STATISTICAL EXPERIMENTS AND OPTIMAL DESIGN

Andrej Pázman
Comenius University, Bratislava, Slovakia

Keywords: Experiment design, linear statistical model, nonlinear regression, least squares, information matrix, entropy, optimality criterion, concentration ellipsoid, distribution of estimators.

Contents

Summary

One part of the cooperation of the statistician with the experimenter is the design of experiments. This should be done in the framework of a statistical model describing the statistically important features of the experiment. This paper starts with a brief overview of the linear regression model, which is universal. Here, the amount of information which can be obtained from an experiment is expressed by the Fisher information matrix, or geometrically by the ellipsoid of concentration. This leads straightforwardly to the basic optimality criteria. In Section 4 Kiefer's conception of design measures is presented briefly, as well as the main consequences of using it: the necessary and sufficient conditions for optimality of a design, the rule for checking approximate optimality, the main idea how to construct iterative methods of computation. In Section

5 the design for the optimum allocation of the times of observations of a random process is considered. In general, the results from a linear model can be extended to nonlinear models when the number of observations is large. However, then the information matrix depends on the unknown parameters, and this is discussed in Section 6. At the end one discuss briefly how to express the amount of information obtained from a nonlinear experiment when the number of observations is small. For that purpose the probability distribution of the estimators and the entropy measure of information are considered.

1. Introduction

A physical (chemical, biological, technological) experiment is a series of activities aimed at obtaining information from observed (measured) data. In a broad sense the design of such an experiment is any well-thought organization of these activities.

Modern experiments cannot avoid the use of statistics while dealing with data. However, each statistical procedure is adequate within a certain model of the experiment under consideration. Although some sciences developed highly sophisticated models (e.g. physics) for their experiments, statistics is reducing them to a restricted class of statistical models, which are relatively simple and universal. It is that universality of statistical modeling, which allows statisticians to deal with very different experimental situations from a common point of view.

The basic model assumption of statistics is that a large amount of nuisance effects, which influence the outcome of the experiment (e.g. the errors of observation, the non-stability of experimental conditions, influence of different side-effects), can be treated as random, and in a well designed and well treated experiment the influence of such effects can be substantially suppressed by probabilistic methods. This is why statistics uses probability theory, and the present success of statistics in applications justifies this approach.

However, the assumption of a probabilistic description of data is not a sufficient model assumption, and further assumptions must be made. This should be done leaving freedom to adapt the model to the data and/or to any prior knowledge about the object investigated in the experiment.

Thus modeling and data processing are often inseparable, dual parts of a statistical approach to the experiment. Without making some prior assumptions on the model, one can not understand the result of the statistical procedures, and conversely, only the results of statistical investigations often justify the use of a statistical model.

But the described approach of the statistician to an experiment would be very passive, if he/she did not consider the possibility also to enter into the preparation of the experiment by giving his/her advice concerning the design of the experiment. A design should be presented before performing the experiment. On the other hand, a good design should be adapted to the choice of the model, and to the choice of statistical procedures for processing the data. Again, strictly speaking, design, modeling, and data processing are related parts of the cooperation of the statistician with the experimenter.

To be able to separate the design activities from modeling and data processing, it is usual to investigate design procedures mainly in the simplest, very universal models, (linear models) and for the simplest and also very universal statistical procedures, like least squares. Progress in other, more sophisticated models is often related to some universal approximations, like those obtained from the asymptotic properties of estimators.

A few words should be said here about the history of experiment design. The concept of a well prepared experiment goes back as to the 14th century however, as a pioneer of what could be called a statistical approach to designing an experiment is often cited A. Young (1741-1820), a specialist in agricultural experiments. It is also in agriculture that "the Newton of contemporary statistics", R.A. Fisher (1890-1962) formulated his conception of design of statistical experiments simultaneously with building the analysis of variance. His approach forms the base of what is called today the combinatorial design of experiments. It requires advanced combinatorial methods, tools like projective geometries, etc. Consequently, this research is followed today not only by statisticians, but also by some "pure mathematicians" interested in combinatorial methods.

