PARAMETRIC ESTIMATION OF COVARIANCES OF REGIONALIZED VARIABLES

Peter K. Kitanidis

ABSTRACT: The usefulness of stochastic models in describing the spatial variability of hydrogeologic quantities, such as permeability, storativity, piezometric head, seepage velocity, and solute concentrations is now widely recognized. In practice, these quantities are represented as the sum of a well-structured component, or drift, and a more erratic fluctuation component which is described statistically through its covariance function. This paper reviews some of the most recent and most promising methods for the estimation of parameters of these covariances from existing data. They are maximum likelihood, restricted maximum likelihood, minimum-variance unbiased quadratic estimation, and minimum-norm (weighted least squares) estimation. The applicability of such methods to conditional and unconditional probability problems is discussed.

(KEY TERMS: spatial processes; parameter estimation; covariance functions.)

INTRODUCTION

Aquifers and aquitards are highly heterogeneous, reflecting the complexity of the natural processes which created them over thousands of years. The spatial variability of hydrogeologic parameters, such as hydraulic conductivities and storativites, can only partially be described through deterministic functions. Much of their variability is highly erratic and can best be described statistically through the use of averages, such as variance and spatial correlation of fluctuation among some mean values. One may then evaluate the effect of spatial variability of hydrogeologic parameters on dependent quantities. A case in point is the spatial variability of hydraulic head and the enhanced mixing of solutes resulting from spatial variability of permeability (Gelhar and Axness, 1983).

A probabilistic description is also useful in making the best use of measurements in predictions. Geologic formations are known to use directly through measurements of properties at some points or indirectly through measurements of their response to given excitations (e.g., pumping tests and geophysical exploration methods). Even when such measurements are available, description of the properties of aquifers and aquitards involves a significant degree of uncertainty. A case in point is the problem of interpolating the values of transmissivity on a fine regular grid, subsequently used to draw contour maps, from measurements of transmissivity at some locations. How can the undesirable effects of uncertainty be minimized so that the most accurate possible estimates can be obtained? And, since the estimates are intended for use in decision making, what is the accuracy of estimation? Such questions can be answered using conditional probabilities methods, particularly geostatistical estimation (see Delhomme, 1970; Kitanidis and Vomvoris, 1983; and Dagan, 1984, for selected applications).

In practice, spatial variability is partially characterized by the first two moments, the mean (or drift) and the covariance function. Before the model can be applied, these first two moments must be determined from data and prior information. For example, for the application of the popular linear minimum-variance unbiased estimation technique known as kriging, the variogram or generalized covariance function must be determined from data. It is well-known in statistical practice that the estimate of covariance functions or power spectra, particularly in the case of relatively few measurements in a fixed size domain, can be fraught with difficulties. Kriging practitioners correctly warn against treating the estimation of covariances as a purely algorithmic problem. However, this caveat should not be interpreted as permission to disregard the important problem of parameter estimation of covariances or, even worse, to use procedures which provide little hope of satisfactory results. Instead, one must recognize that the scientific approach to the development of a parametric model is a three-step procedure:

(a) After preliminary examination of the data, select a parametric model for the mean and the covariance function.
(b) Estimate parameters conditional on the validity of the model.
(c) Confront the model with the data. This process may lead to model modification and/or scrutiny of the data.

In practice, this procedure may have to be repeated several times until an acceptable model is identified. The third step, known as validation or diagnostic checking, involves formal or informal tests of "residuals," splitting the data into a training and a testing set, etc. Before we limit our attention

---

2Civil Engineering, Terman Engineering Center, Stanford University, Stanford, California 94305.
to parametric estimation, which is the topic of this paper, it is important to point out that model validation is equally important. It is the performance of the fitted model in validation tests which will suggest whether it is appropriate for making predictions or just another mathematical curiosity.

This work is a review of methods for parametric estimation of covariances. A disclaimer is in order at this point. The review is not meant to be comprehensive. Instead, this work presents the methods the author is most familiar with and, perhaps not coincidentally, considers the most promising. Section 2 summarizes the commonly used "linear model" and introduces the problem of parameter estimation. The problem is presented in the most general terms possible so that its results can be applicable to a wide variety of cases, including kriging, universal kriging, and cokriging. Section 3 reviews maximum likelihood estimation and Section 4 restricted maximum likelihood estimation of covariance-function parameters. Section 5 reviews minimum-variance unbiased quadratic and Section 6 minimum norm (weighted least squares) estimation. Section 6 also presents a discussion of the normality assumption. Section 7 extends the results of Sections 3 and 4 to account for prior information on drift coefficients and covariance-function parameters. Section 8 discusses the applicability of such methods in the context of conditional and unconditional probabilities, data availability, and computational cost.

2. THE GENERAL LINEAR MODEL AND LINEAR ESTIMATORS

The common assumption is that the regionalized variable, such as the logarithm of transmissivity, is a realization of a random function given by the general linear model:

\[ y(x) = \sum_{i=1}^{p} f_i(x) \beta_i + e(x) \] (1)

where \( x \) is the vector of spatial coordinates of the point where \( y \) is sampled; \( \beta_i, i=1, \ldots, p \) are (generally unknown) parameters; \( f_i(x), i=1, \ldots, p \) are known functions of the spatial coordinates, such as polynomials; and \( e(x) \) is a zero-mean spatial random function. Note that, throughout this paper, underbar letters stand for matrices and underbar lower case letters stand for vectors. The first term on the right-hand side represents a drift (or trend or prior mean of the function) and the second term represents a zero-mean random field ("stochastic part"). Random measurement error, if it exists, may be represented through a term which may be absorbed in \( e \). The stochastic part has covariance function \( \mathbb{R}(u,v|\theta) \), where \( \theta \) is an n×1 vector of parameters, defined through

\[ \mathbb{E}[e(u)e^T(v)] = \mathbb{R}(u,v|\theta) \] (2)

From the general representation given by (1) and (2), one may obtain as special cases most of the useful and commonly assumed models. For example, for \( p=1, f_1(x) = 1 \) and \( \mathbb{R}(u,v) = R(u,v|\theta) \), Equation (1) represents a stationary isotropic field with mean equal to \( \beta_1 \). In other cases, polynomials or periodic functions may be used to represent the drift. The joint analysis of more than one field, such as piezometric head and logtransmissivity may be achieved as described in Journel and Huijbregts (1978), who use a purely statistical approach, or in Kitanidis and Vomvoris (1983) and Hoeksema and Kitanidis (1984, 1985b), who invoke the pertinent flow equations.

