CONF-970241--1

PLACEMENT OF SENSORS FOR CORRELATED OBSERVATIONS

Valerii V. Fedorov

Oak Ridge National Laboratory
Computer Science and Mathematics Division
Mathematical Sciences Section
P. O. Box 2008, Bldg. 6012
Oak Ridge, TN 37831-6367

The study was motivated by the necessity to construct monitoring networks in various applied areas. Environmetry, meteorology, and seismology are the most notable examples. The approach is based on the expansion of covariance kernels with respect to their eigenfunctions and subsequent use of numerical algorithms based on convex design theory ideas.

Research sponsored by the Laboratory Directed Research and Development Program, Oak Ridge National Laboratory, for the Office of Energy Research, U.S. Department of Energy under contract number DE-AC05-96OR22464 with Lockheed Martin Energy Research Corporation.

DISCLAIMER

PLACEMENT OF SENSORS FOR CORRELATED OBSERVATIONS

Valerii Fedorov
Oak Ridge National Laboratory
P. O. Box 2008
Oak Ridge, Tennessee 37831, U.S.A.

Abstract

The study was motivated by the necessity to construct monitoring networks in various applied areas. Environmetry, meteorology, and seismology are the most notable examples. The approach is based on the expansion of covariance kernels with respect to their eigenfunctions and subsequent use of numerical algorithms based on convex design theory ideas.

1 Model

Consider the regression model

$$y_{ij} = \eta(x_i, \theta) + u_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, k. \quad (1)$$

The function $\eta(x, \theta)$ is assumed to be linear, i.e., $\eta(x, \theta) = \theta^T f(x), \theta \in R^m$ and the basis functions $f(x)$ are given. Variables $x \in X \subset R^l$ are usually coordinates of sensors (observing stations, sampling sites). The operability region X is normally a compact set in R^2 (for instance, geographical region). Examples of other dimensions include one dimensional allocations (for instance, sampling along highways) and three dimensional cases (elevation of sensors may be controlled). The subscript "i" corresponds to the number of a sensor, "j" stands for the time when observation is done. We assume that u_{ij} and ε_{ij} are random variables. The same characters are used both for random variables and their realization, if it does not lead to confusion. The objective of an experiment may be prediction of y, estimation of $\eta(x, \theta)$, or θ, or some functions of them.

The random term $v = u + \varepsilon$ is partitioned to trace two different "generators" of randomness. The first random variable u describes deviations of the observed response from $\eta(x, \theta)$ due to some causes, which can be common for various sites. For instance, it can be weather fluctuation on the scale of the whole region X. That is why observations at different sites may be dependent. This dependence is described solely through the covariance kernel of u:

$$E(u_{ij}, u_{ij'}) = K(x_i, x_{i'})\delta_{jj'}, \quad E(u_{ij}) = 0. \quad (2)$$

The second variable describes "observational" errors. These errors can depend upon the observational technique or by the length of the observational interval. We assume that these errors be (at least partly) controlled by an observer and are specific for every particular site or sensor. The latter statement may possibly assume that

$$E(\varepsilon_{ij}, \varepsilon_{i'j'}) = \sigma^2 \delta_{ii'}\delta_{jj'}, \quad E(\varepsilon_{ij}) = 0. \quad (3)$$

Probably, the reader has noticed the presence of Kronecker's symbol $\delta_{jj'}$ in (2) and (3), i.e., there is no correlation in time. As it was mentioned before, "j" frequently indicates the time of observation. We skip the index when it does not cause any confusion.

Let the kernel $K(x, x')$ defined by (2) exist on $Z \times Z$, where Z is compact in R^l and $X \subset Z$. All eigenvalues of $K(x, x')$ are positive. The series (see for instance, Kanwal (1971) for more rigid mathematics and details): $K(x, x') = \sum_{\alpha=1}^{\infty} \lambda_\alpha \varphi_\alpha(x)\varphi(x')$ is uniformly and absolutely convergent, and the series $\sum_{\alpha=1}^{\infty} \lambda_\alpha$ is convergent. Obviously λ_α must diminish not slower than α^{-1}. In many cases the decay is much faster. This fact allows us to hope that for practical needs we can use the approximation

$$K(x, x') \simeq K_p(x, x') = \sum_{\alpha=1}^{p} \lambda_\alpha \varphi_\alpha(x)\varphi_\alpha(x') \quad (4)$$

with some moderate p. The opportunity to use (4) is essential for our approach.

