Kriging in Perspective

CRS

9 October 2015

Abstract

Linear smoothing of spatial or spatio-temporal data rejoices in the name of "kriging", after the mining engineer D. G. Krige, who first realized that the problem could be tackled by the method of least squares. In these notes, I will try to explain what kriging is, how it works, and why it works well (when it does). My aim here is to combine a somewhat abstract mathematical approach with a minimum of modeling assumptions. I therefore begin with a general treatment of optimal linear prediction with dependent variables, after which the specialization to spatial or spatio-temporal prediction is basically trivial. The symmetry assumptions about covariance functions, beloved of geostatisticians, are treated as ways of trading increased bias for reduced variance of estimation.

1 Least-Squares Optimal Linear Prediction

Suppose we have a scalar variable Y, and we wish to predict it from a vector of covariates Z. The covariates may be observations of the same physical quantity at other times or places, or variables of a different sort altogether.

We make three debatable assumptions.

1. We want a point prediction of Y, so our prediction $m(Z)$ will be real-valued.
2. We will measure the quality of the point prediction by expected squared error, $\mathbb{E}\left[(Y - m(Z))^2\right]$.
3. We will limit ourselves to affine functions of Z, so $m(Z) = a + b \cdot Z$ for some scalar a and vector of coefficients b.

I shall return to these assumptions later.

We seek the optimal vector of coefficients β.

$$(\alpha, \beta) = \underset{a,b}{\operatorname{argmin}} \mathbb{E}\left[(Y - (a + b \cdot Z))^2\right] \tag{1}$$

As usual, we find this by doing some algebra on the expected squared error, and then some calculus.

$$\begin{aligned}
\mathbb{E}\left[(Y - (b \cdot Z))^2\right] &= \mathbb{E}\left[Y^2\right] + a^2 + \mathbb{E}\left[(b \cdot Z)^2\right] && (2)\\
&\quad -2\mathbb{E}\left[Y(b \cdot Z)\right] - 2\mathbb{E}\left[Ya\right] + 2\mathbb{E}\left[ab \cdot Z\right] \\
&= \mathbb{E}\left[Y^2\right] + a^2 + b \cdot \mathbb{E}\left[Z \otimes Z\right] b && (3)\\
&\quad -2a\mathbb{E}\left[Y\right] - 2b \cdot \mathbb{E}\left[YZ\right] + 2ab \cdot \mathbb{E}\left[Z\right]
\end{aligned}$$

Taking the gradients with respect to a and b, and setting it to zero at the optimum,

$$\begin{aligned}
-2\mathbb{E}\left[Y\right] + 2\beta\mathbb{E}\left[Z\right] + 2\alpha &= 0 && (4)\\
-2\mathbb{E}\left[YZ\right] + 2\mathbb{E}\left[Z \otimes Z\right]\beta + 2\alpha\mathbb{E}\left[Z\right] &= 0 && (5)\\
\alpha &= \mathbb{E}\left[Y\right] - \beta \cdot \mathbb{E}\left[Z\right] && (6)\\
\mathbb{E}\left[YZ\right] - \alpha\mathbb{E}\left[Z\right] &= \mathbb{E}\left[Z \otimes Z\right]\beta && (7)\\
\mathbb{E}\left[YZ\right] - (\mathbb{E}\left[Y\right] - \beta \cdot \mathbb{E}\left[Z\right])\mathbb{E}\left[Z\right] &= \mathbb{E}\left[Z \otimes Z\right]\beta && (8)\\
\mathrm{Cov}\left[Y, Z\right] &= \mathrm{Var}\left[Z\right]\beta && (9)\\
\beta &= (\mathrm{Var}\left[Z\right])^{-1}\mathrm{Cov}\left[Y, Z\right] && (10)
\end{aligned}$$

Let me repeat the key results from that.

$$\begin{aligned}
\beta &= (\mathrm{Var}\left[Z\right])^{-1}\mathrm{Cov}\left[Y, Z\right] && (11)\\
\alpha &= \mathbb{E}\left[Y\right] - \beta \cdot \mathbb{E}\left[Z\right] && (12)
\end{aligned}$$

The coefficients β depend on the covariance between Y and the different components of Z, "discounted" by the covariances between those components of Z. The intercept α is a nuisance to make sure the expectation value comes out right.

