VERSITA

TATRA MOUNTAINS
Mathematical Publications

DOI: 10.2478/v10127-012-0014-z
Tatra Mt. Math. Publ. **51** (2012), 141–149

OPTIMALITY CRITERIA FOR DESIGN IN NONLINEAR MODELS WITH CONSTRAINTS

Andrej Pázman

ABSTRACT. We shall present different expressions for optimality criteria in nonlinear regression models, and compare them with corresponding expressions in models without constraints. We also present how to formulate the equivalence theorem in models with constraints.

1. Introduction

For the estimation of parameters and testing of hypotheses in models with parameter constraints we have the classical results of Rao (1965) for linear models and of Silvey (1959, 1975) for nonlinear models, but it seems that even Silvey did not put attention to design in such models. Probably because he used the method of Lagrange multipliers, which is difficult. We prefer here to use some geometry (projectors) and the implicit function theorem instead of that. By that we complete and correct the presentation in Pázman (2002), using modified proofs.

The considered model is the nonlinear regression model

$$\begin{aligned} y_x &= \eta(x,\theta)+\varepsilon_x; \quad x\in\mathcal{X}, \qquad \mathrm{Var}(\varepsilon_x)=\sigma^2, \\ \theta &\in \Theta\subset\mathbb{R}^p, \quad C(\theta)=\big(C_1(\theta),\dots,C_q(\theta)\big)^T=0. \end{aligned} \tag{1}$$

The equations $C_1(\theta)=0,\dots,C_q(\theta)=0$ are the constraints. Here $\mathcal{X}$ is a (compact) design space. The functions $\eta(x,\theta)$ and $C(\theta)$ are twice continuously differentiable with respect to θ, and $\eta(x,\theta)$ is continuous on $\mathcal{X}$ for every θ.

2010 Mathematics Subject Classification: 62K05, 62F10.
Keywords: design of experiments, nonlinear regression, equivalence theorem.
Supported by the grant VEGA no. 2/0038/12.

We shall also use the following *notations, assumptions and matrix identities*:
$[\bar{\theta}]$ is a fixed point of int (Θ),
$\left[\xi\left(x\right)\right]$ is the frequency of replications of observations at a point $x \in \mathcal{X}$,

$$L \equiv \left[\frac{\partial C\left(\theta\right)}{\partial \theta^T}\right]_{\bar{\theta}},$$

and we shall suppose that the constraints $C_i\left(\theta\right)$ are locally linearly independent on $\bar{\theta}$, that is, L has a full rank $q < p$,

$$\begin{aligned} f\left(x\right) &\equiv \left[\frac{\partial \eta\left(x,\theta\right)}{\partial \theta}\right]_{\bar{\theta}}, \\ M &\equiv M\left(\xi\right) \equiv \sum_{x \in \mathcal{X}} f\left(x\right) f^T\left(x\right) \xi\left(x\right), \\ H &\equiv H\left(\xi\right) \equiv M\left(\xi\right) + L^T L, \\ P_L^A &\equiv L^T \left[LA^{-1}L^T\right]^- LA^{-1}, \\ P_L &\equiv P_L^I = L^T \left[LL^T\right]^- L \end{aligned}$$

with A some positive definite $p \times p$ matrix, and with arbitrary g-inverses,

$$\tilde{V}\left(A\right) \equiv A^{-1}\left[I - P_L^A\right] = A^{-1} - A^{-1}L^T\left[LA^{-1}L^T\right]^- LA^{-1}, \tag{2}$$

$\beta \in \mathcal{B} \subset \mathbb{R}^{p-q}$, is an auxiliary parametrization of the model (1) (For a justification of this parametrization see Proposition A2),
$\phi\left(\beta\right)$ is a mapping of $\mathcal{B}$ onto a subset of int (Θ),
$\bar{\beta} \in \mathcal{B}$ is a point such that $\phi\left(\bar{\beta}\right) = \bar{\theta}$ (see Proposition A2),

$$D \equiv \left[\frac{\partial \phi\left(\beta\right)}{\partial \beta^T}\right]_{\bar{\beta}}.$$

Notice that we did not denote the dependence on $\bar{\theta}$ in $f\left(x\right)$ and $M\left(\xi\right)$, and the same will be done in other expressions, since $\bar{\theta}$ is fixed in advance. So, the notation is similar to that in linear models.