2 Prediction without trend

Optimality criteria. Let us start with the very simple but quite popular model in applications

$$y_{ij} = u_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, k, \quad (5)$$

which is a particular case of (1) with $\eta(x,\theta) \equiv 0$.

Let points $\xi_n = (x_1, \ldots, x_n)$ be selected from X. It is convenient to introduce the following matrices and vectors (recall our promise to skip "j"):

$$\begin{aligned} \{K(\xi_n)\}_{ii'} &= K(x_i, x_i'), \\ \{K(x,\xi_n)\}_i &= K(x, x_i) \\ Y^T &= (y_1, \ldots, y_n). \end{aligned} \tag{6}$$

It is easy to check that

$$E(Y) = 0, \quad E(YY^T) = V(\xi_n) = \sigma^2 I + K(\xi_n). \tag{7}$$

Let the prediction of $y(x)$ on a given set X_{pr} be an immediate goal for a practitioner. It is known (see, Rao (1973), Ripley (1981)) that

$$\hat{y}(x) = V^T(x,\xi_n)\, V^{-1}(\xi_n) Y \tag{8}$$

is the best linear unbiased predictor for $y(x)$. If x does not coincide with any supporting point of ξ_n, i.e., $x \neq x_i \in \operatorname{supp} \xi_n$, then the variance of this predictor equals

$$\begin{aligned} & Var\,(y(x) - \hat{y}(x)) \\ &= V(x,x) - V^T(x,\xi_n) V^{-1}(\xi_n) V(x,\xi_n) \\ &= \sigma^2 + K(x,x) \\ &\quad - K^T(x,\xi_n) \left[\sigma^2 I + K(\xi_n)\right]^{-1} K(x,\xi_n) \\ &= \sigma^2 + C(x,\xi_n). \end{aligned} \tag{9}$$

Otherwise it obviously diminishes. The definition of $V(x,x), V(x,\xi_n)$ and $V(\xi_n)$ is evident from (6) and (7). Through the whole paper the use of any inverse matrix automatically assumes the existence of this matrix. The most commonly used objective functions related to (9) are

$$Q_1(\xi_n) = \max_{x \in X_{pr}} Var\,(y(x) - \hat{y}(x)), \tag{10}$$

and

$$Q_2(\xi_n) = \int_{X_{pr}} Var\,(y(x) - \hat{y}(x))\, dx. \tag{11}$$

Subsequently the design problem may be stated as

$$\xi_n^* = \arg\min_{\xi_n} Q(\xi_n), \tag{12}$$

where Q stands for either Q_1, or Q_2. To avoid the discussion of singular cases it is assumed that $\int_{X_{pr}} dx > 0$.

Most publications are concerned with the criterion $Q_2(\xi_n)$; see Cambanis (1985) for a survey of main results and further references. Standardly it is assumed that the observational error is negligible, i.e. $\sigma^2 = 0$, and n is large enough to use asymptotical ($n \to \infty$) results. The proposed approaches are strongly based on ideas developed in the classical theory of function and integral approximation. Our intention is to use the ideas developed in the convex design theory. We want to emphasize at this stage that it is essential for the whole approach that $\sigma^2 \neq 0$ and approximation (4) is valid, i.e., $\sigma^2 \gg \sum_{\alpha=p+1}^{\infty} \lambda_\alpha \varphi_\alpha(x)\varphi_\alpha(x')$, for all $x \in X$.

