Andrej Pázman
Information contained in design points of experiments with correlated observations

Kybernetika, Vol. 46 (2010), No. 4, 771--783

Persistent URL: http://dml.cz/dmlcz/140783

KYBERNETIKA — VOLUME 46 (2010), NUMBER 4, PAGES 771–783

INFORMATION CONTAINED IN DESIGN POINTS OF EXPERIMENTS WITH CORRELATED OBSERVATIONS

Andrej Pázman

A random process (field) with given parametrized mean and covariance function is observed at a finite number of chosen design points . The information about its parameters is measured via the Fisher information matrix (for normally distributed observations) or using information functionals depending on that matrix. Conditions are stated, under which the contribution of one design point to this information is zero. Explicit expressions are obtained for the amount of information coming from a selected subset of a given design. Relations to some algorithms for optimum design of experiments in case of correlated observations are indicated.

Keywords: optimal sampling design, spatial statistics, random process, nonlinear regression, information matrix

Classification: 62K05, 62M30, 62B15

1. INTRODUCTION

We observe a random process (or field)

$$\{y_x : x \in \mathcal{X}\} \,, \tag{1}$$

which has a parametrized mean $E(y_x) = \eta(x, \theta)$ and a covariance function $Cov(y_x, y_z) = C(x, z; \gamma)$, respectively. The symbol x and/or z denotes the time in case of a process, or the vector of space coordinates in spatial problems. Observations (without replications) are performed at a finite number of points from $\mathcal{X}$. The aim is to estimate the unknown parameters $\theta = (\theta_1, \ldots, \theta_p)^T \in \Theta$, and $\gamma = (\gamma_1, \ldots, \gamma_q)^T \in \Gamma$. If N is the prescribed number of observations, and $D = \{x_1, \ldots, x_N\}$ is the set of these points, then D is called a design (of size N), and points of D are design points. A standard problem is to find a design with a fixed size N that yields the most precise estimators of θ and γ. One can refer to [2, 6, 13] for a motivation in spatial statistics. In this paper, we would like to contribute indirectly to better understanding of the problem.

In fact, when the design D is fixed we have to consider a regression experiment with a finite number of scalar observations $y_{x_i} : i = 1, \ldots, N$, which are correlated. This makes the detection of the influence of individual observations y_{x_i} to the final

amount of information about θ and γ much more complicated than in the uncorrelated case. So one way to clarify what happens when we try to compute the optimal design, is to find this amount of information. This in fact is the main aim of this paper.

In order to make clear that the situation is much simpler when the observations are uncorrelated, consider a linear experiment with normal uncorrelated observations $y_x = g(x)\,\theta + \varepsilon_x$ with $g(x)$ a given coefficient depending on the design point x and with $\theta \in \mathbb{R}^1$ the unknown parameter. The experiment is performed according to a design D, and $C(x,x) = Var(\varepsilon_x) = \gamma \in \mathbb{R}^+$, $C(x,z) = 0$ if $x \neq z$. The Fisher information matrix is diagonal (see Section 2), with diagonal elements

$$\begin{aligned} M_{\mathrm{I}}(D) &= \frac{1}{\gamma}\sum_{x\in D} g^2(x)\ , \\ M_{\mathrm{II}}(D) &= \sum_{x\in D}\frac{1}{2\gamma^2}\ . \end{aligned}$$

So, the information content of the observation at one point x is equal to $\frac{1}{\gamma}g^2(x)$ or to $\frac{1}{2\gamma^2}$, and it is independent on the design D. Moreover, in experiments with uncorrelated observations we can consider replication of observations at individual points $x \in \mathcal{X}$, which leads to well known convex methods in optimum experimental design, which are not applicable when observing a random process.

2. MEASURES OF INFORMATION AND NOTATION

To measure the amount of information obtained from a design $D = \{x_1, \ldots, x_N\}$ we use the (Fisher) information matrix based on normal density $f(y \mid D, \theta, \gamma)$ with mean $(\eta(x_1,\theta), \ldots, \eta(x_N,\theta))^T$ and covariance matrix $\{C(x_i, x_j; \gamma)\}_{x_i\in D, x_j \in D}$. It is equal to

$$\begin{aligned} \mathbf{M}(D) &\equiv \mathbf{M}(D,\theta,\gamma) \\ &= -E_{\theta,\gamma}\left\{ \begin{pmatrix} \frac{\partial^2 \ln f(y|D,\theta,\gamma)}{\partial\theta\partial\theta^T} & \frac{\partial^2 \ln f(y|D,\theta,\gamma)}{\partial\theta\partial\gamma^T} \\ \frac{\partial^2 \ln f(y|D,\theta,\gamma)}{\partial\gamma\partial\theta^T} & \frac{\partial^2 \ln f(y|D,\theta,\gamma)}{\partial\gamma\partial\gamma^T} \end{pmatrix} \right\} \\ &= \begin{pmatrix} \mathbf{M}_{\mathrm{I}}(D) & 0 \\ 0 & \mathbf{M}_{\mathrm{II}}(D) \end{pmatrix} \end{aligned}$$

with

$$\begin{aligned} \mathbf{M}_{\mathrm{I}}(D) &= \sum_{x,z\in D} \mathbf{f}(x)\left\{\mathbf{C}^{-1}(D)\right\}_{x,z}\mathbf{f}^T(z) \\ &= \sum_{x,z\in D} \mathbf{f}(x)\left\{\mathbf{G}\right\}_{x,z}\mathbf{f}^T(z)\ , \\ \left\{\mathbf{M}_{\mathrm{II}}(D)\right\}_{ij} &= \frac{1}{2}tr\left\{\mathbf{C}^{-1}(D)\frac{\partial\mathbf{C}(D)}{\partial\gamma_i}\mathbf{C}^{-1}(D)\frac{\partial\mathbf{C}(D)}{\partial\gamma_j}\right\} \\ &= \frac{1}{2}tr\left\{\mathbf{G}\frac{\partial\mathbf{C}}{\partial\gamma_i}\mathbf{G}\frac{\partial\mathbf{C}}{\partial\gamma_j}\right\}\ . \end{aligned}$$

