Cokriging Interpolation

Nargess Memarsadeghi *
Advisor: Dr. David Mount

Abstract

Interpolating noisy or scattered data points is of interest in many applications where gathering data at specific locations is either impractical or expensive. In particular, in geological and mining applications, where measurements of interest are available only at scattered locations, these interpolation methods are of special value. One of the popular interpolation methods used in these situations is *cokriging* [3, 2]. A special characteristic of cokriging is that it can utilize data of different nature to model and interpolate a particular attribute [1, 3, 2]. In this paper, we present the cokriging method and its mathematical derivation. Mainly, we discuss cokriging's assumptions, objectives, properties, and implementation.

1 Introduction

In this document we present a popular interpolation method called cokriging [3, 2]. Historically, this interpolation method has been used in geological and mining applications [5]. This method is among the basic methods in the field of *geostatistics*. The term"geostatistics" was first used by G. Matheron (1962) who defined it as "the application of the formalism of random functions to the reconnaissance and estimation of natural phenomena" [5]. In order to understand cokriging interpolation, we must have a good understanding of the mathematical and geostatistical definitions and properties on which it is based.

Cokriging involves the solution of an optimization problem with an equality constraint. It also involves calculations of some statistical quantities such as covariance matrices, variograms, etc. This paper is organized by first reviewing the necessary background material. Section 2 goes over the mathematics involved in solving an optimization problem with equality constraints. Then, we go over some statistical definitions and properties in section 3. In section 4, we define what is meant by kriging and cokriging, how the problem is formalized mathematically, and how it is solved algorithmically. This paper concludes by explaining why this problem is computationally expensive as well as introducing existing implementations of cokriging.

*This paper is submitted on December 06, 2004 to Computer Science Department of University of Maryland at College Park as a scholarly paper to partially satisfy requirements for Masters degree in Fall 2004 academic semester. Email:nargess@cs.umd.edu.

2 Mathematical Background for Solving Linear Systems

Problem: Let x be a vector in $\mathbb{R}^d$, and suppose we would like to minimize a function f subject to a linear constraint $Ax = b$, where x and b are d-vectors and A is a $d \times d$ matrix.

Approach 1: Let Z be a basis for null space of A. Then, it is well known that if $\overline{x}$ is a solution to $Ax = b$, so is $\overline{x} + Zv$, where v is any d-vector. Thus, we can restate our optimization problem with an equality constraint as an unconstrained optimization problem as follows: $\min_v \ f(\overline{x} + Zv)$.

Approach 2: Solving this minimization problem is equivalent to solving the Lagrangian equation, which is defined as follows:

$$L(x, \lambda) = f(x) - \lambda^T c(x),$$

where

$$c(x) = Ax - b.$$

Note that we are not minimizing the Lagrangian function. Rather we are finding a saddle point of this function. We will go over optimality conditions for a solution to our original problem and the meaning of Lagrange multipliers.

2.1 Optimality Conditions for Functions with Linear Equality Constraint

In terms of the first approach, necessary conditions for optimality are that the reduced gradient be zero and that the reduced Hessian be positive semidefinite [8]. That is: $Z^T \bigtriangledown f(x) = 0$ and $Z^T \bigtriangledown^2 f(x) Z$ is positive semidefinite. $Z^T \bigtriangledown f(x) = 0$ is also the sufficient condition for optimality.

In terms of the second approach, partial derivatives of the Lagrangian with respect to both x and λ must be zero. That is, first order necessary conditions for optimality are

$$\begin{aligned} \bigtriangledown_x L &= \bigtriangledown f(x) - A^T \lambda = 0, \text{ and} \\ -\bigtriangledown_\lambda L &= Ax - b = 0. \end{aligned}$$

2.2 Meaning of the Lagrange Multipliers

Suppose the solution to the above minimization problem is x^*. Also, suppose we have a point $\hat{x}$ very close to x^* so that $||x^* - \hat{x}|| \leq \epsilon$ and $A\hat{x} = b + \delta$, where both ϵ and $||\delta||$ are very small.

Then, we can approximate $f(\hat{x})$ using Taylor series expansion.

Note that since $A\hat{x} = b + \delta$ and $Ax^* = b$, we have $A(\hat{x} - x^*) = \delta$.

$$\begin{aligned} f(\hat{x}) &= f(x^*) + (\hat{x} - x^*)^T g(x^*) + O(\epsilon^2) \\ &= f(x^*) + (\hat{x} - x^*)^T A^T \lambda^* + O(\epsilon^2) \\ &= f(x^*) + \delta^T \lambda^* + O(\epsilon^2). \end{aligned}$$

This means that if we perturb b_i by δ_i, then optimal value is changed by $\delta_i \lambda_i^*$. Thus, λ_i is the change in the optimal objective per unit change in b_i. We say that, λ_i is the *sensitivity* of f to b_i [8]. For this reason Lagrange multipliers are also called *shadow prices* or *dual variables* [8].