Regression model with random parameters as an approximation of a random field. If (4) is valid, then model (5) can be replaced by its approximate version

$$y_i = \gamma^T \varphi(x_i) + \varepsilon_i, \tag{13}$$

where $\varphi^T(x) = \{\varphi_\alpha(x)\}_1^p$, the parameters γ are random, and

$$E(\gamma) = 0,\ E(\gamma\gamma^T) = \Lambda,\ \Lambda_{\alpha\beta} = \lambda_\alpha \delta_{\alpha\beta}. \tag{14}$$

In the frame of (13)

$$\begin{aligned} E(YY^T) &= V(\xi_n) \\ &= \sigma^2 I + K(\xi_n) \\ &= \sigma^2 I + \Phi^T(\xi_n)\Lambda\Phi(\xi_n), \end{aligned} \tag{15}$$

where $\Phi(\xi_n) = (\varphi(x_1), \ldots, \varphi(x_n))$.

Let us try to predict $y(x)$ using the regression analysis technique. The best linear predictor for γ is

$$\begin{aligned} \hat{\gamma} &= \sigma^{-2} \left(\underline{M}(\xi) + \Lambda^{-1}\right) \Phi(\xi_n) Y, \\ \underline{M}(\xi_n) &= \sigma^{-2} \Phi(\xi_n)\Phi^T(\xi_n). \end{aligned} \tag{16}$$

The dispersion matrix of the difference $\hat{\gamma} - \gamma$ is

$$D(\xi_n) = E\left[(\hat{\gamma} - \gamma)(\hat{\gamma} - \gamma)^T\right] = (\underline{M}(\xi) + \Lambda^{-1})^{-1}. \tag{17}$$

Let us select $\tilde{y}(x) = \varphi^T(x)\hat{\gamma}$ as a predictor for $y(x)$. On an intuitive level it is obvious that $\hat{y}(x)$ and $\tilde{y}(x)$ must coincide in the frame of approximation (4). Indeed, using the identity

$$(A^{-1} + BB^T)^{-1} = A - AB(B^T AB + I)^{-1} BA. \tag{18}$$

one can check that for $x \neq x_i \in \operatorname{supp} \xi_n$:

$$\begin{aligned} & Var\,(\hat{y}(x) - y(x)) = Var\,(\tilde{y}(x) - y(x)) \\ &= \sigma^2 + \varphi^T(x) D(\xi_n) \varphi(x) \\ &= \sigma^2 + K_p(x,x) \\ &\quad - K_p(x,\xi_n) \left(\sigma^2 I + K_p(\xi_n)\right)^{-1} K_p(x,\xi_n) \\ &= \sigma^2 + C_p(x,\xi) \end{aligned} \tag{19}$$

Combining (19) and (10) we have

$$Q_1(\xi_n) = \sigma^2 + \max_{x \in X_{pr}} \varphi^T(x) D(\xi_n) \varphi(x). \tag{20}$$

For the average variance of prediction, i.e. for case (11),

$$\begin{aligned} Q_2(\xi_n) &= \sigma^2 \textstyle\int_{X_{pr}} dx + \operatorname{tr} A D(\xi_n), \\ A &= \textstyle\int_{X_{pr}} \varphi(x)\varphi^T(x) dx. \end{aligned} \tag{21}$$

3 Optimal designs for prediction

In this section we pursue a very transparent and simple idea. First, we formulate results for optimization problems (20) - (22) in terms of approximation (4), i.e. all calculations are based on our knowledge of Λ and $\varphi(x)$. At the second stage we translate our findings to the language of covariance kernels and linear unbiased predictors. Thus our intention is to use (4) mostly as an intermediate step.

Continuous optimal designs. Let us admit the possibility of repeated observation and let $N = \sum_{i=1}^{n} r_i$, where r_i is a number of observations at point x_i. For sufficiently large N we can introduce continuous designs

$$\xi = \{p_i, x_i\}_1^n, \quad 0 \le p_i \le 1, \quad x_i \in X,$$

or to define a design as a probability measure $\xi(dx)$ on X; see Ermakov (1983), Fedorov (1972), or Pukelsheim (1993).