Here we used the abbreviated notation

$$\begin{aligned}
\mathbf{f}(x) &\equiv \mathbf{f}(x,\theta) = \frac{\partial \eta(x,\theta)}{\partial \theta} \text{ (a column vector from } \mathbb{R}^p) \,, \\
\mathbf{C}(A,B) &= \{C(x,z,\gamma)\}_{x\in A,\, z\in B}\,;\; A\subset D,\; B\subset D \,, \\
\mathbf{C}(A) &\equiv \mathbf{C}(A,A) \,, \\
\mathbf{C} &\equiv \mathbf{C}(D) \,, \\
\mathbf{G} &\equiv \mathbf{G}(D) = [\mathbf{C}(D)]^{-1} \,.
\end{aligned}$$

We omitted θ or γ explicitly, since the information is expressed for fixed θ and γ. We also assume that $\eta(x,\theta)$ and $C(x,z,\gamma)$ are differentiable with respect to θ and γ, and that $\mathbf{C} = \mathbf{C}(D)$ is nonsingular. Hence $\mathbf{C}(A)$ is nonsingular for every $A \subset D$.

The importance of the information matrix in statistics is undoubtfully justified. Moreover, although replications are not allowed, and so asymptotic considerations cannot be done, in case that the error variances $Var(\varepsilon_x)$ are small, the variance matrix of the MLE for θ and γ is well approximated by the inverse of $\mathbf{M}(D,\theta,\gamma)$ (cf. [10]).

We also need measures of information that are one dimensional (scalars). Like in experiments with uncorrelated observations we shall consider information functionals, which are concave, monotone, real-valued functions defined on the set of positive definite matrices (cf. [11] for a justification of these properties in case of uncorrelated observations). The gradient of such a functional Φ is the matrix $\nabla\Phi[\mathbf{M}]$ of the same dimension as M, with components

$$\{\nabla\Phi[\mathbf{M}]\}_{ij} = \frac{\partial\Phi[\mathbf{M}]}{\partial\{\mathbf{M}\}_{ij}} \,.$$

Well known examples are the D-optimality functional $\Phi[\mathbf{M}] = \ln\det(\mathbf{M})$ with $\nabla \ln\det(\mathbf{M}) = \mathbf{M}^{-1}$, or the A-optimality functional $\Phi[\mathbf{M}] = -tr\left(\mathbf{M}^{-1}\right)$ with $\nabla\left[-tr\left(\mathbf{M}^{-1}\right)\right] = \mathbf{M}^{-2}$.

3. DESIGN POINTS GIVING ZERO INFORMATION

Even when the covariance function is known and the model is linear in θ, $y_x = \mathbf{f}^T(x)\,\theta + \varepsilon_x$, it is not quite transparent which design points give zero information. Intuitively, one can perhaps argue that $\mathbf{f}(x) = \mathbf{0}$ implies that y_x is not influenced by the value of θ, hence should give no information about θ. This intuitive approach fails unless the observations are uncorrelated. To see this, consider an example. Suppose that $\theta \in \mathbb{R}$, take $D = \{x,z\}$ a two point design such that $f(x) = 0$, $f(z) = 1$, and suppose that $C(x,x) = C(z,z) = 1$, but $C(x,z) \neq 0$. Then

$$\begin{aligned}
M_{\mathrm{I}}(D) &= (0,1)\begin{pmatrix} 1 & C(x,z) \\ C(x,z) & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0 \\ 1 \end{pmatrix} \\
&= \frac{1}{1-[C(x,z)]^2} > 1 = M_{\mathrm{I}}(\{z\}) \,.
\end{aligned}$$

So although $f(x) = 0$, by deleting the point x we can lose much information. The contribution of the point x to $M_{\mathrm{I}}(D)$ is very large when y_x and y_z are highly correlated.

An alternative approach is indicated in the following lemma, which is valid for the general model (1).

Lemma 1. (cf. Näther [9]) *If (for some fixed θ and γ) there is a set $A \subset D$ and some vectors $\mathbf{a}(z) \in \mathbb{R}^p$; $z \in A$ such that for every $x \in D$*

$$\mathbf{f}(x) = \sum_{z \in A} C(x,z)\, \mathbf{a}(z) \ . \tag{2}$$

then all points belonging to $D - A$ provide zero contribution to $\mathbf{M}_{\mathrm{I}}(D)$, i. e.

$$\mathbf{M}_{\mathrm{I}}(A) = \mathbf{M}_{\mathrm{I}}(D) \ .$$

A simple p r o o f is at hand. We have

$$\begin{aligned}
\mathbf{M}_{\mathrm{I}}(D) &= \sum_{x,z \in D} \mathbf{f}(x) \{\mathbf{G}\}_{x,z} f^T(z) \\
&= \sum_{u,v \in A} \sum_{x,z \in D} \mathbf{a}(u)\, C(u,x) \left\{\mathbf{C}^{-1}\right\}_{x,z} C(z,v)\, \mathbf{a}^T(v) \\
&= \sum_{u,v \in A} \mathbf{a}(u)\, C(u,v)\, \mathbf{a}^T(v) \\
&= \sum_{u,v \in A} \mathbf{f}(u) \left\{C^{-1}(A)\right\}_{u,v} \mathbf{f}^T(v) = \mathbf{M}_{\mathrm{I}}(A) \ .
\end{aligned}$$

