Kybernetika

Andrej Pázman

Criteria for optimal design of small-sample experiments with correlated observations

Kybernetika, Vol. 43 (2007), No. 4, 453--462

Persistent URL: http://dml.cz/dmlcz/135787

KYBERNETIKA — VOLUME 43 (2007), NUMBER 4, PAGES 453 – 462

CRITERIA FOR OPTIMAL DESIGN OF SMALL–SAMPLE EXPERIMENTS WITH CORRELATED OBSERVATIONS

Andrej Pázman

We consider observations of a random process (or a random field), which is modeled by a nonlinear regression with a parametrized mean (or trend) and a parametrized covariance function. Optimality criteria for parameter estimation are to be based here on the mean square errors (MSE) of estimators. We mention briefly expressions obtained for very small samples via probability densities of estimators. Then we show that an approximation of MSE via Fisher information matrix is possible, even for small or moderate samples, when the errors of observations are normal and small. Finally, we summarize some properties of optimality criteria known for the noncorrelated case, which can be transferred to the correlated case, in particular a recently published concept of universal optimality.

Keywords: optimal design, correlated observations, random field, spatial statistics, information matrix

AMS Subject Classification: 62K05, 62M10

1. INTRODUCTION

We consider a regression model of the form

$$y\left(x_i\right) = \eta\left(\theta, x_i\right) + \varepsilon\left(x_i\right) \tag{1}$$

with the points $x_1, \ldots, x_N$ (=the design) taken from a set $\mathcal{X}$ (= the design space), and with an unknown vector parameter $\theta = (\theta_1, \ldots, \theta_p)^T$. The model is supposed to be without systematic errors (i. e. $\mathrm{E}(\varepsilon(x_i)) = 0$), and the variance-covariance structure of the observed variables $y(x_i)$

$$\mathrm{Cov}\left(y\left(x_i\right), y\left(x_j\right)\right) = C\left(x_i, x_j, \beta\right)$$

may depend on another unknown vector parameter $\beta = (\beta_1, \ldots, \beta_q)^T \in B$. We suppose that $\eta(\theta, x_i)$ and $C(x_i, x_j, \beta)$ are twice continuously differentiable on the interiors of the parameter spaces Θ or B.

The problem of optimal choice of the design $x_1, \ldots, x_N$ within such a model appears in several domain of applications: discretization of random processes, spatial statistics [4], computer experiments [13]. The aim is either to obtain a good prediction of the process or good estimates of parameters.

We concentrate upon the second aim when designs are compared according to some optimality criteria, which are functions of the mean squares error matrix of the estimators $\hat{\theta}, \hat{\beta}$,

$$\mathrm{MSE}_{\theta,\beta} = \mathrm{E}_{\theta,\beta}\left\{\left[\begin{pmatrix}\hat{\theta}\\ \hat{\beta}\end{pmatrix} - \begin{pmatrix}\theta\\ \beta\end{pmatrix}\right]\left[\begin{pmatrix}\hat{\theta}\\ \hat{\beta}\end{pmatrix} - \begin{pmatrix}\theta\\ \beta\end{pmatrix}\right]^T\right\}.$$

So the first problem is to express $\mathrm{MSE}_{\theta,\beta}$ in a computationally feasible form.

The problem is easy to solve in the particular case of a linear model

$$\eta(\theta, x_i) = f^T(x_i)\,\theta \tag{2}$$

with $\Theta = \mathbb{R}^p$, and with error covariances and variances not depending on β. The MSE of the minimum variance unbiased estimator $\hat{\theta} = M^{-1}F^T y$ is equal to its variance, $\mathrm{MSE}_\theta = \mathrm{Var}(\hat{\theta}) = M^{-1}$, where $M = F^T C^{-1} F$ is the information matrix, $F^T = (f(x_1), \ldots, f(x_N))$, $\{C\}_{ij} = \mathrm{Cov}(y(x_i), y(x_j))$, and y is the vector of observed variables. An optimality criterion is usually expressed as a function Φ of M (e. g. $\Phi(M) = -\ln\det(M)$ for the D-optimality criterion, $\Phi(M) = \mathrm{tr}(M^{-1})$ for the A-optimality criterion, etc.), and it does not depend on θ.

In the *nonlinear* regression model with *uncorrelated observations* it is standard to express the optimality criteria again as functions of the information matrix. This is justified by the fact, that in the uncorrelated case *replications* of observations are allowed, and *asymptotically* (for large numbers of replications), under some regularity conditions, the maximum likelihood estimators are asymptotically normally distributed, unbiased, and their variance matrix is equal to the inverse of the Fisher information matrix.