Assume that there are \( n \) measurements arranged, for the sake of notational convenience, in an \( n \times 1 \) vector \( y \). These may be measurements at points, weighted averages over given areas, gradients of the function, or any other linear combinations of one or more spatial functions. In all these cases, Equation (1) yields the following general relation

\[ y = X\theta + e \] (3)

where \( X \) is a known \( n \times p \) matrix of the known functions \( f(x) \) of the spatial coordinates; \( \theta \) is the \( p \times 1 \) vector of drift coefficients; and \( e \) is a random vector with zero mean and covariance matrix \( Q_{yy}(\theta) \) which is a known function of the parameter vector \( \theta \). Note that if \( y = c + X\theta + e \), where \( c \) is a known \( n \times 1 \) vector, the problem may be still reduced to the form of Equation (3) by subtracting \( c \) from the data. This case is illustrated in Kitanidis and Vomvoris (1983).

If the drift is not linear in the \( \theta \) coefficients, then an iterative linearization approach may be used, as in Hoeksema and Kitanidis (1985b). Summarizing the basic assumptions associated with (3), the deterministic effects are linear in the drift parameters and the measurements are linearly related to the spatial function (or functions). These commonly made assumptions constitute the "general linear model."

One could devise many procedures, to be called estimators, for utilizing data to assign numerical values to the parameters \( \theta \) and \( \theta \). Superior estimators are those which have desirable properties conditional on the truth of the model. One such property is that on the average they should neither overestimate nor underestimate the values of the parameters. If the average error is zero, the estimator is called unbiased. It is also desired that the estimate have a small squared difference from the actual value. If the mean squared error is the smallest possible, the estimator is called minimum variance. If the estimate tends to the true value of the parameters as more measurements are obtained, the estimator is called consistent.

Before embarking on the problem of estimating \( \theta \) consider the estimation of \( \hat{\theta} \) given \( \theta \). In this case, there is a widely acceptable estimator, \( \hat{\theta} \), which is given from the solution of

\[ (X^TQ_{yy}^{-1}X)\hat{\theta} = X^TQ_{yy}^{-1}y \] (4)
for which the inverse of covariance matrix of estimation errors is

$$P = X^T Q_{yy}^{-1} X$$

(5)

The solution of (4) is the one which minimizes the weighted sum of squares criterion $(y - X\hat{\beta})^T Q_{yy}^{-1} (y - X\hat{\beta})$. The estimator is unbiased and, among all linear estimators, is the one with the smallest possible estimation variables. The estimator is also consistent. However, note that consistency refers to the "effective size" of the sample. If all measurements are taken from the same region and are highly correlated, the effective size of the sample cannot exceed a certain value and the drift coefficients do not tend to their true values no matter how many measurements are obtained.

3. MAXIMUM LIKELIHOOD ESTIMATION

A general approach for the simultaneous estimation of $\hat{\beta}$ and $\hat{\theta}$ is maximum likelihood (ML) estimation. The common assumption is that the data is Gaussian so that the joint pdf of the $n$ measurements is

$$p(y | \theta, \beta) = (2\pi)^{-n/2} |Q_{yy}|^{-1/2} \exp[-\frac{1}{2} (y - X\theta)^T Q_{yy}^{-1} (y - X\theta)]$$

(6)

where $|\cdot|$ denotes determinant. The problem then becomes one of nonlinear optimization. In practice, the negative log-likelihood

$$L(y | \theta, \beta) = -\ln p(y | \theta, \beta)$$

$$= \frac{1}{2} \ln |Q_{yy}| + \frac{1}{2} (y - X\theta)^T Q_{yy}^{-1} (y - X\theta)$$

(7)

where the constant term is neglected, is minimized subject to constraints on the parameters imposed to make the model physically and mathematically plausible.

A general discussion of ML parameter estimation is given in Kitanidis and Lane (1985). For the general linear model and assuming that no constraint is binding, the necessary conditions for minimum of $L$ are

$$(g_{\beta})_j = \frac{\partial L}{\partial \theta_j} = \frac{1}{2} \ln |Q_{yy}| + \frac{1}{2} (y - X\theta)^T Q_{yy}^{-1} (y - X\theta) = 0, \quad j = 1, \ldots, p$$

(8)

and

$$(g_{\theta})_j = \frac{\partial L}{\partial \theta_j} = \frac{1}{2} \ln |Q_{yy}| + \frac{1}{2} (y - X\theta)^T Q_{yy}^{-1} (y - X\theta)$$

(9)

where we introduce the notation $Q_i = \frac{\partial Q_{yy}}{\partial \theta_i}$. Note that given $\theta$, one may obtain from Equation (8) the familiar estimator of $\beta$ already given in Equation (4).