Hence points from D, which are outside A, give zero information about the mean of the process. □

The following example is given in [9]: Let $y_x = \theta_1 + \theta_2 x + \varepsilon_x$; $\quad x \in \langle -1, 1 \rangle$ be a linear model with covariance function

$$\begin{aligned}
C(x,z) &= 1 - |x - z| \quad \text{if } |x - z| < 1 \ , \\
C(x,z) &= 0 \qquad \text{if } |x - z| \geq 1 \ .
\end{aligned}$$

The 3-points design $A = \{-1, 0, 1\}$ is shown to give the same information as if the whole process were observed at all points of $\langle -1, 1 \rangle$.

For further use we reformulate the result of Lemma 1. Define for every $x \in D$ a column vector $\mathbf{a}(x)$

$$\mathbf{a}(x) = \sum_{z \in D} \{\mathbf{G}\}_{x,z} \mathbf{f}(z) \ . \tag{3}$$

Then (2) holds if and only if $\mathbf{a}(x) = \mathbf{0}$ for every $x \in D - A$. So we obtain

Corollary of Lemma 1. *If* $\mathbf{a}(x)$ *is defined by (3) and* $\mathbf{a}(x_o) = 0$, *then* x_o *provides zero contribution to* $\mathbf{M}_{\mathrm{I}}(D)$.

Now we present a result similar to Lemma 1, but for $\mathbf{M}_{\mathrm{II}}(D)$.

First, for every $x, z \in D$ define column vectors $\alpha(x,z) \in \mathbb{R}^q$

$$\alpha(x,z) = -\frac{\partial \{\mathbf{G}\}_{x,z}}{\partial \gamma} = \{\mathbf{G}\}_{x,\cdot} \frac{\partial \mathbf{C}}{\partial \gamma} \{\mathbf{G}\}_{\cdot,z} \quad .$$

Then we have

$$\frac{\partial C(x,z)}{\partial \gamma} = \sum_{u,v \in D} C(x,u)\, \alpha(u,v)\, C(v,z) \ .$$

Lemma 2. *If there is a set* $A \subset D$ *such that* $\alpha(t,z) = 0$ *for every* $t \in D - A$, $z \in D$, *then for every* $x \in D$, $z \in D$ we have

$$\frac{\partial C(x,z)}{\partial \gamma} = \sum_{u,v \in A} C(x,u)\, \alpha(u,v)\, C(v,z) \tag{4}$$

and

$$\mathbf{M}_{\mathrm{II}}(A) = \mathbf{M}_{\mathrm{II}}(D) \ .$$

Proof.

$$\begin{aligned}
\mathbf{M}_{\mathrm{II}}(D) &= \frac{1}{2} tr \left\{ \mathbf{G} \frac{\partial \mathbf{C}}{\partial \gamma} \mathbf{G} \frac{\partial \mathbf{C}}{\partial \gamma^T} \right\} \\
&= \frac{1}{2} tr \left\{ \sum_{u,v,t,s \in A} \mathbf{G} C(.,u)\, \alpha(u,v)\, C(v,.)\, \mathbf{G} C(.,t)\, \alpha^T(t,s)\, C(s,.) \right\} \\
&= \frac{1}{2} \sum_{u,v,t,s \in A} \alpha(u,v)\, C(v,t)\, \alpha^T(t,s)\, C(s,u) \\
&= \frac{1}{2} tr \left\{ [\mathbf{C}(A)]^{-1} \frac{\partial \mathbf{C}(A)}{\partial \gamma} [\mathbf{C}(A)]^{-1} \frac{\partial \mathbf{C}(A)}{\partial \gamma^T} \right\} = \mathbf{M}_{\mathrm{II}}(A) \ .
\end{aligned}$$

□

Corollary of Lemma 2. *If* $\alpha(x_0, x) = 0$ *for every* $x \in D$, *then the point* x_o *gives zero contribution to* $\mathbf{M}_{\mathrm{II}}(D)$.

Conjecture. *The vector*

$$\mathbf{a}(x_o) = \{\mathbf{G}\}_{x_o,\cdot} \frac{\partial \eta(.,\theta)}{\partial \theta}$$

and the vectors

$$\alpha(x_0, z) = \{\mathbf{G}\}_{x_o,\cdot} \frac{\partial \mathbf{C}}{\partial \gamma} \{\mathbf{G}\}_{\cdot,z} \ ; \quad z \in D$$

are fundamental for evaluating the amount of information contained in the observation at the point x_o.

We want to demonstrate the validity of this conjecture in the rest of the paper.