How bad is this optimal linear model? Let's first ask for the bias, i.e., the expected prediction error:

$$\begin{aligned}
\mathbb{E}\left[Y - (\alpha + \beta \cdot Z)\right] &= \mathbb{E}\left[Y - \mathbb{E}\left[Y\right] + \beta \cdot \mathbb{E}\left[Z\right] - \beta \cdot Z\right] && (13)\\
&= \mathbb{E}\left[Y - \mathbb{E}\left[Y\right]\right] - \beta \cdot \mathbb{E}\left[Z - \mathbb{E}\left[Z\right]\right] && (14)\\
&= 0 && (15)
\end{aligned}$$

It does not, of course, follow that $\mathbb{E}\left[Y|Z\right] = \alpha + \beta \cdot Z$; just that the deviations from this linear model *average out* to zero, as Z varies randomly.

With this in hand, the expected squared error is just the variance of the error:

$$\begin{aligned}
\mathrm{Var}\left[Y - (\alpha + \beta \cdot Z)\right] &= \mathrm{Var}\left[Y - \beta \cdot Z\right] && (16)\\
&= \mathrm{Var}\left[Y\right] + \mathrm{Var}\left[\beta \cdot Z\right] - 2\mathrm{Cov}\left[Y, \beta \cdot Z\right] && (17)\\
&= \mathrm{Var}\left[Y\right] + \beta \cdot \mathrm{Var}\left[Z\right]\beta - 2\beta \cdot \mathrm{Cov}\left[Y, Z\right] && (18)\\
&= \mathrm{Var}\left[Y\right] + \mathrm{Cov}\left[Y, Z\right] \cdot \mathrm{Var}\left[Z\right]^{-1}\mathrm{Cov}\left[Y, Z\right] && (19)\\
&\quad -2\mathrm{Cov}\left[Y, Z\right] \cdot \mathrm{Var}\left[Z\right]^{-1}\mathrm{Cov}\left[Y, Z\right] \\
&= \mathrm{Var}\left[Y\right] - \mathrm{Cov}\left[Y, Z\right] \cdot \mathrm{Var}\left[Z\right]^{-1}\mathrm{Cov}\left[Y, Z\right] && (20)
\end{aligned}$$

1.1 Extension to Vectors

If Y is a vector, α must also be a vector, and β must be a matrix. Fortunately, if we use the squared (L_2) error measure, we may simply find the optimal linear predictor of each coordinate of Y separately.

Note that it may not be altogether reasonable to use the L_2 error. If some coordinates of Y are known (or believed) to have larger variance, we should, perhaps, not expend so much effort in trying to predict them. Similarly, if some are correlated, this should be discounted when adding up our prediction error. A reasonable loss function might be

$$\mathbb{E}\left[(Y - m(Z)) \cdot \Omega(Y - m(Z))\right] \tag{21}$$

where Ω might be the inverse covariance matrix of Y. This gives a generalized least squares problem, which also has an analytical solution (Exercise 4).

1.2 Estimation

Given consistent estimators of $\mathbb{E}[Y]$, $\mathbb{E}[Z]$, $\mathrm{Var}[Z]$ and $\mathrm{Cov}[Y, Z]$, consistent estimators of α and β follow by the plug-in principle.

If multiple observations are available, one can also employ the method of least squares, which leads to plugging in the sample versions of all the expectations and covariances. If observations are uncorrelated with each other, the sample versions are consistent estimators.

1.3 Stronger Probabilistic Assumptions

The three main assumptions — point predictions, squared error, and linear predictors — are really more *design choices* than *assumptions*. We are always free to make them; the results might be undesirable in other respects. The only *probabilistic* assumptions were that all the first and second moments invoked in the argument did, in fact, exist.

Some people are unhappy with making these design choices without further justification. They prefer to add the probabilistic assumption that Y and Z are jointly Gaussian, and to estimate by maximum likelihood. These assumptions buy a number of things:

- Rather than just point predictions, we can predict conditional distributions.
- The least-squares estimate becomes efficient.
- The linear model is correct, so the bias conditional on Z is zero.
- The variance conditional on Z becomes calculable (by the law of total variance and the correctness of the linear model).
- There are straightforward sampling distributions for all estimators, with consequent inferential statistics.

