Prediction: Design of Experiments Based on Approximating Covariance Kernels

V. Fedorov
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831, U.S.A.

Abstract

Using Mercer's expansion to approximate the covariance kernel of an observed random function we transform the prediction problem to the regression problem with random parameters. The latter one is considered in the framework of convex design theory. First we formulate results in terms of the regression model with random parameters, then present the same results in terms of the original problem.

1 Model and optimality criteria

Let

\[ y(x) = u(x) + \varepsilon(x), \]

where variables \( x \in X \subseteq \mathbb{R}^d \) are usually coordinates of sensors (observing stations, sampling sites). We assume that \( u(x) \) and \( \varepsilon(x) \) are random variables which are not correlated with each other. 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 the prediction of \( u \), estimation of \( \theta \) or some functions of them. In (1) the first term \( u(x) \) describes deviations of the observed response from the mean 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 \( X \). We assume that \( u(x) \) and \( \varepsilon(x) \) are not correlated with each other. The same characters are used both for random variables and their realization, if it does not lead to confusion. The dependence between observations at different sites described solely through the covariance kernel of \( u \):

\[ E[u(x), u(x')] = K(x, x'), \quad E[u(x)] = 0. \]
DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.
DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.
The term \( \varepsilon(x) \) describes "observational" errors. We assume that these errors are specific for every particular site and moment. Thus, if there are \( r_i \) and \( r_i' \) observations at points \( x_i \) and \( x'_i \) correspondingly, then \( E(\varepsilon_{ij}, \varepsilon_{i'j'}) = \sigma^2 \delta_{i'j'} \delta_{ij} \), \( E(\varepsilon_{ij}) = 0 \), where \( i = 1, \ldots, n \) and \( j = 1, \ldots, r_i \). We skip the indices when it does not lead to confusion.

Let the kernel \( K(x, x') \) defined by (2) exist on \( Z \times Z \), where \( Z \) is compact in \( R^d \) and \( X \subset Z \). Let the eigenfunctions \( \varphi_{\alpha}(x) \) and the eigenvalues \( \lambda_{\alpha} \) of \( K(x, x') \) be defined as

\[
\lambda_{\alpha} \varphi_{\alpha}(x) = \int_Z K(x, x') \varphi_{\alpha}(x') dx'.
\]

All eigenvalues of \( K(x, x') \) are positive according to Mercer's Theorem (see, for instance, Kanwal (1971) for more rigid mathematics and details). The series \( K(x, x') = \sum_{\alpha=1}^\infty \lambda_{\alpha} \varphi_{\alpha}(x) \varphi(x') \) is uniformly and absolutely convergent, and the series \( \sum_{\alpha=1}^\infty \lambda_{\alpha} \) is convergent. Obviously \( \lambda_{\alpha} \) must diminish at least faster than \( \alpha^{-1} \). In many cases the decay is much faster. This fact allows us to hope that for practical needs we can use the approximation

\[
K(x, x') \approx K_m(x, x') = \sum_{\alpha=1}^m \lambda_{\alpha} \varphi_{\alpha}(x) \varphi_{\alpha}(x')
\]

with some moderate \( m \). The opportunity to use approximation (3) is essential for our approach.

Let the design \( \xi = \{p_i, x_i\}_i^n \), \( p_i = r_i/N \), \( N = \sum_{i=1}^n r_i \) describes the allocation of our observational efforts, and let

\[
\{K(\xi)\}_i = K(x_i, x_i), \{K(x, x_i)\}_i = K(x, x_i), \quad \bar{y}^T = (\bar{y}(x_1), \ldots, \bar{y}(x_n)),
\]

where \( \bar{y}(x_i) = \sum_{j=1}^{r_i} y_{ij}/r_i \). One can verify that

\[
E(\bar{Y}) = 0, \quad E(\bar{Y}^T \bar{Y}) = V(\xi) = \frac{\sigma^2}{N} W^{-1}(\xi) + K(\xi),
\]

where the matrix \( W(\xi) \) is diagonal and \( W_{ii}(\xi) = p_i \). Let the prediction of \( u(x) \) on a given set \( X_{pr} \) be an immediate goal for a practitioner. The estimator (compare with the similar exercises in Ripley (1981))

\[
\hat{u}(x) = K^T(x, \xi) V^{-1}(\xi) \bar{Y}
\]

minimizes

\[
\text{Var} \ (u(x) - \hat{u}(x)) = E \ [(u(x) - \hat{u}(x))^2] = K(x, x) - K^T(x, \xi) V^{-1}(\xi) K(x, \xi),
\]
given that $E(u(x) - \hat{u}(x)) = 0$, and where expectation is taken with respect to $u$ and $\varepsilon$. 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 (4) are $Q_1(\xi) = \max_{x \in \mathcal{X}} \operatorname{Var}(u(x) - \hat{u}(x))$ and $Q_2(\xi) = \int_{\mathcal{X}} \operatorname{Var}(u(x) - \hat{u}(x)) \, dx$. In general the design problem may be stated as

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

where $Q$ stands for either $Q_1$, or $Q_2$. In (5) we assume the weights $p_i$ may change continuously.