An evident matrix identity is

$$\left[I - P_L^A\right] H = \left[I - P_L^A\right] M \tag{3}$$

since P_L^A is a projector onto the range of L^T. Less evident is the following identity

$$A^{-1}\left[I - P_L^A\right] = \left[\left(I - P_L\right) A \left(I - P_L\right)\right]^+, \tag{4}$$

where $^+$ denotes the Moore-Penrose g-inverse matrix. See Proposition A1 for the proof.

2. Estimability and the variance matrix of $\hat{\theta}$

If an exact design $x_1, \ldots, x_N$ (with $x_i \in \mathcal{X}$) is used, the corresponding observations $y_{x_1}, \ldots, y_{x_N}$ are supposed to be independent, and the L.S. estimator of θ is equal to

$$\hat{\theta} = \arg \min_{\substack{\theta \in \Theta, \\ C(\theta)=0}} \sum_{k=1}^{N} \left[y_{x_k} - \eta \left(x_k, \theta \right) \right]^2.$$

As it is standard, we consider the (approximate or asymptotic) design ξ, which is a probability measure having a finite support and defined on the design space $\mathcal{X}$. For each $x \in \mathcal{X}$ the value $\xi(x)$ (or in a more correct notation $\xi(\{x\})$) is interpreted as the approximate relative frequency of independently replicated observations at x.

By $\bar{\theta}$ we denote the true value of θ (in theory) or the point of localization of θ (in locally optimal design). It is supposed that $\bar{\theta} \in \operatorname{int}(\Theta)$.

Notice that many results presented in this and the following sections are based on the proofs in Section 5.

In a model without constraints the information matrix is $M(\xi)$, and the estimate of θ can be unique only if $M(\xi)$ is nonsingular. On the other hand, in models with constraints this holds only if the matrix $H(\xi)$ is nonsingular (see Corollary 1 to Proposition A2). However, there is no reason to interpret $H(\xi)$ as the information matrix in the model with constraints. In linear models with constraints linear in θ the nonsingularity of $H(\xi)$ is also sufficient for the uniqueness of $\hat{\theta}$, but in nonlinear models we must add the condition of asymptotic identifiability: $\bar{\theta}$ is supposed to be the unique minimizer in

$$\min_{\substack{\theta \in \Theta, \\ C(\theta)=0}} \sum_{x \in \mathcal{X}} \left[\eta(x, \theta) - \eta\left(x, \bar{\theta}\right) \right]^2 \xi(x)$$

(see Corollary 2 of Proposition A2).

Now we give some alternative formulae for the (asymptotic) variance matrix of $\hat{\theta}$. If the model (1) is without constraints we have, up to the multiplicative term $\left(\frac{\sigma^2}{N}\right)$

$$\operatorname{Var}_{M(\xi)}\left(\hat{\theta}\right) = M^{-1}(\xi)$$

but when we have constraints, we have several, seemingly different expressions. We can write

$$\operatorname{Var}_{M(\xi)}\left(\hat{\theta}\right) = D \left[D^T M(\xi) D \right]^{-1} D^T$$

(see Corollary 3 of Proposition A2), or alternatively

$$\operatorname{Var}_{M(\xi)}\left(\hat{\theta}\right) = \tilde{V} \left[M(\xi) + L^T L \right] \tag{5}$$

with $\tilde{V}(.)$ defined by (2) (see Corollary 3 to Proposition A2), or

$$\mathrm{Var}_{M(\xi)}\big(\hat{\theta}\big) = \big[(I - P_L)\, M(\xi)\, (I - P_L)\big]^{+}. \tag{6}$$

Notice that the equality of (5) with (6) follows from (4) when we take $A = M(\xi) + L^T L$, and from (3) when we take $A = I$.

Moreover, if $(L^*)^T$ is any matrix with the same column space as L^T, we are allowed to put L^* instead of L into (5). This fact follows from the equality of projectors $P_L = P_{L^*}$, hence the right-hand size of (6) remains unchanged, and from (4) it follows that it is equal to $\tilde{V}\big[M(\xi) + (L^*)^T L^*\big]$.

However, it seems that from all these alternatives the alternative (5) is preferable from the practical point of view.