The price, of course, is that the Gaussian assumption has to be correct in its entirety.

Note that the correctness of the linear model is a strictly weaker assumption than Gaussianity.

1.4 Nonlinearity?

As every school-child knows, if we do not limit ourselves to linear functions, the optimal predictor of Y from Z is

$$r(z) = \mathbb{E}[Y|Z = z]$$

or, at least, this minimizes the expected squared error. Again, no assumptions of Gaussian distributions, additive and homoskedastic noise, etc., are needed to derive this, just the existence of all the (conditional) moments invoked. To the extent that $r(z) \neq \alpha + \beta \cdot z$, the optimal linear predictor will be biased, though (by Eq. 15) this bias must average out to 0 over z.

If we observe many (Y, Z) pairs, this may be estimated by any of the usual non-parametric approaches. If we do not, estimation becomes substantially trickier.

2 Application to Spatial and Spatio-Temporal Data

Consider some field, or fields, spread out over space and time. Pick the value of one field at one point[1] This will play the role of Y. We observe the value of some fields — the same one, or others — at various other points; the vector of all our observations plays the role of Z.

Kriging is simply the linear prediction of Y, the value of one field at one point, from Z, the value of various fields at various points. The optimal coefficients thus depend on $\mathbb{E}[Y]$, $\mathbb{E}[Z]$, $\mathrm{Var}[Z]$ and $\mathrm{Cov}[Y, Z]$. Once they are found, we may calculate both the optimal prediction, and the expected squared error around it.

If we need to predict a field at many points at once, we turn Y into a vector, as above. The same trick will work for predicting multiple fields, too.

2.1 Special Case: One Scalar Field

We consider a single, scalar-valued random field $Y(x)$, where the coordinate vector x may range over space or time or both. We have observations at coordinates $x_1, x_2, \ldots x_n$, and desire a prediction at the point x_0. Thus

$$Y : Z \;::\; Y(x_0) : (Y(x_1), Y(x_2), \ldots Y(x_n)) \tag{22}$$

[1] I will use "point" indifferently to refer to a point in space, or to a point in space and time.

What about the covariances? We define the covariance function

$$\gamma(x, x') = \text{Cov}\left[Y(x), Y(x')\right] \tag{23}$$

At this level of generality, this is an almost-arbitrary function of two arguments; there is absolutely no need to presume that $\gamma(x, x') = \gamma(x - x', 0)$ (stationarity), or $\gamma(x, x') = \gamma(\|x - x'\|)$ (isotropy), or any separation along the different coordinates of x, etc. This function does, however, have to be symmetric, and any matrix of the form $\gamma(x_i, x_j)$ does need to be non-negative-definite.

With this function, we may say that

$$\text{Var}\left[Z\right]_{ij} = \gamma(x_i, x_j) \tag{24}$$

while

$$\text{Cov}\left[Y, Z\right]_i = \gamma(x_0, x_i) \tag{25}$$

(See also Exercise 5.)

Given such a covariance function, we are not quite ready to calculate the kriging coefficients; we also need $\mathbb{E}\left[Z\right]$ and $\mathbb{E}\left[Y\right]$. We thus require a mean function $\mu(x)$, so that $\mathbb{E}\left[Y\right] = \mu(x_0)$ and $\mathbb{E}\left[Z\right] = (\mu(x_1), \ldots \mu(x_n))$.

Once we have those functions, everything is a matter of conceptually-trivial calculation.

2.2 Role of Symmetry Assumptions

It is common, in applications, to make various symmetry assumptions, such as stationarity (of the covariance function), isotropy (ditto), separability (ditto), or stationarity (of the mean function), linear trends (ditto), etc. The point of these assumptions is not that kriging is somehow ill-defined or impossible without them. If we have *some* reliable source of knowledge about the covariance and mean functions, we're fine.

