A quick derivation of geostatistical Kriging

William S. Harlan

July 2013

Introduction

A casual encounter with geostatistics can be baffling because of some non-standard terminology. As it turns out, "Kriging," the core numerical method of geostatistics, can be derived on a napkin, if you are already familiar with some standard least-squares methods.

If you want to appreciate geostatistical applications, then try the popular book "An introduction to Applied Geostatistics," by Isaaks and Srivastava [3]. If you want a clean mathematical notation, then try "Multivariate Geostatistics: An introduction with Applications," by Wackernagel [5];

Kriging is just the least-squares solution of a purely under-determined linear inverse problem. Once you see this equivalence, you can see some simple ways to improve distance-weighted interpolation methods.

Posing the problem

Geostatistics poses a useful problem that you are unlikely to have encountered elsewhere in least-squares. The solution looks very familiar, but it has unique elements that make it very useful.

Assume that you have a continuous function $v(\mathbf{x})$ of a spatial vector $\mathbf{x}$. Usually $\mathbf{x}$ has two dimensions, but nothing about the derivation limits us to two dimensions.

The actual function $v(\mathbf{x})$ is unknown and needs to be reconstructed from a collection of samples $\{v_i\}$ at n arbitrarily chosen locations.

$$v_i \equiv v(\mathbf{x}_i) \text{ for } i{=}1, n. \tag{1}$$

We will interpolate a value v_0 at an unsampled location $\mathbf{x}_0$ as a linearly

weighted sum of the n known values $\{v_i\}$, at sampled locations $\{\mathbf{x}_i\}$:

$$\begin{aligned} v(\mathbf{x}_0) &= \sum_{i=1}^{n} w_i \cdot v(\mathbf{x}_i), \\ v_0 &= \sum_{i=1}^{n} w_i v_i, \text{ or} \\ v_0 &= \mathbf{w}^\intercal \cdot \mathbf{v}, \\ \text{where } \mathbf{w}^\intercal &\equiv [w_1, w_2, ..., w_n], \text{ and } \mathbf{v} \equiv [v_1, v_2, ..., v_n]^\intercal. \end{aligned} \tag{2}$$

In the last version we switch to vector notation.

Our problem: how do we optimally estimate weights $\mathbf{w}$ given what we know about the distribution of $\mathbf{x}$?

Assumptions

We will make a few more assumptions appropriate to a linear least-squares solution.

We will treat all our values $\{v_i\}$ as random variables. Without loss of generality (changing variables if necessary), we will assume these variables have zero expected mean:

$$E(v_i) \equiv 0, \text{ for } i{=}0, n. \tag{3}$$

Assume a constant stationary variance σ_v^2 for all samples

$$\sigma_v^2 \equiv E(v_i v_i), \text{ for } i{=}0, n. \tag{4}$$

Notice that we have included the unmeasured value v_0 in our assumptions.

We will find it convenient to construct a covariance $\underset{\sim}{\mathbf{C}}$ matrix for our measured values:

$$\begin{aligned} C_{ij} &\equiv E(v_i v_j), \text{ for } i{=}1, n \text{ and } j{=}1, n, \text{ or} \\ \underset{\sim}{\mathbf{C}} &\equiv E(\mathbf{v}\mathbf{v}^\intercal). \end{aligned} \tag{5}$$

Note that the covariance is symmetric: $C_{ij} = C_{ji}$.

Define a covariance vector $\mathbf{c}$ for our unknown value relative to the measured values:

$$\begin{aligned} c_i &\equiv E(v_0 v_i), \text{ for } i{=}1, n, \text{ or} \\ \mathbf{c} &\equiv E(v_0 \mathbf{v}). \end{aligned} \tag{6}$$

Much of geostatistics concentrates on the estimation of these covariances. For now we assume they are known.

Notice that none of our assumptions require that our interpolated value share the same physical units as the values we are interpolating from. Co-kriging differs from Simple Kriging largely by combining values with mixed units. (See Wackernagel for a demonstration of their equivalence.)

Least-squares solution

Next we estimate our weights as a least-squares optimization, and take advantage of our assumptions.