In parallel, another approach has been developed, which is called the optimum experiment design. Its origin is in problems of getting a good "response" to a given "stress" (say in biological experiments) or to estimate an important physical parameter with the highest precision possible. It is also related to the traditional mathematical problem of finding the optimal allocation of nodes in polynomial interpolation and extrapolation. The start of a systematic research in optimum design is related to the name of J. Kiefer (1924-1981), although he had several predecessors, among which the most famous is G. Elfving (1908-1984).

The work of J. Kiefer is a sophisticated mathematical work, but his important idea, which opened the later development, is in fact a simple reformulation and approximation of the concept of experiment design (See Section 4.1). This changed the design problem to a specific problem of convex analysis, although it seems that only later the convex aspects of the whole theory have been emphasized and fully exploited. Today there are several books on the subject.

On the present research level, besides bringing this theory to a higher mathematical perfection, interest is also in problems which go outside the scope of the convex theory: design for the observation of random processes, design on non-asymptotic criteria in nonlinear statistical models, design for nonparametric estimation, but still there are new and sophisticated results on classical problems, as in the polynomial inter- and extrapolation.

2. Linear Models

Statistical models are often used to model the influence of some (input) factors on the output of an experiment. In many cases such models are linear in some unknown parameters, and are called linear regression models.

Probably the most widely used are the ANOVA models and the factorial models, which are presented here as examples.

In ANOVA (= analysis of variance) models one describes the influence of, say two treatments A, B on the outcome of the experiment (say the yield of some plant). In each trial one can choose levels of A and of B which are denoted symbolically by *i*, *j*. The outcome of the experiment y_{ij} is modeled by

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad (1)$$

where α_i and β_j are the contribution of the treatments A and B respectively to the outcome y_{ij}, μ is some "mean effect", and ε_{ij} is the component of the outcome, which is supposed to be random (in the term ε_{ij} are included many secondary effects which the experimenter did not take into account). Further parameters may appear in the model when it is supposed that the treatments A, B may interact, or when there are further treatments. Also some of the parameters may have different interpretations (e.g. block parameters), but the model remains as the simplest linear model, very universal, as proved in practice.

Other universal linear models are factorial models. They are used when *k* qualitative factors $x_1,\ldots, x_k$ influence the outcome of the experiment modeled by some function $h(x_1,\ldots, x_k)$, where however, $h(\cdot)$ is either unknown or is too complicated to be used. Therefore one can proceed as follows: one finds some reference level $x^o = (x_1{}^o,\ldots, x_k{}^o)$ of the vector of factors $x = (x_1,\ldots, x_k)$, and by the Taylor formula applied on $h(x)$ one obtains an approximate model. The "order" of the factorial model depends on the number of terms used in the Taylor formula. For example, the second order factorial model has the form

$$\theta_o + \sum_{i=1}^{k} \theta_i (x_i - x_i^o) + \sum_{i,\, j=1}^{k} \theta_{ij} (x_i - x_i^o)(x_j - x_j^o)$$

where $\theta_o = h(x^o),\ \ \theta_i = \partial h(x)/\partial x_i\big|_{x=x^o}$, $\theta_{ij} = \partial^2 h(x)\big/\partial x_i \partial x_j\big|_{x=x^o}$ are unknown parameters, which are to be estimated from the results of observations. Evidently, the factorial model remains linear in the unknown parameters also if higher order terms are considered.

In many physical or chemical experiments the outcomes of the observation depend on some important physical parameters, denoted here by $\theta_1,\ldots, \theta_m$, according to some "physical law". So we know that the observed value y_i in the *i*th trial is

$$y_i = \eta_i(\theta_1,\ldots,\theta_m) + \varepsilon_i$$

with known functions $\eta_i(\cdot)$ of the (unknown) vector of parameters $\theta = (\theta_1,\ldots, \theta_m)^T$. The functions $\eta_i(\cdot)$ may be linear, but also nonlinear. In the latter case, a procedure often used (e.g. in geodesy) is to specify a value of θ, say θ^o, and to use the approximate linear model obtained again by the Taylor formula

$$y_i = \eta_i(\boldsymbol{\theta}^o) + \sum_k \left.\frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^o} (\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^o) + \varepsilon_i.$$

Here the derivatives are known, and the parameters θ_k, which are to be estimated, enter linearly.