We notice also that there is a certain algebraic analogy between $\mathbf{a}(x_o)$ and $\alpha(x_0, z)$ since

$$\begin{aligned} \{\mathbf{a}(x_o)\}_i &= \{\mathbf{G}\}_{x_o,\cdot} \frac{\partial \eta(\cdot,\theta)}{\partial \theta_i} , \\ \{\alpha(x_o, z)\}_i &= \{\mathbf{G}\otimes\mathbf{G}\}_{x_o,z,\cdot,\cdot}\, vec\left(\frac{\partial \mathbf{C}}{\partial \gamma_i}\right) . \end{aligned}$$

4. PARTIAL SUPPRESSION OF INFORMATION AT DESIGN POINTS

We may measure the information content at x by adding a white noise at x, i. e. by slightly destroying this information. (We can consider this as an analogy to what is done in physics, where an object is measured by destroying it slightly.) Hence instead of the process (or field) (1) we consider the model

$$y_x = \eta(x,\theta) + \varepsilon_x + \varepsilon_x^* \tag{5}$$

where ε_x^* is an additive independent (virtual) white noise with $Var(\varepsilon_x^*) = \sigma^2(x) < \sigma^2$ where σ^2 is considered small in Propositions 3 and 4. Denote $\mathbf{\Sigma} = diag\left\{\sigma^2(x_1), \ldots, \sigma^2(x_N)\right\}$, the variance matrix of the white noise, and by

$$\begin{aligned} \mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma}) &= \sum_{x,z\in D} \mathbf{f}(x)\left\{\left[\mathbf{C}(D)+\mathbf{\Sigma}\right]^{-1}\right\}_{x,z} \mathbf{f}^T(z) , \\ \{\mathbf{M}_{\mathrm{II}}(\mathbf{\Sigma})\}_{ij} &= \frac{1}{2} tr\left\{\left[\mathbf{C}(D)+\mathbf{\Sigma}\right]^{-1} \frac{\partial \mathbf{C}(D)}{\partial \gamma_i}\left[\mathbf{C}(D)+\mathbf{\Sigma}\right]^{-1} \frac{\partial \mathbf{C}(D)}{\partial \gamma_j}\right\} , \end{aligned}$$

the information matrices for model (5) under the design D. We denote further

$$\|\mathbf{a}(u)\|_\Phi^2 = \mathbf{a}^T(u)\,\nabla\Phi\left[\mathbf{M}_{\mathrm{I}}(D)\right]\mathbf{a}(u) ,$$

which is a (pseudo)norm, since concavity of Φ implies that the gradient $\nabla\Phi\left[\mathbf{M}_{\mathrm{I}}(D)\right]$ is a positive (semi)definite matrix (cf. [8], p. 427). For example, if $\Phi(M) = \ln\det(M)$, we have $\|\mathbf{a}(u)\|_\Phi^2 = \mathbf{a}^T(u)\left[\mathbf{M}_{\mathrm{I}}(D)\right]^{-1}\mathbf{a}(u)$.

Proposition 3.

$$\begin{aligned} \mathbf{M}_{\mathrm{I}}(D) - \mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma}) &= \sum_{u\in D} \sigma^2(u)\,\mathbf{a}(u)\,\mathbf{a}^T(u) + \text{higher order terms in } \mathbf{\Sigma} , \\ \Phi\left[\mathbf{M}_{\mathrm{I}}(D)\right] - \Phi\left[\mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma})\right] &= \sum_{u\in D} \sigma^2(u)\,\|\mathbf{a}(u)\|_\Phi^2 + o\left(\sigma^2\right) \end{aligned}$$

with $\lim_{\sigma\to 0} o\left(\sigma^2\right)/\sigma^2 = 0$.

Proof.

$$\begin{aligned}(\mathbf{C}+\mathbf{\Sigma})^{-1} &= (\mathbf{I}+\mathbf{G\Sigma})^{-1}\mathbf{G} = (\mathbf{I}-\mathbf{G\Sigma})\mathbf{G} + \text{higher order terms in } \mathbf{\Sigma} \\ &= \mathbf{G} - \sum_u \sigma^2(u)\{\mathbf{G}\}_{\cdot,u}\{\mathbf{G}\}_{u,\cdot} + \text{higher order terms in } \mathbf{\Sigma}\end{aligned}$$

and we put this into $\mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma})$.

From the Taylor expansion we obtain

$$\begin{aligned}\Phi[\mathbf{M}_{\mathrm{I}}(D)] - \Phi[\mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma})] &= tr\{\nabla\Phi[\mathbf{M}_{\mathrm{I}}(D)][\mathbf{M}_{\mathrm{I}}(D) - \mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma})]\} \\ &\quad + \mathcal{O}\left(\|\mathbf{M}_{\mathrm{I}}(D) - \mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma})\|^2\right)\end{aligned}$$

and for $\mathbf{M}_{\mathrm{I}}(D) - \mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma})$ we insert the (above) derived expression. □

Proposition 4.

$$\begin{aligned}&\mathbf{M}_{\mathrm{II}}(D) - \mathbf{M}_{\mathrm{II}}(\mathbf{\Sigma}) \\ =\ &\frac{1}{2}\sum_{u\in D}\sigma^2(u)\sum_{x,z\in D}\alpha(u,x)\,C(x,z)\,\alpha^T(u,z) + \text{higher order terms in } \mathbf{\Sigma}\,, \\ &\Phi[\mathbf{M}_{\mathrm{II}}(D)] - \Phi[\mathbf{M}_{\mathrm{II}}(\mathbf{\Sigma})] \\ =\ &\frac{1}{2}\sum_{u\in D}\sigma^2(u)\sum_{x,z\in D}C(x,z)\,\alpha^T(u,x)\,\nabla\Phi[\mathbf{M}_{\mathrm{II}}(D)]\,\alpha(u,z) + o\left(\sigma^2\right)\,.\end{aligned}$$

Proof. The proof is similar as above, but using the definition of $\mathbf{M}_{\mathrm{II}}(\mathbf{\Sigma})$. □

Now we consider briefly the case when the suppression of information by the white noise is important.