3. Global optimality criteria

In the model (1), but without constraints, the classical optimality criteria are based on convex functions of the information matrix: -ln det $M(\xi)$ for D-optimality, $\mathrm{tr}\big\{M^{-1}(\xi)\big\}$ for A-optimality, $\max_{u:\|u\|=1} u^T M^{-1}(\xi)\, u$ for E-optimality. Since in a model without constraints $M^{-1}(\xi)$ is the variance matrix of $\hat{\theta}$, these criteria can be expressed equivalently as functions of $\mathrm{Var}_{M(\xi)}\big(\hat{\theta}\big)$. Evidently, this indicates the way how to obtain optimality criteria in models with constraints. However, more complicated functions of the matrix $M(\xi)$ are then obtained

$$\Phi\big[M(\xi)\big] = \mathrm{tr}\,\tilde{V}\big(M(\xi) + L^T L\big) \qquad \text{for A-optimality,}$$

$$\Phi\big[M(\xi)\big] = \max_{u:\|u\|=1} u^T \Big[\tilde{V}\big(M(\xi) + L^T L\big)\Big] u \qquad \text{for E-optimality.}$$

Here $\tilde{V}(.)$ is defined in (2).

The situation is more complicated in case of D-optimality, because to take $\ln\det \mathrm{Var}_{M(\xi)}\big(\hat{\theta}\big)$ is a nonsense since $\mathrm{Var}_{M(\xi)}\big(\hat{\theta}\big)$ is singular for any ξ. This is because by the presence of constraints the model is overparametrized, so we have to consider the D-optimality criterion in the equivalent model (8) introduced in Proposition A2, which is without constraints. In principle we have to choose the parameters denoted by β, introduced in the equivalent model (8), and the D-optimality criterion should be given by

$$\Phi\big[M(\xi)\big] = -\ln\det\big[D^T M(\xi)\, D\big]$$

(see Proposition A2). However, in Proposition A2 we present no construction of such parameters. Fortunately, as is well known, in models without constraints the criterion of D-optimality is invariant to a reparametrization of the model. More precisely, the ordering of designs according to the optimality criterion does

not depend on this choice of parametrization. Hence, instead of D we can use any matrix with the same column space. Since $LD = 0$ (see Proposition A2), we can follow the recommendation of P á z m a n (2002), and take the QR decomposition of L^T

$$L^T = (T, Q) \begin{pmatrix} R \\ 0 \end{pmatrix}.$$

The columns of the matrix T form an orthogonal basis of the column space of L^T, and the columns of the matrix Q form an orthogonal basis of its orthogonal complement, hence of the column space of D, and we can take for the D-optimality criterion the function

$$\Phi\big[M(\xi)\big] = -\ln\det\left[Q^T M(\xi) Q\right].$$

But curiously, as it is shown in Section 4, for the "equivalence theorem" we do not need the computation neither of Q nor of D.

Are the new criteria given by so complicated expressions still convex functions of $M(\xi)$ or of ξ? The answer is yes. It follows from the formula $\mathrm{Var}_{M(\xi)}\big(\hat{\theta}\big) = D\left[D^T M(\xi) D\right]^{-1} D^T$ (see Corollary 3 of Proposition A2).

4. The equivalence theorem

In general, for any convex and differentiable optimality criterion Φ we have the well known "equivalence theorem" (cf., e.g., P á z m a n (1986), Proposition IV.2.7):

A design μ is Φ-optimal if and only if

$$\left\{\max_{x\in\mathcal{X}} f^T(x)\left[-\nabla_M \Phi(M)\right] f(x)\right\}_{M=M(\mu)} = \left\{\mathrm{tr}\left[-M\nabla_M \Phi(M)\right]\right\}_{M=M(\mu),} \quad (7)$$

where the gradient $\nabla_M \Phi(M)$ is a $p \times p$ matrix with components

$$\left\{\nabla_M \Phi(M)\right\}_{ij} = \frac{\partial \Phi(M)}{\partial M_{ij}}.$$

In a model without constraints, we have $\nabla_M\left[-\ln\det(M)\right] = -M^{-1}$ for D-optimality, $\nabla_M\left[\mathrm{tr}\left(M^{-1}\right)\right] = -M^{-2}$ for A-optimality. In models with constraints we obtain following.