This argumentation can not be used in case of correlated observations except for some very special covariance functions (cf. [1]), since replication as a rule are not allowed. It fails totally when asymptotic approximations are not justified. So for small samples we have to proceed differently. Notice that we do not consider here the case of designing independent replications of the whole realization of the random process.

2. VERY SMALL SAMPLES: THE MSE BASED ON THE DENSITY OF ESTIMATORS

Here we consider a situation when the density of the MLE, $f(\hat{\theta}, \hat{\beta} \mid \theta, \beta)$, is known or well approximated. Then

$$\mathrm{MSE}_{\theta,\beta} = \int_\Theta \int_B \left[\begin{pmatrix}\hat{\theta}\\ \hat{\beta}\end{pmatrix} - \begin{pmatrix}\theta\\ \beta\end{pmatrix}\right]\left[\begin{pmatrix}\hat{\theta}\\ \hat{\beta}\end{pmatrix} - \begin{pmatrix}\theta\\ \beta\end{pmatrix}\right]^T f\left(\hat{\theta}, \hat{\beta} \mid \theta, \beta\right) \mathrm{d}\hat{\beta}\, \mathrm{d}\hat{\theta}.$$

The A-optimality criterion can be expressed as

$$tr\,(\mathrm{MSE}_{\theta,\beta}) = \int_{\Theta\times B} \left[\left\|\hat{\theta} - \theta\right\|^2 + \left\|\hat{\beta} - \beta\right\|^2\right] f\left(\hat{\theta}, \hat{\beta} \mid \theta, \beta\right) \mathrm{d}\hat{\beta}\, \mathrm{d}\hat{\theta}.$$

In [5] it is shown that also the D-optimality criterion, $\det(\mathrm{MSE}_{\theta,\beta})$, can be expressed as one multivariate integral, however with a much higher dimension. Such integral representations of optimality criteria are necessary to use methods of stochastic optimization for finding optimum designs numerically. However, because of complexity of such a procedure, it can be used only when the number of parameters and the number of observations are very small.

This approach is also restricted by the necessity to know $f(\hat{\theta}, \hat{\beta} \mid \theta, \beta)$. Until now we have realistically applicable expressions only for the case that β is known (i. e. $C(\beta) = C$) and that the errors are normal. Then for small dimensions of θ, the density of $\hat{\theta}$ on $\mathrm{int}(\Theta)$ is very well approximated by the expression (cf. [8] or [9])

$$q\left(\hat{\theta} \mid \theta\right) = \frac{\det\left[Q\left(\hat{\theta}, \theta\right)\right]}{(2\pi)^{p/2} \det^{1/2}\left[M\left(\hat{\theta}\right)\right]} \exp\left\{-\frac{1}{2}\left[\eta\left(\hat{\theta}\right) - \eta\left(\theta\right)\right]^T C^{-1} P^{\hat{\theta}}\left[\eta\left(\hat{\theta}\right) - \eta\left(\theta\right)\right]\right\}$$

where $\eta(\theta) = (\eta(\theta, x_1), \ldots, \eta(\theta, x_N))^T$, $M(\theta)$ is the Fisher information matrix, P^θ is a projector, and $Q\left(\hat{\theta}, \theta\right)$ is a modification of the observed Fisher information matrix:

$$\begin{aligned} P^\theta &= \frac{\partial \eta(\theta)}{\partial \theta^T} M^{-1}(\theta) \frac{\partial \eta^T(\theta)}{\partial \theta} C^{-1} \\ \left\{Q\left(\hat{\theta}, \theta\right)\right\}_{i,j} &= M\left(\hat{\theta}\right) + \left[\eta\left(\hat{\theta}\right) - \eta\left(\theta\right)\right]^T C^{-1}\left[I - P^{\hat{\theta}}\right] \frac{\partial^2 \eta(\theta)}{\partial \theta_i \partial \theta_j}\Big|_{\hat{\theta}}. \end{aligned}$$

Cf. [11] for the use of $q(\hat{\theta} \mid \theta)$ for obtaining A-optimal designs via the Kiefer–Wolfowitz stochastic optimization. The observations have been supposed uncorrelated, but for a correlated case the method is exactly the same. A method for dealing with that part of the probability distribution of $\hat{\theta}$ which is located on the boundary of Θ is explained in [11].

In case that the dimension of θ is higher, the expression $q(\hat{\theta} \mid \theta)$ must be corrected, in that instead of $\det[Q(\hat{\theta}, \theta)]$ we write an expression which is a polynomial in the components of $Q(\hat{\theta}, \theta)$ and of the components of the Riemannian curvature tensor of the expectation surface $\{\eta(\theta) : \theta \in \Theta\}$ (cf. [9]). In this more complicated case accelerated stochastic optimization methods must be applied (cf. [5]).