Proposition 5. *The decrease of information caused by the white noise is equal to*

$$\mathbf{M}_{\mathrm{I}}(D) - \mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma}) = \sum_{x,z\in D}\mathbf{a}(x)\left\{\left[\mathbf{G}-\mathbf{\Sigma}^{-1}\right]^{-1}\right\}_{x,z}\mathbf{a}^T(z)$$

and similarly,

$$\begin{aligned}\mathbf{M}_{\mathrm{II}}(\mathbf{\Sigma}) - \mathbf{M}_{\mathrm{II}}(D) &= tr\left\{\alpha(\cdot,\cdot)\,\mathbf{C}\alpha^T(\cdot,\cdot)\left(\mathbf{G}+\mathbf{\Sigma}^{-1}\right)^{-1}\right\} \\ &\quad -\frac{1}{2}tr\left\{\alpha(\cdot,\cdot)\left(\mathbf{G}+\mathbf{\Sigma}^{-1}\right)^{-1}\alpha^T(\cdot,\cdot)\left(\mathbf{G}+\mathbf{\Sigma}^{-1}\right)^{-1}\right\}\,.\end{aligned}$$

Proof. We have

$$(\mathbf{C}+\mathbf{\Sigma})^{-1} = (\mathbf{I}+\mathbf{G\Sigma})^{-1}\mathbf{G}$$

and according to [5] lemmas 18.2.1 and 18.2.3 we have

$$\begin{aligned}(\mathbf{I}+\mathbf{G\Sigma})^{-1} &= \mathbf{I} - \mathbf{G}(\mathbf{I}+\mathbf{\Sigma G})^{-1}\mathbf{\Sigma} \\ &= \mathbf{I} - \mathbf{G}\left(\mathbf{\Sigma}^{-1}+\mathbf{G}\right)^{-1}\,.\end{aligned}$$

Hence, we have

$$\begin{aligned}(\mathbf{C}+\mathbf{\Sigma})^{-1} &= \mathbf{G}\left\{\mathbf{C}-\left[\mathbf{G}-\mathbf{\Sigma}^{-1}\right]^{-1}\right\}\mathbf{G}\\ &= \mathbf{G}-\mathbf{G}\left[\mathbf{G}-\mathbf{\Sigma}^{-1}\right]^{-1}\mathbf{G}\end{aligned}$$

and the results follow from

$$\begin{aligned}\mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma}) &= \sum_{x,z\in D}\mathbf{f}(x)\left[\mathbf{C}+\mathbf{\Sigma}\right]^{-1}\mathbf{f}^{T}(z)\ ,\\ \mathbf{M}_{\mathrm{II}}(\mathbf{\Sigma}) &= \frac{1}{2}tr\left\{\frac{\partial\mathbf{C}}{\partial\gamma}(\mathbf{C}+\mathbf{\Sigma})^{-1}\frac{\partial\mathbf{C}}{\partial\gamma^{T}}(\mathbf{C}+\mathbf{\Sigma})^{-1}\right\}\ .\end{aligned}$$

□

5. LARGE (OR COMPLETE) SUPPRESSION OF INFORMATION AT ONE DESIGN POINT

Proposition 6. *Suppose that the noise is added just at one point x_o, i. e. that $\mathbf{\Sigma}=\mathbf{s}\mathbf{s}^{T}$ where $\mathbf{s}_{x_o}=\sigma(x_o)$, $\mathbf{s}_x=0$ if $x\neq x_o$. Then*

$$\begin{aligned}\mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma}) &= \mathbf{M}_{\mathrm{I}}(D)-\frac{\sigma^{2}(x_o)\,\mathbf{a}(x_o)\,\mathbf{a}^{T}(x_o)}{1+\sigma^{2}(x_o)\{\mathbf{G}\}_{x_o,x_o}}\ ,\\ \mathbf{M}_{\mathrm{I}}(D)-\mathbf{M}_{\mathrm{I}}(D-\{x_o\}) &= \frac{\mathbf{a}(x_o)\,\mathbf{a}^{T}(x_o)}{\{\mathbf{G}\}_{x_o,x_o}}\ .\end{aligned}$$

Proof.

$$\mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma})=\sum_{x,z\in D}\mathbf{f}(x)\left\{\left[\mathbf{C}+\mathbf{s}\mathbf{s}^{T}\right]^{-1}\right\}_{x,z}\mathbf{f}^{T}(z)\ .$$

According to [5], lemmas 18.2.1 and 18.2.3, we have

$$\left[\mathbf{C}+\mathbf{s}\mathbf{s}^{T}\right]^{-1}=\mathbf{G}-\frac{\mathbf{G}\mathbf{s}\mathbf{s}^{T}\mathbf{G}}{1+\mathbf{s}^{T}\mathbf{G}\mathbf{s}}$$

hence multiplying by $\mathbf{f}(x)$ and $\mathbf{f}^{T}(z)$ we obtain

$$\mathbf{M}_{\mathrm{I}}(\mathbf{\Sigma})=\mathbf{M}_{\mathrm{I}}(D)-\frac{\sigma^{2}(x_o)\,\mathbf{a}(x_o)\,\mathbf{a}^{T}(x_o)}{1+\sigma^{2}(x_o)\{\mathbf{G}\}_{x_o,x_o}}\ .$$

The desired result is obtained by taking the limit for $\sigma^{-2}(x_o)\to 0$. □

Proposition 7.