Equation (9) must usually be solved iteratively to calculate $\theta$ for given fitted values of $\hat{\beta}$. The Hessian, i.e., the matrix of the second derivatives needed for the application of Newton-type iterative methods (see Kitanidis and Lane, 1985) can be partitioned into

$$H = \begin{bmatrix} H_{\beta\beta} & H_{\theta\beta} \\ H_{\beta\theta}^T & H_{\theta\theta} \end{bmatrix}$$

(10)

with elements

$$(H_{\beta\beta})_{ij} = (X^T Q_{yy}^{-1} X)_{ij}$$

(11)

$$(H_{\theta\beta})_{ij} = (-X^T Q_{yy}^{-1} Q_{yy}^{-1} (y - X\theta))_i$$

(12)

$$(H_{\theta\theta})_{ij} = \frac{1}{2} \ln |Q_{yy}| + \frac{1}{2} (y - X\theta)^T Q_{yy}^{-1} (y - X\theta)$$

(13)

where

$$F_{\beta\theta} = E(H_{\beta\theta}) = \frac{1}{2} \ln |Q_{yy}| + \frac{1}{2} (y - X\theta)^T Q_{yy}^{-1} (y - X\theta)$$

(14)

$$F_{\theta\theta} = E(H_{\theta\theta}) = 0$$

(15)

and the i,j element of $F_{\theta\theta} = E[H_{\theta\theta}]$ is

$$(F_{\theta\theta})_{ij} = \frac{1}{2} \ln |Q_{yy}| + \frac{1}{2} (y - X\theta)^T Q_{yy}^{-1} (y - X\theta)$$

(16)

\[\text{WATER RESOURCES BULLETIN} \]
The Fisher information matrix can be used to advantage in the iterative Gauss-Newton method to estimate \( \hat{\theta} \) and \( \theta \). Gauss-Newton is Newton optimization in which the Hessian is replaced by the Fisher information matrix. When applied to maximum-likelihood estimation, the approach is widely known as the method of scoring (Rao, 1973). Mardia and Marshall (1984) suggest updating

\[
\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - F_{\theta\theta}^{-1} e_{\theta}
\]

(17)

where \( e_{\theta} \) is calculated using estimate \( \hat{\theta}^{(k)} \). Then \( \hat{\theta}^{(k+1)} \) is recalculated from Equation (4).

This iteration, which is a special case of the more general Gauss-Newton method described in Kitanidis and Lane (1985), has been found by Mardia and Marshall to be more efficient than other iterative procedures. A steepest descent procedure is described in Hartley and Rao (1967). The simulations and analysis of Kitanidis and Lane (1985) suggest that Gauss-Newton can be the most efficient computational procedure, particularly if an appropriate statistical model has been selected. In particular, they present examples which show that the Gauss-Newton method can converge faster than Newton’s method (i.e., the one which uses the Hessian instead of the Fisher information matrix). Note that calculation of \( F_{\theta\theta} \) does not require calculation of second derivatives of \( Q_{yy} \) w.r.t. \( \theta \) and is much more efficient than calculation of \( H_{\theta\theta} \).

The inverse of the Fisher information matrix provides a lower-bound approximation to the estimation error covariance matrix of \( \theta \) and \( \hat{\theta} \). \( \nabla_{\theta\theta} \) and \( \nabla_{\theta\theta} \), respectively.

\[
\nabla_{\theta\theta} \geq (x^T Q_{yy}^{-1} x)^{-1}
\]

and

\[
\nabla_{\theta\theta} \geq (E_{\theta\theta})^{-1}
\]

(18)

These matrices can be used to approximate the mean squared error of, and the correlation between, parameter estimates.

ML is a general and practical methodology, first mentioned in the context of estimation of parameters of covariance functions of regionalized variables in Kitanidis and Vomvoris (1983), to obtain estimates of drift and covariance function parameters and to evaluate their accuracy. In Section 6 we will argue that the assumption of normality is not as crucial as it may first appear and can be safely made in most cases.

4. RESTRICTED MAXIMUM LIKELIHOOD

Maximum likelihood estimates are known to be asymptotically unbiased, minimum-variance, consistent, and normally distributed with covariance matrix given by the inverse of the Fisher information matrix. These properties hold for effectively large samples as discussed in Hartley and Rao (1967) and illustrated in Mardia and Marshall (1984).

Unfortunately, these favorable properties may not materialize in hydrogeology, where measurements are often highly correlated so that the effective number of measurements may be much smaller than the total number of measurements. A common case is that of spatially correlated measurements obtained from a region of finite size where the correlation length is of the same order of magnitude as the longest distance between measurement locations. In this case: (i) the covariance matrix of estimation of drift coefficients cannot be reduced below a certain limit, no matter how many additional measurements are obtained from the same limited-size domain; and (ii) use of fitted estimates of the drift coefficients results in biased estimates of the covariance function parameters. Typically both variances and correlation lengths tend to be underestimated, and nonparametric estimates of the covariance function can exhibit an artificial “hole effect” (Kitanidis and Lane, 1985).

The bias in the estimates of \( \theta \) can be quite serious and has motivated a number of procedures in the analysis of variance and the geostatistical literature. Note that this problem is not by any means particular to maximum likelihood estimation but plagues all methods which use fitted estimates of the drift to estimate covariance-function parameters. If the random part has “large-scale” components, some of the variability of the random part is fitted by the drift. It is obvious that it is impossible to distinguish on the basis of measurements between the drift and the large-scale components of the random term.

Motivated by these and other considerations, Matheron (1971) developed procedures which do not require the fitted drift coefficients. Such procedures work well with kriging, which is linear minimum-variance estimation unbiased for any value of the drift coefficients. These procedures employ generalized increments, i.e., linear combinations of the data which are invariant to the values of the drift coefficients. Similar methods, reviewed in Harville (1977) and Searle (1971), are used in the analysis-of-variance literature.

Restricted maximum likelihood (RML) is based on the simple idea of estimating \( \theta \) by maximizing the likelihood of some generalized increments. Since this likelihood does not depend on the values of the drift coefficients, the covariance-function parameter estimates are not distorted as a result of using incorrect estimates of these coefficients. This method was first proposed by Patterson and Thompson (1971) in the context of analysis of variance. In the context of spatial functions, RML has been used in Kitanidis and Vomvoris (1983), Kitanidis (1983), Hoeksema and Kitanidis (1984), 1985a, 1985b), and Kitanidis and Lane (1985).