Although the presented approach of [5] gives very accurate approximations for MSE and for optimality criteria, it can be used only for rather small dimensions of θ (because of difficulties with the density of $\hat{\theta}$), and also for a rather small number of design points (because of the complexity of the stochastic approximation method). So it make sense to consider further approximations of MSE for small or moderate N.

3. THE FISHER INFORMATION MATRIX AND THE EXPONENTIAL REPRESENTATION OF THE MODEL

For a fixed design we write the nonlinear regression model (1) in a vector form

$$\begin{aligned} y &= \eta(\theta) + \varepsilon \\ \varepsilon &\sim \mathcal{N}(0, C(\beta)) \end{aligned} \tag{3}$$

where $y^T = (y(x_1), \ldots, y(x_N))$. We suppose that the mapping $\theta \in \Theta \to \eta(\theta) \in \mathbb{R}^N$ is one-to-one, and the $N \times N$ covariance matrix $C(\beta)$ with entries $C(x_i, x_j, \beta)$ is nonsingular. Suppose also that $\bar{\theta}$ and $\bar{\beta}$, the true values of θ and β, are points of the interiors $int(\Theta)$, resp. $int(B)$. We consider the MLE

$$\left(\hat{\theta}^T, \hat{\beta}^T\right)^T = \arg \max_{\theta \in \Theta, \beta \in B} \ln f\left(y \mid \theta, \beta\right)$$

where

$$-\ln f\left(y \mid \theta, \beta\right) = \frac{1}{2}\left\{\left[y - \eta\left(\theta\right)\right]^T C^{-1}\left(\beta\right)\left[y - \eta\left(\theta\right)\right] + \frac{1}{2}\ln\det\left[C\left(\beta\right)\right] + \frac{N}{2}\ln\left(2\pi\right)\right\}. \tag{4}$$

By taking derivatives we obtain that the Fisher information matrix of model (3) is (cf. [10] for details)

$$\begin{aligned} M\left(\theta, \beta\right) &= \mathrm{E}_{\theta,\beta}\left\{-\begin{pmatrix} \frac{\partial^2 \ln f(y|\theta,\beta)}{\partial\theta\partial\theta^T} & \frac{\partial^2 \ln f(y|\theta,\beta)}{\partial\theta\partial\beta^T} \\ \frac{\partial^2 \ln f(y|\theta,\beta)}{\partial\beta\partial\theta^T} & \frac{\partial^2 \ln f(y|\theta,\beta)}{\partial\beta\partial\beta^T} \end{pmatrix}\right\} \\ &= \begin{pmatrix} \frac{\partial\eta^T(\theta)}{\partial\theta} C^{-1}\left(\beta\right)\frac{\partial\eta(\theta)}{\partial\theta^T} & 0 \\ 0 & \frac{1}{2}tr\left\{C^{-1}\left(\beta\right)\frac{\partial C(\beta)}{\partial\beta} C^{-1}\left(\beta\right)\frac{\partial C(\beta)}{\partial\beta^T}\right\} \end{pmatrix}. \end{aligned} \tag{5}$$

For further analysis in Section 4 we write model (3) in the exponential family form, which will allow us to use standard expressions (6, 7 and 8) for the mean, variances and the Fisher information matrix. We have

$$\begin{aligned} \ln f\left(y \mid \theta, \beta\right) &= y^T C^{-1}\left(\beta\right)\eta\left(\theta\right) - \frac{1}{2}tr\left\{yy^T C^{-1}\left(\beta\right)\right\} \\ &\quad -\frac{1}{2}\eta^T\left(\theta\right) C^{-1}\left(\beta\right)\eta\left(\theta\right) - \frac{1}{2}\ln\det\left[C\left(\beta\right)\right] - \frac{N}{2}\ln\left(2\pi\right). \end{aligned}$$

Let us denote

$$\begin{aligned} t\left(y\right) &= \begin{pmatrix} t_1\left(y\right) \\ t_2\left(y\right) \end{pmatrix} = \begin{pmatrix} y \\ \mathrm{vec}\left(yy^T\right) \end{pmatrix} \\ \gamma\left(\theta, \beta\right) &= \begin{pmatrix} \gamma_1\left(\theta, \beta\right) \\ \gamma_2\left(\theta, \beta\right) \end{pmatrix} = \begin{pmatrix} C^{-1}\left(\beta\right)\eta\left(\theta\right) \\ -\frac{1}{2}\mathrm{vec}\left[C^{-1}\left(\beta\right)\right] \end{pmatrix}. \end{aligned}$$

