Tatra Mt. Math. Publ. **26** (2003), 69–77

TATRA MOUNTAINS
Mathematical Publications

OPTIMUM DESIGN OF NONLINEAR EXPERIMENTS BASED ON LINEAR BAYESIAN ESTIMATORS

Klára Hornišová — Andrej Pázman

ABSTRACT. In a nonlinear regression model, the linear Bayesian estimator is shown to minimize the mean square error matrix (MSE). The inverse of MSE is proved to be a concave function of the design. Thus optimality criteria applied to MSE could lead to convex Bayesian design methods.

1. Introduction

We consider a nonlinear regression model

$$\begin{aligned} E\big[y(x)\big] &= \eta(x,\theta)\,, \quad \theta \in \Theta\,, \\ \operatorname{Var}\big[y(x)\big] &= \sigma^2 w(x)\,, \quad x \in X\,, \end{aligned} \tag{1}$$

where X is the design space, and $w^{-1}(x)$ is the weight of observations at x. Let $x_1,\dots,x_N$ be any choice of points from X where the observations will be observed independently. It is traditional to call it an "exact" design. Related to it there is the design measure on X

$$\xi(x) = r(x)/N\,; \quad x \in X$$

with $r(x)$ equal to the number of independent replications of the points x in the sequence $x_1,\dots,x_N$. Although algorithms for obtaining optimum designs will be considered in terms of the design measures, we shall first consider estimation of θ in terms of the design $x_1,\dots,x_N$.

The weighted least squares estimator of θ is equal to

$$\hat\theta = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} w^{-1}(x_i)\big[y(x_i) - \eta(x_i,\theta)\big]^2.$$

2000 Mathematics Subject Classification: 62K05, 62J02.
Keywords: Bayesian statistics, optimality criteria, equivalence theorem.

Research supported by the Slovak Vega grant No. 1/7295/20.

It is known that under mild regularity conditions we have asymptotically, for large N,

$$E(\hat{\theta}) \doteq \bar{\theta}\,, \qquad \operatorname{Var}(\hat{\theta}) \doteq \sigma^2 M^{-1}(\bar{\theta};\ x_1,\dots,x_N)\,, \tag{2}$$

where $\bar{\theta}$ is the true value of θ, and $M(\bar{\theta};\ x_1,\dots,x_N)$ is the Fisher information matrix (for $\sigma = 1$). Since we want to minimize the variance matrix $\operatorname{Var}(\hat{\theta})$, and we do not know $\bar{\theta}$, a usual way is to look for a design for which the maximum

$$\max_{x_1,\dots,x_N} \int_{\Theta} \Phi\big[M(\theta;\ x_1,\dots,x_N)\big]\,\pi(\theta)\,d\theta \tag{3}$$

is attained, where $\Phi(\cdot)$ is a classical optimality criterion known from linear models (e.g., the D-optimality with $\Phi(\cdot) = \ln\det(\cdot)$), and where $\pi(\theta)$ is a prior density of θ (cf. [2] or [1]). Two features of this approach must be emphasized:

- One needs to know a prior density $\pi(\theta)$.
- In fact, the approach is based on linear methods, since in the linearized approximate model
$$E\big[y(x)\big] = \eta(x,\bar{\theta}) + \frac{\partial\eta(x,\theta)}{\partial\theta^T}\Big|_{\theta=\bar{\theta}}(\theta-\bar{\theta})\,,$$
the LS estimator has the mean and variance given exactly by (2).

Instead of this combination of a prior and of a linear LS estimator, we propose to consider here from the very beginning the linear Bayesian estimator based on the same prior density $\pi(\theta)$. That means, we adapt from the very beginning a Bayesian approach. Notice that Bayesian designs in linear models have been well studied (e.g., in [6]) however, in nonlinear models the topic is only under development.

Such an estimator is of the form

$$Ay + c\,,$$

where the matrix A and the vector c must be chosen so as to attain the minimum of the mean square error

$$E_{\text{tot}}\Big[\|Ay + c - \theta\|^2\Big]\,,$$

where E_{tot} denotes the mean with respect to the joint density $f(y\,|\,\theta)\,\pi(\theta)$ of y and θ.