Consider the transformation

\[
z = \mathbf{T} y
\]

(20)
where \( z \) is a k\( \times \)l vector of generalized increments and \( T \) is a k\( \times \)n matrix. \( T \) is selected so that \( TX = 0 \). Then \( z \) is Gaussian with

\[
E(z) = TX\theta = 0
\]  

(21)

\[
E(zz^T) = TQ_{yy}T^T
\]  

(22)

One may then proceed to apply the equations of the previous case except that \( y \) is substituted by \( z \), \( Q_{yy} \) by \( TQ_{yy}T^T \) and \( X\theta \) is dropped out. For example, one may obtain \( T \) by dropping from \( (q - X(X^TX)^{-1}X^T) \) as many rows as needed to make \( TQ_{yy}T^T \) invertible. Note that at least p rows must be dropped out. It turns out that, as long as one uses the lengthest vector of generalized increments (i.e., the largest k) with invertible covariance matrix, the results of RML do not depend on which transformation \( T \) is used (Kitanidis, 1986a).

5. MINIMUM-VARIANCE UNBIASED QUADRATIC ESTIMATION

The use of quadratic functions of the data to estimate variance components has a rather long history in the analysis of variance literature. In geostatistics, Matheron alludes to universal (i.e., invariant to drift coefficients) quadratic estimators in his 1971 monograph (page 191). Delfiner's (1976) weighted least squares schemes for the estimation of the parameters of the generalized covariance function can be interpreted as quadratic estimators. Rao (1971, 1972) is credited with the systematic development of optimal minimum-variance and minimum-norm unbiased quadratic estimators. In the context of spatial functions, the first use of such techniques was reported in Kitanidis (1983), followed by Stein (1984), Kitanidis (1985), and Marshall and Mardia (1985).

Consider the special case that the data covariance matrix is linear in the covariance function parameters

\[
Q_{yy}(\theta) = \sum_{i=1}^{w} Q_i \theta_i
\]  

(23)

where \( Q_i \) are known n\( \times \)n matrices. Examples of such cases include the nugget effect; the polynomial generalized covariance function; covariance functions with known correlation structure but unknown variance, such as the exponential with known correlation length; polygonal representations; and, of course, covariance functions formed by summing any of the aforementioned functions.

Following Kitanidis (1985), consider the quadratic estimators

\[
\hat{\theta}_j = Y_{F_j}^X, \quad j = 1, \ldots, m
\]  

(24)

where \( F_j \) are n\( \times \)n matrices. In the single-parameter case (m=1) it is intuitively obvious that \( \hat{\theta}_1 \) should be a quadratic function of the data; Equation (24) extends the same idea to the multiparameter case. Once the form of the estimator has been selected, the problem reduces to selecting the \( F_j \) matrices. The following criteria apply:

(i) Invariance to the Drift Coefficients of the Model. The estimates should not depend on the values of the unknown drift coefficients. Consequently, \( \hat{\theta}_j \) should not change when \( y \) is replaced by \( y + \chi b \), where \( b \) is any arbitrary vector of drift coefficients. This criterion introduces the constraints

\[
F_jX = 0
\]  

(25)

and

\[
F_j^TX = 0
\]  

(26)

(ii) Unbiasedness. The average value of \( \hat{\theta}_j \) should be the actual value \( \theta_j \). This requires (see Kitanidis, 1985) that

\[
\text{Tr}(F_jQ_i) = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}
\]  

(27)

Note that conditions (25) through (27) refer to the first two moments of the underlying stochastic process and require no assumptions in addition to those associated with Equations (3) and (23). In the general case, the solution to Equations (25) through (27) is not unique and a third requirement must be introduced to define the "best" solution.

(iii) Minimum-Variance. It is reasonable to look for the most accurate solution in the sense that the mean squared error of estimation

\[
E[|\hat{\theta}_j - \theta_j|^2] = E[\text{Tr}(F_j X_{F_j})] - \theta_j^2
\]  

(28)

is the smallest possible. If the expression of Equation (28) is known, one may proceed with its minimization subject to the constraints of Equations (25) through (27).

This approach follows a pattern familiar from linear minimum variance estimation. However, in the case of quadratic estimation the variances given by Equation (28) involve fourth moments of the stochastic process and consequently require the introduction of additional assumptions and parameters. In principle, one can always follow the usual procedure of assuming a model for the fourth moments (the same way we have assumed a model for the first and second moments), estimate its parameters, and corroborate its validity with data. But to obtain good estimates of fourth moments, knowledge of even higher moments is required. The well-known difficulties associated with estimating higher moments are a good enough reason to seek alternative methods to
close the problem. If the process is Gaussian, there is no need to introduce additional parameters since higher central moments can be expressed as functions of the covariance matrix. Thus, unless there is convincing evidence to the contrary, it is convenient to assume normality. Using the unbiasedness condition, Equation (28) reduces to

\[
E(\hat{\theta}_j - \theta_j)^2 = 2 \text{Tr}(E_j Q_{yy} E_j Q_{yy})
\]

(29)

If a priori guess, \(Q_0\), of the covariance matrix is available, the minimization of (29) subject to constraints (25) through (27) yields the following closed form solution (see Kitanidis, 1985, for the derivation)

\[
E_j = \sum_{k=1}^{m} C_{kj} \left[ Q_0^{-1} - Q_0^{-1} X (X^T Q_0^{-1} X)^{-1} X^T Q_0^{-1} \right] Q_0 \left[ Q_0^{-1} - Q_0^{-1} X (X^T Q_0^{-1} X)^{-1} X^T Q_0^{-1} \right]^{-1} X^T Q_0^{-1} \]

(30)

where \(C_{kj}\) are the elements of an \(m \times m\) matrix \(C\) calculated as the inverse of another \(m \times m\) matrix \(M\) with \(k^{th}\) element

\[
M_{kk} = \text{Tr} \left\{ \left[ Q_0^{-1} - Q_0^{-1} X (X^T Q_0^{-1} X)^{-1} X^T Q_0^{-1} \right] Q_0 \left[ Q_0^{-1} - Q_0^{-1} X (X^T Q_0^{-1} X)^{-1} X^T Q_0^{-1} \right]^{-1} X^T Q_0^{-1} \right\}
\]

(31)

Exactly the same results can be obtained by first transforming the data to authorized increments (Equation (20)) and then using them to obtain minimum-variance unbiased quadratic (MVUQ) estimators. This approach, which has certain computational advantages, is illustrated in Kitanidis (1983).