Let us redefine the matrix $D(\xi)$ as

$$D(\xi) = \left[\sigma^{-2} N M(\xi) + \Lambda^{-1}\right]^{-1}, \qquad (22)$$

where

$$M(\xi) = \int_X \varphi(x)\varphi^T(x)\xi(dx). \qquad (23)$$

Similar to (12) we can define

$$\xi^* = \arg\min_{\xi} Q(\xi), \qquad (24)$$

where ξ can be any probability measure on X. It is expedient to note that in general ξ^* may depend upon N. This fact makes (24) different from the standard design theory in the continuous case. However, when N is fixed, all the results developed in this theory may be used almost directly. For instance, for the linear criterion (22) we have:

Theorem 1 *The design ξ^* is linear optimal if and only if for all $x \in X$*

$$\psi_1(x,\xi^*) \le \mathrm{tr} M(\xi^*) D(\xi^*) A D(\xi^*), \qquad (25)$$

where $\psi_1(x,\xi) = \varphi^T(x) D(\xi) A D(\xi) \varphi(x)$, and equality takes place at all supporting points of ξ^*.

For criterion (20) in the simplest case, when $X_{pr} = X$ we have an analogue of the Kiefer-Wolfowitz theorem.

Theorem 2 *1. The design ξ^* is minimax if and only if for all $x \in X$*

$$\psi_2(x,\xi^*) \le \mathrm{tr} M(\xi^*) D(\xi^*), \qquad (26)$$

where $\psi_2(x,\xi) = \varphi^T(x) D(\xi) \varphi(x)$, and equality takes place at all supporting points of ξ^*.

2. Minimax designs coincide with D-optimal designs, i.e. $\xi^* = \arg\min_{\xi} |D(\xi)|$.

Let us recall that in the above theorem $D(\xi) = \left(\sigma^{-2} N M(\xi) + \Lambda^{-1}\right)^{-1}$ unlike the classical Kiefer-Wolfowitz theorem, where $D(\xi) = M^{-1}(\xi)$. Both theorems are routine results from convex design theory. They tell us that in the framework of approximation (4) the design of a spatial monitoring network can be embedded in the well developed area of convex design theory.

Let us now move in the reverse direction: having the results stated in the above theorems and using presentation (15) let us try to formulate results in terms of the covariance kernels.

In what follows we consider only designs with a finite number of supporting points. This fact is not very restrictive, because for every design with given matrix $M(\xi)$ there exists a design ξ' with a finite number of supporting points and exactly the same matrix $M(\xi')$, see, for instance, Fedorov (1972).

From identity (18) it follows that

$$D(\xi) = \Lambda - \Lambda\Phi(\xi)(J(\xi) + K_p(\xi))^{-1}\Phi(\xi)^T\Lambda, \qquad (27)$$

where

$$J_{il}^{-1}(\xi) = \sigma^{-2} N_i p_i \delta_{il} \text{ and } \Phi(\xi) = (\varphi(x_1), \ldots, \varphi(x_n)).$$

Combining (25), (27) and introducing (compare with (19)) the function

$$C_p(x,x',\xi) = K_p(x,x') - K_p(x,\xi)\left(J(\xi) + K_p(\xi)\right)^{-1} K_p(x',\xi) \qquad (28)$$

we come to the following result.

Theorem 3 *The design ξ^* minimizes the average variance of prediction if and only if for all $x \in X$*

$$\int_{X_{pr}} C_p^2(x,x',\xi^*)dx' \le \int_X \int_{X_{pr}} C_p^2(x,x',\xi^*)\,\xi^*(dx)dx' \qquad (29)$$

and equality holds at all supporting points of ξ^*.

Obviously, for $x = x'$ the function $C_p(x,x',\xi)$ coincides with the previously introduced function $C_p(x,\xi)$.

For the minimax case we have with $X_{pr} = X$

Theorem 4 *1. The design ξ^* is minimax if and only if for all its supporting points x_i^**

$$C_p(x_i^*,\xi^*) = \max_{x \in X} C_p(x,\xi^*) \qquad (30)$$

2. Minimax designs coincide with D-optimal designs:

$$\xi^* = \arg\max_{\xi} |J(\xi) + K_p(\xi)|\,|J(\xi)|^{-1} \qquad (31)$$

To get (31) one has to notice that

$$|\sigma^{-2}NM(\xi)+\Lambda^{-1}| = \frac{|J(\xi)+K_p(\xi)|}{|\Lambda|\,|J(\xi)|}.$$

Optimization problem (31) may be considered as maximization of the determinant of the variance-covariance matrix of observations. The idea that it can lead to good prediction was probably first stated by Shewry and Wynn (1987) in a different setting.

