Part 4: Spatial Prediction and Kriging

Observe \((Y(s_1), \ldots, Y(s_n))\) from the model:

\[ Y(s) = \mu(s) + U(s) \]

where the covariance structure of \(U(s)\) is unknown. The prediction of \(U(s)\) at a new spatial site \(s_0\) is known as kriging (though the term has been mainly used for the construction of a spatial predictor using a model with known parameters). A generalization is to predict the joint value at several points, or an integral such as \(u(A) = \int_A u(s) \, ds\) doe some set \(A\) (commonly of interest in areal data analysis).

The most common approach to estimating \(U(s)\) is using a linear combination of the observed residuals

\[ \hat{U}(s) = \sum_{i=1}^{n} \lambda_i(s) U(s_i) \]

where \(\lambda(s)\) large when \(|s_i - s|\) small. Many smoothing and interpolation methods (moving averages, nearest neighbor methods, kernel methods) are based on such predictors. Kriging allows for choosing weights \(\lambda_i(s)\) according to the degree of spatial correlation, which can be estimated from the data. Furthermore, accuracy assessments are automatically available.

**Simple kriging.** Assumes \(\mu(s)\) and \(C(s, s')\) known and take \(U(s_i) = Y(s_i) - \mu(s_i)\). Best linear predictor is obtained (mean squared prediction error minimized) by choosing

\[ \lambda(s) = \Sigma^{-1} c(s) \]

where \(\lambda(s)\) is the vector of kriging weights \(\lambda_i(s)\) and \(c(s)\) the vector of covariances \(C(s, s_i)\).

The minimized mean squared prediction error, or the kriging variance,

\[ E(\hat{U}(s) - U(s))^2 = C(0) - c(s)^T \sigma^{-1} c(s). \]

gives an estimate of prediction accuracy. Some difficulties with the ordinary kriging is that mean surface and variogram are usually unknown and if we put estimates of both into the kriging equations we fail to take into account the uncertainty in these estimates. One alternative to overcome this is using GLS estimates. That is,

1. Estimate \(C(s, s')\) from OLS residuals and \(\mu(s)\) by GLS as explained in the spatial regression section.
2. Plug estimates into the simple kriging formulas.
3. Adjust mean square prediction error by adding an extra component arising from estimating trend from the data.

An example under Gaussian assumption is provided below.

**Kriging with Gaussian processes.** Given a set of covariates \(X(s)\) and assume the following model:

\[ Y(s) = X(s)\beta + U(s), \ U \sim N(0, \Sigma). \]

For a spatial covariance structure without nugget effect specify:

\[ \Sigma = \sigma^2 V(\theta) \text{ where } (V(\theta))_{ij} = \rho(\theta; ||s_i - s_j||) \]

where \(\rho\) is a valid correlation function. For a model with a nugget effect, re-write the covariance function as

\[ \Sigma = \sigma^2 V(\theta) + \tau^2. \]
To predict $U$ at a new site $s_0$, we find $f(u)$ to minimize the mean squared prediction error

$$E[(Y(s_0) - f(u))^2|u] = E\{(Y(s_0) - E[Y(s_0)|y])^2|y\} + \{E[Y(s_0)|y] - f(y)\}^2,$$

since the expectation of the cross-product term is 0. Since the second term is nonnegative, we have that

$$E[(Y(s_0) - f(u))^2|u] \geq E\{(Y(s_0) - E[Y(s_0)|y])^2|y\}$$

with equality for $f(y) = E[Y(s_0)|y]$, which must be the predictor $f(y)$ that minimizes the error is the conditional expectation of $Y(s_0)$ given the data. This result is quite intuitive from a Bayesian point of view, since this $f(y)$ is just the posterior mean of $Y(s_0)$.

Under the model framework above, assuming $Y_1 = Y(s_0)$ (to be predicted) and $Y_2 = Y$, and according to the multivariate distribution

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}\right)$$

with $\Gamma_{21} = \Gamma_{12}^T$.

