ON THE CONVERGENCE OF THE KRIGING METHOD

SÁNDOR MOLNÁR

Central Mining Development Institute, H – 1037 Budapest, Mikoviny u. 2 – 4.

(Received June 12, 1983 – revised September 19, 1984)

Abstract. In this paper some mathematical properties of the classical and universal Kriging method will be investigated. Under certain conditions the convergence of the methods is shown if the number of observation points tends to infinite. The speed of the convergence is also estimated, and these estimations can be used in the practice. The effect of the approximating data is also included in the error estimations.

1. Introduction

In this paper the generalized version of the classical Kriging method, the so called universal Kriging method will be examined. The convergence of this method and the speed of the convergence will be investigated under certain assumptions. The error formulas derived in this paper are very useful in the application, since the uncertainty of the estimates can be forecasted before making the actual measurements and therefore they give an important aid in locating the observation points.

2. The universal Kriging method

Let $f(x)$ and $n(x)$ be uncorrelated real-valued functions defined on a domain X in $\mathbf{R}^m$. Suppose $\{(x_i, y_i)\}_{i=1}^{N}$ is a sequence of "noisy" function pairs; that is, suppose

$$y_i = f(x_i) + n(x_i), \quad (1 \le i \le N). \tag{2.1}$$

The interpretation is that $f(x)$ is a function whose values are to be estimated, and $n(x)$ represents a noise if a measurement is taken at position x. We suppose that the noise has zero mean, i.e. $E(n(x)) = 0$ and variance var $(n(x)) = \sigma^2$ not depending on $x \in X$. We discuss below two problems which are central in the Kriging literature.

Problem 1. Let $x^* \in X$ be specified. It may or may not be among the sample pairs. On the basis of the sample pairs $\{(x_i, y_i)\}_{i=1}^N$,

(a) provide an estimate $f_N(x^*)$ of $f(x^*)$, and
(b) provide an estimate of the expected squared error

$$E[(f_N(x^*)-f(x^*))^2|x_1, \ldots, x_n].$$

Remark. Because practitioners desire to estimate piezometric head in oil and water aquifers or the grade of an ore body as a function of positon, the dimension m of the domain X is often 2 or 3.

Problem 2. Let $\{(x_i, y_i)\}_{i=1}^N$ be as above and let D be a subregion of domain X.

(a) estimate the integral $\int_D f(x)\,\mathrm{dx}$, and
(b) provide a formula for the (sample-dependent) expected square error of this estimate.

Remark. An application motivating Problem 2 is that of estimating the total weight of metal which can be extracted from the ore body occupying volume D, given imperfect assay estimates of the grade at distinct locations.

Problems 1 and 2 seem to have their roots in the forestry and geostatistics literature. In fact, it seems that "geostatistics" is almost synonymous with Kriging. We have no doubt that Problems 1 and 2 are important and interesting. For example, in the mining industry $f(x)$ is the thickness of the deposit or the value of a quality parameter of the deposit. In the first case the integral gives the total volume of the deposit under the region D, and in the second case, the average value $\int_D f(x)\,dx/|D|$ give the average value of a quality parameter.

In the Kriging approach, it is presumed that $f(x)$ and $n(x)$ in (2.1) are realizations of stochastic processes uncorrelated from one another with finite second moments. It is further assumed that $f(x)$ is a realization of an *intrinsic random function* (IRF); that is, for some functions $\{\Phi_i(x)\}_{i=1}^J$ known to the user and perhaps unknown constants $a_1, \ldots, a_J$, for all x, h such that $x, x+$ $+h \in X$, one has

$$E[f(x)] = \sum_{j=1}^{J} a_j \Phi_j(x) \tag{2.2}$$

and, independently of x with "var" signifying "variance",

$$1/2 \text{ var } [f(x+h)-f(x)] = \gamma(h).$$

The constants $\{a_j;\ 1 \leq j \leq J\}$ and the function $\gamma(h)$ are quantities which must be inferred from the data $\{(x_i, y_i)\}_{i=1}^N$. In what follows, it is presumed always that $J \leq N$. The function $\gamma(h)$ is called the *variogram.* Even in the case in which the mean $E[f(x)]$ is known to be constant in x (i.e., $J = 1$, $\Phi_1 = 1$), the hypothesis of "intrinsic random function" is weaker than second-order stationarity. For example, Brownian motion is an intrinsic random function, but it is well known to be a nonstationary process.