The mapping $C \to \gamma_2 = -\frac{1}{2}\mathrm{vec}[C^{-1}]$ is one-to-one. So we can define a function

$$\kappa\left(\gamma\right) = \kappa\left(\gamma_1, \gamma_2\right) = \frac{1}{2}\ln\det\left(C\right) + \frac{1}{2}\gamma_1^T C\gamma_1 + \frac{N}{2}\ln\left(2\pi\right)$$

with C depending on γ_2. With this notation we obtain

$$f\left(y \mid \theta, \beta\right) = \exp\left\{t^T\left(y\right)\gamma\left(\theta, \beta\right) - \kappa\left[\gamma\left(\theta, \beta\right)\right]\right\}.$$

Hence $\{f(y \mid \theta, \beta) : \theta \in \Theta, \beta \in B\}$ is an exponential family, $t(y)$ is a sufficient statistics, and $\gamma(\theta, \beta)$ is the canonical function (cf. [3]). Important here are the

following known relations: the mean and the variance of $t(y)$ in an exponential family are equal to

$$\mathrm{E}_{\theta,\beta}\left[t\left(y\right)\right] \equiv \mu\left(\theta,\beta\right) = \left[\frac{\partial\kappa\left(\gamma\right)}{\partial\gamma}\right]_{\gamma=\gamma(\theta,\beta)} \tag{6}$$

$$\mathrm{Var}_{\theta,\beta}\left[t\left(y\right)\right] = \left[\frac{\partial^2\kappa\left(\gamma\right)}{\partial\gamma\partial\gamma^T}\right]_{\gamma=\gamma(\theta,\beta)}. \tag{7}$$

Moreover, the Fisher information matrix (5) can be expressed equivalently in the form

$$M\left(\theta,\beta\right) = \begin{pmatrix} \frac{\partial\gamma^T(\theta,\beta)}{\partial\theta} \\ \frac{\partial\gamma^T(\theta,\beta)}{\partial\beta} \end{pmatrix} \left[\frac{\partial^2\kappa\left(\gamma\right)}{\partial\gamma\partial\gamma^T}\right]_{\gamma=\gamma(\theta,\beta)} \begin{pmatrix} \frac{\partial\gamma(\theta,\beta)}{\partial\theta^T} & \frac{\partial\gamma(\theta,\beta)}{\partial\beta^T} \end{pmatrix}. \tag{8}$$

4. APPROXIMATION OF MLE AND MSE WHEN THE VARIANCES OF THE OBSERVED VARIABLES ARE SMALL

When the variances of the observed variables $y(x_i)$ are small, then the variances of all components of $t(y)$ are small as well. Indeed, we have just to consider the components of $t_2(y)$. In an abbreviated notation we obtain

$$\begin{aligned} \mathrm{Var}\left[y_iy_j\right] &= \mathrm{E}\left[y_iy_j - C_{ij} - \eta_i\eta_j\right]^2 \\ &= \mathrm{E}\left[\varepsilon_i\varepsilon_j + \varepsilon_i\eta_j + \varepsilon_j\eta_i - C_{ij}\right]^2 \\ &= \mathrm{E}\left[\varepsilon_i^2\varepsilon_j^2\right] + C_{ii}\eta_j^2 + C_{jj}\eta_i^2 + C_{ij}\eta_i\eta_j \end{aligned}$$

and by the Schwarz inequality we have $\mathrm{E}^2[\varepsilon_i^2\varepsilon_j^2] \leq \mathrm{E}[\varepsilon_i^4]\mathrm{E}[\varepsilon_j^4] = 9C_{ii}^2C_{jj}^2$, $|C_{ij}|^2 \leq C_{ii}C_{jj}$. So the variances of all components of $t(y)$ tend to zero with the same spead as the variances of the observed y_i.

The MLE can be expressed as a function of the sufficient statistics $t = t\left(y\right)$

$$\begin{pmatrix} \hat{\theta} \\ \hat{\beta} \end{pmatrix} = \arg\max_{\theta\in\Theta,\beta\in B}\left\{t^T\gamma\left(\theta,\beta\right) - \kappa\left[\gamma\left(\theta,\beta\right)\right]\right\}. \tag{9}$$

The domain where this estimator is defined is equal to

$$T = \left\{t = \begin{pmatrix} y \\ \mathrm{vec}\left(yy^T\right) \end{pmatrix} : y \in \mathbb{R}^N\right\}.$$