Under the framework above, with $Y_1 = Y(s_0)$ and $Y_2 = Y$, we find $\Gamma_{11}, \Gamma_{21}$ and $\Gamma_{22}$ as follows

$$\Gamma_{11} = \nabla(Y(s_0)) = \sigma^2 + \tau^2$$

$$\Gamma_{22} = \nabla(Y) = \Sigma = \sigma^2V(\theta) + \tau^2I$$

$$\Gamma_{12} = \gamma^T = (\sigma^2\rho(\theta; d_{01}), \ldots, \sigma^2\rho(\theta; d_{0n})) = \sigma^2w(\theta).$$

Using the properties of a multivariate normal distribution, we obtain that $p(Y_1|Y_2)$ is normal with mean and variance provided by

$$E(Y_1|Y_2) = \mu_1 + \Gamma_{12}\Gamma_{22}^{-1}(Y_2 - \mu_2)$$

$$\nabla(Y_1|Y_2) = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}$$

Next, for a fixed value $x_0$ (for which $\mu_1$ becomes $x_0^T\beta$), we find the predicted value and its variance (i.e. $E(Y(s_0)|Y)$ and $\nabla(Y(s_0)|Y)$) assuming the parameters $\beta, \sigma, \theta$ and $\tau^2$ are known. Substituting $Y(s_0) = x_0^T\beta$ and $Y_2 = y$ in these formulas, we obtain the predicted value and its variance

$$E(Y(s_0)|y) = x_0^T\beta + \gamma^T\Sigma^{-1}(y - X^T\beta)$$

$$\nabla(Y(s_0)|y) = \sigma^2 + \tau^2 - \gamma^T\Sigma^{-1}\gamma.$$

How would you proceed in the more realistic case when $\beta, \sigma, \theta$ and $\tau^2$ are unknown? What if $x_0$ is unobserved also? When the parameters are unknown, we obtain their consistent estimators and denote them $\hat{\beta}, \hat{\sigma^2}, \hat{\theta}$ and $\hat{\tau^2}$ and plug them in the formulas above:

$$E(Y(s_0)|y) = x_0^T\hat{\beta} + \hat{\gamma}^T\hat{\Sigma}^{-1}(y - X^T\hat{\beta})$$

where $\hat{\gamma}^T = \sigma^2w(\hat{\theta}), \hat{\Sigma} = \hat{\sigma^2H(\hat{\theta})}$ and

$$\hat{\beta} = \left(X^T\hat{\Sigma}^{-1}X\right)^{-1}X^T\hat{\Sigma}^{-1}y.$$

Thus the predictor can be written as

$$\hat{f}(y) = \lambda^Ty$$

with $\lambda = \hat{\Sigma}^{-1}\hat{\gamma} + \Sigma^{-1}X\left(X^T\hat{\Sigma}^{-1}X\right)^{-1}(x_0 - X^T\hat{\Sigma}^{-1}\hat{\gamma}).$
Bayesian Kriging. The Bayesian approach generalizes to the case in which the variogram parameters are unknown, whereas the classical approach essentially makes the assumption that these parameters are known and only deals with the question of uncertainty of model parameters in a very peripheral way. We will use similar notations as for Gaussian kriging.

Based on the model discussed above

\[ Y(s) = X(s)\beta + U(s) \text{ or } Y = X\beta + U \]

with the same covariance specifications as above, a common approach is to consider the improper prior (we will consider \( \tau^2 = 0 \) for simplicity here)

\[ \pi(\beta, \sigma^2, \theta) = \frac{\pi(\theta)}{\sigma^2} \]

for some prior \( \pi(\theta) \). The posterior density takes the form

\[ \pi(\beta, \sigma^2, \theta | Y) \approx \frac{\pi(\theta)}{\sigma^2} (\pi \cdots)^{-n/2} \sigma^{-n} |V(\theta)|^{-1/2} \exp \left( -\frac{1}{2\sigma^2} (Y - X\hat{\beta})^T V(\theta)^{-1} (Y - X\hat{\beta}) \right) \]

(1)