The Kriging method is composed of two activities, (i) inferring the variogram from the data, and (ii) assuming that the inferred variogram is exact indeed , providing a best linear unbiased estimator and associated error variance, as required by Problem 1 or Problem 2.

Activity (ii) is a standard least-squares problem, and is consequently by far the best understood of the two facets of Kriging. There are some inconsistencies in the fundamental definitions and results in the Kriging literature. For example, the definitions of "intrinsic random function" given by David [1] and Matheron [3] do not coincide. The discussions of noise and the "nugget effect" have likewise been inconsistent. The equations for Kriging in the presence of noise as given by Rendu [4], for example, agree with our calculations, but differ from formulas offered by other authors (e.g. Journel [2]). In view of these inconsistencies, we have elected to derive the "universal Kriging" equations for prediction with known variogram from first principles.

The task of inferring a covariance function or power spectral density from data is known by experienced statisticians to be somewhat delicate, and and one which furthermore requires a considerable quantity of data. The subtleties of the covariance inference problem translate directly to the task of inferring a variogram from data.

There are some very real difficulties with variogram estimation in the published Kriging applications. To avoid effects of "non-stationarity", practitioners tend to have a single variogram apply only to a relatively small region X of domain points of $f(x)$, or we use different variograms in different directions of h. Moreover, they have not developed procedures to ascertain whether the intrinsic random function hypothesis is tenable for their applications. A particular difficulty is that in the bounded domain case, ergodic theorems are inapplicable to the task of demonstrating consistency. To our knowledge, with the exception of certain extreme cases such as white noise, no methods for inferring the covariance function from sample pairs $\{(x_i, f(x_i)\}$, $f(x)$ a fixed sample function are known to be consistent.

First, we concern ourselves with outlining the present practice with regard to variogram inference. The recommended procedure is to choose a parametric family of variograms from the five or six popular families mentioned in the literature, and then to select the variogram from the chosen family which agrees best, in some sense, with the covariance function constructed from the data $\{(x_i, y_i)\}_{i=1}^{N}$. We list in Table 2.1 some of the prominent variogram families.

Monomial $\qquad \gamma_\Theta(h) = \omega(h)^a$

Spherical $\qquad \gamma_\Theta(h) = \begin{cases} \omega\left[\frac{3}{2}\frac{|h|}{a} - \frac{1}{2}\left(\frac{|h|}{a}\right)^3\right], & |h| \leq a \\ \omega, & |h| > a \end{cases}$

Exponential	$\gamma_\Theta(h) = \omega[1-\exp(-\lvert h\rvert/a)]$
Gaussian	$\gamma_\Theta(h) = \omega[1-\exp(-\lvert h\rvert^2/a^2)]$

where $\Theta = (a, \omega)$.

Table 2.1. A listing of popular variogram families

There seems to be no consensus in the lit rature on methodology for the selection of a parametric family from Table 2.1 on the basis of an observed sample $\{(x_i, y_i)\}_{i=1}^n$. Some heuristic approaches are proposed by David [1]. Concerning the task of selection of the member $\gamma_\Theta(h)$ the foremost criteria seem to be

(i) least squares,
(ii) cross validation and
(iii) a geometric procedure [1].

In the least squares approach, one selects the parameter Θ^* so as to nimimize

$$I_1(\Theta) = \sum_\nu(\gamma_n(h_\nu)-\gamma_\Theta(h_\nu))^2$$

where the index ν is running over some finite collection of arguments h_ν and $\gamma_n(h)$ being some sample approximation to the variogram, such as

$$\gamma_n(h) = 1/(2N(h))\left[\sum_{j=1}^{N(h)}(y_j-y_j(h))^2\right],$$

where $j(h)$ is an index selected so that $\lvert x_{j(h)}-x_j\rvert = h$ and $N(h)$ is the number of such points selected. If "drift" is thought to be present (that is, if Φ_j, $j>1$, in (2.2) is not zero), then this approach entails some serious conceptual difficulties, Matheron [3, Chapter 4] has addressed these difficulties.