For A-optimality

$$\nabla_M \,\mathrm{tr}\left\{\tilde{V}\left[M + L^T L\right]\right\} = -\big(I - P_L^H\big) H^{-2}\big(I - P_L^H\big) = -\left[\mathrm{Var}_M\big(\hat{\theta}\big)\right]^2.$$

Here we used (2), and the equality

$$A(\alpha)\frac{dA^-(\alpha)}{d\alpha}A(\alpha) = -A(\alpha)A^-(\alpha)\frac{dA(\alpha)}{d\alpha}A^-(\alpha)A(\alpha)$$

valid for any g-inverse $A^{-}(\alpha)$ of a square matrix $A(\alpha)$ (cf. H a r v i l l e (2000), Lemma 15.10.5). The resulting formula for the gradient hence does not depend on the matrix D when using (5) for $\mathrm{Var}_M(\hat{\theta})$.

For D-optimality

$$\begin{aligned}\frac{\partial}{\partial M_{ij}}\left\{-\ln\det\left(D^T M D\right)\right\} &= -\operatorname{tr}\left\{\left(D^T M D\right)^{-1}\frac{\partial\left(D^T M D\right)}{\partial M_{ij}}\right\}\\ &= -\left\{D\left(D^T M D\right)^{-1}D^T\right\}_{ij}\\ &= -\left\{\mathrm{Var}_M\left(\hat{\theta}\right)\right\}_{ij}.\end{aligned}$$

Here we used that $\frac{\partial}{\partial A_{ij}}\left\{-\ln\det(A)\right\} = -\{A^{-1}\}_{ji}$ for any nonsingular square matrix A (cf. H a r v i l l e (2000), Eq. (8.7)). So

$$\nabla_M\left\{-\ln\det\left(D^T M D\right)\right\} = -\,\mathrm{Var}_M\left(\hat{\theta}\right),$$

a formula that again does not depend on the matrix D when using (5).

The equivalence theorem for D-optimality then follows, according to (7)

A design μ is D-optimal if and only if

$$\max_{z\in\mathcal{X}} f^T(z)\left[\mathrm{Var}_{M(\mu)}\left(\hat{\theta}\right)\right]f(z) = \sum_{x\in\mathcal{X}} f^T(x)\left[\mathrm{Var}_{M(\mu)}\left(\hat{\theta}\right)\right]f(x)\,\mu(x).$$

As a consequence we have that the D-optimal design is supported only by those points $x\in\mathcal{X}$, where the maximum of $f^T(x)\left[\mathrm{Var}_{M(\mu)}\left(\hat{\theta}\right)\right]f(x)$ is attained.

It is easy to obtain corresponding results for A-optimality.

5. Proofs

Proposition A1. *The identity* (4) *is valid.*

P r o o f. We use here properties of orthogonal projectors. First we verify straightforwardly three equalities

$$\begin{aligned}(I-P_L)A^{-1}\left(I-P_L^A\right) &= A^{-1}\left(I-P_L^A\right),\\ (I-P_L)\left(I-P_L^A\right) &= (I-P_L),\\ \left(I-P_L^A\right)(I-P_L) &= \left(I-P_L^A\right).\end{aligned}$$

We used here that P_L and P_L^A are projectors onto the same space (= the column space of L^T), and that $LA^{-1}L^T\left(LA^{-1}L^T\right)^{-}L = L$ since L and $LA^{-1}L^T$ have the same column space. Then, after denoting $C = (I-P_L)A(I-P_L)$,

$B = A^{-1}\left[I - P_L^A\right]$ we verify that $CBC = C,\ BCB = B,\ CB = I - P_L = BC$. Hence B is the Moore-Penrose g-inverse of C. □

(We notice that similar properties have been used for estimators in singular models without constraints or for quadratic estimators in K u b á č e k, K u b á č k o v á, V o l a u f o v á (1995), pp. 116 and 441).