It is essential for our approach that $\sigma^2 \neq 0$ and approximation (3) is valid. On an intuitive level it also means that $\sigma^2 \gg \sum_{a=p+1}^{\infty} \lambda_a \phi_a(x) \phi_a(x')$ for all $x \in \mathcal{X}$.

2 Regression model with random parameters as an approximation of a random field.

If (3) is valid, then model (4) can be replaced by its approximate version

$$y = \gamma^T \varphi(x) + \varepsilon(x),$$

where the vector $\varphi^T(x) = \{\phi_a(x)\}_{1}^{m}$, the parameters $\gamma$ are random, and $E(\gamma) = 0$, $E(\gamma \gamma^T) = \Lambda$, $\Lambda_{a\beta} = \lambda_a \delta_{a\beta}$. In the framework of (6)

$$E(\tilde{Y} \tilde{Y}^T) = V(\xi_n) = \frac{\sigma^2}{N} W^{-1}(\xi) + \Phi^T(\xi) \Lambda \Phi(\xi),$$

where $\Phi(\xi) = (\varphi(x_1), \ldots, \varphi(x_n))$. We do not introduce subscript $m$ for vectors $\varphi$ and $\gamma$ because from now on only the first $m$ eigenfunctions are involved in all considerations. In what follows we consider only designs with a finite number of supporting points. This fact is not very restrictive, because for any design there exists a design with a finite number of supporting points and exactly the same information matrix (see, for instance, Fedorov and Hackl (1997)).

The best linear unbiased estimator for $\gamma$ is

$$\hat{\gamma} = \frac{N}{\sigma^2} \left( \frac{N}{\sigma^2} M(\xi) + \Lambda^{-1} \right)^{-1} \Phi(\xi) W(\xi) \tilde{Y},$$

$$M(\xi) = \Phi(\xi) W(\xi) \Phi^T(\xi).$$
This estimator minimizes the dispersion matrix of the difference \( \hat{y} - y \), and

\[
D(\xi_n) = E \left[ (\hat{y} - y)(\hat{y} - y)^T \right] = \left( \frac{N}{\sigma^2} M(\xi) + \Lambda^{-1} \right)^{-1}.
\]

Unbiasedness of \( \hat{y} \) means here that \( E(\hat{y} - y) = 0 \), where expectation is taken with respect to \( u \) and \( \epsilon \) (compare, for instance, with Pilz (1991)).

Let us select \( \hat{u}(x) = \varphi^T(x)\hat{y} \) as a predictor for \( u(x) \). On an intuitive level it is obvious that \( \hat{u}(x) \) and \( \hat{u}(x) \) must coincide in the framework of approximation (6). Indeed, using the identity \( (A^{-1} + BB^T)^{-1} = A - AB(B^T AB + I)^{-1}B^T A \), one can verify that

\[
\begin{align*}
V ar (\hat{u}(x) - u(x)) &= \varphi^T(x)D(\xi)\varphi(x) \\
&= K_m(x,x) - K_m^T(x,\xi) \left( \frac{\sigma^2}{N} W^{-1}(\xi) + K_m(\xi) \right)^{-1} K_m(x,\xi) \\
&= C_m(x,\xi) = V ar (\hat{u}(x) - u(x))
\end{align*}
\]

where the matrix \( K_m(\xi) \) and the vector \( K_m(x,\xi) \) are obviously defined parts of \( K(\xi) \) and \( K(x,\xi) \), respectively. The function \( K_m(x,x') \) is defined in (3). Consequently

\[
Q_1(\xi) = \max_{x \in X_{pr}} \varphi^T(x)D(\xi)\varphi(x)
\]

and

\[
Q_2(\xi) = tr AD(\xi), \quad A = \int_{X_{pr}} \varphi(x)\varphi^T(x)dx.
\]