2. Properties of the linear Bayesian estimator

We shall use the notation $\eta = \big(\eta(x_1,\theta),\dots,\eta(x_N,\theta)\big)^T$. We denote further by $E_\pi, \operatorname{Var}_\pi, \operatorname{Cov}_\pi$ the mean, variance and covariance with respect to the prior density. If a function l on X is given, then by $\operatorname{diag}\big[l(x_i)\big]$ we shall denote a diagonal matrix with $l(x_i)$ as its ith diagonal element.

PROPOSITION 1. *The linear Bayesian estimator has the form*

$$\tilde{\theta}(y) = A\big(y - E_\pi(\eta)\big) + E_\pi(\theta)$$

with

$$A = \mathrm{Cov}_\pi\big(\theta, \eta^T\big)\Big[\sigma^2 \,\mathrm{diag}\big[w(x_i)\big] + \mathrm{Var}_\pi(\eta)\Big]^{-1}.$$

Its mean square error $E_{\mathrm{tot}}\Big[\big\|\tilde{\theta}(y) - \theta\big\|^2\Big]$ *is equal to*

$$\mathrm{tr}\Big[\mathrm{Var}_\pi(\theta) - \mathrm{Cov}_\pi\big(\theta, \eta^T\big)\big[\sigma^2 \,\mathrm{diag}\big[w(x_i)\big] + \mathrm{Var}_\pi(\eta)\big]^{-1} \mathrm{Cov}_\pi\big(\eta, \theta^T\big)\Big].$$

A further, even more strong justification of the use of this estimator is given by the following.

PROPOSITION 2. *Let $h \in \mathbb{R}^m$ be a given vector, and $h^T\theta$ a corresponding linear function of the parameters. Let $\tau(y)$ be any linear estimator of $h^T\theta$. Then*

$$E_{\mathrm{tot}}\Big[\big|\tau(y) - h^T\theta\big|^2\Big] \geq E_{\mathrm{tot}}\Big[\big|h^T\tilde{\theta}(y) - h^T\theta\big|^2\Big].$$

Proof of Propositions 1 and 2. Let S be any positive semidefinite matrix and a a vector. We shall use the notation $\|a\|_S^2 = a^T S a$.

Any linear estimator of θ can be written in the form

$$T(y) = A\big(y - E_\pi(\eta)\big) + E_\pi(\theta) + d$$

for some matrix A and some vector d. Then

$$E_{\mathrm{tot}}\Big[\big\|T(y) - \theta\big\|_S^2\Big] = E_{\mathrm{tot}}\Big[\big\|A(y-\eta) + A\big(\eta - E_\pi(\eta)\big) - \big(\theta - E_\pi(\theta)\big)\big\|_S^2\Big] + \|d\|_S^2.$$

Taking $C = \mathrm{Cov}_{\mathrm{tot}}\big(\eta, \theta^T\big) = \mathrm{Cov}_\pi\big(\eta, \theta^T\big)$, $V = \mathrm{Var}_{\mathrm{tot}}(\theta) = \mathrm{Var}_\pi(\theta)$, $W = \mathrm{Var}_{\mathrm{tot}}(y) = \big[\sigma^2 \,\mathrm{diag}\big[w(x_i)\big] + \mathrm{Var}_\pi(\eta)\big]$, we obtain

$$\begin{aligned} E_{\mathrm{tot}}\Big[\big\|T(y) - \theta\big\|_S^2\Big] &= \mathrm{tr}\, S\Big\{\big[A - C^T W^{-1}\big] W \big[A - C^T W^{-1}\big]^T + V - C^T W^{-1} C\Big\} + \|d\|_S^2 \\ &\geq \mathrm{tr}\, S\big\{V - C^T W^{-1} C\big\} \end{aligned} \tag{4}$$

with equality if $d = 0$ and $A = C^T W^{-1}$, i.e., if $T(y) = \tilde{\theta}(y)$.

To prove Proposition 1 take $S = I$.

