A short note on Kriging

Yen-Chi Chen
University of Washington
March 3, 2022

This short note presents a gentle introduction on the basic concept of kriging, a popular method in spatial statistics.

1 Setup

Kriging considers the problem where we have a random field $Y(s)$ that $s \in \mathcal{S} \subset \mathbb{R}^2$. In reality, we do not observe the entire random field but only a few sampling locations (known as sites) $s_1, \cdots, s_n \in \mathcal{S}$. Namely, our data is

$$Y_n = (Y(s_1), \cdots, Y(s_n))^T \in \mathbb{R}^n.$$

We often assume that the random field is centered, i.e., $\mathbb{E}(Y_s) = 0$ for all s. However, for any pairs of points s_1, s_2, their covariance is often non-zero, i.e. $\mathrm{Cov}(Y(s_1), Y(s_2)) = C(s_1, s_2) \neq 0$. Note that $2C(s_1, s_2)$ is also called the *variogram* is spatial statistics. Note also that a random field is called stationary if $C(s_1, s_2) = C(s_1 - s_2)$.

For an unobserved location s, we are interested in the realization of the random field $Y(s)$. Namely, we want to predict $Y(s)$ using our data Y_n (and the site information $s_1, \cdots, s_n$). Our predictor $\widehat{Y}(s)$ can be written as

$$\widehat{Y}(s) = g_s(Y_n, s_1, \cdots, s_n).$$

The key question is: *how do we construct a good predictor?*

To answer this question, we need a measure of success/loss. A common way to measure the prediction accuracy is the mean squared error (MSE). In this case, we measure the error of predictor g_s via

$$R(g_s) = \mathbb{E}\left((g_s(Y_n, s_1, \cdots, s_n) - Y(s))^2 \right). \tag{1}$$

We want to find a good predictor that minimizes the above MSE.

Note. One may notice the similarity of kriging problem to the usual regression problem (with a fixed design). The key difference is the heterogenous error versus homogeneous error; see the discussion in Section 4.

2 Linear interpolation

A very popular approach in kriging is the linear interpolation method. Namely, instead of searching for all possible predictor g_s, we consider the predictor

$$\widehat{Y}(s) = g_s(Y_n, s_1, \cdots, s_n) = \lambda(s)^T Y_n, \tag{2}$$

where $\lambda(s) \in \mathbb{R}^n$ is the vector of parameters (weights) of predicting $Y(s)$ using Y_n.

Under the linear interpolation method, one can easily show that the MSE

$$\begin{aligned} R(g_s) &= R(\lambda(s)) \\ &= \mathbb{E}\left((\lambda(s)^T Y_n - Y(s))^2\right) \\ &= \lambda(s)^T \mathbb{E}(Y_n Y_n^T)\lambda(s) - 2\lambda(s)^T \mathbb{E}(Y_n Y(s)) + \mathbb{E}(Y(s)^2). \end{aligned}$$

Thus, the minimizer will be

$$\lambda^*(s) = \mathbb{E}(Y_n Y_n^T)^{-1}\mathbb{E}(Y_n Y(s)) = \Sigma_n^{-1} C_n(s), \tag{3}$$

where $\Sigma_n = \mathbb{E}(Y_n Y_n^T) = \mathsf{Cov}(Y_n, Y_n) \in \mathbb{R}^{n\times n}$ and $C_n(s) = \mathbb{E}(Y_n Y(s)) \in \mathbb{R}^n$ has the j-th element being $C_{n,j}(s) = \mathsf{Cov}(Y(s_j), Y(s)) = C(s_j, s)$.

Therefore, if we can estimate both the covariance matrix Σ_n and the vector $C_n(s)$, we can apply the linear interpolation method.

However, a challenge is that both Σ_n and $C_n(s)$ has too many free parameters that scales with respect to the sample size n, so we cannot obtain a reliable estimate without making additional assumptions.

To resolve this problem, a common approach is to assume *isotropic random field*:

$$\mathsf{Cov}(s_1, s_2) = C(s_1, s_2) = \omega(\|s_1 - s_2\|) \tag{4}$$

for some smooth function ω. Note that very often we will assume that $\omega(t) = 0$ for all $t \geq \bar{t}$, i.e., no long distance correlation.

With the isotropic assumption, we just need to estimate ω and then plug-in it into estimating Σ_n and $C_n(s)$.

2.1 Uniform sites/design

We first consider a simple scenario where the sampling sites $s_1, \cdots, s_n$ forms a uniform grid over the region $\mathcal{S}$.

In this case, the pairwise distance $d_{ij} = \|s_i - s_j\| \in \{t_1, t_2, \cdots, t_M\}$ only takes value among a finite set of possible values. This makes the estimation of Σ_n a lot easier. For instance, the (i, j) element of Σ_n can be estimated by

$$\widehat{\Sigma}_{n,i,j} = \frac{1}{2n(d_{ij})} \sum_{\ell,m} (Y(s_\ell) - Y(s_m))^2 I(d_{\ell m} = d_{ij}), \quad n(d_{ij}) = \sum_{\ell,m} I(d_{\ell m} = d_{ij}).$$

For the quantity $C_n(s)$, while we may not have pairs of observations with the exact distance to $\|s - s_i\|$, we may apply an interpolation method to estimate it. Using the idea of estimating $\widehat{\Sigma}_n$, we can estimate ω via estimating

$$\widehat{\omega}(t_\ell) = \frac{1}{2n(t_\ell)} \sum_{\ell,m} (Y(s_\ell) - Y(s_m))^2 I(d_{\ell m} = t_\ell)$$

for each $\ell = 1, \cdots, M$.