Defining \( \hat{\beta}(\theta) = (X^T V(\theta)^{-1}X)^{-1} X^T V(\theta)^{-1}Y \) and ignoring constants, the equation above together with \( G^2(\theta) = (Y - X\hat{\beta}(\theta))^T V(\theta)^{-1} (Y - X\hat{\beta}(\theta)) \) leads to

\[ \pi(\beta, \sigma^2, \theta | Y) \approx \frac{\pi(\theta)}{\sigma^2} (\pi \cdots)^{-n/2} \sigma^{-n} |V(\theta)|^{-1/2} \exp \left( -\frac{G^2(\theta)}{2\sigma^2} \right) \exp \left( -\frac{1}{2\sigma^2} (\beta - \hat{\beta})^T V(\theta)^{-1} (\beta - \hat{\beta}) \right) \]

(2)

Under the Gaussian assumption, the predictor at a new location \( s_0 \) has the following distribution

\[ Y_0 | Y = y, \beta, \sigma^2, \tau^2, \theta \sim N(x_0^T \beta + \gamma^T \Sigma^{-1}(y - X^T \beta), \sigma^2 + \tau^2 - \gamma^T \Sigma^{-1} \gamma) \]

(3)

where \( x_0 = X(s_0) \) and \( Y_0 = Y(s_0) \). We will now reduce the dimensionality by successively removing the conditioning on \( \beta, \sigma^2 \) and \( \theta \) (take \( \tau^2 = 0 \) here). We write \( \pi(x|y) \) for the generic density of a variable \( X \) conditioned on another variable \( Y \). To remove the conditioning on \( \beta \), we write

\[ \pi(y_0|Y, \sigma^2, \theta) = \int \pi(y_0|Y, \beta, \sigma^2, \theta) \pi(\beta|Y, \sigma^2, \theta) d\beta \]

(4)

where the last density function within the integral is derived from (1). It follows from (1) that the distribution of \( \beta \) given \( Y, \sigma^2 \) and \( \theta \) is multivariate normal with mean \( \beta = \hat{\beta}(\theta) \) and covariance matrix \( \sigma^2 (X^T V(\theta)^{-1}X)^{-1} \). Combining this with (3), we find that the conditional distribution of \( Y(s_0) \) given \( \sigma^2 \) and \( \theta \) is multivariate normal with mean

\[ \hat{Y}(s_0) = (x_0 - X^T \Sigma^{-1} \gamma)^T \hat{\beta} + \gamma^T \Sigma^{-1}Y = x_0 - X^T V(\theta)^{-1}w(\theta))^T \hat{\beta} + w(\theta)^T V(\theta)^{-1}Y \]

(5)

and the covariance matrix

\[ (x_0 - X^T \Sigma^{-1} \gamma)^T (X^T \Sigma^{-1} X)^{-1} (x_0 - X^T \Sigma^{-1} \gamma) + \sigma^2 - \gamma^T \Sigma^{-1} \gamma = \]

(6)

\[ \sigma^2 [(x_0 - X^T V(\theta)w(\theta))^T (X^T V(\theta)^{-1}X)^{-1} (x_0 - X^T V(\theta)^{-1}w(\theta)) + 1 - w(\theta)^T V(\theta)^{-1}w(\theta)] = \sigma^2 V_0(\theta) \]

The next step is to remove the conditioning on \( \sigma^2 \). Similarly we integrate it out

\[ \pi(y_0|Y, \sigma^2, \theta) = \int \pi(y_0|Y, \sigma^2, \theta) \pi(\sigma|Y, \theta) d\sigma. \]

(7)
The posterior distribution of $\sigma$ given $Y$ and $\theta$ may be obtained from (2). The result is that $G^2(\theta)/\sigma^2$ has a $\chi^2_{n-q}$ distribution. Define

$$\hat{\sigma}^2(\theta) = \frac{G^2(\theta)}{n - q}.$$ 

Then with slight abuse of notation, we have

$$\sigma^2|Y, \theta \sim \hat{\sigma}^2(\theta) \frac{n - q}{\chi^2_{n-q}}.$$ 

Conditionally on $Y$ and $\theta$, we then have

$$\frac{Y(s_0) - \bar{Y}(s_0)}{\sqrt{\hat{\sigma}^2(\theta)V_0(\theta)}} \sim \frac{N(0, 1)}{\chi^2_{n-q}/(n - q)} \sim t_{n-q}.$$ 