One may readily verify that the invariance and unbiasedness requirements are satisfied. However, the minimum variance requirement has been met only to the extent that \(Q_0\) is a good estimate of \(Q_{yy}\) and the data are Gaussian. In practice, \(Q_0\) is only an approximation to the actual covariance matrix, \(Q_{yy}\), and the results can be called locally minimum variance. Sensitivity analysis indicates that they are not very sensitive to \(Q_0\) and the single parameter case (\(m=1\)) does not depend on \(Q_0\) at all. As a final resort, one can always adopt the gross approximation \(Q_0 = 1\), the unit matrix, and still obtain a reasonable solution at a reduced computational cost from simplified versions of (30) and (31). However, neglecting correlations and unequal variances defies the purpose of introducing such advanced statistical techniques and does not result in better estimates than the unweighted least squares approach (Kitanidis, 1985; Stein, 1986). In practice, one can often come up with a reasonable \(Q_0\) which represents the anticipated spatial correlation.

Another approach to specifying \(Q_0\) is to apply the procedure iteratively by updating at the end of each iteration the estimate of \(Q_0\) based on the new values of the parameters. In principle, this procedure may yield biased estimates. However, the numerical simulations of Kitanidis (1985) indicate no appreciable bias even for moderately large samples while at the same time giving results which are practically as good as MVUQ with \(Q_0\) calculated with the actual parameters. The procedure converged in all cases and the results did not depend on the values of initial \(Q_0\). Actually, a rather interesting result is that iterative MVUQ and restricted maximum likelihood yield identical results, a fact first shown in Kitanidis (1983). This equivalence holds for any sample size.

6. MINIMUM-NORM (LEAST SQUARES) UNBIASED QUADRATIC ESTIMATION

For the case examined in Section 5, one may develop quadratic estimators of \(\theta_1, \ldots, \theta_m\) which satisfy invariance and unbiasedness (conditions i and ii) but, instead of seeking minimum variance (condition iii), minimize a reasonable fitting criterion (or norm). Such methods have been used by Rao (1972) and Marshall and Mardia (1985). Their approach parallels the development of minimum-variance unbiased quadratic estimators. Here, we will follow the approach of Kitanidis (1983, 1985) (which may be seen as a generalization of Delfiner’s 1976 method) which gives the same results and, at the same time, provides additional insight.

Consider the linear transformation of the data

\[
y' = W y
\]

(32)

where \(W\) is a given transformation matrix selected so that \(W X = 0\). Thus, \(y'\) is a vector of authorized increments. According to the assumed model, the covariance of \(y'\) as defined by (32) is \(W Q_{yy} W^T\). But the covariance matrix of \(y'\) is the expected value of \(W y' y' W^T\). Consequently, a reasonable approach is to fit the parameters of \(Q_{yy}\) so that the sum of the squares of all elements of \((W y' y' W^T - W Q_{yy} W^T)\) is minimized. This criterion may be written

\[
\min L = \sum_{i} \sum_{k} (W (Q_{yy} - y y^T) W^T)_{ik}^2
\]

(33)

\[
= \text{Tr} [W^T (W (Q_{yy} - y y^T) W^T) W (Q_{yy} - y y^T)]
\]

For the case of covariance matrix linear in the parameters (Equation (23), this is a quadratic optimization problem with solution given by the system of \(m\) linear equations with \(m\) unknowns

\[
\sum_{j=1}^{m} \text{Tr}(W^T W Q_j W^T W) \theta_j = y^T W^T W Q_{yy} W^T W y,
\]

(34)

\[i = 1, \ldots, m\]
One may readily verify that the solution is a quadratic estimator of the form of Equation (24) and satisfies invariance and unbiasedness (Equations 25-27) while at the same time minimizing the weighted squares criterion.

The same idea can be extended to the general case where $Q_{yy}(\theta)$ is a nonlinear function of the parameters. Then the minimization problem is not quadratic but the solution can be obtained iteratively using the Gauss-Newton method. Starting with an initial estimate of the covariance parameters $\hat{\theta}(0)$, iterate according to:

$$\hat{\theta}(k+1) = \hat{\theta}(k) - M^{-1}g$$  \hspace{1cm} (35)

where $g$ is the m x 1 vector with $i^{th}$ element

$$g_i = 2 \text{Tr}[W^T Q_i W (Q_{yy} - \mu y^T)]$$  \hspace{1cm} (36)

$M$ is m x m symmetric matrix with $ij^{th}$ element

$$M_{ij} = 2 \text{Tr}[W^T Q_i W_i W_j]$$  \hspace{1cm} (37)

and $Q_{ij} = \frac{\partial Q_{yy}}{\partial \theta_i}$ is calculated using $\hat{\theta}(k)$. $M^{-1}$ also serves in an approximation to the estimation error covariance matrix. For more details, see Kitanidis (1986b).

Minimum norm solutions are intuitively appealing and do not invoke prior estimates of the covariance matrix or the normality assumption. The estimates are optimal in the sense of minimizing a reasonable fitting criterion.