$$\begin{aligned}\mathbf{M}_{\mathrm{II}}(D)-\mathbf{M}_{\mathrm{II}}(\mathbf{\Sigma}) &= \frac{1}{\sigma^{-2}(x_o)+\{\mathbf{G}\}_{x_o,x_o}}\sum_{x,z\in D}\alpha(x_o,x)\,\mathbf{C}(x,z)\,\alpha^{T}(z,x_o)\\ &\quad-\frac{1}{2}\left[\frac{1}{\sigma^{-2}(x_o)+\{\mathbf{G}\}_{x_o,x_o}}\right]^{2}\alpha(x_o,x_o)\,\alpha^{T}(x_o,x_o)\end{aligned}$$

and the expression for $\mathbf{M}_{\mathrm{II}}(D) - \mathbf{M}_{\mathrm{II}}(D - \{x_o\})$ *is obtained when* $\sigma^{-2}(x_o)$ *tends to zero.*

P r o o f . is obtained from

$$\begin{aligned}\{\mathbf{M}_{\mathrm{II}}(\boldsymbol{\Sigma})\}_{ij} &= \frac{1}{2} tr\left\{\frac{\partial \mathbf{C}}{\partial \gamma_i}\left[\mathbf{C} + \mathbf{s}\mathbf{s}^T\right]^{-1}\frac{\partial \mathbf{C}}{\partial \gamma_j}\left[\mathbf{C} + \mathbf{s}\mathbf{s}^T\right]^{-1}\right\} \\ &= \frac{1}{2} tr\left\{\mathbf{G}\frac{\partial \mathbf{C}}{\partial \gamma_i}\mathbf{G}\left[\mathbf{C} - \frac{\mathbf{s}\mathbf{s}^T}{1 + \mathbf{s}^T\mathbf{G}\mathbf{s}}\right]\mathbf{G}\frac{\partial \mathbf{C}}{\partial \gamma_j}\mathbf{G}\left[\mathbf{C} - \frac{\mathbf{s}\mathbf{s}^T}{1 + \mathbf{s}^T\mathbf{G}\mathbf{s}}\right]\right\} \\ &= \frac{1}{2} tr\left\{\alpha_i(\cdot,\cdot)\left[\mathbf{C} - \frac{\mathbf{s}\mathbf{s}^T}{1 + \mathbf{s}^T\mathbf{G}\mathbf{s}}\right]\alpha_j(\cdot,\cdot)\left[\mathbf{C} - \frac{\mathbf{s}\mathbf{s}^T}{1 + \mathbf{s}^T\mathbf{G}\mathbf{s}}\right]\right\}\end{aligned}$$

and from $\{\mathbf{M}_{\mathrm{II}}(D)\}_{ij} = \frac{1}{2} tr\{\alpha_i(\cdot,\cdot)\mathbf{C}\alpha_j(\cdot,\cdot)\mathbf{C}\}$. □

6. THE DECOMPOSITION OF $\mathbf{M}_{\mathrm{I}}(D)$

The symmetry, which appears in the following statement seems to be interesting.

Proposition 8. *Let* $A \subset D$ *be an arbitrary subdesign of* D*. Then*

$$\mathbf{M}_{\mathrm{I}}(D) = \sum_{x,z \in A} \mathbf{f}(x)\left\{[\mathbf{C}(A)]^{-1}\right\}_{x,z}\mathbf{f}^T(z) + \sum_{x,z \in A^c} \mathbf{a}(x)\left\{[\mathbf{G}(A^c)]^{-1}\right\}_{x,z}\mathbf{a}^T(z)$$

where $A^c = D - A$*, and* $\mathbf{G}(A^c)$ *is a submatrix of* $\mathbf{G} = \mathbf{G}(D)$*. The fist term on the right-hand side is the information obtained when the design* A *is used, and the second term is the information which is contained in the points of* A^c *when the design* D *is used.*

Remark. Since $\mathbf{G}$ is the inverse of $\mathbf{C}$, we can write

$$[\mathbf{G}(A^c)]^{-1} = \mathbf{C}(A^c) - \mathbf{C}(A^c, A)[\mathbf{C}(A)]^{-1}\mathbf{C}(A, A^c)$$

(cf. theorem 8.5.1 in [5]), which is the conditional variance matrix of the vector of observations $(y_x : x \in A^c)$, given the vector $(y_x : x \in A)$. So the information contained in points of A^c is small if either the norms of the vectors $\mathbf{a}(x)$, $x \in A^c$ are small, or if the conditional variance $[\mathbf{G}(A^c)]^{-1}$ is small, i. e. if from the observation of y_x at $x \in A$ we can predict quite precisely the variables $y_x : x \in A^c$, and so observation of these variables is not necessary.

P r o o f of Proposition 8. Without loss of generality suppose that $D = \{x_1, \ldots, x_N\}$, $A = \{x_1, \ldots, x_k\}$, $k < N$. Denote $\mathbf{S} = \left(\mathbf{e}^{(k+1)}, \ldots, \mathbf{e}^{(N)}\right)$ where $\mathbf{e}^{(i)}$ is the ith canonical vector of $\mathbb{R}^N$. Take $\boldsymbol{\Sigma} = \sigma^2\mathbf{S}\mathbf{S}^T$, that means the virtual noise with variance σ^2 is applied at all points of A^c. Then

$$\begin{aligned}(\mathbf{C} + \boldsymbol{\Sigma})^{-1} &= \left[I + \sigma^2\mathbf{G}\mathbf{S}\mathbf{S}^T\right]^{-1}\mathbf{G} = \left[I - \sigma^2\mathbf{G}\mathbf{S}\left(I + \sigma^2\mathbf{S}^T\mathbf{G}\mathbf{S}\right)^{-1}\mathbf{S}^T\right]\mathbf{G} \\ &= \mathbf{G} - \mathbf{G}(D, A^c)\left[\sigma^{-2}I + \mathbf{G}(A^c)\right]^{-1}\mathbf{G}(A^c, D)\end{aligned}$$

