Kriging and Cross-validation for Massive Spatial Data

Hao Zhang
Department of Statistics
Purdue University
West Lafayette, IN 47907
U.S.A
Phone: (765)496-9548
zhanghao@purdue.edu

Yong Wang
Department of Statistics
Purdue University
West Lafayette, IN 47907
U.S.A
wang49@purdue.edu
Kriging and Cross-validation for Massive Spatial Data

February 27, 2009

Abstract

Spatial prediction such as kriging involves the inversion of a covariance matrix. When the number of locations is very large as in many studies, inversion of the covariance matrix may not be practical. Covariance tapering, predictive process models and low rank kriging are some methods for overcoming the large matrix problem, all of which can be regarded as approximations to the underlying stationary process. Therefore, efficient cross-validation is very helpful for spatial prediction with large data to assess how well an approximation works. This work studies the calculation of drop-one prediction and various prediction scores. Estimators are constructed to minimize some prediction scores. One advantage of this approach is that it integrates estimation and cross-validation and does not treat them as two separate procedures. Further simplification of calculation is studied that is based on the infill asymptotic theory. The methods are illustrated through the analysis of a US precipitation dataset.

Keywords: cross-validation; infill asymptotics; low rank kriging; predictive process model; predictive scores; Woodbury formula;

1 Introduction

Due to the technology advancement, massive data are often observed at a large number of spatial locations in environmental, ecological and climate studies. These massive data present a challenge to the adoption of the well-studied inferential methods such as the maximum likelihood estimation, the Bayesian methods and best linear prediction (kriging) because these methods require the inversion of a large covariance matrix. In recent years, there have been several approaches to overcoming this large matrix problem. Some take the computational advantage of sparse matrices as in the covariance tapering (Furrer et al., 2006; Zhang and Du, 2008) and some put a low rank structure on the covariance matrix that enables us to calculate the inverse of a large covariance by inverting a matrix of lower dimension according to the Woodbury formula (e.g., Cressie and Johannesson, 2008; Banerjee et al., 2008). Some adopts the spectral method for parameter estimation (e.g., Fuentes, 2007) although this method does not overcome the difficulty in prediction with massive data.
The following low rank model has been employed in different forms (e.g., Higdon, 1998; Cressie and Johannesson, 2008; Stein, 2008; Banerjee et al., 2008)

\[ Y(s) = x(s)' \beta + a(s)' Z + \epsilon(s), \]  

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

Model (1) includes the Gaussian process convolution (Higdon, 1998) for which \( a_j(s) = \exp(-\|s - u_j\|^2/\phi)/\sqrt{2\pi} \) is given by the Gaussian kernel where \( u_j \)'s are some known spatial locations and \( \phi \) is the bandwidth parameter. It also includes the predictive process model in Banerjee et al. (2008) for which \( a(s) = \text{Var}(Z) - 1 \text{Cov}(Z, Z(s)) \) where \( Z(s) \) is a Gaussian stationary process with mean 0 and some parametric covariogram, and the model for fixed rank kriging (Cressie and Johannesson, 2008) for which \( a_j(s) = \) a certain form involving \( \|s - u_j\|^2/\phi \) for some known basis function \( b \) and \( Z = (Z(u_1), \ldots, Z(u_m))' \) is a realization of a latent process at locations \( u_j, j = 1, \ldots, m \).

When the process \( Y(s) \) is observed at \( n \) locations, say, \( s_1, \ldots, s_n \), the low rank 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 \( n \times n \) identity matrix and \( V_z = \text{Var}(Z) \), and \( A \) is \( n \times m \). 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 (Harville, 1997),

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

(2)

Therefore, one would be able to compute the kriging predictor for \( Y(s) \), which can be expressed as

\[ \hat{Y}(s) = x(s)' \hat{\beta} + a(s)' \hat{Z} \]

where the \( \hat{Z} = V_z A' V^{-1}(Y - X\hat{\beta}) \) is the best linear predictor for \( Z \). It follows (2) that we can calculate \( \hat{Z} \) without inverting the \( n \times n \) covariance matrix:

\[ \hat{Z} = (r^2V_z^{-1} + A'A)^{-1}A'Y \]

which involves the inverse of an \( m \times m \) matrix. The computational simplification results from the fact that only \( m \) variables \( Z_1, \ldots, Z_m \) are used in the representation of the process \( Y(s) \).

For parameter estimations in the low rank model (1), Bayesian methods are often used with the Markov chain Monte Carlo techniques (Higdon, 1998; Banerjee et al., 2008). Cressie and Johannesson (2008) used
moments estimators to for the variance parameters in $K$ and $\tau^2$. Stein (2008) discussed a computational method for calculating the Gaussian likelihood function.