Proposition A2. *Let L be of full rank (=q). Then there is an open set $\mathcal{B} \in \mathbb{R}^{p-q}$ and a mapping ϕ of $\mathcal{B}$ onto a subset of Θ such that*

1) *there is a $\bar{\beta} \in \mathcal{B}$ such that $\phi\left(\bar{\beta}\right) = \bar{\theta}$,*
2) *$C\left[\phi\left(\beta\right)\right] = 0$ for every $\beta \in \mathcal{B}$,*
3) *$\left[\frac{\partial \phi(\beta)}{\partial \beta^T}\right]_{\bar{\beta}}$ is full rank $= p - q$,*
4) *$L\left[\frac{\partial \phi(\beta)}{\partial \beta^T}\right]_{\bar{\beta}} = LD = 0$.*
5) *The original model with constraints* (1) *is locally (hence asymptotically) equivalent to the model without constraints*

$$\begin{aligned} y_x &= \eta\left(x, \phi\left(\beta\right)\right) + \varepsilon_x; \quad \beta \in \mathcal{B} \in \mathbb{R}^{p-r}, \\ x &\in \mathcal{X}, \quad \mathrm{Var}\left(\varepsilon_x\right) = \sigma^2. \end{aligned} \tag{8}$$

P r o o f. Suppose, without loss of generality, that the first q rows of the $q \times p$ matrix L are linearly independent. Denote $\alpha = \left(\theta_1, \ldots, \theta_q\right)^T$, $\beta = \left(\theta_{q+1}, \ldots, \theta_p\right)^T$. Then from $C\left(\alpha, \beta\right) = 0$ we obtain by the implicit function theorem that there is a neighborhood $\mathcal{B}$ of $\bar{\beta} = \left(\bar{\theta}_{q+1}, \ldots, \bar{\theta}_p\right)^T$ and a mapping g from $\mathcal{B}$ onto a neighborhood of $\bar{\alpha} = \left(\bar{\theta}_1, \ldots, \bar{\theta}_q\right)^T$ such that $g\left(\bar{\beta}\right) = \bar{\alpha}$, $C\big(g\left(\beta\right), \beta\big) = 0$ for $\beta \in \mathcal{B}$, and that

$$\frac{\partial g\left(\beta\right)}{\partial \beta^T} = -\left[\frac{\partial C\left(\alpha, \beta\right)}{\partial \alpha^T}\right]^{-1} \frac{\partial C\left(\alpha, \beta\right)}{\partial \beta^T}. \tag{9}$$

Denote

$$\phi\left(\beta\right) = \big(g\left(\beta\right), \beta\big)^T.$$

Evidently, $C\left[\phi\left(\beta\right)\right] = 0$ for every $\beta \in \mathcal{B}$, and $\phi\left(\bar{\beta}\right) = \bar{\theta}$. Moreover,

$$\frac{\partial \phi\left(\beta\right)}{\partial \beta^T} = \left(\frac{\partial g\left(\beta\right)}{\partial \beta^T},\ I\right)$$

has rank $p - q$ because I is here the identity $(p - q) \times (p - q)$ matrix. For $\beta \in \mathcal{B}$ we have from $C\left[\phi\left(\beta\right)\right] = 0$

$$0 = \frac{\partial C\left[\phi\left(\beta\right)\right]}{\partial \beta^T} = L\left[\phi\left(\beta\right)\right] \frac{\partial \phi\left(\beta\right)}{\partial \beta^T}; \qquad \beta \in \mathcal{B}.$$

Since for the asymptotics only some neighborhoods of $\bar{\beta}$ and of $\bar{\theta}$ are of importance, statement 5) follows from $C\left[\phi\left(\beta\right)\right] = 0$ which is valid for every $\beta \in \mathcal{B}$. □

Corollary 1. *The information matrix in the model* (8) *is equal to* $D^T M(\xi) D$ *and it is nonsingular if and only if* $H(\xi) = M(\xi) + L^T L$ *is nonsingular.*

P r o o f. The information matrix in the model (8) is

$$\sum_{x\in\mathcal{X}} \frac{\partial \eta(x,\phi(\beta))}{\partial \beta} \frac{\partial \eta(x,\phi(\beta))}{\partial \beta^T} \xi(x) = D^T M(\xi) D.$$