The final result agrees with equation (3.1) of Handcock and Stein (1993), except that they have a factor $n/(n - q)$ multiplying $\hat{\sigma}^2(\theta)$, which results from slightly different definition of the latter quantity. Finally, following Handcock and Stein (1993), we integrate over $\theta$ to obtain

$$\pi(y_0|Y) = \int \pi(y_0|Y, \theta)\pi(\theta|Y) d\theta$$ 

(8)

where the second factor is derived from (2) integrating out $\sigma^2$:

$$\pi(\theta|Y) \approx \pi(\theta|V(\theta)|^{-1/2}G^2(\theta)^{-(n-q)/2}|X^TV(\theta)^{-1}X|^{-1/2}.$$ 

(9)

This last part has to be carried out numerically. Handcock and Stein (1993) provide several examples.

**Frequentist corrections for unknown covariance structure.** When $\theta$ is known, the predictor $\bar{Y}_0(\theta)$ defined in (5) is the best linear unbiased predictor of $Y(s_0) = Y_0$, with mean squared prediction error $m(\theta)$ given by (6). For the purpose of the present discussion, in the case that $\sigma^2$ is also unknown, we include it in $\theta$.

When $\theta$ is unknown and estimated by $\hat{\theta}$, using any of the estimation methods, we have discussed, the obvious strategy is to use $\bar{Y}_0(\hat{\theta})$ as a predictor of $y_0$ and to use $m(\hat{\theta})$ as its MSPE. However, this approach makes no allowance for the error difference between $\hat{\theta}$ and $\theta$, and may therefore be expected to underestimate the true MSPE. There are essentially three strategies one can adopt in response to this shortcoming.

(i) Ignore it, i.e. use $\bar{Y}_0(\hat{\theta})$ as the predictor of $y_0$ and $m(\hat{\theta})$ as its MSPE even when $\theta$ is unknown (very common);

(ii) Estimate the difference between $m(\hat{\theta})$ and the true MSPE, and use that to derive a corrected MSPE;

(iii) Adopt the Bayesian approach discussed in this notes in the hope that it will have good properties from a frequentist as well as Bayesian point of view.

Strategies of the form (ii) have been considered by a number of authors, in particular Prasad and Rao (1990) and Harville and Jeske (1992) in the case of variance components models, and Zimmerman and Cressie (1992) in the more general cases arising from spatial covariance matrices. Next we will have a closer look at the difference between $m(\theta)$ and $m(\hat{\theta})$.

Denote $e_1 = Y_0(\theta) - Y_0$ (the prediction error due to estimating $Y_0$) and $e_2 = Y_0(\theta) - \bar{Y}_0(\theta)$ (the prediction error due to estimating $\theta$). The results below are based on the assumption of multivariate normal. Assuming that the two errors are independent, we further obtain that

$$\mathbb{E}(Y_0(\hat{\theta}) - Y_0)^2 = \mathbb{E}(e_1 + e_2)^2 \leq \mathbb{E}(e_1^2) = m(\theta).$$
Therefore, \( m(\theta) \) does under-estimate the true MSPE when \( \theta \) is unknown. Moreover, if we approximate

\[
e_2 \approx (\hat{\theta} - \theta)^T \nabla \hat{Y}(\theta)
\]

(with \( \nabla \) denoting the gradient) and apply the same argument a second time, we deduce

\[
\mathbb{V}(e_2) \approx \mathbb{E} \left( \nabla \hat{Y}_0(\theta)^T (\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \nabla \hat{Y}_0(\theta) \right) = \mathbb{E} \left( \text{tr} \{ \nabla \hat{Y}_0(\theta)^T (\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \nabla \hat{Y}_0(\theta) \} \right) = \mathbb{E} \left( \text{tr} \{ \nabla \hat{Y}_0(\theta)^T \nabla \hat{Y}_0(\theta) \} \right)
\]

where the middle of this argument uses the fact that \( \text{tr}(AB) = \text{tr}(BA) \). Therefore, it appears that we ought to estimate the mean squared prediction error by

\[
\mathbb{V}(\hat{Y}_0(\hat{\theta})) \approx m(\theta) + \text{tr} \{ \text{Cov}(\hat{\theta}) \text{Cov}(\nabla \hat{Y}_0(\theta)) \}, \tag{10}
\]

which will improve on the initial approximation \( m(\theta) \).