The cross-validation apprach to parameter selection is as follows. Let $P(x_j, \Theta)$ be the universal Kriging estimate of $f(x_j)$ on the basis of the sample points $\{(x_i\ y_i)\}_{i\neq j}$ and parametric variogram $\gamma_\Theta(h)$. One then chooses Θ^* to minimize the squared error of the predicted values , which is

$$I_2(\Theta) = \sum_{j=1}^n(y_j-(\Theta,x_j))^2.$$

Practitioners insist, quite rightly, that one should not select a variogram entirely algorithmically, but with attention also to past experience with similar geological data.

Next, the linear estimation for $f(x^*)$ will be examined supposing that the variogram is already known.

To begin with, suppose the noise term in (2.1) is zero. Let us assume that the variogram $\gamma(h)$ and the mean function components $\{\Phi_i(x)\}$. of the expectation (2.2) are given. The assumption that one of these functions, say Φ_1, is 1, seems to be a universal and perhaps unavoidable assumption which we will also adopt. To begin with, let us discuss the solution of Problem 1. The objective is to choose the parameters $\{\lambda_i\}_{i=1}^{N}$ so that the linear estimator

$$f_N(x^*) = \lambda_1 y_1 + \ldots + \lambda_N y_N \tag{2.3}$$

minimizes

$$E[(f(x^*) - f_N(x^*))^2]$$

subject to

$$E[f_N(x^*)] = E[f(x^*)]. \tag{2.4}$$

In view of the assumed form (2.2) of the mean value function, a sufficient (but not necessary) condition for the unbiasedness equation (2.4) to hold is that

$$\sum_{i=1}^{n} \lambda_i \Phi_j(x_i) = \Phi_j(x^*), \quad 1 \leq j \leq J \tag{2.5}$$

Equation (2.5) with $\Phi_1 = 1$, implies that

$$\sum_{i=1}^{n} \lambda_i = 1.$$

Use this fact, with the unbiasedness of the estimator $f_N(x^*)$ of $f(x^*) = f^*$, to conclude that, with "cov" signifying "covariance",

$$\begin{aligned} E\left[\left(f^* - \sum_{i=1}^{N} \lambda_i y_i\right)^2\right] &= \operatorname{var}\left(f^* - \sum_{i=1}^{N} \lambda_i y_i\right) = \operatorname{var}\left(\sum \lambda_i (f^* - y_i)\right) = \\ &= \sum_i \sum_j \lambda_i \lambda_j \operatorname{cov}\left[(f^* - y_i),\ (f^* - y_j)\right]. \end{aligned} \tag{2.6}$$

Now observe that

$$\begin{aligned} \operatorname{cov}\ [(f^* - y_i), (f^* - y_j)] &= 1/2[-\operatorname{var}((f^* - y_i) - (f^* - y_j)) + \\ &+ \operatorname{var}(f^* - y_i) + \operatorname{var}(f^* - y_j)] = -\gamma(x_i - x_j) + \gamma(x^* - x_i) + \gamma(x^* - x_j). \end{aligned}$$

One substitutes these into (2.6) and after some easy calculus, sees that the Lagrange multiplier technique for minimizing $E[(f(x^*) - f_N(x^*))^2]$ subject to (2.5) yields

$$\sum_{k=1}^{N} \lambda_k \gamma(x_i - x_k) = 2\gamma(x_i - x^*) + \sum_{j=1}^{J} \mu_j \Phi_j(x_i), \quad 1 \leq i \leq N \tag{2.7}$$

$$\sum_{i=1}^{N} \lambda_i \Phi_j(x_i) = \Phi_j(x^*), \quad 1 \leq j \leq J. \tag{2.8}$$

The variables μ_j are the Lagrange multipliers, Journel [2] calls the above linear equation the *universal Kriging system*.

