bar{x} = \Theta$ и была нормальна к собственному вектору e_1 матрицы **A**, соответствующему наибольшему собственному числу $\lambda_1 = \max\limits_{1 \leq j \leq n} (\lambda_j)$.

Следствие 1. Гиперплоскость L_a^* является единственной, если выполнено условие $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$.

Положим дисперсии всех координат случайного вектора x равными 1, т. е. $Dx_1 = \ldots = Dx_k = \ldots = Dx_n = 1$ и обозначим поперечник по Колмогорову множества $S(c)$ в задаче 3а через $\pi = \min\limits_{L} \sup\limits_{\xi \ni V} \inf\limits_{\eta \ni L} [\varrho(\xi, \eta)\, f(\xi)]$, тогда максимальная ошибка аппроксимации эллипсоида $S(c)$ гиперплоскостью L_a^* не будет превышат π. Пусть длина собственного вектора $e_1 = (e_{11}, \ldots, e_{1n})$ будет равна единице, т. е. $\|e_1\| = 1$, а функция $x_1 = x_1(x_2, \ldots, x_n)$ не тривиальна, т. е. хотя бы одна из координат $e_{ij} \neq 0$. Тогда имеет место следствие.

Следствие 2. При использовании для линейного прогноза функции x_1 по аргументам $x_2, \ldots, x_n$ гиперплоскость L_a^* максимальная ошибка прогноза не превышает величины π/e_{11}.

3.° Рассмотрим класс эллипсоидов $\{P(c)\}$ типа (6) в предположении (5) при $c = 1$ имеющих один и тот же поперечник $\pi_1 = \frac{1}{\sqrt{\lambda_1}} L_a^*$ в данном случае будет иметь вид $y_1 = 0$, что очевидно).

Наряду с абсолютной максимальной ошибкой аппроксимации — поперечником π_2 — можно ввести и относительную ошибку аппроксимации r, связанную с поперечником Π_2 проекции эллипсоида $P(c)$ на гиперплоскость $P_1 = 0$, следующим образом $r = \frac{\pi_1}{\pi_2}$, где $\pi_2 = \frac{1}{\sqrt{\lambda_2}}$. В случае $\lambda_1 \gg \lambda_2$ имеем сильно «сжатый» по оси y_1 эллипсоид $P(c)$, при этом $\pi_1 \ll \Pi_2$ и, следовательно, гиперплоскость $y_1 = 0$ достаточно хорошо аппроксимирует множество $P(c)$ ($r = \sqrt{\frac{\lambda_2}{\lambda_1}}$); если же $\lambda_1 = \lambda_2 \ldots = \lambda_\nu$, то проекцией эллипсоида $P(c)$ на подпространство L размерности ν, натянутое на систему собственных векторов $e_1, \ldots, e_\nu$ является шаром и поэтому существует пучок гиперплоскостей $\{L_a^*\}$ ранга $\nu - 1$ одинаково аппроксимирующих эллипсоид $P(c)$ с абсолютной ошибкойв метрике задачи За не превосходящей числа $\frac{1}{\sqrt{\lambda_1}}$; относительная ошибка в этом случае $r = 1$, т. е. принимает максимальное значение.

Из последних рассуждений ясна невозможность построения линейного регрессионного уравнения $x_1 = x_1(x_2, \ldots, x_n)$.

Малые коэффициенты парной корреляции ещё не означают слабой групповой связи случайных координат $x_2, \ldots, x_0$ с функцией x_1. Это показывает следующий пример.

Пример. Рассмотрим поле рассеяния в виде 3-х мерного эллипсоида для случайного вектора $x = (x_1, x_2, x_3) \in E_3$ распределённого нормально. Пусть в базисе (y_1, y_2, y_3) эллипсоид P представлен каноническим уравнением $y'\mathbf{Л}y = 100\, y_1^2 + y_2^2 + y_3^2 = 1$, где $\mathbf{Л}$-диагональная матрица, а в базисе (x_1, x_2, x_3) соответственно $x'\mathbf{A}x$, где $\mathbf{A} = \mathbf{TЛT}^{-1}$, при этом матрица перехода от базиса (y_1, y_2, y_3) к базису (x_1, x_2, x_3) имеет вид

$$\mathbf{T} = \begin{bmatrix} \frac{2}{\sqrt{12}} & 0 & -\frac{2}{\sqrt{6}} \\ \frac{2}{\sqrt{12}} & \frac{3}{\sqrt{18}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{12}} & -\frac{3}{\sqrt{18}} & \frac{1}{\sqrt{6}} \end{bmatrix}, \text{ обратная к ней } \mathbf{T}^{-1} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