3 Geostatistics Background

3.1 Assumptions

In this section we go over various statistical definitions. In order to do so, assume we are dealing with two random variables X and Y, such that X can take on the values $\{x_1, \ldots x_n\}$ and Y can take on the values $\{y_1 \ldots y_m\}$. Also let

$$\begin{aligned} \mu_x &= E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ and} \\ \mu_y &= E(Y) = \frac{1}{m}\sum_{i=1}^{m} y_i \end{aligned}$$

denote the expected values of these variables.

3.2 h-Scatterplots

An *h-scatterplot* is a graph that shows all possible pairs of data values separated in location by a certain distance in a particular direction. It is can be used to evaluate the continuity of data values over a certain distance in a particular direction. If data points at locations separated by h are similar, then they will plot close to line $x = y$. As the similarity among the data decreases, the clouds of points on the h-scatterplot becomes more diffuse. This feature is calculated by a *correlation coefficient*. The more similar the pairs of points are, the higher their correlation coefficient.

3.3 Spatial Analysis

When performing geostatistical modelling, some assumptions are usually made about the data. Some of the main assumptions made are defined below:

- *Stationarity Assumption* means that the statistics of a random function are invariant under translation [2, 3, 4].
- *Isotropic Assumption* means that data statistics are independent of direction. Thus, when calculating data statistics only the distance between the pairs of data points needs to be taken into account and not their orientation [2, 3, 4].
- *Anisotropic Assumption* indicates that variability of data changes as a function of direction. Thus, for computation of data statistics both the distance and orientation between pairs of data points needs to be taken into account [2, 3].
- *Intrinsic Hypothesis* indicates that variance may be unbounded [4].
- *Quasi-stationarity* implies that stationarity applies to a neighborhood of the data and not to the entire domain of data [4].

3.4 Covariance

Generally, the covariance between two random variables $x_i \in X$ and $y_j \in Y$ is defined as:

$$\begin{aligned} Cov(X,Y) &= E\left[(X-\mu_x))(Y-\mu_y\right] \\ &= E(XY) - \mu_x\mu_y. \end{aligned}$$

In geostatistics, two random variables z and z' of the same distribution are usually location dependant, that is, they are a function, lets say Z, of their locations. Let us denote these locations by u and u' respectively. Thus, $z = Z(u)$ and $z' = Z(u')$.

In geostatistics, $C(u,u')$ is a shorthand for $Cov(Z(u), Z(u'))$, where $Z(u)$ and $Z(u')$ are values of a random function in locations u and u'. The covariance between two random variables $Z(u)$ and $Z(u')$ is defined as follows:

$$\begin{aligned} C(u,u') &= Cov(Z(u), Z(u')) \\ &= E\left[(Z(u) - E(Z(u)))(Z(u') - E(Z(u')))\right] \\ &= E(Z(u)Z(u')) - E(Z(u))E(Z(u')). \end{aligned}$$

The above notation can be confusing if one needs to calculate random variables from different distributions at locations u and u', and thus one needs to specify random functions which are being used at each location.

The *stationary covariance*, $C(h)$, is defined as the covariance between two random variables $Z(u)$ and $Z(u+h)$, separated in location by vector h:

$$C(h) = Cov(Z(u+h), Z(u)) = E(Z(u+h)Z(u)) - (E(Z(u)))^2, \forall u, u+h \in A.$$

In other words, $C(h)$, is the covariance between pairs of random variables that are separated by vector h from each other. Thus, we can also represent $C(h)$ as follows:

$$C(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} x_i y_i - m_{-h} m_{+h}.$$

3.5 Variance

Variance is defined for a single random variable. For a variable X, variance is defined as follows:

$$\begin{aligned} Var(X) &= Cov(X,X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 \\ &= E\left[X - \mu_x\right]^2 \\ &= E(X^2) - \mu_x^2. \end{aligned}$$

Recall the definition of stationary covariance function $C(h)$ from 3.4. Notice that when $h = 0$ we have:

$$C(0) = Cov(Z(u), Z(u)) = Var(Z(u)).$$

Thus, for a stationary random variable $Z(u)$, we have $C(0) = Var(Z(u))$.

3.6 Standard Deviation

Standard deviation of a variable is defined as square root of its variance: $\delta(X) = \sqrt{Var(X)}$.