One possible source of reliable knowledge would be multiple replications of the same situation. If we had many independent replicas of the (Y, Z) pair, we could calculate everything we needed from sample moments. (Indeed, independence is more than is really needed; lack of correlation across replicates would suffice.) However, we often have only a single realization of the process, so we cannot calculate any useful sample moments.

The point of the symmetry assumptions is that they say certain moments are all equal, so we can pool data, within a single realization of the process, to estimate them. If the covariance is isotropic,

$$\gamma(x, x') = \gamma(\|x - x'\|) \tag{26}$$

then we can pool all pairs of observations which are separated by a distance h in order to estimate $\gamma(h)$; similarly for all the other symmetry assumptions.

Imposing a parametric form on γ or μ, in addition to or instead of symmetries, is also about data pooling. If $\gamma(x, x') = \gamma_0 e^{-\|x-x'\|/\lambda}$ then we do not need to estimate $\gamma(h)$ separately for every separation h; we can just estimate

the (assumed-constant) variance γ_0 and the correlation length λ. If those can be consistently estimated, then, by the plug-in principle, we get covariances between the field at our observation points and the field at our prediction point, from which the coefficients follow.

Notice that if we have a parametric form for the functions γ and μ, we can estimate the parameters even when we *don't* also assume symmetries.

There are some situations, primarily in physics when dealing with homogeneous substances, where there are genuine scientific reasons for symmetry assumptions; maybe a somewhat wider range of situations where one might justify specific parametric functional forms. Otherwise, their use is really about bias-variance trade-offs: by allowing for more data pooling, stronger assumptions lead to less variance in the estimates, at the cost of more bias when the assumptions are false. (Note that it is senseless to try to assess the bias from mis-specification from *within* a parametric model; from the premise that the model is completely right, one concludes that the model is completely right.)

2.3 A Worked Example

Pretend we're working with a single scalar field on the plane. Take $\mu(x) = 0$, and $\gamma(x, x') = e^{-\|x-x'\|}$, so that we are working in units (for the field Y) where the variance is 1, and in units (for the coordinates x) where the correlation length is also 1. We wish to predict the value of the field at the origin, and have observations on a square grid, where the grid spacing is (by coincidence) also 1. What are the coefficients?

First off, $\alpha = 0$. (See Exercise 1.)

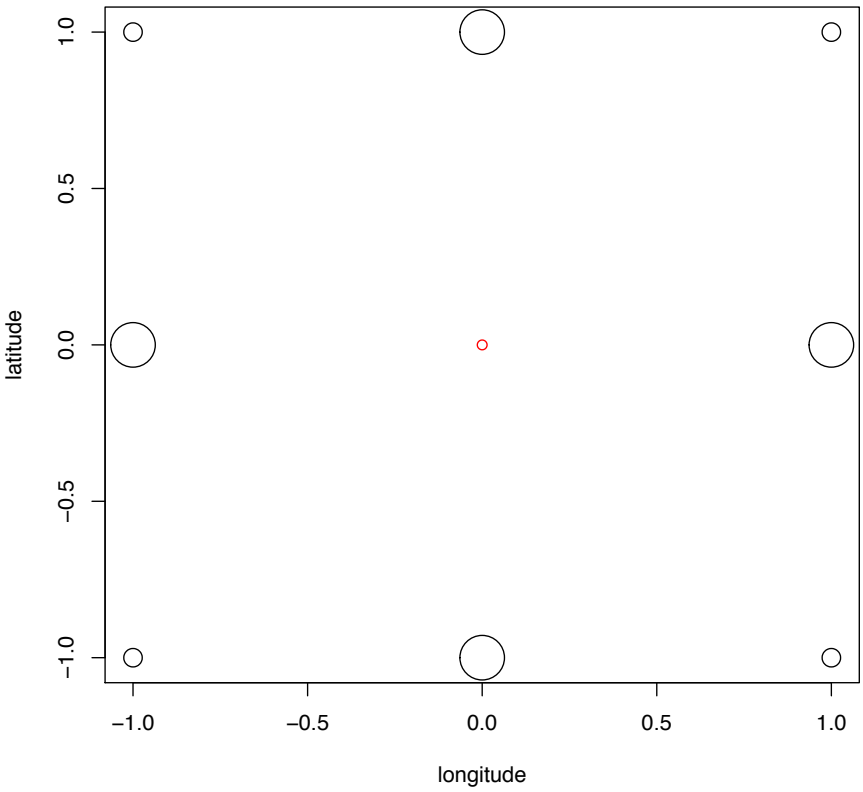
To find β, we will need to evaluate $e^{-\|x-x'\|}$ for every pair of points involved — distances between the origin and the measurement points, and between all the measurement points. This means we will need the matrix of inter-point distances, and so we might as well start with a matrix of coordinates.

```
# Create the 9x2 matrix of coordinates
  # Slight notation clash to call the coordinates "x" and "y", but this
  # will simplify later plotting
coords <- expand.grid(x=c(-1,0,1), y=c(-1,0,1))
# Keep track of which row is the origin (where we want to predict at)
predict.pt <- which(coords$x==0 & coords$y==0)
# Use the built-in function to create distance matrices
distances <- dist(coords) # Euclidean matrix by default
# That returns a special structure w/ just lower-triangular half
distances <- as.matrix(distances)
# Create the matrix of all covariances
covars <- exp(-distances)
# Break it into Cov(Y,Z) and Var(Z)
Cov.YZ <- covars[predict.pt, - predict.pt]
Var.Z <- covars[-predict.pt, -predict.pt]
# Find the coefficients
beta <- solve(Var.Z) %*% Cov.YZ
```

```r
# Print: for each coordinate, what is the coefficient?
signif(data.frame(coords[-predict.pt,], coef=beta),3)
```

```
##    x  y   coef
## 1 -1 -1 0.0358
## 2  0 -1 0.2060
## 3  1 -1 0.0358
## 4 -1  0 0.2060
## 6  1  0 0.2060
## 7 -1  1 0.0358
## 8  0  1 0.2060
## 9  1  1 0.0358
```