In vector notation the linear model is described by

$$\mathbf{y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$
$$E(\boldsymbol{\varepsilon}) = \mathbf{0}. \qquad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{W} \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the vector of observed values, **F** is a given $N \times m$ matrix, $\boldsymbol{\varepsilon}$ is the error vector which is random, the variance parameter σ and the vector $\boldsymbol{\theta} \in \mathbb{R}^m$ are unknown, and the matrix **W** is supposed to be known. Usually **W** = **I**, which means that the observations are uncorrelated and with constant variance, but sometimes, e.g. when a sample from a random process with a given covariance function is observed, one has **W** ≠ **I**. Notice that in general we denote by *E* the mean, and by Var the variance, or the variance (covariance) matrix.

3. How to Measure the Information Obtained from an Experiment Modeled Linearly?

3.1. Information Matrices

Statistics is an information science, so the problem of evaluating the amount of information obtained from an experiment is a crucial problem. In linear models this is done straightforwardly as follows.

We consider the estimator of the parameter vector $\boldsymbol{\theta}$ in (2), which is unbiased and which has the minimum variance matrix (i.e. which is the most precise in a wide sense). It is obtained by minimizing a "weighted sum of squares"

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^m} [\mathbf{y} - \mathbf{F}\boldsymbol{\theta}]^T \mathbf{W}^{-1} [\mathbf{y} - \mathbf{F}\boldsymbol{\theta}]$$

which in case of a non-singular matrix

$$\mathbf{M} = \mathbf{F}^T \mathbf{W}^{-1} \mathbf{F}$$

is equal to

$$\hat{\boldsymbol{\theta}} = \mathbf{M}^{-1} \mathbf{F}^T \mathbf{W}^{-1} \mathbf{y}$$

($\hat{\boldsymbol{\theta}}$ is called the Gauss-Markov estimator or, in case of normal errors $\boldsymbol{\varepsilon}$, the maximum likelihood estimator). The experiment is considered to give much information about $\boldsymbol{\theta}$ if the variance matrix of $\hat{\boldsymbol{\theta}}$

$$\text{Var}\,(\hat{\boldsymbol{\theta}}) = \sigma^2 \mathbf{M}^{-1}$$

is "small" , that means if the matrix $\mathbf{M}$ is "large" in a certain sense. So, in a linear model (2) measures of information should be based on the matrix $\mathbf{M}$. Even if $\mathbf{M}$ is not full rank, and consequently the estimator $\hat{\boldsymbol{\theta}}$ is not specified uniquely, one can still estimate some parameter functions of the form $\mathbf{h}^T\boldsymbol{\theta}$, and, in that case $\mathbf{h}^T\hat{\boldsymbol{\theta}}$ is unique, and

$$\text{Var}(\mathbf{h}^T\hat{\boldsymbol{\theta}}) = \sigma^2 \mathbf{h}^T \mathbf{M}^{-}\mathbf{h}.$$

Here $\mathbf{M}^{-}$ is any g-inverse of $\mathbf{M}$, i.e. any solution of the equation
$\mathbf{M}\mathbf{M}^{-}\mathbf{M} = \mathbf{M}$.

Again, a large $\mathbf{M}$ means more information obtained from the experiment. Further arguments in favor of the fact that the matrix $\mathbf{M}$ expresses the information obtained from the experiment are:

1. the matrix $\mathbf{M}$ is known generally in statistics as the Fisher information matrix (for $\sigma = 1$)
2. $\sigma^2\mathbf{h}^T\mathbf{M}^{-}\mathbf{h}$ is the lower bound of the variance of any unbiased (linear or nonlinear) estimator of the parametric function $\mathbf{h}^T\boldsymbol{\theta}$ (the Rao-Cramer bound).
3. asymptotically, i.e. for a large amount N of observations, $\sigma^2\mathbf{M}^{-1}$ specifies the approximate variance matrix for many other estimators of $\boldsymbol{\theta}$ in model (2) (the nonparametric estimators, the Bayesian estimator, the ridge estimator).