3.7 Correlation Coefficient

Correlation coefficient, ρ, is a measure of linear relationship between two variables, or how close the values come to falling into a straight line. For two variables X and Y we have

$$\rho(x, y) = \frac{Cov(x, y)}{\delta(x)\delta(y)}$$

, where $\delta(x)$ and $\delta(y)$ are standard deviations of variable X and Y respectively.

stationary correlation coefficient, $\rho(h)$, is defined as the correlation coefficient between values of random function Z at locations u and $u + h$:

$$\rho(h) = \rho(Z(u + h), Z(u)) \quad \forall u, u + h.$$

3.8 Variogram

Variogram is a measure of calculating spatial variability or dissimilarity between values of a random variables approximately separated by a vector h. This measure can be used as an alternative to measure $C(h)$, described above. For a set of points, variogram, or $2\gamma(h)$, is defined as follows ([1, 3, 2]):

$$2\gamma(h) = \frac{1}{N(h)} \sum_{h_{ij}=h} (x_i - x_j)^2.$$

- $N(h)$ is number of pairs separated by vector h.
- x_i and x_j are values of variables at two ends of the vector.
- $h_{ij} = loc(x_i) - loc(x_j)$, where $loc(x_i)$ is the location where x_i is calculated, and similarly $loc(x_j)$ is the location where x_j is measured.

Similarly, *semivariogram*, is defined as half of the average squared difference between two attributes separated by vector h:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{h_{ij}=h} (x_i - x_j)^2.$$

We expect spatial variability of values of a random variable, ie. its variogram, to increase as the distance between locations of those values increase. However, after reaching a certain distance, this increase in variogram function stops. The distance at which variogram function stops increasing is called *range* of variogram, and the value that variogram has at distance equal to range is called the *sill*.

The sill value of the variogram is also the variance of the random function [3]. That is, $C(0) = \gamma(\infty)$. In other words the maximum variability of the random function values whose locations are far enough from each other is the same as maximum similarity among values of a random function evaluated at the same location (see 3.8 for more information).

3.8.1 Important Notes

1. Variogram and semivariogram are used interchangeably in practice. In fact, in most cases, a semivariogram as defined above is calculated, while for convenience it is referred to as variogram.

2. For a *stationary* random function variogram is defined as the variance of the increment between two random variables separated by vector h:

$$2\gamma(h) = Var(Z(u+h) - Z(u)), \quad \forall u.$$

   Where Z is a random function taking a location as its parameter.

3.9 Cross-Variograms

The definition of variogram can be extended to the following *cross-variogram* equation where we are dealing with values of two random variables from different distributions ([3, 2]).

$$2\gamma_{XY}(h) = \frac{1}{N(h)} \sum_{h_{ij}=h} (x_i - x_j)(y_i - y_j),$$

where

- $N(h)$ is number of pairs whose locations are separated by vector h.
- x_i, y_i and x_j, y_j are values of two variables at two ends of the vector.
- $h_{ij} = loc(x_i) - loc(x_j) = loc(y_i) - loc(y_j)$, where $loc(x_i)$ is the location where x_i is calculated, and similarly for other values.

3.9.1 Important Notes

1. Similar to semivariograms, we also have notion of *semi-cross-variogram*, and the same comment mentioned about variograms and semivariograms applies to cross-variograms and semi-cross-variograms.

2. For a stationary random function cross-variogram is defined as the covariance of the increment between two random variables separated by vector h:

$$\begin{aligned} 2\gamma(h) &= Cov\left[(Z_i(u) - Z_i(u+h)), (Z_j(u) - Z_j(u+h))\right] \\ &= E\left[(Z_i(u) - Z_i(u+h)).(Z_j(u) - Z_j(u+h))\right], \quad \forall i, j, u, \end{aligned}$$

where Z_i and Z_j are two random functions, each taking a location as their parameter.

3.10 Variogram Modeling

As we will see in section 5, in order to perform cokriging (see 4.1), we need to model variograms. That is, we need to fit variogram values (see 3.8) as function of distance h to a function which best fits it. Variograms are usually modelled so that we be able to model pairwise covariances as a function of distance h. For stationary data, having a variogram model $\gamma(h)$ allows us to come up with a covariance model $C(h)$, as a function of distance, using relation $C(h) = C(0) - \gamma(h)$ (see proof of this equation in 3.11, 3a). In other words, once we have a model for variogram, the covariance model for distance h can be calculated by subtracting $\gamma(h)$ from variogram's sill value.

As we will see in section 4, to perform kriging/cokriging it is important that the involved covariance matrix C be positive definite. Thus, only models which will result in positive definite covariance matrices are considered for modelling variograms. Some of the most frequently used models are as follows [2, 3]:

- **Nugget effect model**:

$$\gamma(h) = \begin{cases} 0 & \text{if} \quad h = 0, \\ 1 & \text{otherwise.} \end{cases}$$

- **Spherical model with range a**: This is the mostly used variogram model, where a represents the range of the variogram. This model is defined as follows:

$$\gamma(h) = \begin{cases} 1.5\frac{h}{a} - 0.5(\frac{h}{a})^3 & \text{if} \quad h \leq a, \\ 1 & \text{otherwise.} \end{cases}$$

- **Exponential model with practical range a**:

$$\gamma(h) = 1 - \exp(-\frac{3h}{a}).$$

- **Gaussian model with practical range a**: This model is mostly used for extremely continuous values, and is defined as follows:

$$\gamma(h) = 1 - \exp(-\frac{3h^2}{a^2}).$$

Important Note:

1. Practical range a for Gaussian and exponential models is defined as the distance where variogram reaches 95% of its sill value.
2. In some geostatistical literature, Gaussian and exponential models are defined without the factor 3, and thus in those cases variable a would be $\frac{1}{3}$ of the practical range. [3].
3. Linear combination of acceptable variogram models is also an acceptable model.