Substituting (2.7) into (2.6), one concludes that the means square prediction error is given by

$$E[(f^*-f_N(x^*))^2] = \sum_{i=1}^{N} \lambda_i \gamma(x^*-x_i) - \sum_{j=1}^{J} \mu_j \Phi(x^*).$$

If the noise term $n(x)$ in (2.1) has zero mean, one accounts for its presence by noting that, because it is presumed uncorrelated from the f-process,

$$\operatorname{cov}((f^*-y_i),(f^*-y_j)) = \operatorname{cov}((f^*-f_i-n_i),(f^*-f_j-n_j)) =$$
$$= \operatorname{cov}((f^*-f_i),(f^*-f_j)) + \operatorname{cov}(n_i, n_j).$$

In the above equation n_i and f_j denote $n(x_i), f(x_j)$. As a result one readily sees that in the presence of noise (2.7) should be replaced by the following relations:

$$\sum_{k=1}^{N} \lambda_k \gamma(x_i-x_k) - 2\operatorname{cov}(n(x_i),\ n(x_k)) = \gamma(x_i-x^*) +$$
$$+\sum_{j=1}^{J} \mu_j \Phi_j(x_i), \quad 1 \le i \le N.$$

Let us now investigate the modifications necessary for solution of Problem 2 described before. Assume $\int_D dx = 1$. In this case, we replace the objective (2.6) by the task of minimizing

$$E\left[\left(\int_D f(x)dx - \sum \lambda_i y_i\right)^2\right]$$

subject to

$$E[\sum \lambda_i y_i] = E\left[\int_D f(x)dx\right].$$

The preceding Kriging analysis leads, in the integral estimation case, to the following universal Kriging system;

$$(2.9)\quad \begin{aligned} &\sum_{k=1}^{N} \lambda_k \gamma(x_i-x_k) = \int_D \gamma(x_i-x)\,dx + \sum_{j=1}^{J} \mu_j \Phi_j(x_i), \quad 1 \le i \le N, \\ &\sum_{i=1}^{N} \lambda_i \Phi_j(x_i) = \int_D \Phi_j(x)dx. \end{aligned}$$

The expected square error of the integral estimate is given by

$$E\left[\left(\int_D f(x)dx - \sum_{i=1}^{N} \lambda_i y_i\right)^2\right] = \sum_{i=1}^{N} \lambda_i \int_D \gamma(x_i - x)dx - \tag{2.10}$$
$$- \sum_{j=1}^{J} \mu_j \int_D \Phi_j(x)dx - \int_D \int_D \gamma(x-x')dxdx'.$$

Note, that equations (2.7), (2.8) and (2.9) can be solved by using standard techniques (Szidarovszky and Yakowitz, [5]).

3. Convergence and consistency

As has been noted earlier, there is no consistent variogram estimator based on observations $\{(x_i, \bar{f}(x_i))\}^N$ for x_i in a bounded domain X and $\bar{f}$ a fixed sample of an intrinsic random function f. Note that $f(x_i) = y_i$ in short, the variogram cannot be consistently inferred, even if it is known to be a member of a given family such as listed in Table 2.1. On the other hand, as we will later demonstrate, under certain circumstances, the Kriging estimate will converge, with increasing number of samples, to the correct value, even when the variogram is not correct. An interpretation of these remarks is that the Kriging method can be effective for estimating values on the basis of noisy samples, but that the associated error estimate need not be consistent.

The fact that the estimate of the square error need not become more accurate with increasing data is significant because Kriging practitioners and their clients place great value on the error estimation feature.

Let us begin our analysis of convergence of Kriging estimate under the simplest conditions by assuming that

(i) The observations are noiseless $(n(x_i) = 0)$.
(ii) $\gamma(0) = 0$, and γ is continuous in a neighborhood of the origin.
(iii) There is no "drift", that is, $J = 1$ and $\Phi_1 = 1$.
(iv) The "true" variogram is known.