since $\mathbf{GS} = \mathbf{G}(D, A^c)$, a submatrix of $\mathbf{G}$, and similarly $\mathbf{S}^T\mathbf{GS} = \mathbf{G}(A^c)$. Hence

$$\begin{aligned}\mathbf{M}_{\mathrm{I}}(\boldsymbol{\Sigma}) &= \sum_{x,z\in D} \mathbf{f}(x)\left\{[\mathbf{C}+\boldsymbol{\Sigma}]^{-1}\right\}_{x,z}\mathbf{f}^T(z) = \mathbf{M}_{\mathrm{I}}(D) \\ &\quad - \sum_{x,z\in D} \mathbf{f}(x)\left\{\mathbf{G}(D,A^c)\right\}_{x,\cdot}\left[\sigma^{-2}I + \mathbf{G}(A^c)\right]^{-1}\left\{\mathbf{G}(A^c,D)\right\}_{\cdot,z}\mathbf{f}^T(z) \\ &= \mathbf{M}_{\mathrm{I}}(D) - \sum_{u,v\in A^c}\mathbf{a}(u)\left\{\left[\sigma^{-2}I + \mathbf{G}(A^c)\right]^{-1}\right\}_{u,v}\mathbf{a}^T(v)\end{aligned}$$

since $\sum_{x\in D}\mathbf{f}(x)\left\{\mathbf{G}(D,A^c)\right\}_{x,u} = \mathbf{a}(u)$ for $u \in A^c$. The required result is then obtained by taking $\sigma^{-2} \to 0$. □

Notice that the decomposition of $\mathbf{M}_{\mathrm{II}}(D)$ can be obtained in a similar way, but the result is not so symmetrical.

7. RELATIONS TO SOME PUBLISHED SEARCH ALGORITHMS

Let us compare briefly the presented exposition with some known algorithms for the search of an optimal $\Phi[\mathbf{M}_{\mathrm{I}}(D)]$ on finite design spaces.

Besides the pioneering paper [12], which had a rather specific set-up, probably the oldest algorithm on a finite design space is the *exchange algorithm* of [1]. In this algorithm they start by a k-point design D, and alternatively add the most informative design point $x_o \notin D$, and reject the less informative design point $x_o \in D$. For that, besides the second formula presented in Proposition 6, which allows to find the less informative point of D, one needs also the expression for finding the most informative point outside D

$$\mathbf{M}_{\mathrm{I}}(D\cup\{x_o\}) = \mathbf{M}_{\mathrm{I}}(D) + \frac{\mathbf{f}^*(x_o)\,\mathbf{f}^{*T}(x_o)}{C(x_o,x_o) - \mathbf{C}(x_o,D)\,\mathbf{GC}(D,x_o)}$$

where

$$\mathbf{f}^*(x_o) = \mathbf{f}(x_o) - \sum_{u\in D}\mathbf{a}(u)\,C(u,x_o)\ .$$

Notice that this can be proven directly using (cf. [5] theorem 8.5.11)

$$\begin{aligned}[\mathbf{C}(D\cup\{x_o\})]^{-1} &= \begin{pmatrix}\mathbf{C} & \mathbf{C}(D,x_o)\\ \mathbf{C}(x_o,D) & C(x_o,x_o)\end{pmatrix}^{-1} \\ &= \begin{pmatrix}\mathbf{G} + \mathbf{GC}(D,x_o)\,\mathbf{Q}^{-1}C(x_o,D)\,\mathbf{G} & -\mathbf{GC}(D,x_o)\,q^{-1}\\ -q^{-1}\mathbf{C}(x_o,D)\,\mathbf{G} & q^{-1}\end{pmatrix}\end{aligned}$$

where

$$q = \mathbf{C}(x_o,x_o) - C(x_o,D)\,\mathbf{C}^{-1}C(D,x_o)\ .$$

Another class of algorithms is *based on the virtual noise*, as described in Section 4. The basic idea is to start with a sufficiently large reference design D and consider model (5) with

$$\boldsymbol{\Sigma}^{(0)} = diag\left\{\sigma^2,\ldots,\sigma^2\right\}\ ,$$

and to modify iteratively the virtual noise, i. e. to compute subsequently a sequence of matrices

$$\mathbf{\Sigma}^{(1)}, \mathbf{\Sigma}^{(2)}, \ldots, \mathbf{\Sigma}^{(n)}, \ldots,$$

until arriving to a stable $\mathbf{\Sigma}^{(n)}$ which has large components at $N-k$ points and almost zero components at k points, where k is the prescribed number of design points in the target design. If the algorithm is successful, this k points correspond to the optimal design A_{opt} ,

$$A_{opt} = \arg \min_{A,\#(A)=k} \Phi \left[\mathbf{M}_{\mathrm{I}}(A)\right] \ .$$

In [8] the choice of $\mathbf{\Sigma}^{(n)}$ was

$$\mathbf{\Sigma}^{(n)} = \delta \ diag \left\{ \ln \left[\frac{\xi_{\max}^{(n)}}{\xi^{(n)}(x_1)} \right], \ldots, \ln \left[\frac{\xi_{\max}}{\xi^{(n)}(x_n)} \right] \right\}$$

where $\delta > 0$ is a small smoothing parameter, $\xi^{(n)}$ is a probability measure on the reference design D, and

$$\xi_{\max}^{(n)} = \max_{x \in D} \xi^{(n)}(x) \ .$$

This has been smoothly approximated by

$$\left\{ \sum_{x \in D} \left[\xi^{(n)}(x) \right]^{1/\delta+1} \right\}^{\delta} \ .$$

