I-OPTIMAL DESIGN FOR AFFINE MULTIPLE REGRESSION

Mukha V. S.
Department of Information Technologies of Automated Systems,
Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
E-mail: mukha@bsuir.by

This article is devoted to the development of the experiment design theory. The new so called I-optimality criterion is substantiated and I-optimal design is developed which do not inferior to the known designs in the sense of known-optimality and has no restrictions on the number of the observations.

Introduction

The optimal designs of experiment use widely for identification of the subjected control objects in many technical and technological systems and processes [1]. This implies the further development of the theory in this area. We will consider the regression experiment for the affine multiple regression.

It is now traditionally accepted to represent the mathematical model of the observations for the affine multiple regression function in the following form [2]:

$$y_{o,i} = \theta_* + x_i^T\Theta + \epsilon_i,\ \ i = 1, 2, \dots, n,\ \ m+1 \le n,$$

or

$$y_{o,i} = H_i^T\Theta_1 + \epsilon_i,\ \ i = 1, 2, \dots, n,\ \ m+1 \le n, \quad (1)$$

where $x_i^T = (x_{i,1}, x_{i,2}, \dots, x_{i,m}) \in E^m$, $i = 1, 2, \dots, n,$, are input vector variables, so called factors, $\Theta^T = (\theta_1, \theta_2, \dots, \theta_m)$, $\Theta_1^T = [\theta_*, \Theta^T]$ is the vector-rows of the unknown coefficients, $H_i^T = [1, x_i^T] \in E^{m+1}$ is the vectors of the basis functions, ϵ_i are uncorrelated random errors with zero mean value and variation σ^2 , $y_{o,i}$ is observed values of the output variable, n is the size of the sample. The factors $x_{i,j}$ are subjected to the constraints in the form $-1 \le x_{i,j} \le 1$. Such factors are called normalized. The expression (1) can be written also in the vector-matrix form as

$$Y = F\Theta_1 + E,$$

where

$$F = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,m} \\ 1 & x_{2,1} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,m} \end{pmatrix}, Y = \begin{pmatrix} y_{o,1} \\ y_{o,2} \\ \vdots \\ y_{o,n} \end{pmatrix},$$

$$E = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \Theta_1 = \begin{pmatrix} \theta_* \\ \theta_1 \\ \vdots \\ \theta_m \end{pmatrix},$$

The vector $\hat{\Theta}_1$ of the LSE-estimations of the vector Θ_1 is defined by the expression

$$\hat{\Theta}_1 = (F^TF)^{-1}F^TY, \quad (2)$$

and the covariance matrix of the estimation (2) is defined by the following expression [2]:

$$V_{\hat{\Theta}_1} = \sigma^2(F^TF)^{-1}.$$

The set of the points $x_i^T = (x_{i,1}, x_{i,2}, ..., x_{i,m}) \in E^m$, $i = 1, 2, ..., n$, is called a design of the experiment and is denoted $\xi_m(n)$. The design $\xi_m(n)$ is represented in the form of the following $(n \times m)$-matrix:

$$\xi_m(n) = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix}. \quad (3)$$

The number of the rows of the matrix $\xi_m(n)$ (3) not less than the number of the columns. The matrix $A = F^TF$ is called the information matrix of the design $\xi_m(n)$ [2]. The inverse matrix $A^{-1} = (F^TF)^{-1}$ is called the covariance matrix of the design $\xi_m(n)$. The task of the optimal experiment designing consists of the finding the design $\xi_m(n)$ satisfying the given constraints and optimal in some sense.

There are a number of optimality criteria for the designs [2]. A wide set of the optimal designs there exists, but they have the constraints on the numbers of the experiments (points). It is desirable that the researcher has the complete freedom in choosing the number of the experiments.

We introduce in this work the new optimality criterion which allows us to develop the designs for arbitrary number of observations.

I. The model of the observations and the I-optimal design

We will consider the observations model for the affine multiple regression function in opposition to the model (1) in the following form:

$$y_{o,i} = \theta_0 + (x_i - \overline{x})^T\Theta + \epsilon_i,\ \ i = 1, 2, ..., n,\ \ m+1 \le n, \quad (4)$$

where $x_i^T = (x_{i,1}, x_{i,2}, ..., x_{i,m}) \in E^m$, $y_{o,i}, \theta_0, \epsilon_i \in R$, $\Theta^T = (\theta_1, \theta_2, ..., \theta_m) \in E^m$,

$$\overline{x}^T = \frac{1}{n}\sum_{i=1}^{n} x_i^T = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_m),$$