Theorem 3.1. *Let X be the domain of the intrinsic random function $f(x)$ and assume the conditions above are satisfied. If the infinite sequence $\{x_i\}$ is dense in X, then for any $x^* \in X$ and for $f_N(x^*)$ as in* (2.3),

$$E[(f(x^*) - f_N(x^*))^2] \to 0 \textit{ as } N \to \infty.$$

Proof. In view of assumption (iii), for every i, $y_i = f(x_i)$ is itself an unbiased linear estimator of $f(x^*)$ and so for $N \geq 1$

$$E[(f(x^*) - f_N(x_i))^2] \leq E[(f(x^*) - f(x_i))^2] = 2\gamma(x^* - x_i).$$

Let $x^*(N)$ denote the member of $\{x_i\}_{i=1}^N$ which is closest to x^*. By the assumption that $\{x_i\}$ is dense, $x^*(N) \to x^*$ as $N \to \infty$, and therefore

$$E[f((x^*) - f_N(x^*))^2] \leq E[(f(x^*) - f(x^*(N)^2] = 2\gamma(x^* - x^*(N)). \tag{3.1}$$

The proposition follows by observing that, in light of property (ii), $\gamma(x^*-$ $-x^*(N))$ must converge to 0. □

The bound given by (3.1) is of some practical interest in itself.

The Browian motion process affords an example ef a situation in which the best estimate is not consistent unless x^* is an accumulation point of the sample points $\{x_i\}$. For Brownian motion is Markov, and the best estimate of $f(x^*)$ will depend only on the points $(x_a, f(x_a))$ and $(x_b, f(x_b))$. where x_a is the largest domain sample less than x^* and x_b the smallest sample greater than x^*.

There are many common situations in which the hypothesis that $\{x_i\}$ is dense in X will be satisfied. One important case is that in which the x_i' s are selected independently according to a measure that assigns positive probability to every open set (such as when the probability density function exists and is positive).

Corollary. *Let $\{x_i\}$ be the above dense sequence in X and let*

$$\varepsilon_N = \max_{x\in D}\min_{1\le i\le N} |x-x_i|,$$

then (3.1) *implies that*

$$E[(f(x^*)-f_N(x^*))^2]\le 2\gamma(\varepsilon_N)$$

where it is assumed that $\gamma(h)$ depends only on the length $|h|$ of vector h.

Consider now the general case, when $J\ge 1$. Let $x_i^*(N)$ denote the mearest point to x^* selected from the finite set $x_1, \dots, x_N$. Assume again that the infinite set $\{x_i\}$ is dense in D, furthermore the matrix

$$\Phi_N = (\Phi_j(x_i(N)))_{j,\,i=1}$$

is nonsingular. Let λ_N be the solution of the equation

$$\Phi_N\lambda_N = \Pi$$

where

$$\Pi = (\Phi_1(x^*), \dots, \Phi_J(x^*))^T.$$

Then the components of λ_N obviously satisfy the conditions (2.5), and by using the notation

$$\lambda_N = (\lambda_N(\lambda_1^N, \dots, \lambda_J^N)$$

and the Cauchy inequality we get

$$E[(f(x^*)-f_N(x))^2]\le E\left[\left(f(x^*)-\sum_{j=1}^{J}\lambda_j^N f(x_j(N))\right)^2\right] =$$

$$= E\left[\left(\sum_{j=1}^{J}\lambda_i^N f(x^*)-f(x_j(N))\right)^2\right]\le 2\sum_{j=1}^{J}|\lambda_j^N|^2\gamma(x^*-x_j(N)).$$

Observe that inequality (3.2) implies the following theorem.

Theorem 3.2. *Assume that the assumptions* (i), (ii), (iv) *of the previous theorem hold, the matrix Φ_N is nonsingular, the infinite sequence $x_1, x_2, \ldots$ is dense in D, furthermore the lengths of the vectors λ_N are bounded by a constant not depending on N. Then for $N \to \infty$,*

$$E[(f(x^*)-f_N(x^*))^2] \to 0.$$

The effect of the noise term $n(x_i)$ will next be discussed. Let the estimate obtained from the exact functional values be

$$f_N(x^*) = \sum_{i=1}^{N} \lambda_i f(x_i),$$

where λ_i is the solution of the Kriging equations. The estimate obtained from the noisy data has the similar form:

$$\tilde{f}_N(x^*) = \sum_{i=1}^{N} \lambda_i [f(x_i) + n(x_i)].$$

The error is defined by the expression

$$\varepsilon = \tilde{f}_N(x^*) - f_N(x^*) = \sum_{i=1}^{N} \lambda_i n(x_i).$$

The expected value of this error is by supposition

$$E[\varepsilon] = \sum_{i=1}^{N} \lambda_i E[n(x_i)] = 0,$$

and the variance of the error can be given as follows:

$$E[\varepsilon^2] = E\left[\left(\sum_{i=1}^{N} \lambda_i n(x_i)\right)^2\right] = \sum_{i=1}^{N} |\lambda_i|^2 \delta^2.$$

In evaluating the convergence statements concerning Kriging in the previous discussions, it should be emphasized that they are valid only if $f(.)$ really is an intrinsic random function and the variogram and drift functions are *known perfectly*.