On the other hand, it is often the case that for prediction at a location, observations nearer to the prediction location have the largest impact on the kriging predictor. This has long been known in geostatistics literature as the screening effect. Stein (2002) justified the screening effect by the infill or fixed-domain asymptotic theory. Therefore, one may use only a subset of nearest neighbors to make the prediction. This is common practice in geostatistics and has implemented in ArcGIS. Haas (1990, 1995) called it the moving window kriging.

Although the local kriging effectively overcomes the large matrix problem, one difficulty of moving window kriging is the repeated estimation of parameters within each window or neighborhood. Haas (1990, 1995) used the least squares method to estimate the variogram within each neighborhood although some numerical studies have shown that the maximum likelihood estimation usually has superior predictive performance (Zimmerman and Zimmerman, 1991; Zhang and Zimmerman, 2007). However, repeated estimation by maximizing the likelihood function greatly increases the computational time. On the other hand, if these parameters are not re-estimated and the same estimates are used in all neighborhoods as in the ArcGIS software, the local kriging may not achieve the best predictive performance it could. The other reason for estimation within each neighborhood is that the underlying process may not be stationary globally but is more likely to be approximately stationary within a neighborhood. By estimating the parameters within a neighborhood under the assumption of stationarity, we could arrive at prediction results that are better than those obtained using all data.

The purpose of this paper is to further simplify the computation for parameter estimation for massive spatial data without scarifying predictive performance. For this purpose, we first develop an efficient computing algorithm for calculating different predictive scores that are measures of the predictive performance. We introduce an estimation method that minimizes some predictive scores and apply the method to the lower rank model and the local kriging. We illustrate this method through a real dataset on the US precipitation.

2 Kriging and predictive scores

Our aim is to develop an efficient algorithm for both parameter estimation and the evaluation of predictive performance. For now let us assume that the underlying process is stationary with a covariogram $C(h) = \sigma^2 \rho(h/\phi)$ where $\rho$ is the correlation function and could depend on an additional parameter.

Empirical studies have shown that in most cases the parameter $\phi$ only slightly affects the interpolation and it is a function of $\sigma^2$ and $\phi$ that primarily affects the interpolation. Theoretical results explain why $\phi$ does not significantly affect the interpolation when the covariogram is in the Matérn class. Indeed, for the
Matérn covariogram

\[ C(h) = \frac{\sigma^2 (h/\phi)^\nu}{2^{\nu - 1} \Gamma(\nu)} K_\nu(h/\phi), \quad h \geq 0, \]  

(3)

by employing the infill asymptotic theory (e.g., Stein, 1999), Zhang (2004) showed theoretically and illustrated numerically that it is the ratio \( \sigma^2 / \phi^{2\nu} \) that matters more to interpolation. Indeed, a misspecified Matérn model with the correct ratio yields an asymptotically optimal interpolator. Furthermore, we can fix \( \phi \) at an arbitrary value \( \phi_1 \) and estimate the variance \( \sigma^2 \) by maximizing the likelihood function. Zhang (2004) showed that \( \hat{\sigma}^2 / \phi_1^{2\nu} \) is a consistent estimator of \( \sigma^2 / \phi^{2\nu} \). Recently, Du, Zhang, and Mandrekar (2009) showed that \( \hat{\sigma}^2 / \phi_1^{2\nu} \) is asymptotically efficient. All these theoretical results imply that we can fix \( \phi \) and estimate other parameters and will not suffer any substantial loss of predictive performance. Zhang and Zimmerman (2007) used exactly this idea and proposed hybrid estimators by fixing \( \phi \) at the weighted least squares estimate. Their numerical results show that the hybrid estimators have a predictive performance that is comparable with that of the exact MLE but superior than that of the least squares method even for some non-Matérn covariograms for which the theoretical results are not yet available.

When the covariogram has a nugget effect \( \tau^2 \), i.e., \( C(h) = \sigma^2 \rho(h/\phi) + \tau^2 1_{\{h=0\}} \) where \( 1_{\{h=0\}} \) is the indicator function and take the value 1 if \( h = 0 \) and 0 otherwise, the similar conclusion still holds, i.e., what matters more to prediction is the nugget effect \( \tau^2 \) and the ratio \( \sigma^2 / \phi^{2\nu} \) for the Matérn model. The hybrid estimator works well in the presence of a nugget effect as shown by the results of Zhang and Zimmerman (2007).

What is useful though would be an efficient algorithm that evaluates the predictive performance. Predictive performance is usually measured by some predictive scores. Gneiting and Raftery (2007) and Gneiting et al. (2007) discussed several predictive scores that are reviewed below, all of which are based on the drop-one prediction.