Define an estimation error σ_E^2 as the expected error between the true value and our interpolated value.

$$\begin{aligned}
\sigma_E^2 &\equiv E[(v_0 - \sum_{i=1}^{n} w_i v_i)^2] && (7)\\
&= E(v_0 v_0) - 2\sum_{i=1}^{n} E(v_i v_0) w_i + \sum_{i=1}^{n}\sum_{j=1}^{n} E(v_i v_j) w_i w_j \\
&= \sigma_v^2 - 2\sum_{i=1}^{n} c_i w_i + \sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij} w_i w_j. && (8)
\end{aligned}$$

Naturally we want this error to be small. Since this expression is a quadratic of the weights $\mathbf{w}$, we can minimize σ_E^2 by differentiating with respect to the weights and setting the value to zero.

$$\begin{aligned}
&\frac{\partial \sigma_E^2}{\partial w_i} = -2c_i + 2\sum_{j=1}^{n} C_{ij} w_j = 0 && (9)\\
\Rightarrow\ &\sum_{j=1}^{n} C_{ij} w_j = c_i && (10)\\
\Rightarrow\ &\underset{\sim}{\mathbf{C}} \cdot \mathbf{w} = \mathbf{c} \\
\Rightarrow\ &\mathbf{w} = \underset{\sim}{\mathbf{C}}^{-1} \cdot \mathbf{c}. && (11)
\end{aligned}$$

This result is entirely equivalent to Simple Kriging. The optimum weights are intuitive when you see them in this form.

Naive weights $\mathbf{w}$ would simply use the covariance vector $\mathbf{c}$ between values at the known and the unknown locations. Because the known locations are also correlated with each other, we must remove that effect by multiplying by their inverse covariance $\underset{\sim}{\mathbf{C}}^{-1}$.

Once we have our estimated weights $\mathbf{w}$, we can explicitly quantify the error in our estimate. This will allow us to put Gaussian error bars on each interpolated value. Substitute our solution (10) into our estimation error (8) for

$$\begin{aligned}
\sigma_E^2 &= \sigma_v^2 - 2\sum_{i=1}^{n} c_i w_i + \sum_{i=1}^{n} c_i w_i \\
&= \sigma_v^2 - \sum_{i=1}^{n} c_i w_i \\
&= \sigma_v^2 - \mathbf{c}^\intercal \cdot \mathbf{w}. && (12)
\end{aligned}$$

Notice that this expected error is independent of correlations between measured samples. If it were not, then the estimate could be improved.

Choosing covariances

Geostatistics usually concentrates on the estimation of covariances (or variograms, discussed later).

Here is the most geostatistical assumption of all. The covariance between values depends only on the vector distance between their sampled positions:

$$C_{ij} \equiv E[v_i v_j] \equiv E[v(\mathbf{x}_i)v(\mathbf{x}_j)] \equiv C(\mathbf{x}_i - \mathbf{x}_j) \quad \text{for } i{=}0,n \text{ and } j{=}0,n. \tag{13}$$

In the absence of other information, it would be appealing to choose a scale-invariant version of a covariance. Scale-invariance allows covariances to be independent of the units chosen for spatial distance. Covariance would similar at any scale, like a fractal.

Unfortunately, this assumption leads to a degenerate solution. Scale-invariance requires covariances to go to infinity at a distance of zero, which should have a finite variance σ_v^2. As we approach infinity, the covariance matrix $\underset{\sim}{\mathbf{C}}$ becomes diagonal like an identity matrix, and so the inverse covariance has no effect on weights.

We see this degeneracy with a popular interpolation known as Shepard's "Inverse Distance Weighting":

$$\begin{aligned} v(\mathbf{x}_0) &= \sum_{i=1}^{n} w_i \cdot v(\mathbf{x}_i) \text{ where} \\ w_i &= \|\mathbf{x}_i - \mathbf{x}_0\|^{-2} / \sum_{j=1}^{n} \|\mathbf{x}_j - \mathbf{x}_0\|^{-2}. \end{aligned} \tag{14}$$

This weight decreases with distance squared, like gravity. The denominator is just a normalization over all samples.

If we are assuming a specific correlation between our measured and interpolated values, then we ought to be able to assume the same correlation between measured values.