3.11 Properties

Above mentioned statistical quantities have several properties which we will go over and show them in this section [1, 3, 2].

1. Variance of a random variable created as linear combination of other random variables, $V_1 \dots V_n$, is estimated as follows (see [3], p. 216):

$$Var(\sum_{i=1}^{n} w_i V_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \, Cov(V_i V_j),$$

   where $w_1 \dots w_n$ are the weights associated with $V_1 \dots V_n$ respectively.
2. It is trivial to see for variograms and cross-variograms that $\gamma(h) = \gamma(-h)$.
3. For a stationary random function, we have
   (a) $\gamma(h) = C(0) - C(h)$.
   Proof:

$$\begin{aligned}
2\gamma(h) &= Var\left[Z(u+h) - Z(u)\right] \\
&= E\left[Z(u+h) - Z(u)\right]^2 - \left[E\left(Z(u+h) - Z(u)\right)\right]^2 \\
&= E\left[Z(u+h) - Z(u)\right]^2 - \left[E\left(Z(u+h)\right) - E\left(Z(u)\right)\right]^2 \\
&= E\left[Z(u+h) - Z(u)\right]^2 - 0 \qquad \text{(by stationarity)} \\
&= E\left(Z(u+h)\right)^2 + E\left(Z(u)\right)^2 - 2E\left(Z(u+h)Z(u)\right) \\
&= 2E\left(Z(u)\right)^2 - 2E\left(Z(u+h)Z(u)\right) \Longleftrightarrow \\
\gamma(h) &= E\left(Z(u)\right)^2 - E\left(Z(u+h)Z(u)\right).
\end{aligned}$$

   We also have:

$$\begin{aligned}
C(0) &= E(Z(u))^2 - \left[E(Z(u))\right]^2, \text{and} \\
C(h) &= E(Z(u+h)Z(u)) - \left[E(Z(u))\right]^2 \Longleftrightarrow \\
C(0) - C(h) &= E(Z(u))^2 - E(Z(u+h)Z(u)) \;=\; \gamma(h).
\end{aligned}$$

   Note that derivation of second definition for semivariogram requires Z be a stationary random function. That is, the mean, or expected value of its values are invariant under transformation of variables passed to it and is always constant.

(b) $\rho(h) = \dfrac{C(h)}{C(0)}$.

Proof: Recall that

$$\begin{aligned} C(h) &= Cov(Z(u+h), Z(u)), \text{ and} \\ C(0) &= Cov(Z(u), Z(u)) = Var(Z(u)). \end{aligned}$$

Then, we have:

$$\begin{aligned} \rho(h) &= \rho(Z(u+h), Z(u)) \\ &= \frac{Cov(Z(u+h), Z(u))}{\delta(Z(u+h))\delta(Z(u))} \\ &= \frac{Cov(Z(u+h), Z(u))}{\delta(Z(u))^2} \qquad \text{(by stationarity)} \\ &= \frac{Cov(Z(u+h), Z(u))}{Var(Z(u))} \\ &= \frac{C(h)}{C(0)}. \end{aligned}$$

(c) $\rho(h) = 1 - \dfrac{\gamma(h)}{C(0)}$.

Proof: Using the property we just proved, we have

$$\begin{aligned} \rho(h) &= \frac{C(h)}{C(0)} \\ &= \frac{C(0) - \gamma(h)}{C(0)} \\ &= 1 - \frac{\gamma(h)}{C(0)}. \end{aligned}$$

4 Cokriging

4.1 Definition

Cokriging is defined as multivariate version of kriging. "A method for estimation that minimizes the variance of the estimation error by taking into consideration the spatial correlation between the variables of interest and the secondary variables."

In other words, a function U at location 0 is estimated as a *linear* combination of both the variable of interest and the secondary variables. That is, to estimate $\hat{u_0}$, the estimate of U at location 0, as mentioned in [3], is given by

$$\hat{u_0} = \sum_{i=1}^{n} a_i u_i + \sum_{j=1}^{m} b_j v_j.$$

The following holds for above error estimation equation:

- $u_1, u_2, \ldots, u_n$ are primary data at n nearby locations.
- $v_1, v_2, \ldots, v_n$ are secondary data at m nearby locations.
- $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_m$ are cokriging weights which are needed to be found and calculated.
- Auxiliary information does not need to be collected at the same data points as the variable of interest.
- Estimation error, R, is calculated as

$$R = \hat{U_0} - U_0 = w^t Z,$$

where $w^t = (a_1, \ldots, a_n, b_1, \ldots, b_m, -1)$, and $Z^t = (U_1, \ldots, U_i, V_1, \ldots, V_m, U_0)$.