The passage from $\xi^{(n)}$ to $\xi^{(n+1)}$ is done according to

$$\xi^{(n+1)} = \frac{n}{n+1} \xi^{(n)} + \frac{1}{n+1} \xi_{x_n} \tag{6}$$

where ξ_{x_n} is the probability measure concentrated at one point x_n, and

$$x_n = \arg \min_{x \in D} \frac{1}{\xi^{(n)}(x)} \left[\|\mathbf{a}(x)\|_{\Phi}^2 - \frac{\mathcal{I}_{B_n}(x)}{N_{B_n}} \sum_{z \in D} \|\mathbf{a}(z)\|_{\Phi}^2 \right]$$

where $B_n = \left\{ x \in D : \xi^{(n)}(x) = \xi_{\max}^{(n)} \right\}$, $\mathcal{I}_{B_n}.(x) = 1$ if $x \in B_n, \mathcal{I}_{B_n}.(x) = 0$ if $x \notin B_n$, and N_{B_n} is the number of points in B_n. The choice of x_n corresponds to the steepest descent method for the limit case $\delta \to 0$.

A similar method has been used in [7], however a more complicated expression for $\Sigma^{(n)}$, which respects the size k of the target design

$$\mathbf{\Sigma}^{(n)} = \delta \ diag \left\{ \ln \left[\frac{\max\left\{\xi_{\max}^{(n)} - \frac{1}{k}, 0\right\}}{\max\left\{\xi^{(n)}(x_1) - \frac{1}{k}, 0\right\}} \right], \ldots, \ln \left[\frac{\max\left\{\xi_{\max}^{(n)} - \frac{1}{k}, 0\right\}}{\max\left\{\xi^{(n)}(x_N) - \frac{1}{k}, 0\right\}} \right] \right\}$$

and this expression has been smoothed by some approximations. In this method the correcting point x_n in (6) is chosen according to

$$x_n \in \arg \min_{\{x \in D: \xi^{(n)}(x) > 1/k\}} \frac{1}{\xi^{(n)}(x) - 1/k} \left[\left\| \mathbf{a}^{(n)}(x) \right\|_{\Phi}^2 - \frac{\mathcal{I}_{B_n}(x)}{N_{B_n}} \sum_{z \in D} \left\| \mathbf{a}^{(n)}(z) \right\|_{\Phi}^2 \right]$$

if $\xi_{\max}^{(n)} - 1/k > 0$, or according to a modified expression if not. Here (compare with (3))

$$\mathbf{a}^{(n)}(x) = \sum_{u \in D} \left\{ \left[\mathbf{C} + diag \left\{ \ln \frac{1/k}{\min\left\{\xi^{(n)}(x), 1/k\right\}}; x \in D \right\} \right]^{-1} \right\}_{x,u} \mathbf{f}(u) \ .$$

So in both methods the vectors $\mathbf{a}(x)$ are important.

Notice that a method based on the Mercer's expansion of the covariance function of the process (1), which was proposed in [3], can be also considered as a special case of the method of virtual noise, with the variance at the nth step $\sigma^2(x) = \frac{c}{\xi^{(n)}(x)}$. However, this choice of the virtual variance gives rather more an optimum design for the Bayesian estimation of the (random) coefficients of the Mercer's decomposition than the optimum design for the parameters of the mean, as demonstrated by examples in [4].

ACKNOWLEDGMENT

This work was supported by the VEGA-grant No. 1/0077/09.

(Received August 31, 2009)

REFERENCES

[1] U. N. Brimkulov, G. V. Krug, and V. L. Savanov: Numerical construction of exact experimental designs when the measurements are correlated. (In Russian.) Industr. Laboratory *36* (1980), 435–442.

[2] N. A. C. Cressie: Statistics for Spatial Data. Second edition. Wiley, New York 1993.

[3] V. V. Fedorov: Design of spatial experiments: model fitting and prediction. In: Handbook of Statistics (S. Gosh and C. R. Rao, eds.), Vol. 13, Elsevier, Amsterdam 1996, pp. 515–553.

[4] V. V. Fedorov and W. G. Müller: Optimum design for correlated fields via covariance kernel expansions. In: Model Oriented Data and Analysis *8*, (J. Lopez-Fidalgo, J. H. Rodriguez-Diaz, and B. Torsney, eds.), Physica-Verlag, Heidelberg 2007, pp. 57–66.

[5] D. A. Harville: Matrix Algebra from a Statistician's Perspective. Springer, New York 1997.

[6] W. G. Müller: Collecting Spatial Data. Third edition. Springer, Heidelberg 2007.

[7] W. G. Müller and A. Pázman: An algorithm for the computation of optimum designs under a given covariance structure. Comput. Statist. *14* (1999), 197–211.

[8] W. G. Müller and A. Pázman: Measures for designs in experiments with correlated errors. Biometrika *90* (2003), 423–445.

[9] W. Näther: Exact designs for regression models with correlated errors. Statistics *16* (1985), 479–484.

[10] A. Pázman: Criteria for optimal design of small-sample experiments with correlated observations. Kybernetika *43* (2007), 453–462.

[11] F. Pukelsheim: Optimal Design of Experiments. Wiley, New York 1993.

[12] J. Sacks and D. Ylvisaker: Design for regression problems with correlated errors. Ann. Math. Statist. *37* (1966), 66–84.

[13] D. L. Zimmerman: Optimum network design for spatial prediction, covariance parameter estimation and empirical prediction. Environmetrics *17* (2006), 635–652.

Andrej Pázman, Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská dolina, 842 48 Bratislava. Slovak Republic.

e-mail: pazman@fmph.uniba.sk