Notice however, that one can loose the justification for using the matrix $\mathbf{M}$ to measure the information if one considers so called robust estimators of $\boldsymbol{\theta}$. In fact, robust estimators are used only if it is supposed that the model assumptions in model (2) are violated (by the presence of so called "outliers").

3.2. Information Matrices for Parameter Subsets

The matrix $\mathbf{M}$ is the information matrix for all parameters cc. If we are interested only in a parameter subset, say $\theta_1,...,\theta_s$ with $s < m$, the corresponding information matrix is expressed by

$$\mathbf{M}^{(s)} = \mathbf{M}_{11} - \mathbf{M}_{12}(\mathbf{M}_{22})^{-}\mathbf{M}_{21} \qquad (3)$$

where the used g-inverse is arbitrary, and where

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$$

is a block decomposition of **M**, with $\mathbf{M}_{11}$, being a $s \times s$ matrix. The fact that the matrix $\mathbf{M}^{(s)}$ is the information matrix is justified by the equality

$$Var\ \left[(\hat{\theta}_1,\ldots,\hat{\theta}_s)^T\right] = \sigma^2(\mathbf{M}^{(s)})^{-1}.$$

3.3. Geometrical Presentation of the Information Contained in the Experiment

It is useful to consider the set

$$\mathcal{E} = \left\{ \mathbf{z} \in \mathbb{R}^m : \mathbf{z}^T\mathbf{M}\mathbf{z} \le 1 \ \right\}$$

which is called the ellipsoid of concentration of model (2). It is an *m*-dimensional ellipsoid if **M** is non-singular. It is an ellipsoidal cylinder when **M** is singular. The set $\mathcal{E}$ characterizes the concentration of the probability distribution of the estimator $\hat{\boldsymbol{\theta}}$ around the true value of $\boldsymbol{\theta}$. A small $\mathcal{E}$ means a concentrated distribution of the estimator $\hat{\boldsymbol{\theta}}$. The set $\mathcal{E}$ depends exclusively on the information matrix **M**, so to consider $\mathcal{E}$ is in the spirit of the previous section.

Moreover, one knows some important and interesting properties of $\mathcal{E}$: The volume of $\mathcal{E}$ is proportional to $\left[\det(\mathbf{M})\right]^{1/2}$. For any $\mathbf{h} \in \mathbb{R}^m$ one has that $Var\left(\mathbf{h}^T\ \hat{\boldsymbol{\theta}}\right)$ is proportional to $\max_{\mathbf{z}\in\mathcal{E}}\left(\mathbf{h}^T\mathbf{z}\right)^2$. In particular, when $\|\mathbf{h}\| = 1$ one has that $\sigma^{-1}\left[Var\left(\mathbf{h}^T\ \hat{\boldsymbol{\theta}}\right)\right]^{1/2}$ is simply the length of the projection of the ellipsoid $\mathcal{E}$ onto the straight line defined by the vector **h**. As a consequence of this property we have also further properties of $\mathcal{E}$: The radius *r* of the smallest ball containing the set $\mathcal{E}$ is equal to

$$r^2 = \sigma^{-2}\max_{h\in\mathbb{R}^m,\ \|h\|=1} Var(\mathbf{h}^T\hat{\boldsymbol{\theta}}).$$

The half-diagonal *d* of the parallelepiped having faces parallel to the coordinate planes is equal to

$$d^2 = \sigma^{-2}\sum_{i=1}^{m} Var(\hat{\theta}_i)$$

etc. So many geometrical features of the ellipsoid of concentration express in some way how good the experiment under question is.

3.3.1. The Ellipsoid for Parameter Subsets

One can construct the ellipsoid of concentration for any parameter subset, say for $\theta_1,...,\theta_s$ with $s < m$,

$$\mathcal{E}^{(s)} = \left\{ \mathbf{z} \in \mathbb{R}^{(s)} : \mathbf{z}^T \mathbf{M}^{(s)} \mathbf{z} \le 1 \right\}$$

where $\mathbf{M}^{(s)}$ is the information matrix for the parameter subset given by (3). The interpretation is the same as for $\mathcal{E}$, but for $\theta_1,...,\theta_s$ instead of $\theta_1,...,\theta_m$.