4.2 Problem

- The goal of cokriging is to find weights, vector w^t above, such that:
  - Variance of the error is minimized.
  - Estimate for $\hat{U_0}$ be unbiased. That is, we try to have the mean residual or error equal to 0.
- Constraints that are imposed on the system's cokriging coefficients in order to ensure unbiasedness of the interpolant, distinguishes among various types of cokriging methods (see [2], p. 204, and [3], ch. 17):

  1. ***Simple Cokriging:*** No constraints are imposed on the weights. Means of primary and secondary data are required. Simple cokriging considers that local means are known and constant through the study area.
  2. ***Ordinary Cokriging:*** Imposes the following two constraints on coefficients: $\sum_{i=1}^{n} a_i = 1$ and $\sum_{j=1}^{m} b_j = 0$. This method limits the influence of the secondary variables greatly. As we will see, these conditions indicate that ordinary cokriging considers local means to be constant but unknown.
  3. ***Standardized Ordinary Cokriging***: is performed by creating new secondary variables so that they have the same mean as the primary variables. The constraint is that coefficients should add up to one: $\sum_{i=1}^{n} a_i + \sum_{j=1}^{m} b_j = 1$.

  In section 4.6 we show how the above conditions on coefficients of the system ensures unbiasedness of the interpolant for each type of cokriging.

4.3 Solution

In this section we illustrate how weight coefficient for a cokriging system is found. In order to do so, we illustrate solving the ordinary cokriging, and it should be clear how to set up and solve other varieties of cokriging.

From above equation, and definition of variance, we have

$$\begin{aligned} Var(R) &= w^t C_Z w \\ &= \sum_i^n \sum_j^n a_i a_j Cov(U_i U_j) + \sum_i^m \sum_j^m b_i b_j Cov(V_i V_j) \\ &+ 2\sum_i^n \sum_j^m Cov(U_i V_j) - 2\sum_i^n a_i Cov(U_i U_0) \\ &- 2\sum_j^m b_j Cov(V_j U_0) + Cov(U_0 U_0). \end{aligned}$$

One way of ensuring unbiasedness for our estimation $\hat{U_0}$ is to require $\sum_{i=1}^n a_i = 1$ and $\sum_{j=1}^m b_j = 0$ (we will show why in section 4.6). So now we have an optimization problem with two constraints. This is where we take advantage of Lagrange multiplies 2. Let our Lagrange multipliers be μ_1 and μ_2. Then, we are trying minimize $Var(R)$ subject to two mentioned constraints by solving for coefficients $a_1 \ldots a_n, b_1 \ldots b_m, \mu_1, \mu 2$, where

$$Var(R) = w^t C_Z w + 2\mu_1(\sum_{i=1}^n a_i - 1) + 2\mu_2(\sum_{j=1}^m b_j).$$

The next step is taking partial derivatives of the above equation with respect to all $n + m$ cokriging variables and the two Lagrange multipliers and setting them to zero. Then, we will get the following $n + m + 2$ equations to solve:

$$\begin{aligned} \sum_{i=1}^n a_i Cov(U_i U_j) + \sum_{i=1}^m b_i Cov(V_i U_j) + \mu_1 &= Cov(U_0 U_j) \quad (j = 1...n), \\ \sum_{i=1}^n a_i Cov(U_i V_j) + \sum_{i=1}^m b_i Cov(V_i V_j) + \mu_2 &= Cov(U_0 V_j) \quad (j = 1...m), \\ \sum_{i=1}^n a_i &= 1, \text{ and} \\ \sum_{i=1}^m b_i &= 0. \end{aligned}$$

Equivalently:

$$\begin{pmatrix} C_{u_1u_1} & \dots & C_{u_nu_1} & C_{v_1u_1} & \dots & C_{v_mu_1} & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & 1 & 0 \\ C_{u_1u_n} & \dots & C_{u_nu_n} & C_{v_1u_n} & \dots & C_{v_mu_n} & 1 & 0 \\ C_{u_1v_1} & \dots & C_{u_nv_1} & C_{v_1v_1} & \dots & C_{v_mv_1} & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & 0 & 1 \\ C_{u_1v_m} & \dots & C_{u_nv_m} & C_{v_1v_m} & \dots & C_{v_mv_m} & 0 & 1 \\ 1 & \dots & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ \dots \\ a_n \\ b_1 \\ \dots \\ b_m \\ \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} C_{u_0u_1} \\ \dots \\ C_{u_0u_n} \\ C_{u_0v_1} \\ \dots \\ C_{u_0v_m} \\ 1 \\ 0 \end{pmatrix}.$$

Once the above system of equations is solved, we have necessary coefficients $a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_m$ to estimate function U at location 0.

Note that the above mentioned method works only for point estimation.

4.4 Generalized Cokriging System

One can see that instead of having one set of secondary variables $V_1 \dots V_m$, we may use multiple sets of secondary variables. Each additional set of secondary variables $W_1 \dots W_k$ will introduce a new set of coefficients $c_1 \dots c_k$ and a new lagrange multiplier μ_w.