Unfortunately, this covariance approaches infinity at a zero distance:

$$\begin{aligned} \underset{\sim}{\mathbf{C}}(\mathbf{x}_i - \mathbf{x}_j) &\propto \|\mathbf{x}_i - \mathbf{x}_j\|^{-2} \\ \Rightarrow \quad \sigma_v^2 &\approx \infty \qquad (15) \\ \Rightarrow \quad \underset{\sim}{\mathbf{C}} &\propto \underset{\sim}{\mathbf{I}} \\ \Rightarrow \quad \mathbf{w} &\approx \mathbf{c}. \qquad (16) \end{aligned}$$

Sure enough, this degenerate solution tells us to ignore covariances between our known values.

As it turns out, the degeneracy is easily remedied if we sacrifice scale invariance below the density of our samples. Choose a small distance Δx below which the correlation does not increase any further.

$$\underset{\sim}{\mathbf{C}}(\mathbf{x}_i - \mathbf{x}_j) \quad \propto \quad \min(\Delta x^{-2}, \|\mathbf{x}_i - \mathbf{x}_j\|^{-2}). \tag{17}$$

Covariances at any smaller distance will all equal the marginal σ_v^2.

Let's see how this assumption works with a simple example.

A simple example

Let's say we have two measured values to interpolate along a one-dimensional dimension x.

One value v_2 is twice as far as v_1 from the interpolated point v_0. Specifically $\|x_1 - x_0\| = 1$ and $\|x_2 - x_0\| = 2$.

If we are using Shepard's interpolation, then we already know enough to estimate the weights as $\mathbf{w}^\intercal = [0.8, 0.2]$.

However, it should matter very much if the point we are interpolating is between the two measured points, or on the same side of both.

For a non-degenerate covariance, we will assume $\Delta x = 0.5$ in our covariance (17).

On opposite sides of the interpolated point, the two measured points are weakly correlated with $\|x_2 - x_1\| = 3$. We get adjusted weights of $\mathbf{w}^\intercal = [0.82, 0.18]$, which is very close to the Shepard's interpolation weights.

On the same side of the interpolated point, the two measured points are strongly correlated with $\|x_2 - x_1\| = 1$. The value v_2 is actually better correlated with v_1 than it is with the interpolated v_0. We expect the nearer value v_1 to override and diminish the contribution of v_2. And sure enough, the adjusted weights are actually $\mathbf{w}^\intercal = [1, 0]$, which ignores the second point entirely.

Variograms

Most geostatistics literature prefers "variograms" to covariances. This is one of the biggest and most unnecessary obstacles to newcomers. The two quantities have a simple equivalence:

$$\gamma_v(\mathbf{x}_i, \mathbf{x}_j) \quad \equiv \quad \frac{1}{2} E\{[v(\mathbf{x}_i) - v(\mathbf{x}_j)]^2\} = \sigma_v^2 - C_{ij}. \tag{18}$$

Variograms have the minor convenience that values no longer need a zero mean.

The geostatistical assumption (13) can be rewritten

$$\gamma_v(\mathbf{x}_i - \mathbf{x}_j) = \sigma_v^2 - C(\mathbf{x}_i - \mathbf{x}_j) \tag{19}$$

Usually variograms are defined in terms of an offset $\mathbf{h}$:

$$\begin{aligned} \mathbf{h}_{ij} &= \mathbf{x}_i - \mathbf{x}_j; \\ \gamma_v(\mathbf{h}_{ij}) &= \sigma_v^2 - C(\mathbf{h}_{ij}). \end{aligned} \tag{20}$$

Considering only offsets and ignoring sampling, we can write the equivalence this way:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}). \tag{21}$$

If we assume values are uncorrelated at infinite distance, we find

$$\begin{aligned} C(\boldsymbol{\infty}) &= 0 \\ \Rightarrow \quad \gamma(\boldsymbol{\infty}) &= C(\mathbf{0}) \\ \Rightarrow \quad C(\mathbf{h}) &= \gamma(\boldsymbol{\infty}) - \gamma(\mathbf{h}). \end{aligned} \tag{22}$$

The most popular of standard analytic variograms is an exponential:

$$\gamma_{\exp}(\mathbf{h}) \propto 1 - \exp(-3\|\mathbf{h}\|/a), \quad \text{or} \tag{23}$$

$$C_{\exp}(\mathbf{h}) \propto \exp(-3\|\mathbf{h}\|/a). \tag{24}$$

The constant a is the distance over which about 95% of the correlation occurs. Unlike our distance weighting, this function is not scale invariant over any range of offsets.