Harville and Jeske (1992) argued that even the approximation above is not much improved as compared to \( m(\theta) \) because although it adjusts for the difference between \( m(\theta) \) and the true MSPE, there is still a bias due to the underestimation of \( m(\theta) \). Harville and Jeske (1992) showed that this additional bias is asymptotically equivalent with the bias of (10), and therefore, an improved approximation is to double the correction term:

\[
\mathbb{V}(\hat{Y}_0(\hat{\theta})) \approx m(\theta) + 2\text{tr} \{ \text{Cov}(\hat{\theta}) \text{Cov}(\nabla \hat{Y}_0(\theta)) \}. \tag{11}
\]

However, these corrections have been proposed under specific conditions and the asymptotic results may not hold under strong spatial correlation.

**Kriging Under Transformations.** Within this formulation, we assume that \( Y(s) \) is a spatial process measured with errors. That is,

\[
\mathbb{E}(Y(s)) = Z(s)
\]

where \( Z(s) \) is a spatial process and \( Y(s) \) is a noisy version of \( Z(s) \). The connection with the previous set-up is that for

\[
Y(s) = Z(s) + \epsilon(s)
\]

the error measurement does not contain any spatial information and it is commonly assumed independent and identically distributed (\( \epsilon \sim N(0, \tau^2 I) \)), and therefore, the nugget effect is interpreted as measurement error. Under this set up, the focus is on prediction of the spatial process \( Z(s) \). In general, target for prediction is a functional \( T(Z(h)) \) of the signal process, and the minimum mean square error predictor of \( T \) is its conditional expectation given the data

\[
\hat{T}(y) = \mathbb{E}(T|y).
\]

Under the Gaussian example above, signal process \( Z \) can be a Gaussian process with \( \mathbb{E}(Z(s)) = \mu(s) \) \( (\mu(s) = X(s)\beta) \), \( \mathbb{V}(Z(s)) = \sigma^2 \) and correlation function \( \rho \), which is usually assumed continuous at origin (with nugget effect \( \tau^2 > 0 \)).

**Linear Transformation.** Suppose we are interested in predicting the areal average over \( A \subset S \), or more generally, a linear functional of \( Z(\cdot) \):

\[
T = \int_A w(s)Z(s)\,ds
\]
for some prescribed weight function $w(s)$. Assuming $Y(s)$ is a Gaussian process (i.e. $Z(s)$ is also Gaussian) and $Y(s_i)$ for $i = 1, \ldots, n$ are mutually independent given $Z(\cdot)$, then

$$
\mathbb{E}(T|y) = \int_{A} w(s)\mathbb{E}(Z(s)|y) \, ds
$$
$$
\text{Var}(T|y) = \int_{A} \int_{A} w(s)w(s')\text{Cov}(Z(s), Z(s')|y) \, dsds'.
$$

In other words, linear functional of the signal process can be predicted by the corresponding linear combination of the predicted values of $Z(s)$. However, this can be done computationally much more effectively through block kriging equations involving just one matrix inversion.

**Non-linear targets.** If $T$ is a non-linear functional of $Z(\cdot)$ then

$$
\mathbb{E}(T(Z(\cdot))|y) \neq T(\mathbb{E}(Z(\cdot)|y)).
$$

A generally applicable way to calculate a kriging predictor for a nonlinear target $T = T(Z(s_0))$ is by simulation:

- Evaluate $\mathbb{E}(Z(s_0)|y)$ and $\text{Var}(Z(s_0)|y)$ by conventional kriging.
- Simulate $m$ realisations $z^*_1, z^*_2, \ldots, z^*_m$ from the multivariate Gaussian conditional distribution of $Z(s_0)|y$.
- Approximate the minimum mean square error predictor by

$$
\mathbb{E}(Z(s_0)|y) \approx \frac{1}{m} \sum_{i=1}^{m} T(z^*_i).
$$

In particular, for the transformed Gaussian model, back-transform the simulations before taking the averages.

1 **Homework**

For the next lecture, you will all read three papers:


**References**