We define

$$T^* = \left\{t = \begin{pmatrix} y \\ \mathrm{vec}\left(Z\right) \end{pmatrix} : y \in \mathbb{R}^N, Z \in \mathbb{R}^{N\times N} \text{ and positive semidefinite}\right\}$$

and we denote by $\begin{pmatrix} \tilde{\theta}(t) \\ \tilde{\beta}(t) \end{pmatrix}$ the extension of $\begin{pmatrix} \hat{\theta}(t) \\ \hat{\beta}(t) \end{pmatrix}$ from T to T^*

$$\begin{aligned} \begin{pmatrix} \tilde{\theta}(t) \\ \tilde{\beta}(t) \end{pmatrix} &= \arg \max_{\theta\in\Theta,\beta\in B} \{y^T C^{-1}(\beta)\eta(\theta) - \frac{1}{2} tr\left[ZC^{-1}(\beta)\right] \\ &\quad -\frac{1}{2}\eta^T(\theta) C^{-1}(\beta)\eta(\theta) - \frac{1}{2}\ln\det\left[C(\beta)\right]\} \\ &= \arg \max_{\theta,\beta} \left\{t\gamma(\theta,\beta) - \kappa\left[\gamma(\theta,\beta)\right]\right\}; \quad t \in T^*. \end{aligned} \tag{10}$$

Notice that this is just a mapping, not an estimator. The idea is to express it as a Taylor expansion around the point

$$\bar{\mu} = \mu\left(\bar{\theta},\bar{\beta}\right) = \begin{pmatrix} \mathrm{E}_{\bar{\theta},\bar{\beta}}(y) \\ \mathrm{vec}\left[\mathrm{E}_{\bar{\theta},\bar{\beta}}\left(yy^T\right)\right] \end{pmatrix} = \begin{pmatrix} \eta\left(\bar{\theta}\right) \\ \mathrm{vec}\left[C\left(\bar{\beta}\right) + \eta\left(\bar{\theta}\right)\eta^T\left(\bar{\theta}\right)\right] \end{pmatrix}.$$

So we have

$$\begin{aligned} \tilde{\theta}(t) &= \tilde{\theta}\left[\bar{\mu}\right] + \frac{\partial\tilde{\theta}(t)}{\partial t^T}\Big|_{t=\bar{\mu}}(t-\bar{\mu}) \\ &\quad + \frac{1}{2}(t-\bar{\mu})^T \left[\frac{\partial^2\tilde{\theta}(t)}{\partial t\partial t^T}\right]_{t=q}(t-\bar{\mu}) \end{aligned}$$

and similarly for $\tilde{\beta}(t)$. Here q is a point between t and $\bar{\mu}$. Since $\tilde{\theta}(t)$ is an extension of $\hat{\theta}(t)$ we can write for $t \in T$

$$\hat{\theta}(t) \doteq \tilde{\theta}\left[\bar{\mu}\right] + \frac{\partial\tilde{\theta}(t)}{\partial t^T}\Big|_{t=\bar{\mu}}(t-\bar{\mu}).$$

We neglected the term quadratic in t since the variances of the components of the statistics t are small. Similarly

$$\hat{\beta}(t) \doteq \tilde{\beta}\left[\bar{\mu}\right] + \frac{\partial\tilde{\beta}(t)}{\partial t^T}\Big|_{t=\bar{\mu}}(t-\bar{\mu}).$$

We can prove that

$$\tilde{\theta}\left[\bar{\mu}\right] = \bar{\theta}, \quad \tilde{\beta}\left[\bar{\mu}\right] = \bar{\beta} \tag{11}$$

$$\begin{pmatrix} \frac{\partial\tilde{\theta}(t)}{\partial t^T} \\ \frac{\partial\tilde{\beta}(t)}{\partial t^T} \end{pmatrix}_{t=\bar{\mu}} = M^{-1}\left(\bar{\theta},\bar{\beta}\right) \begin{pmatrix} \frac{\partial\gamma^T}{\partial\theta} \\ \frac{\partial\gamma^T}{\partial\beta} \end{pmatrix}_{\bar{\theta},\bar{\beta}}. \tag{12}$$

Indeed, using the notation $\delta^T = (\theta^T, \beta^T)$ we write (10) in the form

$$\tilde{\delta}(t) = \arg\max_{\delta} \left\{t\,\gamma(\delta) - \kappa\left[\gamma(\delta)\right]\right\}.$$