A user must prepare scatterplots and histograms to scan possible values for the constant a, then choose a best fitting value before kriging can begin. For this sort of analysis, variograms have proven more popular than covariances.

Kriging as a posterior estimate

It is also possible to recognize Kriging as a special case of a least-squares inverse problem, without postulating the specific weighted solution (2). I thank Dave Hale [1] for explaining this equivalence to me. Kriging is a purely underdetermined inverse problem, which makes it a bit unusual and less familiar.

We assume that we have a model vector $\mathbf{m}$ and a data vector $\mathbf{d}$. The data are assumed to be a linear transform (matrix multiplication) $\underset{\sim}{\mathbf{F}}$ of the model, plus a vector $\mathbf{n}$ of noise:

$$\mathbf{d} = \underset{\sim}{\mathbf{F}}\mathbf{m} + \mathbf{n}. \tag{25}$$

We assume that our model and noise are selected from stationary correlated Gaussian random processes. The expected value of a model given the data (a posterior estimate) should minimize the following objective function:

$$S(\mathbf{m}) = (\mathbf{d} - \underset{\sim}{\mathbf{F}}\mathbf{m})^\intercal \underset{\sim}{\mathbf{C}}_n^{-1}(\mathbf{d} - \underset{\sim}{\mathbf{F}}\mathbf{m}) + \mathbf{m}^\intercal \underset{\sim}{\mathbf{C}}_m^{-1}\mathbf{m}. \tag{26}$$

The covariance $\underset{\sim}{\mathbf{C}}_m$ is the Gaussian covariance of the model $\mathbf{m}$, and covariance $\underset{\sim}{\mathbf{C}}_n$ is the Gaussian covariance of the noise. We also assume that the noise is uncorrelated to the model. The definition (25) determines the data covariance $\underset{\sim}{\mathbf{C}}_d$:

$$\underset{\sim}{\mathbf{C}}_d = \underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal + \underset{\sim}{\mathbf{C}}_n. \tag{27}$$

Since this objective function is a quadratic in $\mathbf{m}$, we can find an optimum $\hat{\mathbf{m}}$ by setting the derivative to zero:

$$\begin{aligned}
\frac{\partial S}{\partial \mathbf{m}^\intercal}(\hat{\mathbf{m}}) &= -\underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_n^{-1}(\mathbf{d} - \underset{\sim}{\mathbf{F}}\hat{\mathbf{m}}) + \underset{\sim}{\mathbf{C}}_m^{-1}\hat{\mathbf{m}} = 0; && (28)\\
\Rightarrow \quad \hat{\mathbf{m}} &= (\underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_n^{-1}\underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{C}}_m^{-1})^{-1}\underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_n^{-1}\mathbf{d}; && (29)\\
\Rightarrow \quad \hat{\mathbf{m}} &= \underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal(\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal + \underset{\sim}{\mathbf{C}}_n)^{-1}\mathbf{d} = \underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_d^{-1}\mathbf{d}. && (30)
\end{aligned}$$

The first solution (29) is directly solved from the derivative (28). The second solution (30) was shown by Tarantola [4] (Appendix 6.30) to be entirely equivalent. If (29) is equivalent to (30), then

$$\begin{aligned}
(\underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_n^{-1}\underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{C}}_m^{-1})^{-1}\underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_n^{-1} &= \underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal(\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal + \underset{\sim}{\mathbf{C}}_n)^{-1}; && (31)\\
\Rightarrow \quad \underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_n^{-1}(\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal + \underset{\sim}{\mathbf{C}}_n) &= (\underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_n^{-1}\underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{C}}_m^{-1})\underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal && (32)\\
&= \underset{\sim}{\mathbf{F}}^\intercal \underset{\sim}{\mathbf{C}}_n^{-1}\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{C}}_m\underset{\sim}{\mathbf{F}}^\intercal + \underset{\sim}{\mathbf{F}}^\intercal. && (33)
\end{aligned}$$

Both sides of (32) are trivially equal to (33).