Then we apply a linear interpolation to form an estimator $\widehat{\omega}(t)$ for all t. Specifically, suppose $t \in [t_a, t_{a+1}]$, we use

$$\widehat{\omega}(t) = \frac{t - t_a}{t_{a+1} - t_a}\widehat{\omega}(t_a) + \frac{t_{a+1} - t}{t_{a+1} - t_a}\widehat{\omega}(t_{a+1}).$$

2.2 Irregular sites/design

When the sampling sites do not form a uniform grid, the pair d_{ij} may take a total of $\binom{n}{2}$ possible values. This makes the above procedure infeasible. However, we can still estimate ω by a nonparametric approach such as kernel smoothing.

Let $K(\cdot)$ be a smoothing kernel such as a Gaussian and $h > 0$ be a smoothing bandwidth. We estimate $\omega(t)$ via

$$\widehat{\omega}_h(t) = \frac{\sum_{i,j}(Y(s_i) - Y(s_j))^2 K\left(\frac{\|s_i - s_j\| - t}{h}\right)}{\sum_{\ell,m} K\left(\frac{\|s_\ell - s_m\| - t}{h}\right)}.$$

With this estimator, we further estimate

$$\widehat{\Sigma}_{n,i,j} = \widehat{\omega}_h(\|s_i - s_j\|)$$
$$\widehat{C}_{n,j}(s) = \widehat{\omega}_h(\|s_j - s\|).$$

and construct our final estimator.

Remark. One can derive the convergence rate of $\widehat{\omega}_h(t)$ using techniques from U-Statistics with concentration inequalities of kernel smoothing.

3 Optimal predictor

In the above analysis, we have been focusing on the linear interpolation method. If we want to directly search for the minimizer of equation (1), what will we obtain?

Fact: the minimizer of equation (1) is

$$g_s^*(Y_n) = \mathbb{E}(Y(s)|Y_n) = \mu_s(Y_n). \tag{5}$$

To see why equation (5) is the optimal predictor, we expand the MSE as follows:

$$\begin{aligned}
R(g_s) &= \mathbb{E}\left((g_s(Y_n, s_1, \cdots, s_n) - Y(s))^2\right) \\
&= \mathbb{E}\left((g_s(Y_n, s_1, \cdots, s_n) - \mu_s(Y_n) + \mu_s(Y_n) - Y(s))^2\right) \\
&= \mathbb{E}\left(\underbrace{\mathbb{E}\left((g_s(Y_n, s_1, \cdots, s_n) - \mu_s(Y_n) + \mu_s(Y_n) - Y(s))^2 | Y_n\right)}_{=(A)}\right) \\
(A) &= (g_s(Y_n, s_1, \cdots, s_n) - \mu_s(Y_n))^2 + 2(g_s(Y_n, s_1, \cdots, s_n) - \mu_s(Y_n))\underbrace{\mathbb{E}(\mu_s(Y_n) - Y(s)|Y_n)}_{=0} + \mathbb{E}(Y(s)^2|Y_n).
\end{aligned}$$

Thus, the quantity

$$(A) = (g_s(Y_n, s_1, \cdots, s_n) - \mu_s(Y_n))^2 + \mathbb{E}(Y(s)^2|Y_n).$$

The second term is independent of g_s and the first term is non-negative. So the only way to minimize (A) is choosing $g_s(Y_n, s_1, \cdots, s_n) = \mu_s(Y_n)$, which is the choice g_s^*.

Therefore, the optimal predictor of $Y(s)$ is $g_s^*(Y_n) = \mathbb{E}(Y(s)|Y_n)$, the conditional mean of $Y(s)$ given the data Y_n. However, similar to the linear interpolation problem, estimating this conditional mean is infeasible since when the sample size n increases, we are having more random variables being conditioned on. So the function's argument is also expanding.

Gaussian random field. Note that in all of the above analysis, we did not assume the random field belongs to any particular family. In the case that the random filed is Gaussian random field, the response $(Y(s_1), \cdots, Y(s_n), Y(s))$ follows a multivariate Gaussian distribution. It is well-known that in this case, the condition mean $\mathbb{E}(Y(s)|Y_n)$ will be a linear function of Y_n. Thus, using the derivation in Section 2, we conclude that the optimal predictor under Gaussian random field is

$$g_s^*(Y_n) = \mathbb{E}(Y(s)|Y_n) = \mathbb{E}(Y_n Y_n^T)^{-1}\mathbb{E}(Y_n Y(s)) = \Sigma_n^{-1} C_n(s),,$$

the same as equation (3).

4 Relation to regression and time series

The setup of the kriging looks very similar to the fixed design regression problem where sites are the covariates. The optimal predictor g_s^* in equation (5) also shows similarity to the regression function.

In fact, you can view the kriging problem as *fixed design regression problem with correlated error.* The key is **correlated error**.

Suppose there is no correlated error, $Y(s_1)$ and $Y(s_2)$ are independent. Then you can easily see that the optimal predictor $g_s^*(Y_n) = \mathbb{E}(Y(s)|Y_n) = \mathbb{E}(Y(s))$ is just the usual regression function and there is no need to interpolate information from other observations into predicting $Y(s)$. When the random field is centered, i.e., $\mathbb{E}(Y(s)) = 0$, the optimal predictor is trivially 0.

Because of correlated error, the observation $Y(s_j)$ will provide information on the actual value of $Y(s)$. And this is why even when we assume the random field is centered, our predictor will not be the trivial predictor (always predicting 0).

Kriging is related to the *time series* problem in the sense that kriging considers 2D correlated errors while time series consider 1D correlated errors. However, a key difference between the two problems is that time series models often have a generative form, i.e., we often start with a model on how the current (and past) value of response variable affects the value of the next response variable. This types of model often has a nice economics interpretation. Such model is often unavailable in the spatial statistics and kriging problems.