3 Optimal designs for prediction

In this section we pursue a very transparent and simple idea. First, we formulate results for criteria (10) and (11) in terms of approximation (3), i.e., using our knowledge of \( \Lambda \) and \( \varphi(x) \). At the second stage we translate our findings to the language of covariance kernels and best linear predictors. It is expedient to note that, in general, the optimal design \( \xi^* \) depends upon \( N \) and \( \sigma^2 \) similar to some problems in the Bayesian approach to experimental design theory, e.g., Pilz (1991). When \( N \) is fixed, all the results developed in the convex design theory may be used almost directly. For instance, for the linear criterion (11) we have:

**Theorem 1** The design \( \xi^* \) is linear optimal (i.e. it minimizes (11)) if and only if for all \( x \in X \)

\[
\psi_1(x,\xi^*) \leq tr M(\xi^*)D(\xi^*)AD(\xi^*),
\]

\[
(12)
\]
where \( \psi_1(x, \xi) = \varphi^T(x)D(\xi)\varphi(x) \), and equality takes place at all supporting points of \( \xi^* \).

For criterion (10) in the simplest case when \( X_{pr} = X \), we have the following analogue to the Kiefer-Wolfowitz theorem.

**Theorem 2.** The design \( \xi^* \) is minimax if and only if for all \( x \in X \)
\[
\psi_2(x, \xi^*) \leq \text{tr} M(\xi^*)D(\xi^*),
\]
where \( \psi_2(x, \xi) = \varphi^T(x)D(\xi)\varphi(x) \), and equality takes place at all supporting points of \( \xi^* \).

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

Both theorems are routine results following from convex design theory. Thus, in the framework of approximation (3) the design for correlated observations cases can be embedded in the well developed area of convex design theory. Having the results stated in the above theorems and using presentation (7), let us try to re-formulate results in terms of the covariance kernels.

From the same identity, which was used to derive (9) it follows that
\[
D(\xi) = \Lambda - \Lambda \Phi(\xi) \left( \frac{\sigma^2}{N} W^{-1}(\xi) + K_m(\xi) \right)^{-1} \Phi^T(\xi) \Lambda.
\]

Combining (12) and (14) and introducing the function
\[
C_m(x, x', \xi) = K_m(x, x') - K_m(x, \xi) \left( \frac{\sigma^2}{N} W^{-1}(\xi) + K_m(\xi) \right)^{-1} K_m(x', \xi)
\]
we come to the following result.

**Theorem 3.** The design \( \xi^* \) minimizes the average variance of prediction if and only if for all \( x \in X \)
\[
\int_{X_{pr}} C_m^2(x, x', \xi^*) dx' \leq \int_X \int_{X_{pr}} C_m^2(x, x', \xi^*) \xi^*(dx) dx',
\]
and equality holds at all supporting points of \( \xi^* \).
For the minimax case with $X_{pr} = X$ we have the following theorem.

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

$$C_m(x_i^*, \xi^*) = \max_{x \in X} C_m(x, \xi^*).$$

2. Minimax designs coincide with D-optimal designs:

$$\xi^* = \arg \max_{\xi} \left| W^{-1}(\xi) + K_m(\xi) \right| |W(\xi)|.$$

The latter optimization problem 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. The above results remarkably remind the results for the discrete $X$ presented in Batsell et al. (1998). The comparative analysis of two approaches is beyond the scope of the present paper.

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 Fedorov and Hackl (1997). 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

$$\min_{x \in X_s} \zeta(x, \xi_s) \text{ or } \max_{x \in X_s} \zeta(x, \xi_s),$$

where the sensitivity function $\zeta$ may coincide either with $r^{-1} \psi_1$ or $r^{-1} \psi_2$ correspondingly, and $X_s = \text{supp } \xi_s$. The algorithm is simple, but it is necessary to know the eigenfunctions $\varphi(x)$.

Theorems 3 and 4 allow the development of numerical procedures, that “directly” use the covariance kernel $K_m(x, x')$. For instance, the first order exchange algorithm can be written for the minimax criterion with $X_{pr} = X$ as follows:

**Step a.** There is a design $\xi_s$. Find

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

and construct $\xi_{s+} = \xi_s + \alpha_s \delta(x_s^+)$, where $\delta(x)$ is a probability measure atomized at $x$.

**Step b.** Find

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

6
where $X_s = \text{supp} \xi_s^+$, and construct $\xi_{s+1} = \xi_s - \alpha_s \delta(x_s^-)$.

The changes which must be done in the case of linear criterion are evident. To guarantee convergence of the above iterative procedure, the sequence $\{\alpha_s\}$ may be chosen similarly to what was proposed in standard design theory.

5 Acknowledgments

This study was supported in part by the Applied Mathematical Sciences Research Program, Oak Ridge National Laboratory, managed by Lockheed Martin Energy Research Corp. for the U.S. Department of Energy under contract number DE-AC05-96OR22464.

References