However, a well known fact (see, for example, Sorooshian and Dracup, 1980) that is often missed in the practical application of least squares or regression techniques is that minimization of a fitting criterion does not necessarily yield the most accurate estimates of the parameters. For least squares to yield the smallest possible mean squared error, it is required that the terms whose sum of squares is minimized have the same variance and are mutually uncorrelated. These assumptions are not satisfied in the estimation of spatial functions unless the $W$ matrix has been carefully selected. For example, $W = 1 - X(X^T X)^{-1}X^T$ may appear to be a reasonable choice, but generally yields estimates with significantly larger variance than when the covariance of the residuals has been appropriately taken into account. This has been illustrated in the simulations of Kitanidis (1985) and in the analytical results for a special case reported in Marshall and Mardia (1985, page 522). The point is also made in Stein (1986).

One may then ask: for which choice of $W$ are the weighted residuals of the least squares criterion (Equation 33) uncorrelated and with the same variance? Or, for which $W$ are the estimates minimum variance? It turns out that this $W$ is a function of the covariance matrix of the data as well as the fourth moment of the data. The dependence on fourth moments has discouraged some practitioners of geostatistics from seeking the most accurate solution possible. Their argument is that it would be of little use since fourth moments are very difficult to estimate in practice.

For Gaussian data, the minimum-variance choice corresponds to the value of $W$ which satisfies

$$W^T W = Q_o^{-1} - Q_o^{-1} X (X^T Q_o^{-1} X)^{-1} X^T Q_o^{-1}$$  \hspace{1cm} (38)

where $Q_o$ is the best estimate of $Q_{yy}$. However, for this choice of $W$ the minimum-norm solution is the same with the minimum-variance solution of Section 5. That is, one may view the minimum-variance unbiased quadratic solution of Section 5 as application of Delfiner's (1976) method for a particular (and particularly good if the data are nearly normal) selection of generalized increments (Equation 32).

Thus, minimum-variance unbiased quadratic and the related restricted maximum likelihood estimation methods of Sections 3 through 5 are acceptable procedures even for non-Gaussian data since they can be interpreted as minimizations of reasonable fitting criteria.

The normality assumption may not be readily acceptable by geostatisticians who, understandably enough, would rather avoid assumptions about complete probability distributions limiting attention to the first two moments. Not that normality is new to linear geostatistics, where it is assumed whenever needed to determine confidence intervals of estimates or to detect outliers. Variable transformations (such as logarithmic transformation of permeability and transmissivity) are often introduced to make the data more Gaussian-like. Actually, if the data are in gross violation of the normality assumption, the first two moments can be quite inadequate to characterize the probability laws of the estimator and, at the same time, one can find nonlinear estimators which have much smaller mean square estimation errors than the assumed linear estimator. Thus, in a sense, normality should not be such a bad assumption whenever kriging is deemed an appropriate estimation procedure.

Another important point is that minimum-variance estimates depend only partially on moments higher than the second. For example, Rao (1972) considered the case of data with a common kurtosis and showed that MVUQ estimates which presume normality partially minimize the actual estimation variance. Or, going in the opposite direction, one can verify from Rao's results (page 113) that the true minimum-variance estimate is not sensitive to perturbations of the kurtosis from the value which corresponds to the normal. (The value of the measure of reliability, the variance itself, is more sensitive.) Kitanidis (1985) generated data from Equation (3) with non-Gaussian fluctuations and then estimated covariance parameters. It turned out that the restricted maximum likelihood approach (which assumes normality but also accounts for the correlation structure of the data) gave much more accurate results than unweighted least squares or minimum norm with $W = 1 - X(X^T X)^{-1}X^T$. Thus, the methods of Sections 3 through 5 are recommended for
7. USE OF PRIOR INFORMATION ABOUT PARAMETERS

In most analyses prior information is used only in the selection of the functions representing drift and covariance. Prior information about the parameters is not used except in the form of constraints to make the model mathematically or physically plausible. For example, parameters should be selected so that an ordinary covariance function should be positive definite; a parameter which represents variance or correlation length must be nonnegative; and so on.

In some cases there may be knowledge extraneous to the data about the parameters of the model from past studies or based on the judgment of the professional who conducts the study. It is important that this information be used, particularly if the data set is relatively uninformative about some of the parameters. Prior and data information can be combined in a Bayesian framework. In this section we will review only the extension (or "pseudo-Bayesian" equivalent) of maximum likelihood techniques presented in Sections 3 and 4. Extension of other methodologies (such as minimum norm unbiased quadratic estimation) is possible but goes beyond the scope of this review paper.

Let $p''(\theta, \beta)$ be the prior joint pdf of the drift and covariance function parameters. Then the posterior pdf, $p''(\theta, \beta)$, can be determined from Bayes theorem

$$p''(\theta, \beta) = c \cdot p'(\theta, \beta) \cdot p(y | \beta, \theta)$$

where $c$ is a positive coefficient of proportionality and $p(y | \beta, \theta)$ is given by Equation (6). A point estimator of $\theta$ and $\beta$ is defined by the values which maximize $p''(\theta, \beta)$, i.e., the mode of the posterior distribution. This estimator, known as maximum a posteriori probability (MAP) is a generalization of the ML estimator described in Section 3. For diffuse priors (uniformly distributed throughout the range of possible values), MAP reduces to ML.

The generalization of restricted maximum likelihood estimation when prior information is available is somewhat less obvious but very instructive. Following Kitanidis (1986a), assume that the prior conditional pdf of $\beta$ given $\theta$ is Gaussian with mean $b'$ and inverse of covariance matrix $P'$. Then, using $p'(\theta, \beta) = p(\beta | \theta)p'(\theta)$, the posterior joint pdf of $\theta$ and $\beta$ is

$$p''(\theta, \beta) = c \cdot |P'|^{1/2} \exp \left\{ -\frac{1}{2} (\beta - b')^T P'(\beta - b') \right\} p'(\theta)$$

$$\cdot |Q_{yy}|^{-1/2} \exp \left\{ -\frac{1}{2} (y - \beta \theta)^T Q_{yy}^{-1} (y - \beta \theta) \right\}$$

(40)