For the general case where we have s set of variables (as oppose to just 2 sets, one primary and one secondary), our linear system will be as follows:

$$\begin{pmatrix} C & E \\ E^T & 0 \end{pmatrix} \begin{pmatrix} T \\ \mu \end{pmatrix} = \begin{pmatrix} C0 \\ I0 \end{pmatrix}.$$

Where C is the covariance (or its estimate) matrix of all known variables' pair, and C_0 is the vector of pairwise covariances between the unknown variable U_0 and all other known variables.

μ is the vector of all lagrange multipliers $\mu_1 \dots \mu_s$. E is a vector of matrices $I_1 \dots I_s$. Each matrix $I_i, i \in \{1 \dots s\}$ is of size (number of points in i^{th} variable set) $\times s$. All elements in the i^{th} column of I_i are one and all other entries are zero.

T is the vector of all coefficients, and I_0 is a column vector of of size $s \times 1$ of all elements under C_0 on the right hand side of the equation. As we will see later, this vector is made of a 1 on top and all zeros for the rest of entries to insure unbiasedness of our estimator.

4.5 Kriging

Kriging is a special case of cokriging where we estimate value of a variable using values of only one set of primary variables of points around it. That is, to estimate the value of a random function U at location 0, u_0, using values of U at n other locations $(u_1, \dots u_n)$ we need to come up with coefficients $a_1, \dots a_n$ such that

- $\hat{u_0} = \sum_{i=1}^{n} a_i u_i$, and

- Variance of error is minimized, where error is $R = \hat{U_0} - U_0$.

Similarly we have *Simple* and *Ordinary Kriging*, depending on whether or not local means are known and/or constant or not 4.2. Simple kriging does not require any constraints on coefficients while ordinary kriging requires sum of coefficients to add up to one for insuring unbiasedness (see 4.6 for details).

Thus, for ordinary kriging, the solution mentioned in 4.3 simplifies as follows:

$$\begin{aligned} \sum_{i=1}^{n} a_i \, Cov(U_i U_j) + \mu_1 &= Cov(U_0 U_j) \quad (j = 1...n), \\ \sum_{i=1}^{n} a_i &= 1. \end{aligned}$$

Equivalently:

$$\begin{pmatrix} C_{u_1 u_1} & \dots & C_{u_n u_1} & 1 \\ \dots & \dots & \dots & 1 \\ C_{u_1 u_n} & \dots & C_{u_n u_n} & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ \dots \\ a_n \\ \mu_1 \end{pmatrix} = \begin{pmatrix} C_{u_0 u_1} \\ \dots \\ C_{u_0 u_n} \\ 1 \end{pmatrix}.$$

4.6 Unbiasedness Condition

For most interpolation methods used in geostatistics, in particular kriging and cokriging, it is often required that our estimate of a function for a given location be *unbiased.* Unbiasedness of the estimator means that expected value of error should be zero. Now we go over details of what this this condition means mathematically.

Suppose we have a random function $V(x)$, where x is a location. Assume that for points x_1 to x_n we know value of function V, and we would like to estimate V at an unknown location x_0 as a linear combination of n known function values so that our estimate be unbiased (this is the case in kriging).

Let $\hat{V}$ be our estimate function, R the error associated with our estimate, and $v_1 \dots v_n$ be weights given to each known function value in estimation. Then we have

$$\begin{aligned} \hat{V}(x_0) &= \sum_{i=1}^{n} v_i V(x_i), \text{ and} \\ R(x_0) &= \hat{V}(x_0) - V(x_0) = \sum_{i=1}^{n} v_i V(x_i) - V(x_0). \end{aligned}$$

Unbiasedness condition states

$$
\begin{aligned}
E\left(R(x_0)\right) &= 0 \iff \\
E\left(\sum_{i=1}^{n} v_i V(x_i) - V(x_0)\right) &= 0 \iff \\
E(V)\sum_{i=1}^{n} v_i - E(V) &= 0 \iff \\
E(V)\left[\sum_{i=1}^{n} v_i - 1\right] &= 0.
\end{aligned}
$$

This indicates that we need to have $\sum_{i=1}^{n} v_i = 1$ to assure unbiasedness of our estimate, and this is often one of the constraints in optimization problems that we end up solving for our interpolation methods ([3], p. 285).

