In this homework assignment, you will compare the predictive performance of two low rank models: the predictive process model and the process convolution model. The parameters are obtained by minimizing some predictive scores. Some predictive scores introduced in class are reviewed below.

1 Predictive scores

Given observations $Y(s_i)$ made at locations $s_i$ for $i = 1, \cdots, n$, we denote by $\hat{Y}_{-i}(s_i)$ the best linear prediction based on all data $Y(s_j)$ for $j \neq i$. This is commonly called the drop-one prediction. Let $\sigma_{-i}^2 = \sigma_{-i}(s_i)^2$ be the corresponding drop-one prediction variance. Write

$$z_i = \frac{Y(s_i) - \hat{Y}_{-i}(s_i)}{\sigma_{-i}}, \quad i = 1, \cdots, n. \quad (1)$$

One predictive score is the root-mean-square error (RMSE),

$$RMSE = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y(s_i) - \hat{Y}_{-i}(s_i) \right)^2 \right\}^{1/2}. \quad (2)$$

The RMSE is commonly used in cross-validation. Another score that incorporates the prediction variance is the logarithmic score

$$LogS = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{2} \log(2\pi \sigma_{-i}(s_i)^2) + \frac{1}{2} z_i^2 \right) \quad (2)$$

where $z_i$ is defined in (1). The continuous ranked probability score is a function of $y \in (-\infty, \infty)$:

$$CRPS = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left( F_i(y) - 1\{Y(s_i) \leq y\} \right)^2 dy \quad (3)$$

where $F_i$ is the conditional distribution of $Y(s_i)$ given $Y_j, j \neq i$. If $Y = (Y(s_1), \cdots, Y(s_n))'$ has a multivariate normal distribution, CRPS can be calculated based on the $z_i$’s in (1) (see, Gneiting et al., 2007):

$$CRPS = \frac{1}{n} \sum_{i=1}^{n} \sigma_{-i}(s_i) \left( z_i (2\Phi(z_i) - 1) + 2\Phi(z_i) - \frac{1}{\sqrt{\pi}} \right) \quad (4)$$

2 Calculation of drop-one prediction

Let us temporarily assume that $EY(s) = 0$ and write $Y_i = Y(s_i)$ for brevity. Write the observed variables in vector form $Y = (Y_1, \cdots, Y_n)'$ and assume $Var(Y) = V$. Let $q_{ij}$ be the $(i,j)$th element of $V^{-1}$. It is well known that the drop-one prediction for $Y_i$ is

$$\hat{Y}_{-i} = -\sum_{j \neq i} q_{ij} Y_j / q_{ii}.$$ 

Therefore

$$Y_i - \hat{Y}_{-i} = \frac{1}{q_{ii}} \sum_{j=1}^{n} q_{ij} Y_j \quad (5)$$

$$\sigma_{-i}^2 = E(Y_i - \hat{Y}_{-i})^2 = 1/q_{ii}. \quad (6)$$
If we let $\Lambda = \text{diag}(q_{ii}, i = 1, \cdots, n)$, we have

$$MSE = \frac{1}{n} \| \Lambda^{-1/2} V^{-1} Y \|^2,$$  \hspace{1cm} (7)

and the $z_i$'s in (1) can be given in vector form

$$z = \Lambda^{-1/2} V^{-1} Y.$$

When the means are not all zeros, we center the variable by replacing $Y$ by $Y - EY$. For example, the MSE given by (11) now becomes

$$MSE = \frac{1}{n} \| \Lambda^{-1/2} V^{-1} (Y - EY) \|^2.$$ \hspace{1cm} (8)

### 3 Lower rank models

Several people have used the following lower rank models (e.g., Higdon, 1998; Cressie and Johannesson, 2008; Stein, 2008; Banerjee et al., 2008)

$$Y(s) = x(s)' \beta + a(s)' Z + \epsilon(s),$$ \hspace{1cm} (9)

where $x(s)$ is a vector of covariates, $\beta$ is a vector of parameters, $a(s) = (a_1(s), \cdots, a_m(s))'$ is a function of $s$ that may depend on some parameters $\theta$, $Z = Z(u_1), \cdots, Z(u_m)$ is a vector of latent variables at locations $u_1, \cdots, u_m$, and $\epsilon(s)$ is a white noise with variance $\tau^2$ and is independent of $Z$.