3.4. Optimality Criteria

One can read from the properties of the ellipsoid of concentration some classical optimality criteria.

In general, an optimality criterion is a function $\Phi(\mathbf{M})$ of the information matrix $\mathbf{M}$, interpreted so that a small value of $\Phi(\mathbf{M})$ means a good quality of the experiment. This is traditional. Sometimes, it is proposed to take $\Psi(\mathbf{M}) = -\Phi(\mathbf{M})$ instead of $\Phi(\mathbf{M})$, and call $\Psi(\mathbf{M})$ "the information functional".

A particular choice of the function Φ means to consider the information obtained in the experiment from a particular point of view.

Examples of well known optimality criteria are:

$\Phi(\mathbf{M}) = -\ln[\det(\mathbf{M})]$ ····· the D-optimality criterion

$\Phi(\mathbf{M}) = tr(\mathbf{M}^{-1}) \propto \sum_{i=1}^{m} Var(\hat{\theta}_i)$ ····· the A-optimality criterion

$\Phi(\mathbf{M}) = \lambda_{\min}^{-1}(\mathbf{M}) \propto \max_{\mathbf{h} \in \mathbb{R}^m, \|\mathbf{h}\|=1} \mathbf{h}^T \mathbf{M}^{-1} \mathbf{h}$ ·····*the E-optimality criterion.*

Here *tr* denotes the trace of a matrix, $\lambda_{\min}(\mathbf{M})$ denotes the minimum eigenvalue of $\mathbf{M}$.

These criteria are used when all parameters are important. One can call them global criteria. When only the parameters $\theta_1,\ldots,\theta_s$ are important, and the others are considered as nuisance parameters, one uses criteria of the form $\Phi(\mathbf{M}^{(s)})$ called partial criteria. Here the function Φ is defined as above, and $\mathbf{M}^{(s)}$ is the information matrix for the parameters $\theta_1,...,\theta_s$ as defined in (3). So we can consider partial D-optimality, partial A-optimality, etc. Of course, there are many possibilities to construct new optimality criteria. A criterion must be statistically meaningful, and it must express somehow the information obtained from the experiment.

-

-
-

Bibliography

Atkinson, A.C. and Donev, A.N. (1992). *Optimum experimental designs*. Oxford, Clarendon Press [A book explaining the implementation of methods into applications, with many examples.]

Atkinson, A.C., Pronzato, L. and Wynn, H.P. (Eds.) (1998). MODA 5 - Advances in Model -*Oriented Data Analysis and Experimental Design*. Heidelberg, Physica-Verlag, [The last proceedings from regular international conferences on experimental design. They started 1987 (Wartburg, Germany), then 1990 (St. Kyril, Bulgaria), 1992 (St. Petersburg, Russia), 1995 (Spetses, Greece), 1998 (Marseilles, France), and the next will be in 2001 (Austria).]

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science* **10**, 273-304 [A clear, simple exposition of the main features of the subject.]

Dresbeke, J.J., Fine, J. and Saporta, G. (Eds.) (1997). *Plans d'expériences (Applications à 1'entreprise)*. Paris, Editions Technip. [A recent handbook on experimental design. Divided into chapters written by different authors]

Elfving, G. (1952). Optimum allocation in linear regression. *Ann. Math. Statist*. **23**, 255-262. [Contains The "Elfving theorem" where the Elfving set (see Section 4.3) has been used the first time.]

Fedorov, V.V. (1972). *Theory of Optimal Experiments*. New York, Academic Press. [First book on optimum experimental design. The style of the exposition is close to that used in books on physics.]

Fedorov, V.V. (1980) Convex design theory. Math. *Operationsforschung u. Statistik, Ser. Statistics* **11**, 403-413. [It stresses on extensions of the equivalence theorem.]

Fisher, R.A. (1935). *The Design of Experiments*. Edinburgh, Oliver and Boyd. [Contains the classical fundaments of combinatorial experimental design.]