It is easy to similarly derive the necessary conditions for ensuring unbiasedness in ordinary cokriging. In this case, we estimate a random function value as a linear combination of values of more than one random function at each point. Let our secondary random function be W. Also, assume we know values of W at points $x_1 \ldots x_m$. Then,

$$
\begin{aligned}
\hat{V}(x_0) &= \sum_{i=1}^{n} v_i V(x_i) + \sum_{i=1}^{m} w_i W(x_i), \text{ and} \\
R(x_0) &= \hat{V}(x_0) - V(x_0) = \sum_{i=1}^{n} v_i V(x_i) + \sum_{i=1}^{m} w_i W(x_i) - V(x_0).
\end{aligned}
$$

Unbiasedness condition states

$$
\begin{aligned}
E\left(R(x_0)\right) &= 0 \iff \\
E\left(\sum_{i=1}^{n} v_i V(x_i) + \sum_{i=1}^{m} w_i W(x_i) - V(x_0)\right) &= 0 \iff \\
E(V)\sum_{i=1}^{n} v_i + E(W)\sum_{i=1}^{m} w_i - E(V) &= 0 \iff \\
E(V)\left[\sum_{i=1}^{n} v_i - 1\right] + E(W)\sum_{i=1}^{m} w_i &= 0.
\end{aligned}
$$

By requiring $\sum_{i=1}^{n} v_i = 1$ and $\sum_{i=1}^{m} w_i = 0$ we ensure that the above equation and thus our unbiasedness condition holds. Similarly, if we use more additional functions in ordinary cokriging, we need to require sum of coefficients of values in our linear combination obtained from each particular additional function to be equal to zero.

Standardized cokriging assumes that all random functions used in our estimation process have the same mean. That is in above equation we have $E(V) = E(W)$. This condition results in reducing our number of constraints from s to 1, where s is the number of random functions used in our linear interpolation.

$$\begin{aligned} E(V)\left[\sum_{i=1}^{n} v_i - 1\right] + E(W)\sum_{i=1}^{m} w_i &= 0 \Longleftrightarrow \\ E(V)\left[\sum_{i=1}^{n} v_i - 1\right] + E(V)\sum_{i=1}^{m} w_i &= 0 \Longleftrightarrow \\ E(V)\left[\sum_{i=1}^{n} v_i + \sum_{i=1}^{m} w_i - 1\right] &= 0. \end{aligned}$$

Thus, in standardized cokriging we require

$$\sum_{i=1}^{n} v_i + \sum_{i=1}^{m} w_i = 1.$$

Similarly, if more random functions are involved, we require sum of all coefficients in our linear combination be one.

4.7 Positive-Definiteness Condition

A requirement for cokriging system to have a solution is that its covariance/variogram matrix, lets call it K, needs to be positive definite. This is due to two reasons:

1) To insure that the variance of the error of our interpolant is positive ([3], p. 372).

2) To insure invertibility of coefficients in linear system involved [7].

The first part is easy to show. As seen above, the error involved in cokriging result can be written as follows:

$$\begin{aligned} R &= \hat{U_0} - U_0 = w^t Z \\ Var(R) &= w^t C_Z w. \end{aligned}$$

Requiring $Var(R)$ to be positive is the same as requiring $C = C_z$ to be positive-definite: $w^t C_z w > 0$.

As we saw, our cokriging linear system is as follows:

$$\begin{pmatrix} C & E \\ E^T & 0 \end{pmatrix} \begin{pmatrix} T \\ \mu \end{pmatrix} = \begin{pmatrix} C0 \\ I0 \end{pmatrix}.$$

We just showed why in order for variance of error for our interpolation method to be positive we need matrix C to be positive definite [3]. Next, we will show that how this positive-definiteness condition insures our cokriging system to have a solution.

We show the second part by proving its contrapositive statement: If the system does not have a solution, then C not positive definite [7].

Let

$$G = \begin{pmatrix} C & E \\ E^T & 0 \end{pmatrix}.$$

If G is not invertible, it means that there exist non-zero U and V such that

$$\begin{pmatrix} C & E \\ E^T & 0 \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Thus,

1. $CU + EV = 0 \implies U^T CU + U^T EV = 0$
2. $E^T U = 0 \implies U^T E = 0$

(1) and (2) $\implies U^T CU = 0 \implies C$ is not positive definite.

Notice that this positive definiteness condition on C is different from positive definiteness condition for insuring optimality of a minimization/optimization problem. Here we are just insuring that the coefficient matrix is invertible while in an optimization problem we require Hessian of the system to be positive definite as described in section 2.

5 Algorithmic Approach

Estimating an unknown variable via kriging/cokriging involves the following steps:

1. Setting up the linear system as mentioned in its most general form in 4.4 satisfying conditions mentioned in 4.6 and 4.7.
2. Solving the linear system for coefficients.
3. Evaluating the estimation for the unknown variable.

The main task in first step is coming up with matrices C and C_0 as mentioned in 4.4 since other components of the system are well defined in 4.4. Elements of C are pairwise covariances between random variables for which we have only one value. That is, we know values of a primary variable at n points $(u_1, \ldots u_n)$, and we are treating each of these values as an instant of a random variable U_1. Thus, C_{ij} means we are looking for covariance between two random variables U_i and U_j that generated values u_i and u_j at i^{th} and j^{th} points respectively. But how do we go about calculating covariances between two random variables for each of which we only have one value?