To prove Proposition 2 consider $\tau(y) = l^T y + l_o$ an arbitrary linear estimator of $h^T\theta$. Take $c = l_o h$, $A = h l^T / \|h\|^2$ to write $\tau(y) = h^T(Ay + c)$. Finally take $S = hh^T$ to obtain from (4)

$$E_{\mathrm{tot}}\Big\{\big|\tau(y) - h^T\theta\big|^2\Big\} \geq h^T\big\{V - C^T W^{-1} C\big\} h$$

with equality if $\tau(y) = h^T\tilde{\theta}(y)$. $\square$

Let $T(y)$ be an estimator of θ. The matrix

$$\mathrm{MSE}(T) = E_{\mathrm{tot}}\Big\{\big[T(y)-\theta\big]\big[T(y)-\theta\big]^T\Big\}$$

is the mean square error matrix of $T(y)$. In particular

$$\begin{aligned}\mathrm{MSE}(\tilde{\theta}) &= \mathrm{Var}_\pi(\theta) - \mathrm{Cov}_\pi(\theta,\eta^T)\Big[\sigma^2 D\big[w(x_i)\big] + \mathrm{Var}_\pi(\eta)\Big]^{-1}\mathrm{Cov}_\pi(\eta,\theta^T)\\ &= V^T - C^T W^{-1} C\,.\end{aligned}$$

According to Proposition 2,

$$u^T\{\mathrm{MSE}(\tilde{\theta})\}u \le u^T\{\mathrm{MSE}(T)\}u$$

for any $u \in \mathbb{R}^m$ and for any linear estimator $T(y)$.

3. Properties of the mean square error as a function of the design

Let ξ be a design measure, and let $S_\xi = \{x_1,\dots,x_k\}$ be its support. Then the mean square error matrix of the estimator $\tilde{\theta}$ can be rewritten in the form

$$\mathrm{MSE}(\tilde{\theta},\xi) = \mathrm{Var}_\pi(\theta) - \mathrm{Cov}_\pi(\theta,\eta_\xi^T)W^{-1}(\xi)\,\mathrm{Cov}_\pi(\eta_\xi,\theta^T) \tag{5}$$

with $W(\xi) = \Big[\frac{\sigma^2}{N}\,\mathrm{diag}\big[w(x_i)\xi^{-1}(x_i)\big] + \mathrm{Var}_\pi(\eta_\xi)\Big]$, $\eta_\xi = \big(\eta(x_1,\theta),\dots,\eta(x_k,\theta)\big)^T$.

Equation (5) can be justified by the fact that, in terms of second order statistical properties, the model

$$\begin{aligned}E\big[y_j(x_i)\big] &= \eta(x_i,\theta)\,,\\ \mathrm{Var}\big[y_j(x_i)\big] &= \sigma^2 w(x_i)\,, \qquad i=1,\dots,k\,; \quad j=1,\dots,r(x_i)\,,\end{aligned}$$

with $r(x_i)$ replications at each point x_i can be replaced by the model

$$\begin{aligned}E\big[\bar{y}(x_i)\big] &= \eta(x_i,\theta)\,,\\ \mathrm{Var}\big[\bar{y}(x_i)\big] &= \sigma^2 w(x_i)\big/\big[N\xi(x_i)\big]\,, \qquad i=1,\dots,k\,,\end{aligned}$$

where $\bar{y}(x_i) = r^{-1}(x_i)\Big[\sum_{j=1}^{r(x_i)} y_j(x_i)\Big]$. Notice that it depends on σ^2/N. This cannot be avoided, since it expresses the ratio of the prior information to the information obtained from the observations.

We shall denote by $I(\xi)$ the inverse of $\mathrm{MSE}(\tilde{\theta}, \xi)$. The matrix $I(\xi)$ can be considered as a kind of information matrix, since a large $I(\xi)$ means a small $\mathrm{MSE}(\tilde{\theta}, \xi)$, i.e., a good precision of $\tilde{\theta}(y)$. Of course, $I(\xi)$ differs very much from the Fisher information matrix. We extend its definition to the set Ξ of all probability measures ξ having a finite support: $S_\xi = \{x \in X : \xi(x) > 0\}$. Notice, that this does not mean that the set X is finite. Even for an arbitrary X, the set Ξ is a convex set on which we can build the whole design theory in linear models, as shown in [5].