Denote $F^T \equiv \left(f(x_1)\sqrt{\xi(x_1)}, \ldots, f(x_N)\sqrt{\xi(x_N)}\right)$, where $\{x_1, \ldots, x_N\}$ is the support of ξ. Then $M(\xi) = F^T F$, and $M(\xi)$ has the same column space as F^T, which we denote by $\mathcal{E}$. The matrix $D^T M(\xi) D$ is nonsingular if and only if D is full rank and the column space of D is a subset of $\mathcal{E}$. And this is if and only if the matrix $\left(L^T, F^T\right)$ is full rank $(= p)$ since L and D are orthogonal matrices with complementary ranks (Proposition A2). Finally, $L^T L + M(\xi) = \left(L^T, F^T\right)\binom{L}{F}$ has the same rank as $\left(L^T, F^T\right)$. □

Corollary 2. *We have*

$$\begin{aligned} \{\bar\beta\} &= \arg\min_{\beta\in\mathcal{B}} \sum_{x\in\mathcal{X}} \left[\eta\big(x,\phi(\beta)\big) - \eta\Big(x,\phi\big(\bar\beta\big)\Big)\right]^2 \xi(x) \\ &\Leftrightarrow \\ \{\bar\theta\} &= \arg\min_{\substack{\theta\in\Theta,\\ C(\theta)=0}} \sum_{x\in\mathcal{X}} \left[\eta(x,\theta) - \eta\left(x,\bar\theta\right)\right]^2 \xi(x). \end{aligned}$$

Corollary 3. *In the model* (8) *the asymptotic variance is*

$$\mathrm{Var}_{M(\xi)}\big(\hat\beta\big) = \left[D^T M(\xi) D\right]^{-1}.$$

In the model (1) *we have*

$$\mathrm{Var}_{M(\xi)}\big(\hat\theta\big) = D\,\mathrm{Var}_{M(\xi)}\big(\hat\beta\big) D^T = \tilde V\left[M(\xi) + L^T L\right].$$

P r o o f. The first statement follows from Corollary 1. Further, since $\hat\theta = \phi\big(\hat\beta\big)$ we have

$$\mathrm{Var}_M\big(\hat\theta\big) = D\,\mathrm{Var}_M\big(\hat\beta\big) D^T.$$

Hence

$$\mathrm{Var}_M\big(\hat\theta\big) = D\left[D^T M D\right]^{-1} D^T = D\left[D^T H D\right]^{-1} D^T.$$

We used here that $LD = 0$. Hence

$$\mathrm{Var}_M\big(\hat\theta\big) = H^{-1/2} Z H^{-1/2},$$

where $Z \equiv H^{1/2} D\left[D^T H D\right]^{-1} D^T H^{1/2}$ is evidently the orthogonal projector onto the column space of the matrix $H^{1/2} D$. But according to Proposition A2, point 4, $H^{1/2} D$ and $H^{-1/2} L$ are orthogonal matrices with complementary ranks, so Z and P_L^H are complementary orthogonal projectors. In other words, $Z = I - P_L^H$. Hence finally,

$$\mathrm{Var}_M\big(\hat\theta\big) = H^{-1/2}\left[I - P_L^H\right] H^{-1/2} = \tilde V\left[M + L^T L\right]. \qquad \square$$

Acknowledgements. The author thanks the VEGA grant 2/0038/12 for financial support, and to the referee for helpful remarks to the organization of the paper.

REFERENCES

[1] HARVILLE, D. A.: *Matrix Algebra from a Statistician's Perspective* (3rd ed.), Springer, New York, 2000.
[2] KUBÁČEK, L.—KUBÁČKOVÁ, L.—VOLAUFOVÁ, L.: *Statistical Models with Linear Structures.* VEDA, Bratislava, 1995.
[3] PÁZMAN, A.: *Foundations of Optimum Experimental Design.* Reidel (Kluwer Group), Dordrecht, 1986.
[4] PÁZMAN, A.: *Optimal design of nonlinear experiments with parameter constraints,* Metrika **56** (2002), 113–130.
[5] RAO, C. R.: *Linear Statistical Inference and its Applications.* J. Wiley, New York, 1965.
[6] SILVEY, S. D.: *The Lagrangian multiplier test,* Ann. Math. Statist. **30** (1959), 389–407.
[7] SILVEY, S. D.: *Statistical Inference* (3rd ed.), Chapman and Hall, London, 1975.

Received August 30, 2011

Department of Applied Mathematics and Statistics
Faculty of Mathematics, Physics and Informatics
Comenius University
Mlynská dolina
SK–842-48 Bratislava 4
SLOVAKIA
E-mail: pazman@fmph.uniba.sk