Here is where variograms come into picture when performing cokriging/kriging. Variograms between pairwise variables are function of distances between the two particular samples and values of random variables at those samples.

Notice that for a **stationary** random variable, once we know the variogram between variables at a pair of points separated by distance h in location, it is easy to calculate their pairwise covariance as well (see 3.11, 3a).

Also, one can transform anisotropic data to one which is isotropic (stationary) [2, 3]. Thus, once we know how to estimate covariances from variograms as a function of distance between locations where measurements were taken for stationary data, we can do the same for anisotropic data as well.

Usually variograms are modelled as a function of distance between points , and so are the covariances. After fitting various possible models to calculated variograms, the best one which gives the least error is picked and used to model the pairwise variograms. Having a general variogram function $\gamma(h)$, we can obtain a covariance function $C(h)$ using equation in 3.11, 3a).

At this point, we have a modelled variogram and covariance, $\gamma(h)$ and $C(h)$ as a function of distance between two samples of random variables. Then, matrices C and $C0$ in 4.4 are generated as follows: for every element c_{ij} in C or $C0$, calculate distance h_{ij} between point labelled i and the one labelled j. Then simply calculate $c_{ij} = C(h_{ij})$.

In addition to just calculating matrix C, we need to make sure that it is positive definite (see 4.7). Instead of checking for this condition every time a model is picked, we limit our selection of variogram models to only those functions which will lead to a positive definite matrix C (see 3.10).

Also, it is important that the model we choose for our variogram be bounded so that we be able to calculate $C(h)$ given $\gamma(h)$ for a given distance h using equation in 3.11, 3a.

Note that matrix C in 4.4, whose elements are represented as c_{ij}, is different from modelled covariance function $C(h)$.

Now that we are done with first step, and have our linear system set up, we need to solve it to obtain coefficients. There are variety of well-known methods for solving a linear system of equations [8], and once the solution to the system in 4.4 is known, evaluating the unknown variable is just a matter of substituting values and calculating $\hat{u_0}$ as mentioned in 4.1.

6 Computational Challenges

As we discussed in this paper, estimating an unknown value using cokriging/kriging requires us to:

1. Set up the linear system which needs to be solved (see 4.4). This includes the problem of fitting values of random functions involved to best possible variogram model that will result in a positive definite covariance matrix for the linear system we need to solve.

2. Solve the linear system.

3. Evaluate the unknown value as the linear combination of known values. (see 4.1)

While the above steps involve a lot of computations in themselves for estimation of any one unknown value, they are well known and relatively easily solvable math problems (fitting data,

solving a system, evaluating a function). What really makes cokriging/kriging computationally expensive and inefficient is that:

1. After doing the above steps to estimate an unknown value at location i, we cannot use the results (coefficients needed for evaluation step obtained from solving the linear system) to estimate another unknown value at location let's say j. That is, we need to repeat all the above mentioned three steps (see 5 for details) every time we want to estimate a new value.

2. Size of matrices involved in our linear system is usually very large for applications that kriging/cokriging is used for. This makes variogram modelling and solving the linear systems involved very slow.

7 Past Work

The following implementations of cokriging are mostly used among users of this method:.

1. **Fortran Implementation**: This is the code written by Deutsch and Journal for GSLIB library [1], and has support for both kriging and cokriging. See [1] and http://ekofisk.stanford.edu/SCRFweb/GSLIB/gslibhlp.html, for more information.

2. **Matlab Implementation**: This code has support for cokriging [6].

8 Conclusion

In this paper we showed what is meant by cokriging in general and kriging in particular. We went over terminologies and mathematical properties and methods involved in these interpolation methods. We showed how this interpolation technique is formalized mathematically and solved algorithmically. We also went over computational issues involved and introduced references to available implementations of these two methods.

References

[1] Clayton V. Deutsch and Andre' G.Journal. *GSLIB Geostatistical Software Library and User's Guide*, volume second. Oxford University Press, 1998.

[2] Pierre Goovaerts. *Geostatistics for Natural Resources Evaluation.* Oxford University Press, New York, Oxford, 1997.

[3] Edward H. Isaaks and R. Mohan Srivastava. *An Introduction to Applied Geostatistics.* Oxford University Press, New York, Oxford, 1989.

[4] Shine J.A and Krause. Exploration and estimation of North American climatological data. In 32^{nd} *Symposium on the Interface: Computing Science and Statistics*, April 5-8 2000.

[5] A.G. Journel and Ch. J. Huijbregts. *Mining Geostatistics*. Academic Press Inc, New York, 1978.

[6] Denis Marcotte. Cokriging with Matlab. *Computers and Geosciences*, 17(9):1265–1280, 1991.

[7] Donald E. Myers. Pseudo-cross variograms, positive-definiteness, and cokriging. *Mathematical Geology*, 23(6):805–816, 1991.

[8] Stephen G. Nash and Ariela Sofer. *Linear and Nonlinear Programming*. McGraw-Hill Companies, 1996.