This alternate solution (30) is less often used because it requires a matrix inversion in the data space rather than in the model space. Most typically, an overdetermined problem has a greater number of data values than model values. In the case of Kriging, however, the situation is reversed.

Others such as Hansen et al [2] have observed that this alternate solution (30) is equivalent to Kriging. We will elaborate below.

In the case of geostatistical Kriging, we assume that our model $\mathbf{m}$ is a large dense grid of regularly distributed values. The forward operator $\underset{\sim}{\mathbf{F}}$ merely subsamples that grid for the values $\mathbf{d}$. The dimensionality of $\mathbf{d}$ is possibly orders of magnitude smaller than that of the model $\mathbf{m}$. The transform $\underset{\sim}{\mathbf{F}}$ is a rectangular matrix that has many more columns than rows. Each row contains mostly zeros and a single column with a value of 1. Most columns do not have a value of 1 on any row.

Applying the transposed operator $\underset{\sim}{\mathbf{F}}^\intercal$ to subsampled data $\mathbf{d}$ merely repopulates the sampled values of $\mathbf{m}$ and leaves all other values set to 0. This is also the right inverse of the original transform, so that $\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{F}}^\intercal = \underset{\sim}{\mathbf{I}}$.

If we have sampled the model directly, there is no need to assume any noise at all. If we allow $\underset{\sim}{\mathbf{C}}_n = \epsilon \underset{\sim}{\mathbf{I}}; \epsilon \to 0$, then the usual estimate (29) becomes degenerate, and $\hat{\mathbf{m}} \to \underset{\sim}{\mathbf{F}}^\intercal \mathbf{d}$. Rather than interpolating any unsampled values, this estimate sets them all to zero.

The alternate form (30) lets us set the noise variance directly to 0. We can then find the following covariances of our sampled data $\underset{\sim}{\mathbf{C}}_d$:

$$\begin{aligned} \text{Assume } \underset{\sim}{\mathbf{C}}_n &\equiv 0; && (34) \\ \Rightarrow \underset{\sim}{\mathbf{C}}_d &= \underset{\sim}{\mathbf{F}} \underset{\sim}{\mathbf{C}}_m \underset{\sim}{\mathbf{F}}^\intercal && (35) \end{aligned}$$

and then rewrite our alternate estimate (30) as

$$\begin{aligned} \hat{\mathbf{m}} &= (\underset{\sim}{\mathbf{C}}_m \underset{\sim}{\mathbf{F}}^\intercal)(\underset{\sim}{\mathbf{F}} \underset{\sim}{\mathbf{C}}_m \underset{\sim}{\mathbf{F}}^\intercal)^{-1} \mathbf{d} && (36) \\ &= (\underset{\sim}{\mathbf{F}} \underset{\sim}{\mathbf{C}}_m)^\intercal \underset{\sim}{\mathbf{C}}_d^{-1} \mathbf{d} && (37) \\ &= \underset{\sim}{\mathbf{W}}^\intercal \mathbf{d}; && (38) \\ \text{where } \underset{\sim}{\mathbf{W}} &\equiv \underset{\sim}{\mathbf{C}}_d^{-1} (\underset{\sim}{\mathbf{F}} \underset{\sim}{\mathbf{C}}_m) && (39) \end{aligned}$$

These weights (39) for all samples are entirely equivalent to those for a single sample (11). Each measured sample of $\mathbf{d}$ is weighted by its covariance to an interpolated point, then these weights are adjusted by the inverse covariance of the data samples. We now have the advantage of directly relating data covariances to that of the model. This expression is also much easier to generalize to data that do more than merely subsample the model.

References

[1] Dave Hale. Personal communication, 2013.

[2] Thomas Mejer Hansen, Andre Journel, Albert Tarantola, and Klaus Mosegaard. Linear inverse gaussian theory and geostatistics. *Geophysics*, pages 101–111, 2006.

[3] Edward H. Isaaks and Mohan R. Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, USA, January 1990.

[4] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2004.

[5] H. Wackernagel. *Multivariate geostatistics: an introduction with applications*. Springer-Verlag, 2nd edition, 2003.