The following statement is basic for experimental design.

Proposition 3. *The mapping*

$$\xi \in \Xi \mapsto I(\xi)$$

is concave, and the function

$$\xi \in \Xi \mapsto \Phi\big[I(\xi)\big]$$

is concave for any function $\Phi(\cdot)$ which is concave and nondecreasing in the Loewner sense.

Proof. Take $\xi_\lambda = (1-\lambda)\xi + \lambda\mu$. It is sufficient to prove that for any fixed $\xi, \mu \in \Xi$, $u \in \mathbb{R}^m$, the real valued function $\phi(\lambda)$

$$\phi(\lambda) = u^T \big\{ I[\xi_\lambda] \big\} u$$

has a non-positive second order derivative for each $\lambda \in (0,1)$. To show this we have to use several times the rule

$$\frac{d}{d\lambda} A^{-1}(\lambda) = -A^{-1}(\lambda) \left[\frac{d}{d\lambda} A(\lambda) \right] A^{-1}(\lambda)$$

valid for any differentiable matrix function $A(\lambda)$ of λ, and the rule showing how to invert a positive definite matrix U, which is decomposed into blocks:

$$U = \begin{pmatrix} A & B \\ B^T & D \end{pmatrix} \Longrightarrow U^{-1} = \begin{pmatrix} P & Q \\ Q^T & S \end{pmatrix}, \tag{6}$$

where $P = \big(A - BD^{-1}B^T\big)^{-1}$, $S = D^{-1} + D^{-1}B^T\big(A - BD^{-1}B^T\big)^{-1}BD^{-1}$. By interchanging the matrices A and B, B and B^T, it follows also that

$$\big(A - BD^{-1}B^T\big)^{-1} = A^{-1} + A^{-1}B\big(D - B^TA^{-1}B\big)^{-1}B^TA^{-1}. \tag{7}$$

Take the notation similar to that in the proofs of Propositions 1 and 2, $R_\lambda = \mathrm{MSE}(\tilde{\theta}, \xi_\lambda)$, $W_\lambda = W(\xi_\lambda)$, $\bar{V} = \mathrm{Var}_\pi(\eta)$, $V = \mathrm{Var}_\pi(\theta)$, etc. Notice that for $\lambda \in (0,1)$ the set S_{ξ_λ} (the support of ξ_λ) does not depend on λ, so the matrix C does not depend on λ as well. We obtain

$$\frac{d}{d\lambda}\phi(\lambda) = u^T R_\lambda^{-1} C^T W_\lambda^{-1} U_\lambda W_\lambda^{-1} C R_\lambda^{-1} u\,,$$

where

$$U_\lambda = \operatorname{diag}\Big[\frac{\sigma^2}{N} w(x_i)\big(\mu(x_i) - \xi(x_i)\big)\xi_\lambda^{-2}(x_i)\Big].$$

Similarly

$$\frac{d^2}{d\lambda^2}\phi(\lambda) = 2u^T R_\lambda^{-1} C^T W_\lambda^{-1} H_\lambda W_\lambda^{-1} C R_\lambda^{-1} u\,,$$

where

$$H_\lambda = U_\lambda\big[W_\lambda^{-1} C R_\lambda^{-1} C^T W_\lambda^{-1} + W_\lambda^{-1}\big] U_\lambda - Q_\lambda\,,$$

$$Q_\lambda = \operatorname{diag}\Big[\frac{\sigma^2}{N} w(x_i)\big(\mu(x_i) - \xi(x_i)\big)^2 \xi_\lambda^{-3}(x_i)\Big].$$

According to the definition of R_λ and by (7) we obtain

$$\begin{aligned} &W_\lambda^{-1} C R_\lambda^{-1} C^T W_\lambda^{-1} + W_\lambda^{-1} \\ &\qquad = W_\lambda^{-1} + W_\lambda^{-1} C\big(V - C^T W_\lambda^{-1} C\big)^{-1} C^T W_\lambda^{-1} = \big(W_\lambda - C V^{-1} C^T\big)^{-1}. \end{aligned}$$