When the process $Y(s)$ is observed at $n$ locations, say, $s_1, \cdots, s_n$, the model can be written as

$$Y = X\beta + AZ + \epsilon.$$  

The covariance matrix of $Y$ is

$$V = AV_z A' + \tau^2 I_n,$$

where $I_n$ is the identity matrix and $V_z = \text{Var}(Z)$. Due to the this particular structure, the inverse of $V$ can be given by inverting only an $m \times m$ matrix by the Woodbury Formula,

$$V^{-1} = \tau^{-2} (I_n - A(\tau^2 V_z^{-1} + A' A)^{-1} A').$$

When $Z$ has equal variances, write $V_z = \sigma^2 R_z$ and define

$$\Gamma = A(wR_z^{-1} + A' A)^{-1} A' = (AR_z)(wR_z + (AR_z)'(AR_z))^{-1}(AR_z)'.$$ \hspace{1cm} (10)

It follows (7) and the Sherman-Morrison-Woodbury formula that

$$MSE = \frac{1}{n} \| (I_n - D)^{-1} (I_n - \Gamma)(Y - X\beta) \|^2,$$ \hspace{1cm} (11)

where $D = \text{diag}(\Gamma)$ is the diagonal matrix consisting of the diagonal elements of $\Gamma$.

In this homework, consider two particular low rank models.
3.1 Predictive process model

Let \( W(s) \) be a stationary process with mean 0 and covariogram \( \sigma^2 \rho(h, \phi) \), where \( \phi \) is a parameter and could be a vector. Let \( u_1, \ldots, u_m \) be the \( m \) known locations, also called “knots” by Banerjee et al. (2008). The predictive process model is

\[
Y(s) = \mu + k(s)' R^{-1} W + \epsilon(s),
\]

where \( W = (W(u_1), \ldots, W(u_m))' \), \( R \) is the correlation matrix of \( W \) whose \((i, j)\)th element is \( \rho(u_i - u_j, \phi) \), and \( k(s) \) is the correlation of \( W \) and \( Y(s) \), i.e., the \( j \)th element is \( \rho(s - u_j) \).

The \( \Gamma \) matrix defined in (10) is now, because \( AR = K \),

\[
\Gamma = K(wR + K'K)^{-1}K'.
\]

The best linear unbiased estimator for \( \mu \) is

\[
\hat{\mu} = \frac{1}{V} Y = \frac{1}{V} (I_n - \Gamma) Y.
\] (12)

Note \( \hat{\mu} \) depends on \( w \) and \( \phi \), which are estimated by minimizing the MSE (8), i.e.,

\[
(\hat{w}, \hat{\phi}) = \text{ArgMin}_{w, \phi} \left\| (I_n - D)^{-1}(I_n - \Gamma) (Y - \hat{\mu}1) \right\|^2.
\] (13)

We estimate the nugget effect \( \tau^2 \) by minimizing the logarithmic score. Hence,

\[
\hat{\tau}^2 = \frac{1}{n} \left\| (I_n - \hat{D})^{-1/2} (I_n - \hat{\Gamma}) Y \right\|^2,
\] (14)

where \( \hat{D} \) and \( \hat{\Gamma} \) are \( D \) and \( \Gamma \) corresponding to the estimates \( \hat{w} \) and \( \hat{\phi} \), respectively.

3.2 Process convolution model

In the process convolution model, we assume

\[
Y(s) = \mu + a'(s) Z + \epsilon(s)
\]

where \( Z \) consists of white noises with variance \( \sigma^2 \), \( \epsilon(s) \) is a process of white noise and is independent of \( Z \), \( a(s) = (\rho_\nu(s - u_1, \phi), \ldots, \rho_\nu(s - u_m, \phi))' \), where \( \rho_\nu \) is the correlation function of the Matérn model with the smoothness parameter \( \nu \) and range parameter \( \phi \). Since the correlation matrix \( R_z \) is now the identify matrix, the \( \Gamma \) matrix in (10) is

\[
\Gamma = A(wI_m + A'A)^{-1} A'
\]

Then \( w, \phi \) and \( \tau^2 \) are estimated by (13) and (14).

4 Specific problems

You will do the following problems.

1. Given the (7) is true, prove (11).
2. Apply the predictive process model to the US precipitation data anom and find the estimates using the method outlined in Section 3. Use $m = 760$ and the $u_i$ locations are in the file lambdas.txt in the Handouts directory. The correlation function $\rho$ is chosen to be the Matérn model with $\nu = 1.5$, i.e., $\rho(h, \phi) = (1 + h/\phi) \exp(-h/\phi)$.

(a) For each $\phi = 50, 150, 200, 250, 300, 350, 400, 450, 500, 550, 650, 750$, minimize the RMSE and report the minimums in a table.

(b) Find the $\phi$ among the 12 $\phi$ values that yields the lowest RMSE. Calculate other predictive scores (LogS and CRPS) corresponding to this $\phi$.

3. Apply the process convolution model to the same data using the same $u_i$ locations. The correlation function is chosen to be the exponential isotropic model $\rho(h) = \exp(-h/\phi)$. Do the same as in 2(a) and 2(b).

4. Comments on the results and computing time and see if you would prefer one model over the other.

References