We take the derivative of $\{t\,\gamma(\delta) - \kappa[\gamma(\delta)]\}$ and use (6)

$$\left[t - \mu\left(\tilde{\delta}(t)\right)\right]^T \frac{\partial\gamma(\delta)}{\partial\delta^T}\Big|_{\tilde{\delta}(t)} = 0$$

and put $t = \mu\left(\bar{\delta}\right) = \bar{\mu}$ to obtain (11). Taking the derivative once more but with respect to t we obtain

$$\left[I - \frac{\partial \tilde{\delta}^T(t)}{\partial t} \frac{\partial \mu^T(\delta)}{\partial \delta}\Big|_{\tilde{\delta}(t)}\right] \frac{\partial \gamma(\delta)}{\partial \delta^T}\Big|_{\tilde{\delta}(t)} + \left[t - \mu\left(\tilde{\delta}(t)\right)\right]^T \frac{\partial^2 \gamma(\delta)}{\partial \delta \partial \delta^T}\Big|_{\tilde{\delta}(t)} \frac{\partial \tilde{\delta}(t)}{\partial t} = 0.$$

The second term is zero if $t = \mu(\bar{\delta})$. By the implicit function theorem ([14], p. 41) we have that $\frac{\partial \tilde{\delta}^T(t)}{\partial t}$ is the solution of this equation, hence

$$\frac{\partial \tilde{\delta}^T(t)}{\partial t}\Big|_{t=\mu\left(\bar{\delta}\right)} = \frac{\partial \gamma(\delta)}{\partial \delta^T}\Big|_{\bar{\delta}} M^{-1}\left(\bar{\delta}\right)$$

since from (8) it follows that $M(\delta) = \frac{\partial \mu^T(\delta)}{\partial \delta} \frac{\partial \gamma(\delta)}{\partial \delta^T}$. This proves (12) (cf. [10] for more details).

So we obtain that in case of small variances of $y(x_i)$ the approximate expression for the MLE is

$$\begin{pmatrix} \hat{\theta} \\ \hat{\beta} \end{pmatrix} \doteq \begin{pmatrix} \bar{\theta} \\ \bar{\beta} \end{pmatrix} + M^{-1}\left(\bar{\theta}, \bar{\beta}\right) \begin{pmatrix} \frac{\partial \gamma^T}{\partial \theta} \\ \frac{\partial \gamma^T}{\partial \beta} \end{pmatrix}_{\bar{\theta}, \bar{\beta}} (t - \bar{\mu}). \tag{13}$$

This gives

$$\mathrm{E}_{\bar{\theta}, \bar{\beta}}\left[\begin{pmatrix} \hat{\theta} \\ \hat{\beta} \end{pmatrix}\right] \doteq \begin{pmatrix} \bar{\theta} \\ \bar{\beta} \end{pmatrix}$$

$$\begin{aligned} \mathrm{Var}_{\bar{\theta}, \bar{\beta}}\left[\begin{pmatrix} \hat{\theta} \\ \hat{\beta} \end{pmatrix}\right] &\doteq M^{-1}\left(\bar{\theta}, \bar{\beta}\right) \begin{pmatrix} \frac{\partial \gamma^T}{\partial \theta} \\ \frac{\partial \gamma^T}{\partial \beta} \end{pmatrix}_{\bar{\theta}, \bar{\beta}} \mathrm{Var}_{\bar{\theta}, \bar{\beta}}(t) \left(\frac{\partial \gamma}{\partial \theta^T} \frac{\partial \gamma}{\partial \beta^T}\right) M^{-1}\left(\bar{\theta}, \bar{\beta}\right) \\ &= M^{-1}\left(\bar{\theta}, \bar{\beta}\right) \end{aligned}$$

where we used (7) and (8). Hence within this approximation $\mathrm{MSE}_{\bar{\theta}, \bar{\beta}} = M^{-1}(\bar{\theta}, \bar{\beta})$ with $M(\bar{\theta}, \bar{\beta})$ given by (5). Notice that this does not mean that $\hat{\beta}$ is approximately normally distributed, although $\hat{\beta}$ is expressed as a linear function of t, since by definition t is a quadratic function of the observed variables $y(x_i)$.

Summarizing, in case that the errors are normally distributed with sufficiently small variances, the mean square error matrix of MLE is approximately equal to the inverse of the information matrix even for small samples. We can apply criteria functions Φ like in the linear model, just the resulting criteria depend on $\bar{\theta}, \bar{\beta}$. For design purposes we do not interpret $\bar{\theta}, \bar{\beta}$ as the true parameter values, but as some parameter values taken ad hoc, and we suppose that the true parameter values are in a neighborhood of $\bar{\theta}, \bar{\beta}$. As known, this "local" feature of optimality criteria is unavoidable in nonlinear models.