Figure 1 shows this visually.
See Exercise 3 for more.




```
plot(coords, xlab="longitude", ylab="latitude", type="n")
points(coords[predict.pt,], col="red")
points(coords[-predict.pt,], cex=10*sqrt(beta))
```

Figure 1: Kriging coefficients for prediction at the origin (red) from eight points in a square box around it; the assumed covariance function is exponential, and the distance from the origin to the sides of the box is exactly the correlation length. The area (not radius) of each point is proportional to its coefficient.

3 Further Reading

The viewpoint on optimal linear prediction taken in §1 is, or ought to be, a standard one, though I think I find it more commonly in writings on stochastic processes than in statistics proper. Certainly I learned it from Wiener (1949, 1961) and Grimmett and Stirzaker (1992) (also see Bartlett 1955). Wiener, along with the independent parallel work of Kolmogorov (1941), was the first to give a rigorous mathematical formulation of the general problem for dependent random variables; thus with time series, the equivalent of the kriging predictor is called the "Wiener filter" (or predictor).

As far as I can work out from the secondary literature — I admit I haven't gone back to the original papers for the history on this one — Krige, in the 1950s, came up with the idea of applying least squares over space, and this was later properly math-ed up by others in the 1960s.

As Wiener (1949) emphasized, under stationarity assumptions, the covariance function γ can be deduced from the power spectrum, and vice versa. (This equivalence is basically the "Wiener-Khinchin theorem".) This is important, because the power spectrum may be an easier object to estimate than the covariance function itself, and because it can sometimes simplify the calculation of the optimal linear predictor. Many readers find Wiener (1949) notoriously hard to follow; a more user-friendly presentation may be found in (among many other places) Bartlett (1955). For an especially thorough treatment of the connection between stationarity and Fourier representations, see Loève (1955).

On covariance functions in physics, see Chaikin and Lubensky (1995) or Forster (1975); note (again) that these books treat homogeneous systems, which are either carefully contrived or quite small. Symmetry in large systems with heterogeneous parts usually requires the heterogeneity to be (in some sense) sufficiently random, and then holds only approximately, on scales of length or time large compared to the heterogeneities.

Exercises

1. Suppose that $\mathbb{E}[Y] = \mathbb{E}[Z] = \vec{0}$. Show that $\alpha = 0$. Suppose that $\mathbb{E}[Y] = a$, $\mathbb{E}[Z] = a\vec{1}$, for some scalar $a \neq 0$. Is α still zero? If not, what is it?