4 Algorithms

Theorems 1 and 2 lead immediately to numerical procedures, that are well known in experimental design theory and can be found, for instance, in Ermakov (1983). Actually the corresponding algorithms are technically identical to algorithms developed for the Bayesian approach. At every *s-th* iteration of these algorithms one has to find either

$$\min_{x\in X_s}\psi(x,\xi_s) \quad \text{or} \quad \max_{x\in X}\psi(x,\xi_s), \tag{32}$$

where the function ψ may coincide with ψ_1 or ψ_2 correspondingly, and $X_s = \operatorname{supp}\xi_s$. The algorithm is simple, but it is necessary to know the eigenfunctions $\psi(x)$. However, it is not easy to compute this function even for relatively simple kernels $K(x,x')$ and symmetrical Z.

Theorems 3 and 4 allow the development of numerical procedures, that use "directly" the covariance kernel $K_p(x,x')$. For instance, the first order exchange algorithm (compare with Mitchell, 1974) can be written for the minimax criterion with $X'_{pr} = X$ as follows:
Step a. There is a design ξ_s. Find

$$x_s^+ = \arg\max_{x\in X} C_p(x,\xi_s), \tag{33}$$

and construct $\xi_s^+ = \xi_s + \alpha_s\delta(x_s^+)$, where $\delta(x)$ is a probability measure atomized at x.
Step b. Find

$$x_s^- = \arg\min_{x\in X_s} C_p(x,\xi_s), \tag{34}$$

where $X_s = \operatorname{supp}\xi_s^+$, and construct $\xi_{s+1} = \xi_s - \alpha_s\delta(x_s^-)$.

The changes which must be done in the case of linear criterion (see (22)) are evident. To guarantee covergence of the above iterative procedure, the sequence $\{\alpha_s\}$ may be chosen similarly to what was proposed in standard design theory (see, Ermakov, 1983, Silvey, 1980).

Probably the reader has noticed that to use the above numerical procedure one has to find $K_p(x,x')$ and that may be rather difficult. However, if approximation (4) is valid, then $K_p(x,x')$ may be replaced by $K(x,x')$, which in this case is assumed to be known. As soon as it is done the iterative procedure becomes a simple tool for "optimal" design construction. The word "optimal" is used in quotation marks to emphasize that a rigorous mathematical analysis of the limit behavior of ξ^* defined by (24) when $p\to\infty$ is still an open problem, which hopefully will attract the attention of statisticians. In particular, the following questions must be answered:

Are limit versions of Theorems 3 and 4, i.e., when $C_p(x,x',\xi)$ is replaced by $C(x,x',\xi)$, valid?

Does this algorithm converge to an optimal design when $C_p(x,\xi)$ is replaced by $C(x,\xi)$?

Our computational exercises seem to confirm positive answers to both questions.

5 Spatial trends

The ideas considered in Sections 2-4 can also be applied in the case of a non-zero trend. Instead of the matrix $D(\xi)$ defined in (22) one has to introduce the extended covariance matrix

$$D(\xi) = \left[\sigma^{-2}N\,M(\xi) + \begin{pmatrix} 0 & 0 \\ 0 & \Lambda^{-1}\end{pmatrix}\right]^{-1}, \tag{35}$$

where now

$$M(\xi) = \begin{pmatrix} M_{f\psi}(\xi) & M_{f\psi}(\xi) \\ M^T_{f\psi}(\xi) & M_{\psi\psi}(\xi)\end{pmatrix},$$

$$\begin{aligned} M_{ff}(\xi) &= \textstyle\int_X f(x)f^T(x)\xi(dx), \\ M_{f\psi}(\xi) &= \textstyle\int_X f(x)\psi^T(x)\xi(dx), \end{aligned} \tag{36}$$

and $M_{\psi\psi}(\xi)$ is identical to $M(\xi)$ previously defined in (23).