θ_0, Θ are unknown coefficients of the regression function, ϵ_1 is the uncorrelated random error with zero mean value and variation σ^2 , x_i is the point of the experiment, n is the size of the sample (the number of the experiments, the number of the design points, the number of runs), $y_{o,i}$ is the observed value of the output variable. The LSE-estimations $\hat{\theta}_0, \hat{\Theta}$ of the coefficients θ_0, Θ are defined by the expressions

$$\hat{\theta}_0 = (\sum_{i=1}^{n} y_{o,i}^T)/n = \overline{y}, \quad \hat{\Theta} = \mu_{xx}^{-1}\mu_{xy}, \qquad (5)$$

where

$$\mu_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^T,$$

$$\mu_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})y_{o,i}.$$

If we denote $\dot{x}_i = x_i - \overline{x}$ the centered vector x_i , $\dot{y}_i = y_{o,i} - \overline{y}$ the centered observation $y_{o,i}$,

$$\dot{\xi}_m(n) = \\ = \begin{pmatrix} x_{1,1} - \overline{x}_1 & x_{1,2} - \overline{x}_2 & \cdots & x_{1,m} - \overline{x}_m \\ x_{2,1} - \overline{x}_1 & x_{2,2} - \overline{x}_2 & \cdots & x_{2,m} - \overline{x}_m \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} - \overline{x}_1 & x_{n,2} - \overline{x}_2 & \cdots & x_{n,m} - \overline{x}_m \end{pmatrix} \qquad (6)$$

the centered design $\dot{\xi}_m(n)$, then

$$\mu_{xx} = \frac{1}{n}\dot{\xi}_m^T(n)\dot{\xi}_m(n), \quad \mu_{xy} = \frac{1}{n}\dot{\xi}_m^T(n)\dot{y}, \qquad (7)$$

and the expression (5) for the estimation $\hat{\Theta}$ takes the following form:

$$\hat{\Theta} = (\dot{\xi}_m^T(n)\dot{\xi}_m(n))^{-1}\dot{\xi}_m^T(n)\dot{y}$$

by analogy with the expression (2).

We will call the matrix (5) (or (7)) the information matrix of the centered design (6) (or of the design $\xi_m(n)$ (3)). The variation $V_{\hat{\theta}_o}$ of the estimation θ_0 and the covariance matrix $V_{\hat{\Theta}}$ of the estimation $\hat{\Theta}$ are defined by the following expressions [3]:

$$V_{\hat{\theta}_o} = \frac{\sigma^2}{n}, \quad V_{\hat{\Theta}} = \frac{\sigma^2}{n}\mu_{xx}^{-1}.$$

The estimations $\hat{\theta}_0$ and $\hat{\Theta}$ are uncorrelated [3], in opposite to the coefficients of the traditional model (1), what is the advantage of the model (4) .

We will call the design $\xi_m(n)$ (**??**) I-optimal if it provides the maximal value of the trace of the information matrix μ_{xx}. It is can to show that the trace of the information matrix μ_{xx} is represented in the following form:

$$tr(\mu_{xx}) = \frac{1}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}||x_i - x_j||^2. \qquad (8)$$

The problem of building of I-optimal design is formulated as follow: find the design $\xi_m(n)$ (3), providing a maximum for the function $tr(\mu_{xx})$ (8), or for the function $I = n^2 tr(\mu_{xx})$:

$$I = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}||x_i - x_j||^2. \qquad (9)$$

provided the points of the design are pairwise distinct, do not lie on the same linear variety in E^m, satisfy the constrains

$$a \le x_i \le b, \quad i = 1, 2, ..., n,$$

$a = (a_1, a_2, ..., a_m) \in E^m$, $b = (b_1, b_2, ..., b_m) \in E^m$, and the number of the points $n \ge m + 1$.

As a result, we propose the following formalized algorithm for synthesis of the -optimal design.

Let k be the remainder of dividing n by 4. If $n < 4$ $(m < n - 1 = 3)$, then the saturated nonsingular design $S_m(m + 1)$ (10) below augmented with $n - m - 1$ points in accordance with criterion I-optimality (9) is used:

$$S_m(m{+}1) = \begin{pmatrix} b_1 & b_2 & \cdots & b_m \\ b_1 & a_2 & \cdots & a_m \\ \vdots & \vdots & \vdots & \vdots \\ a_1 & a_2 & \cdots & b_m \end{pmatrix} = (s_{i,j}), \ (10)$$

$$s_{1,j} = b_j, \quad i = 1, \quad j = 1, 2, ..., m,$$

$$s_{i,j} = \begin{pmatrix} b_j, & j = i - 1 \\ a_j, & j \neq i - 1 \end{pmatrix}, i = 2, 3, ..., m + 1.$$

If $n \ge 4$, then provided $n - k \ge m + 1$ and $n - k = o \bmod 4$ the Plackett–Burman design $PB_m(n - k)$ is used [4], otherwise the design $S_m(m+1)$ (10) augmented with $n - m - 1$ points in accordance with criterion I -optimality (9) is used.

II. References

1. Rachel T. Johnson, Douglas C. Montgomery, Bradley A. Jones. An Expository Paper on Optimal Design. Quality Engineering, 23:287–301, 2011. DOI: 10.1080/08982112.2011.576203.
2. Ermakov S. M., Zhigljavsky A. A. Mathenatical Theory of the Optimal Design. Moscow, Science, 1987, 320 p. (in Russian).
3. Mukha V. S. Multidimensional-matrix Linear Regression Analysis: Distributions and Properties of the Parameters Estimations. Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series. 2014, no. 2, pp. 71-81 (in Russian).
4. Plackett R. L., Burman J.P. The Design of Optimum Multifactor Experiments. Biometrika, 1946, 33, № 4, pp. 305-325. DOI: 10.2307/2332195.