5. SOME BASIC PROPERTIES OF OPTIMALITY CRITERIA AND THE CRITERION OF UNIVERSAL OPTIMALITY

Optimality criteria in linear models can be derived from geometrical properties of the confidence ellipsoid for θ. This is not possible here, since $\hat{\beta}$ is not distributed normally, even within the considered approximation, and confidence regions are not ellipsoids. However, still remains the interpretation through the variance matrices of $\hat{\theta}$ and $\hat{\beta}$, and according to the results of Section 4, a criterion can be still be expressed as a function $\Phi[M]$ of the information matrix $M = M(\bar{\theta}, \bar{\beta})$. Since this matrix depends also on the design, say $A = \{x_1, \ldots, x_N\}$, we write it sometimes as $M(A; \bar{\theta}, \bar{\beta})$.

The aim of this section is to summaries known properties of criteria functions Φ which can be transferred (eventually after some minor changes) from the linear model (2) with uncorrelated observations and with allowed replications, to model (3) allowing no replications.

A good design should give a small variance matrix, therefore traditionally, in most books on experimental design, the function Φ is related to the variance matrix, and it is antiisotonic, i. e. if $M^* - M$ is p.s.d., then $\Phi[M^*] \leq \Phi[M]$ (since the variances are $[M^*]^{-1}$ and $[M]^{-1}$). Alternatively, as pointed out in [12], criteria should be "information criteria" i. e. they should have following properties:

i) nonnegativity: $\Phi(M) \geq 0$,

ii) isotonicity $M^* - M = p.s.d. \Rightarrow \Phi[M^*] \geq \Phi[M]$

iii) positive homogeneity: $\Phi[kM] = k\Phi[M]$; $k > 0$

iv) superadditivity: $\Phi[M + M^*] \geq \Phi[M] + \Phi[M^*]$.

For example, $\Phi[M] = -\ln\det[M]$, or $\Phi[M] = \mathrm{tr}[M^{-1}]$ are antiisotonic forms of the criteria of D- or of A-optimality, $\Phi[M] = \ln\det[M]$ is an isotonic form of the criterion of D-optimality, which is not homogeneous, and $\Phi[M] = [\det(M)]^{1/(p+q)}$ or $\Phi[M] = 1/\mathrm{tr}[M^{-1}]$ are isotonic, homogeneous and concave (superadditive) forms of the criteria of D- or A-optimality. Notice that we consider the two functions $\ln\det[M]$, and $[\det(M)]^{1/(p+q)}$ as different forms of the same criterion, since they induce the same ordering of information matrices.

A direct consequence of these properties is that Φ is concave (cf. [12]). The properties i) – iv) are important to define with a proper scaling the relative efficiency of an experiment (or a design with the matrix M) with respect to another reference experiment with M^*

$$eff_\Phi[M \mid M^*] = \frac{\Phi[M]}{\Phi[M^*]}. \tag{14}$$

The information matrix M^* is used to be "the largest in the given situation". Standardly one takes in the linear model with replications

$$M^* = \arg\max_M \Phi[M] \tag{15}$$

where M^* is computed by convex methods. We can not do this in model (3), so we propose to take $M^* = M(\mathcal{X}; \bar{\theta}, \bar{\beta})$, since the largest possible information is obtained when we observe the whole process. (Technical problems connected with the definition of $M(\mathcal{X}; \bar{\theta}, \bar{\beta})$ evidently disappear when $\mathcal{X}$ is a finite set.)

The choice of a suitable optimality criterion is sometime ambiguous, and we would like to have designs which are "quite good" with respect to a class of optimality criteria. One can speak about "universal optimality", when this class is very large. Such a class is evidently the class $\mathcal{K}$ of all criteria Φ which have properties i) – iv), and which are orthogonally invariant, i. e. such that

$$\Phi(M) = \Phi\left(UMU^T\right)$$

for every orthogonal matrix U. Not only the D- and A-optimality criterion belongs to this class, but also all criteria commonly used in case that we want to estimate all parameters θ_i and β_j. The "criterion of universal optimality" related to the class $\mathcal{K}$ is equal to "the worst efficiency in the class $\mathcal{K}$"

$$\Psi\left[M\left(A; \bar{\theta}, \bar{\beta}\right)\right] = \inf_{\Phi \in \mathcal{K}} \frac{\Phi\left(M\left(A; \bar{\theta}, \bar{\beta}\right)\right)}{\Phi\left(M\left(\mathcal{X}; \bar{\theta}, \bar{\beta}\right)\right)}.$$