From these equations, one can derive the posterior marginal pdf of $\theta$ (Kitanidis, 1986a, Equation 33)

$$p''(\theta) = \int p''(\theta, \beta) d\beta$$

$$= c \cdot p(\theta)|Q_{yy}|^{-1/2} |P'|^{1/2} p' + X^T Q_{yy}^{-1} x |^{-1/2}$$

$$\exp \left\{ -\frac{1}{2} (y^T Q_{yy}^{-1} - Q_{yy}^{-1} x (P' + X^T Q_{yy}^{-1} x)^{-1} x^T Q_{yy} y \right.$$}

$$+ b^T P' b' - b^T P' (P' + X^T Q_{yy}^{-1} x)^{-1} p' b') \}$$

(41)

while the posterior conditional pdf of $\beta$ given $\theta$ is Gaussian with mean

$$b'' = (P' + X^T Q_{yy}^{-1} x)^{-1} [X^T Q_{yy}^{-1} x + p' b'])$$

(42)

and inverse of covariance matrix

$$P'' = P' + X^T Q_{yy}^{-1} x$$

(43)

One may then proceed as follows:

(1) Estimate $\hat{\theta}$ which maximizes $p''(\theta)$, i.e., determine the mode of the posterior marginal pdf of $\theta$ given by Equation (41). This can be achieved using the Gauss-Newton method to minimize the negative logarithm of $p''(\theta)$.

(2) Once $\hat{\theta}$ has been determined, $b''$ and $P''$ can be calculated from Equations (42) and (43), if appropriate.

Note that $p''(\theta)$ does not involve drift coefficients fitted from the data. Estimates of $\theta$ depend only on prior estimates of the drift coefficient.

Consider the case of no prior information about the drift coefficients. This can be examined as a special case of Equation (41) by taking the limit $P' \rightarrow 0$.

$$p''(\theta) = c \cdot p(\theta)|Q_{yy}|^{-1/2} |Q_{yy}^{-1} x |^{-1/2}$$

$$\exp \left\{ -\frac{1}{2} y^T Q_{yy}^{-1} x \right.$$}

$$(X^T Q_{yy}^{-1} x)^{-1} x^T Q_{yy} y \}$$

(44)

where, again, $c$ is a coefficient of proportionality. An alternate expression is given in Harville (1974). One may verify that if an arbitrary drift were introduced in the data, i.e. $y$ were replaced by $y + Xb$ where $b$ is any pxd vector representing drift coefficients, the argument of the exponent would
Parametric Estimation of Covariances of Regionalized Variables

not be affected. Consequently, \( p''(y \mid y_0) \) depends only on combinations of the measurements which are unaffected by drift coefficients. Such increments are known in geostatistics as "generalized or authorized increments" and in the analysis of variances as "contrasts." An important conclusion is that only generalized increments appear in the posterior marginal pdf of \( \theta \) when no prior information is available about the drift coefficients.

Furthermore, it can be shown (see Kitanidis, 1986a) that \( p''(\theta) \) is proportional to \( p'(\theta) \) times the likelihood of any complete set of generalized increments. By complete we mean a set whose variance matrix has the maximum possible rank, which cannot be larger than \( n-p \). Thus, the MAP procedure which uses the mode of \( p''(\theta) \) as an estimator is a natural generalization of RML estimation. The analysis is also very useful in that it provides theoretical support to the method of RML estimation.

Note that at the other extreme (a priori known drift coefficients or \( P^{p-1} \rightarrow 0 \)), the posterior marginal of \( \theta \) tends to

\[
p''(\theta) \propto Q_{yy}^{-1/2} \exp\left[-\frac{1}{2} (y - Xb')Q_{yy}^{-1}(y - Xb')\right] p'(\theta)
\]

The marginal likelihood thus has the same form as Equation (6) except that \( b' \) represents a priori known drift coefficients, not parameters estimated from the sample.

8. CHOOSING AMONG PARAMETRIC ESTIMATION METHODS

Up to this point, we have dealt with parametric estimation methods that have favorable properties. However, the most appropriate among these methods must be selected on a case-by-case basis because it depends on the intended use of the developed models, data availability, and computational capabilities. We will address, in general terms, some of these points.

Regarding intended use, an important distinction is between conditional and unconditional probabilities:

1. Unconditional Probabilities. Examples include the derivation of the mean and covariance function of the piezometric head or the concentrations given that Equation (1) describes the variability of log-permeability using the pertinent flow and mass transport equations (e.g., see Gelhar and Axness, 1983). In this case, the data are used only to identify (parametrically estimate) and validate the model. Subsequently, only the fitted model is used with no further reference to the measurements.

2. Conditional Probabilities. In many prediction problems, knowledge of the values of permeability, head, or concentration at some points may be used to advantage to reduce the mean squared estimation error of prediction at nearby locations. Thus, the data may be used first to determine covariances and then to estimate unknown quantities given these covariances and the data. This is the main theme of kriging and co-kriging or the work of Dagan (1984).

The method of unconditional probabilities often allows analytical treatment and yields results which depend on a few parameters of the drift and the covariance function or the spectrum. An implicit assumption associated with this method is that the drift is perfectly known (i.e., \( P'' \) of Equation 43 tends to infinity). In this case, the maximum likelihood estimation method of Section 3 gives practically unbiased covariance parameters and is recommended for use. The drift can be known either a priori or can be estimated from the data, provided that measurements are obtained over a region of size much larger than the integral scale of the stationary random function.

The method of conditional probabilities gives predictions which depend on some of the parameters of the model as well as the location and outcomes of measurements. This method is not well-suited for analytical results or the development of relations which can be transferred from one aquifer to another. However, it can yield more accurate predictions than the unconditional probabilities method by using measurements which are correlated with the variables to be estimated. Another advantage of this approach is that predictions are rather insensitive to model parameters. Before discussing what parameter estimation procedures are appropriate for this problem, we will review the conditional probabilities problem in the context of the linear model of Equations (1) and (3).