Our next discussion of Kriging convergence is directed to Problem 2 of the previous section, i.e., the integral estimation problem. For Problem 2, as has been observed earlier, one must modify the universal Kriging equation development by replacing f^* in (2.6) by $\int_D f(x)\, dx$. The effect of this substitution is that $\gamma(x_i - x^*)$ and $\Phi_j(x^*)$ are replaced by $\int_D \gamma(x_i - x)\, dx$ and $\int_D \Phi_j(x) dx$ in (2.7) and (2.8), respectively.

Let $I(f)$ denote the universal Kriging estimate of $\int_D f(x) dx$ obtained by the modifications just mentioned, and let $f_N(x^*)$ denote the Kriging estimate of $f(x^*)$. Recall our assumption that $\int_D dx = 1$. Tben we have the following result.

Theorem 3.3.

$$I(f) = \int_D f_N(x)dx.$$

Proof. One may express (2.7), (2.8) in matrix form as

$$\boldsymbol{\lambda}(x^*) = A^{-1}C(x^*)$$

where $\boldsymbol{\lambda}(x^*) = (\lambda_1, \ldots, \lambda_N, \mu_1, \ldots, \mu_J)^T$,

$$c_i(x^*) = 2\gamma(x_i - x^*), \quad 1 \leq i \leq N,$$

$$c_{j+N}(x^*) = \Phi_j(x^*), \; 1 \leq j \leq J,$$

and

$$A_{ij} = \gamma(x_i - x_j), \; 1 \leq i, \; j \leq N; \; A_{j+N,i} = A_{i,j+N} = \Phi_j(x_i),$$
$$1 < i < N, \; 1 < j < J.$$

From (2.3), we see that if we define $\mathbf{B} = (f(x_1), \ldots, f(x_N), 0, \ldots, 0)$, then

$$f_N(x^*) = \mathbf{BA}^{-1}\mathbf{c}(x^*).$$

Now it is clear from (2.9) that the universal Kriging equation for the integration problem may be represented as

$$I(f) = \boldsymbol{\beta}\mathbf{A}^{-1}\int_D \mathbf{c}(x)dx = \int_D \boldsymbol{\beta}\mathbf{A}^{-1}\mathbf{c}(x)dx = \int_D f_N(x)dx$$

and our proposition is established. □

The predicted mean square error was given in (2.10). But the following evident result is useful:

Corollary.

$$E\left[\left(I(f) - \int_D f(x)dx\right)^2\right] \leq \sup_{x \in D} \text{Var}\,(f_N(x)).$$

Proof. By using the Cauchy inequality we get

$$E\left[\left(I(f) - \int_D f(x)dx\right)^2\right] = E\left[\left[\left(\int_D (f_N(x) - f(x))dx\right)^2\right]\right] \leq$$

$$\leqq \sup_{x \in D} E[(f_N(x) - f(x))^2] = \sup_{x \in D} \text{Var}\,(f_N(x)). \; \square$$

REFERENCES

[1] *David M.:* Geostatistical Ore Reserve Estimation. Elsevier, New York, 1977.

[2] *Journel A. G.:* Kriging in terms of projections. *J. Math. Geol.* **9** (6) (1977), 563 – 586.

[3] *Matheron G.:* The Theory of Regionalized Variables and its Applications, Les Cahiers du CMM. Fasc, no. 5, ENSMP, Paris (1971), 211 p.

[4] *Rendu J.:* Disjunctive Kriging: comparison of theory with actual results. *Mathematical Geology* **12** (4) (1980), 305 – 320.

[5] *Szidarovszky F.* and *Yakovitz S.:* Principles and Procedures of Numerical Analysis. Plenum Press, New York, 1978.