Hence,

$$\begin{aligned} H_\lambda &= U_\lambda \big(W_\lambda - C V^{-1} C^T\big)^{-1} U_\lambda - Q_\lambda \\ &= -\left[\begin{pmatrix} Q_\lambda & U_\lambda \\ U_\lambda & (W_\lambda - C V^{-1} C^T) \end{pmatrix}^{-1}\right]_{11} \end{aligned}$$

where by the subscript 11 we indicated the first diagonal block of the matrix. Using the definitions of W_λ, Q_λ, U_λ, and applying (6) we obtain that the last inverted matrix is positive semidefinite. It follows that the matrix H_λ is negative semidefinite, so

$$\frac{d^2}{d\lambda^2}\phi(\lambda) \le 0\,.$$

To prove the second statement of the proposition we write

$$\begin{aligned} \Phi\Big[I\big(\lambda\xi + (1-\lambda)\mu\big)\Big] &\ge \Phi\big[\lambda I(\xi) + (1-\lambda) I(\mu)\big] \\ &\ge \lambda\Phi\big[I(\xi] + (1-\lambda)\Phi\big[I(\mu)\big]\,, \end{aligned}$$

where the first inequality follows from the concavity of $I(\xi)$ and the monotonicity of Φ, while the second follows from the concavity of Φ. $\square$

4. The equivalence theorem and the iterative computation of optimum designs

Consider the directional derivative

$$\frac{\partial\Phi\big[I(\xi_\lambda)\big]}{\partial\lambda}\bigg|_{\lambda=0} = \left\{\operatorname{tr}\ \nabla\Phi\big[I(\xi)\big]\,\frac{\partial I(\xi_\lambda)}{\partial\lambda}\bigg|_{\lambda=0}\right\},$$

where $\xi_\lambda = (1-\lambda)\xi + \lambda\mu$. By direct computation we obtain

$$\frac{\partial\Phi\big[I(\xi_\lambda)\big]}{\partial\lambda}\bigg|_{\lambda=0} = \sum_{x\in X}\mu(x)\Big[\frac{\sigma^2}{N}w(x)\Big]^{-1} Z^T(x,\xi)\nabla\Phi\big[I(\xi)\big]Z(x,\xi) - c(\xi)$$

with

$$\begin{aligned} Z(x,\xi) &= I(\xi)\big\{\operatorname{Cov}_\pi(\theta,\eta_x) - \operatorname{Cov}_\pi(\theta,\eta_\xi)W^{-1}(\xi)\operatorname{Cov}_\pi(\theta,\eta_x)\big\}\,,\\ \eta_x &= \eta(x,\theta)\,,\\ c(\xi) &= \sum_{x\in X}\xi(x)\Big[\frac{\sigma^2}{N}w(x)\Big]^{-1} Z^T(x,\xi)\nabla\Phi\big[I(\xi)\big]Z(x,\xi)\,. \end{aligned}$$

A design is called Φ-optimal if it maximizes $\Phi\big[I(\xi)\big]$.

PROPOSITION 4. *If ξ^* is a design such that for some $\delta > 0$ and for every $x \in X$*

$$\Big[\frac{\sigma^2}{N}w(x)\Big]^{-1} Z^T(x,\xi^*)\nabla\Phi\big[I(\xi^*)\big]Z(x,\xi^*) \le c(\xi^*) + \delta$$

then

$$\Phi\big[I(\xi^*)\big] \ge \max_{\xi\in\Xi}\Phi\big[I(\xi)\big] - \delta$$

(approximate Φ-optimality).

In particular, the design ξ^* is Φ-optimal if and only if the inequality holds for $\delta = 0$.

The proof is similar as in linear models (cf. [5] Propositions IV.27 and IV.28), and relies on concavity and on the fact that the expression for the directional derivative is linear in μ.