2. Can you re-write Eq. 20 to eliminate all appearances of $\mathrm{Cov}[Y, Z]$ in favor of β?

3. Repeat the calculations from §2.3 under the following circumstances:

   (a) Remove each prediction point in turn, and see how the coefficients of the remaining seven points vary. Add a ninth prediction point mid-way between each pair of the original eight, and see how the coefficients vary.

   (b) Repeat the whole calculation for prediction at the origin, with eight predictor points in a diamond shape around the origin, each two steps

away along a square grid.

(c) Prediction at the origin with eight predictor points in a square around the origin, the sides of the square two distance units away from the origin.

(d) Prediction at the origin, with sixteen predictor points, each one unit away from its neighbor, forming the boundary of a 4×4 square around the origin.

(e) The same geometry as in 3d, but shrink all the disances towards the origin by a factor of $1/2$.

(f) Prediction at the origin with six predictor points equally spaced around the unit circle.

(g) Prediction at the origin with six predictor points on the unit circle, at angles $0°$, $5°$, $90°$, $180°$, $270°$, $359°$ from the horizontal axis.

(h) Prediction at the origin with six predictor points equally spaced around a circle of radius $1/3$.

(i) Prediction at $(0.3, 0.4)$, with the eight predictor points as in the worked example.

(j) Prediction at $(0.3, 0.4)$, with the six predictor points as in 3g.

(k) Prediction at the origin from four points uniformly distributed over the rectangle $[-2, 2] \times [-2, 2]$; with six points; with eight; with 100.

In every case, you should create a visualization (or, if that works better for you, a table) which lets you see at a glance the coefficients associated with each predictor point, and describe, in words, how they vary from one condition to another.

4. Find the function $m(Z) = \alpha + \beta \cdot Z$ which minimizes Eq. 21. Express the argmin in terms of Ω, $\mathbb{E}[Y]$, $\mathbb{E}[Z]$, $\mathrm{Var}[Z]$ and $\mathrm{Cov}[Y, Z]$.

5. Modify the set-up of §2.1 slightly. Suppose that at $x_1, \ldots x_n$, we observe $Z_i = Y(x_i) + \epsilon_i$, where the noise process (ϵ) has $\mathbb{E}[\epsilon_i | Z] = 0$ and $\mathrm{Cov}[\epsilon_i, \epsilon_j | Z] = \sigma_i^2 \delta_{ij}$. Show that instead of Eq. 24, $\mathrm{Var}[Z]_{ij} = \gamma(x_i, x_j) + \delta_{ij}\sigma_i^2$, but Eq. 25 doesn't change. What happens if the ϵ_i are themselves correlated across points?

   You may not assume that ϵ is jointly or even marginally Gaussian.

References

Bartlett, M. S. (1955). *An Introduction to Stochastic Processes, with Special Reference to Methods and Applications*. Cambridge, England: Cambridge University Press.

Chaikin, Paul M. and T. C. Lubensky (1995). *Principles of Condensed Matter Physics*. Cambridge, England: Cambridge University Press.

Forster, Dieter (1975). *Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions*. Reading, Massachusetts: Benjamin Cummings.

Grimmett, G. R. and D. R. Stirzaker (1992). *Probability and Random Processes*. Oxford: Oxford University Press, 2nd edn.

Kolmogorov, Andrei N. (1941). "Interpolation und Extrapolation von stationären Zufälligen Folgen." *Bulletin of the Academy Sciences, USSR, Math.*, **3**: 3–14. In Russian with German summary.

Loève, Michel (1955). *Probability Theory*. New York: D. Van Nostrand Company, 1st edn.

Wiener, Norbert (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. Cambridge, Massachusetts: The Technology Press of the Massachusetts Institute of Technology. "First published during the war [1942] as a classifed report to Section D_2, National Defense Research Council".

— (1961). *Cybernetics: Or, Control and Communication in the Animal and the Machine*. Cambridge, Massachusetts: MIT Press, 2nd edn. First edition New York: Wiley, 1948.