Ford, I., Titterington, D.M. and Kitsos, C. (1989). Recent advances in nonlinear experimental design. *Technometrics* **31**, 49-60 [A short survey on the design in nonlinear models.]

Kiefer, J. (1 985). *Jack Carl Kiefer collected papers* (eds. Brown, L.D., Olkin, Sacks, I.J., Wynn H.P.) Vol III (*Design of experiments*). New York, Springer. [The collected work of the grounder of optimum experimental design theory.]

Müller, Ch. H. (1997). *Robust Planning and Analysis of Experiments*. Lecture Notes in Statistics No. 124, New York, Springer. [The theory of optimum design in the case that the model is violated by the presence of outliers.]

Müller, W. G. and Pázman, A. (1999). An algorithm for the computation of optimum designs under a given covariance structure. *Computational Statistics* **14**, 197-21 1. [The algorithm from Section 4, point c) is described.]

Pázman, A. (1986). *Foundations of Optimum Experimental Design*. Dordrecht, Reidel (Kluwer Group). [A book which contains mathematical details on the properties of optimality criteria, a survey on algorithms for optimal design, which were developed up to 1986, and full proofs of many properties of optimum designs in linear models. An extension to infinite-dimensional models is considered as well.]

Pázman, A. (2000). Concentration sets, Elfving sets and norms in optimum design. In: A.C. Atkinson, A.A. Zhigljavsky (Eds.): *Optimum Design* 2000. Dordrecht, Kluwer Acad. Flubl. (in print). [New points of view to the Elfving sets (Section 4.3.) are developed.]

Pázman, A., Pronzato, L. (1992). Nonlinear experimental design based on the distribution of estimators. *J. Statist. Planning and Inference*. **33**, 385-402 [Here is presented a method how to implement the density of the estimators in the design of nonlinear experiments.]

Pukelsheim, F. (1993). *Optimum Experimental Design*. New York, Wiley. [The book contains an informational approach to optimality criteria in linear models, an extended and strictly mathematical use of convexity in optimum design, and some elements on connecting combinatorial and optimum experimental design.]

Silvey, S.D. (1980). *Optimal Design*. London, Chapman and Hall. [A nice, mathematically complete introductory book on the subject.]

Silvey, S.D., Titterington, D.M. and Torsney, B. (1978). An algorithm for optimal designs on a finite design space. *Communications in Statist. A*, **7**, 1379-89. [Here an algorithm of the type b) from Section 4.5 has been presented for the first time.]

Biographical Sketch

Andrej Pázman, Dr.Sc (Fellow of Royal Statistical Society, member of AMS) was born in Prague in December 1938. He obtained his masters degree from the Comenius University in Bratislava (Physics and Mathematics) in 1961, his Candidate of Sciences (CSc., equivalent of Ph.D.) in 1964, and became Doctor in Sciences (DrSc.) at the Academy of Sciences in Bratislava in 1980.

He has been with the Slovak Academy of Sciences from 1964 to1991, and worked at the "International Institute for Nuclear Research" in Dubna (Russia) from 1966 to 1969. He is full professor at the Comenius University in Bratislava since 1991, where he has been Head of Department from 1992 to 1998 and is now Head of the Physics and Informatics Section of Mathematics of the Faculty of Mathematics. He was visiting professor for one semester at the University of Economics in Vienna in 1995, at the University of Augsburg in 1998/99 and at the Technological University of Vienna in 2000. He had several collaborations with different French institutes and universities (University J. Fourier in Grenoble, "Centre National de la Recherche Scientifique" Paris and Sophia Antipolis, "Institut National de la Recherche Agronomique" Versadies), and with Free University, Berlin.

His main research interests are in optimum experimental design, nonlinear regression, and more generally mathematical statistics. He is the author of three books in Slovak and of the two books: Foundations of Optimum Experimental Design (Reidel/Kluwer, Dordrecht, 1986) and Nonlinear Statistical Models (Kluwer, Dordrecht, 1993). He is on the editorial board of the journals Statistics (1987-1995), Applications of Mathematics (1976-2000), Acta Mathematica Universitatis Comenianae (1999-) and Metrika (2000-).