а $\mathbf{Л} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Легко видеть, что преобразования T и $\mathbf{T}^{-1}$ изомеричны и сохраняют ортонормированный базис. Поэтому для эллипсоида P гиперплоскость L_a^* в базисе (y_1, y_2, y_3) имеет вид $y_1 = 0$, а в базисе (x_1, x_2, x_3) — $x_1 + x_2 + x_3 = 0$; поперечникэллипсоида P: $\pi = \frac{1}{10}$ и относительная ошибка аппроксимации $\varepsilon = \sqrt{\frac{\lambda_2}{\lambda_1}} = \frac{1}{10}$ что показывает довольно тесную групповую связь между координатами x_1, x_2, x_0, при этом величины π и г из-за изометричности преобразования T не зависят от выбора базиса. С другой стороны эллипсоид P проектируется на плоскости $x_1 0 x_2$, $x_1 0 x_3$, $x_2 0 x_3$ в виде эллипсов с малым эксцентриситетом, что свидетельствует о малых коэффициентах парных связей между x_1 и x_2, x_1 и x_3, x_2 и x_3.

Введём понятие «наилучшей группы» аргументов (которые, возможно, являются зависимыми случайными переменными) — $(x_{i_1}, \ldots, x_{i_k})^*$. Будем называть группу $(x_{i_1}, \ldots, x_{i_k})$ наилучшей для прогноза x_1 из всех групп $(x_{j_1}, \ldots, x_{j_e})$, где $i_1, \ldots, i_k, j_1, \ldots, j_e \in N\{1, \ldots, n\}$ и $K, l \leq n$, при этом $i_1 = j_1 = 1$ т. е. в каждой из рассматриваемых групп присутствует координата x_1-изучаемая функция от аргументов $x_2, \ldots, x_n$, если для этой группы

$$\varepsilon(i_1, \ldots, i_k)^* = \min_{j_1, \ldots j_e \in N} \varepsilon(j_1, \ldots, j_e), \tag{7}$$

где $i_1 = j_2 = 1$, а $i_2, \ldots, i_k, j_k, \ldots, j_k \in N$ при этом $r \leq K, l \leq n$. Обозначим через $\hat{R}$ ковариационную матрицу для группы координат $x_{j_1}, \ldots, x_{j_e}$, являющуюся по сути дела подмножеством полной матрицы ковариаций $\mathbf{K}$.

Таким образом в случае, когда дисперсии всех координат x_j центрированной векторной случайной величины $x = (x_1, \ldots, x_n)$ равны 1, т. е.

$$Dx_j = 1 \quad j = 1, \ldots, n \tag{8}$$

очевидно следующее свойство наилучшей группы аргументов:

Предложение. В случае линейной аппроксимации случайной функции $x_1 = x_1(x_2, \ldots, x_n)$, распределение которой удовлетворяет (1), для наилучшего линейного прогноза x_1 по значениям аргументов $x_2, \ldots, x_n$ необходимо выбрать такую группу аргументов $x_{i_1}, \ldots, x_{i_k}$, для которой спектр матрицы $\mathbf{A} = \hat{\mathbf{R}}^{-1}$ таков, что выполнено условие (7).

Замечание 2. В случае достаточно высоких групповых связей, т. е. при малых значениях поперечника π полного эллипсоида рассеяния (т. е. такого

эллипсоида, для которого имеем $\int_V f(x)dx = 0{,}96$ и условии (8) гиперплоскость L_a^* будет близка к гиперплоскости L_c^*, поэтому её можно использовать для прогноза вместо решения задачи методом наименьших квадратов, которое может быть затруднено из-за плохой обусловленности матрицы системы нормальных уравнений.

Замечание 3. Все рассуждения разделов 1°, 2°, 3° относились к случаю нормального распределения вектора x, но по существу использовался только тот факт, что поверхности уровня плотности распределения заданы квадратичной формой (2). Поэтому выводы, полученные в разделах 1°, 2°, 3° легко распространить и на другие типы распределений, имеющих в качестве линий уровня плотности квадратичную форму.

Алгоритм выбора наилучшей группы сводится к вычислению полной ковариационной матрицы $\mathbf{K}$, а затем полного перебора её подматриц в связи с фиксацией определённых групп координат.

Литература

1. Anderson, T. W.: An Introduction to Multivariate Statistical Analysis. New York, 1958
2. Lancaster, P.: Theory of Matrices. Academic Press, New York—London, 1969

КАМИНСКИЙ, В. А., доцент,
МИСИ, Москва, Шлюзовая наб. 8, СССР