Consider the problem of estimating \( \gamma_0 \), an m \times 1 vector of quantities to be determined from data \( y \), and prior information. Like the measurements, \( \gamma_0 \) is assumed to be a realization of the linear model

\[
\gamma_0 = X_0 \theta + \epsilon_0
\]

where \( X_0 \) is a known m \times p matrix; \( \theta \) is the vector of drift coefficients; \( Q_{00} \) is the covariance matrix of \( \epsilon_0 \); and \( Q_{0y} = Q_{y0}^T \) is the cross-covariance matrix between \( \gamma_0 \) and \( \gamma \). Following the assumptions of Section 7, the prediction problem can be defined as that of determining the pdf of \( \gamma_0 \) conditional on the measurements and prior information about the parameters. According to Kitanidis (1986a), for the Gaussian case, this can be given by:

\[
\tilde{P}(y_0 \mid \gamma, \theta) = \int_{\theta} \tilde{P}(y_0 \mid \theta, \gamma) p''(\theta) d\theta
\]

where \( \tilde{P}(y_0 \mid \theta, \gamma) \) is Gaussian with mean

\[
E(y_0 \mid \theta, \gamma) = (X_0 - Q_{0y}Q_{yy}^{-1}X)(P' + X^TQ_{yy}^{-1}X)^{-1}X^TQ_{yy}^{-1}y
\]

\[
+ [Q_{0y}Q_{yy}^{-1}(X_0 - Q_{0y}Q_{yy}^{-1}X)(P' + X^TQ_{yy}^{-1}X)^{-1}X^TQ_{yy}^{-1}y]
\]

\[
(P' + X^TQ_{yy}^{-1}X)^{-1}X^TQ_{yy}^{-1}y
\]

565 WATER RESOURCES BULLETIN
and covariance matrix
\[
\tilde{\Sigma}(\mathbf{x}_0 | \mathbf{y}, \mathbf{x}) = (\Sigma_{00} - \Sigma_{0y} \Sigma_{yy}^{-1} \Sigma_{oy}) \\
+ (\mu_0 - \Sigma_{0y} \Sigma_{yy}^{-1} \mathbf{x})(\mathbf{p}' + \mathbf{x}^T \Sigma_{yy}^{-1} \mathbf{x})^{-1} \\
\times (\mu_0 - \Sigma_{0y} \Sigma_{yy}^{-1} \mathbf{x})^T
\] (49)

and \(p''(\theta)\) is the posterior marginal pdf of \(\theta\) given by Equation (41). Note that the tilde is used to indicate Bayesian distribution, i.e., one that accounts for parameter uncertainty. \(\tilde{\mathbb{E}}(\mathbf{x}_0 | \mathbf{y}, \mathbf{x})\) and \(\tilde{\Sigma}(\mathbf{x}_0 | \mathbf{y}, \mathbf{x})\) are generalizations of the kriging-ckriging equations which may be obtained as a special case for \(\mathbf{p}' = 0\).

In many applications of conditional probabilities, as more measurements are obtained \(\tilde{p}(\mathbf{x}_0 | \mathbf{y}, \mathbf{x})\) becomes much more insensitive to values of \(\theta\) than \(p''(\theta)\). Consequently, in the interest of avoiding the calculation of the multiple integral, one may substitute \(p''(\theta)\) with a unit impulse function (i.e., all the mass of the pdf concentrated at one point) at the value of \(\theta\) which maximizes \(p''(\theta)\). Then
\[
\tilde{p}(\mathbf{x}_0 | \mathbf{y}) \approx \tilde{p}(\mathbf{x}_0 | \hat{\theta}_{MAP}, \mathbf{y})
\] (50)

This, of course, is a generalization of the common procedure of first estimating variograms (or generalized covariance functions) and then using them in linear estimation. Another advantage of this analysis is that it makes clear that whenever kriging (or its more general form given by Equations 48 and 49) is used, it is \(p''(\theta)\) rather than \(p''(\hat{\theta})\) that is relevant. For example, in the common case that all information comes from the data set, restricted maximum likelihood is the appropriate procedure for estimation of covariance function parameters. It is also obvious that the validity of the method suggested by Equation (50) depends on the spread of \(p''(\theta)\) (parameter uncertainty) not in absolute terms but in comparison to the flatness of \(\tilde{p}(\mathbf{x}_0 | \hat{\theta}, \mathbf{y})\) with respect to \(\theta\) (insensitivity of linear estimator to parameters). Thus, even though the estimation error of covariance parameters implied by \(p''(\theta)\) may appear quite sizeable, it can often be neglected in linear estimation if sufficient measurements are available.

Regarding computational feasibility, note that the most accurate procedures involve inversion of the data covariance matrix \(\Sigma_{yy}\), as well as calculation of the trace. In the author's experience, the cost increases approximately with \(n^k\) where \(k\) is between 3 and 4. In many practically encountered cases, the number of the measurements is not too high and these methods can be applied at a reasonable computational cost. However, for large data sets the computational cost may become prohibitively high. Mardia and Marshall (1984) suggest partitioning the data into subsets, calculating the estimates from each subset, and then pooling the results. This idea has been used to advantage in Stein (1986). If the hydrogeologist is fortunate enough to have a large sample at his disposal, approximate methods may suffice.

9. CONCLUDING REMARKS

This paper has presented a review of methods for the parametric estimation of covariance functions of regionalized variables. Their common characteristics are that they employ conventional estimation techniques and seek the best possible solutions consistent with the assumed model. Parametric estimation of covariances of spatial functions remains an important area of research, particularly for cases other than Gaussian and in relevance to the prediction problem.

ACKNOWLEDGMENTS

The material is based upon work supported by the Minnesota Water Resources Center under Grant USDI 14-08-0001-G1025-02, the National Science Foundation under Grant EEC-8517598, and the Graduate College of the University of Minnesota. Comments by M. L. Stein and anonymous reviewers on a draft of the paper were helpful in improving the presentation.

LITERATURE CITED