Obviously we now have a greater variety of possible optimality criteria. To complement the previous considerations let us assume that the trend parameters θ are of prime interest, i.e., we have to search for

$$\xi^* = \arg\min_{\xi}\Psi\left[D_{ff}(\xi)\right], \tag{37}$$

where D_{ff} is the obviously defined submatrix of the matrix D. Again, the optimization problem (37) is a particular case of convex design theory, and the equivalence theorem can be immediately formulated. For instance, if $\Psi(D_{ff}) = \ln|D_{ff}|$, then a necessary and sufficient condition for ξ^* to be optimal is that for all $x\in X$

$$\begin{aligned} &\psi(x,\xi^*) = \psi_3(x,\xi^*) - \psi_2(x,\xi^*) \\ &\le [\operatorname{tr}D(\xi^*)M(\xi^*) - \operatorname{tr}D_{\psi\psi}(\xi^*)M_{\psi\psi}(\xi^*)] \end{aligned} \tag{38}$$

where $\psi_3(x,\xi) = f^T(x)Df(x)$ and $\psi_2(x,\xi)$ is defined in Theorem 2.

Combining (38) and (32) one can implement an iterative procedure based on the function $\psi(x,\xi)$. However, the eigenfunctions $\varphi(x)$ must be computed to use the presentation of $\psi(x,\xi)$ described in (38). Can we replace this presentation by another one similar to (29), i.e., which does not include the functions $\varphi(x)$ explicitly? The answer is positive and

$$\psi(x,\xi) = d(x,\xi) - C_p(x,\xi) + R_p(x,\xi) + q^T(x,\xi)R_p(\xi)q(x,\xi) - 2q^T(x,\xi) \qquad (39)$$

where

$$\begin{aligned}
d(x,\xi) &= d(x,x,\xi),\\
d(x,x',\xi) &= \sigma^2 N f^T(x) M_{jj}^{-1}(\xi) f(x),\\
q(x,\xi) &= J(\xi)\underline{d}(x,\xi),\\
\{\underline{d}_i(x,\xi)\}_i &= d(x,x_i;\xi),\\
R_p(x,x') &= K_p(x,x')\\
&\quad -K_p^T(x,\xi)\left(\beta(\xi) + K_p(\xi)\right)^{-1} K_p(x,\xi),\\
\beta(\xi) &= J(\xi)\left(J(\xi) - d(\xi)\right)^{-1} J(\xi),\\
\{d(\xi)\}_{ij} &= d(x_i,x_j;\xi).
\end{aligned}$$

The matrix $R_p(\xi)$ and the vector $R_p(x,\xi)$ are obviously defined through $R_p(x,x')$; compare with (7). Assuming that $p \to \infty$ and replacing $K_p(x,x')$ by $K(x,x')$ we have an opportunity to construct an algorithm (compare with (33), and (34)) which includes only the basis functions $f(x)$ and the correlation kernel $K(x,x')$. Still the questions similar to those two posted at the concluding part of Section 4 must be answered.

6 Acknowledgements

I thank Darryl Downing and George Ostrouchov for their valuable comments and remarks.

References

[1] Cambanis, S. (1985). Sampling Design for Estimating a Random Process, in *Handbook of Statistics*, in Rao, C. R. (Ed.), Vol. 5, North-Holland, New York, pp. 337-362

[2] Ermakov, S.M. (Ed.) (1983). *Mathematical Theory of Experimental Design*, Nauka, Moscow.

[3] Fedorov, V.V.(1972) *Theory of Optimal Experiments*, Academic Press, New York.

[4] Kanwal, R. P. (1971). Linear Integral Equations, Theory and Technique. Academic Press, New York.

[5] Mitchell, T.J. (1974). An Algorithm for Construction of D-optimal Experimental Design. *Technometrics*, 16, 203-210.

[6] Pukelsheim, F. (1993). *Optimal Design of Experiments*, Wiley, New York.

[7] Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd Edition, Wiley, New York.

[8] Ripley, B. D. (1981). *Spatial Statistics*, Wiley, New York.

[9] Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sampling. *Journal of Applied Statistics*. 14, pp. 165-170.

[10] Silvey, S.D. (1980). *Optimal Design*, Chapman and Hall, London.