An iterative method of computation can be constructed on these results, like in linear models. We start by taking any ξ_0 having a regular $R(\xi_0)$, and at the nth step we correct the design according to

$$\xi_{n+1} = (1-\beta_n)\xi_n + \beta_n\xi_{x_n}\,,$$

where x_n is concentrated in one point x_n which is chosen so as to maximize the directional derivative, i.e.,

$$x_n = \arg\max_{x\in X}\Big[\frac{\sigma^2}{N}w(x)\Big]^{-1} Z^T(x,\xi^*)\nabla\Phi\big[I(\xi^*)\big]Z(x,\xi^*)$$

and

$$\beta_n = \arg\max_{\beta\in[0,1]}\Phi\big[(1-\beta)\xi_n + \beta\xi_{x_n}\big]\,.$$

EXAMPLE 1. We considered the model $\eta(x,\theta) = \exp\{-\theta x\}$ with $\theta \in (0,\infty)$, $x \in X = [0,2]$, $N = 1$, $\sigma^2 = 1$, and with the gamma prior

$$\pi_{a,p}(\theta) = \frac{a^p}{\Gamma(p)} \exp\{-a\theta\}\theta^{p-1}$$

for $a = 2$, $p = 2$. In this model the design which is optimal according to the D-optimality criterion (3) has been obtained in [3]. This design is concentrated in one point $x^* = 1$. By the method of the present paper we obtained that the same design minimizes also $\mathrm{MSE}(\tilde{\theta},\xi)$.

EXAMPLE 2. We considered the model

$$\eta(x,\theta) = \frac{\theta_1}{\theta_1 + \theta_2 - \theta_3}\Big[\exp\{-\theta_3 x\} - \exp\big\{-(\theta_1 + \theta_2)x\big\}\Big]$$

(cf. [4]) with $N = 1$, $\sigma^2 = 1$, and the prior equal to the product of gamma priors

$$\pi(\theta_1,\theta_2,\theta_3) = \pi_{2,2}(\theta_1)\pi_{3,8}(\theta_2)\pi_{2,3}(\theta_3).$$

We considered 2 choices of X. When $X = [0,2]$, we obtained that the D-optimum design is concentrated at the point $x^* = 0.4$. On the other hand, when $X = [-1,1]$, the approximate optimum design obtained by the iterative algorithm is supported by the set

$$\{x_1, x_2, x_3, x_4\} = \{-1, -0.95, -0.7, -0.65\}$$

and is equal to

$$\big(\xi(x_1), \xi(x_2), \xi(x_3), \xi(x_4)\big) = (0.028,\ 0.009,\ 0.07,\ 0.893).$$

Acknowledgment. The authors thank the anonymous referee for helpful advices.

REFERENCES

[1] ATKINSON, A. C.—DONEV, A. N.: *Optimum Experimental Designs*, Clarendon Press, Oxford, 1992.

[2] CHALONER, K.—VERDINELLI, I.: *Bayesian experimental design: a Review*, Statist. Sci. **10** (1995), 273–304.

[3] DETTE, H.: *A note on Bayesian c- and D-optimal designs in nonlinear regression models*, Ann. Statist. **24** (1996), 1225–1234.

[4] FORD, I.—TITTERINGTON, D. M.—KITSOS, C.: *Recent advances in nonlinear experimental design*, Technometrics **31** (1989), 49–60.

[5] PÁZMAN, A.: *Foundations of Optimum Experimental Design*, Reidel, Dordrecht, 1986.

[6] PILZ, J.: *Bayesian Estimation and Experimental Design in Linear Regression Models*, Teubner, Leipzig, 1983, (2nd. ed., Wiley, New York, 1991), p. 296.

Received February 26, 2002

Klára Hornišová
Institute of Measurement Science
Slovak Academy of Sciences
Dúbravská cesta 9
SK–842 19 Bratislava
SLOVAKIA
E-mail: umerhorn@savba.sk

Andrej Pázman
Department of Probability and Statistics
Faculty of Mathematics, Physics and Informatics
Comenius University
Mlynská dolina
SK–842 48 Bratislava
SLOVAKIA
E-mail: pazman@fmph.uniba.sk