However, to deal directly with such a complex criterion is impossible. Surprisingly, we have the following fundamental result

$$\inf_{\Phi \in \mathcal{K}} \frac{\Phi\left(M\left(A; \bar{\theta}, \bar{\beta}\right)\right)}{\Phi\left(M\left(\mathcal{X}; \bar{\theta}, \bar{\beta}\right)\right)} = \min_{1 \le k \le p+q} \frac{\Phi_{\mathrm{E}_k}\left(M\left(A; \bar{\theta}, \bar{\beta}\right)\right)}{\Phi_{\mathrm{E}_k}\left(M\left(\mathcal{X}; \bar{\theta}, \bar{\beta}\right)\right)} \tag{16}$$

where

$$\Phi_{\mathrm{E}_k}(M) = \sum_{i=1}^{k} \lambda_i(M)$$

is the sum of k minimal eigenvalues of the matrix M. (We remind that $M(A; \bar{\theta}, \bar{\beta})$ is a $(p+q) \times (p+q)$ matrix.) As a consequence, instead of considering the extremely large class $\mathcal{K}$ we have to consider a finite number of criteria $\Phi_{\mathrm{E}_k}(M)$,where evidently $\Phi_{\mathrm{E}_1}(M)$ is the well known criterion of E-optimality. Such a result has been first time proved in [6], Theorem 6, in the context of design in linear experiments with uncorrelated observation, i. e. using the definition (15). However, if we go carefully through the proof of the "auxiliary" Theorem 5 in [6], we see that it works for any positive definite matrix M, so the inner structure of the information matrix is irrelevant, and the result (16) is obtained straightway from [6] also in a model without replications and with correlated observations.

We end by a brief remark about potential possibilities to compute a design which is (nearly) optimum with respect to a given criterion. Since replications of observations are not allowed, we can not apply convex methods of optimal design, which are known from experiments with uncorrelated observations. But it seems that we can apply without essential difficulties some methods known for linear models with correlated observations, like the method of [2] (cf. [15] for a corresponding exchange method) or the method of virtual noise (cf. [7]). More details on the last one extended to the setup of the present paper are given in [10].

ACKNOWLEDGEMENT

This work was supported by the VEGA-grant No. 1/3016/06 and by the project APVV No. SK-AT-01206.

(Received February 1, 2006.)

REFERENCES

[1] M. Apt and W. J. Welch: Fisher information and maximum likelihood estimation of covariance parameters in Gaussian stochastic processes. Canad. J. Statist. *26* (1998), 127–137.
[2] U. N. Brimkulov, G. K. Krug, and V. L. Savanov: Design of Experiments in Investigating Random Fields and Processes. Nauka, Moscow 1986.
[3] L. D. Brown: Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. (Vol. 9 of Institute of Mathematical Statistics Lecture Notes – Monograph Series.) Institute of Mathematical Statistics, Hayward 1986.
[4] N. A. C. Cresie: Statistics for Spatial Data. Wiley, New York 1993.
[5] J. P. Gauchi and A. Pázman: Design in nonlinear regression by stochastic minimization of functionals of the mean square error matrix. J. Statist. Plann. Inference *136* (2006), 1135–1152.
[6] R. Harman: Minimal efficiency of designs under the class of orthogonally invariant information criteria. Metrika *60* (2004), 137–153.
[7] W. G. Müller and A. Pázman: An algorithm for computation of optimum designs under a given covariance structure. Comput. Statist. *14* (1999), 197–211.
[8] A. Pázman: Probability distribution of the multivariate nonlinear least squares estimates. Kybernetika *20* (1984), 209–230.
[9] A. Pázman: Nonlinear Statistical Models. Kluwer, Dordrecht – Boston 1993.
[10] A. Pázman: Correlated Optimum Design with Parametrized Covariance Function: Justification of the Use of the Fisher Information Matrix and of the Method of Virtual Noise. Research Report No. 5, Institut für Statistik, WU Wien, Vienna 2004.
[11] A. Pázman and L. Pronzato: Nonlinear experimental design based on the distribution of estimators. J. Statist. Plann. Inference *33* (1992), 385–402.
[12] F. Pukelsheim: Optimal Design of Experiments. Wiley, New York 1993.
[13] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn: Design and analysis of computer experiments. Statist. Sci. *4* (1989), 409–435.
[14] M. Spivak: Calculus on Manifolds. W. A. Benjamin, Inc., Menlo Park, Calif. 1965.
[15] D. Uciński and A. C. Atkinson: Experimental design for time-dependent models with correlated observations. Stud. Nonlinear Dynamics & Econometrics *8* (2004), Issue 2, Article 13.

Andrej Pázman, Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská dolina, 842 48 Bratislava. Slovak Republic.
e-mail: pazman@fmph.uniba.sk