2.1 Predictive score

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 \( \sigma_{-i}^2 \) be the corresponding drop-one prediction variance. Write

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

(4)

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

\[ RMSE = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y(s_i) - \hat{Y}_{-i}(s_i) \right)^2 \right\}^{1/2}. \]
The RMSE is commonly used in cross-validation. Another score that incorporates the prediction variance is the logarithmic score

\[ \text{LogS} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{2} \log(2\pi \hat{\sigma}_{-i}(s_i)^2) + \frac{1}{2} z_i^2 \right) \]  

(5)

where \( z_i \) is defined in (4).

The following two scores utilize the explicit predictive distribution. The Brier score is defined as (Brier, 1950)

\[ BS(y) = \frac{1}{n} \sum_{i=1}^{n} \left( F_i(y) - 1\{Y(s_i) \leq y\} \right)^2. \]  

(6)

where \( F_i(y) \) is the predictive cdf of \( Y(s_i) \), namely, \( F_i(y) = P(Y(s_i) \leq y | Y(s_j), j \neq i) \), and \( 1\{Y(s_i) \leq y\} \) is the indicator function and takes the value 1 if \( Y(s_i) \leq y \) and 0 otherwise. When the underlying process is Gaussian, the predictive distribution for \( Y_i(s_i) \) is \( N(\hat{Y}_{-i}(s_i), \sigma_{-i}(s_i)^2) \). The continuous ranked probability score is the integration of \( BS(y) \) over the \( y \in (-\infty, \infty) \):

\[ \text{CRPS} = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} (F_i(y) - 1\{Y(s_i) \leq y\})^2 dy = \int_{-\infty}^{\infty} BS(y) dy. \]  

(7)

If \( Y = (Y(s_1), \ldots, Y(s_n))^\prime \) has a multivariate normal distribution, CRPS can be calculated based on the \( z_i \)'s in (4) (see, Gneiting et al., 2007):

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

(8)

All these predictive scores require the computation of the drop-one predictions \( \hat{Y}_{-i}(s_i) \) and prediction variance \( \sigma_{-i}(s_i)^2 \). We could of course calculate iteratively all these drop-one predictions but this would be computationally inefficient. Therefore for large spatial data, we should avoid the iterative calculation of the drop-one prediction. Next we show how they can be calculated efficiently without iteration. Then it is feasible to minimize a predictive score to obtain parameter estimates for superior predictive performance.

### 2.2 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 \( Y = (Y_1, \ldots, Y_n)^\prime \) and assume \( \text{Var}(Y) = V \). 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

\[ Y_i - \hat{Y}_{-i} = \frac{1}{q_{ii}} \sum_{j=1}^{n} q_{ij} Y_j \]  

(9)

\[ \sigma_{-i}^2 = E(Y_i - \hat{Y}_{-i})^2 = 1/q_{ii}. \]  

(10)
If we let $\Lambda = \text{diag}(q_{ii}, i = 1, \cdots, n)$, we have

$$MSE = \frac{1}{n} \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n} \| \Lambda^{-1} V^{-1} Y \|^2,$$  \hspace{1cm} (11)

and the $z_i$'s in (4) can be given in vector form

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

Let us apply these results to the low rank model (1) for which $V$ is usually a large matrix and its inverse is computed by the Woodbury formula. We assume that the latent process $Z(s)$ has a constant variance $\sigma^2$ so that $\text{Var}(Z) = \sigma^2 R_Z$. Write $w = \tau^2 / \sigma^2$, and

$$\Gamma = A(wR_z^{-1} + A' A)^{-1} A' = (wR_z + (AR_z)'(AR_z))^{-1}(AR_z)'.$$ \hspace{1cm} (12)

Then

$$V^{-1} = \tau^{-2} (I_n - \Gamma), \text{ and }$$

$$MSE = \frac{1}{n} \| (I_n - \gamma)^{-1}(I_n - \Gamma)Y \|^2$$ \hspace{1cm} (13)

where $\gamma$ is a diagonal matrix consisting of the diagonal elements of $\Gamma$.

When the means are not all zeros, we center the variable by replacing $Y$ by $Y - EY$. For example, the MSE given by (13) now becomes

$$MSE = \frac{1}{n} \| (I_n - \gamma)^{-1}(I_n - \Gamma)(Y - EY) \|^2.$$ \hspace{1cm} (14)

In reality, the means have to be estimated. Let us consider the simple case that the mean is a constant as in the ordinary kriging. In this case, the best linear unbiased estimator of $\mu$ is

$$\hat{\mu} = \frac{1}{V^{-1} Y}.\frac{1}{V^{-1} 1}.$$

Then the drop-one prediction error is now

$$Y_i - \hat{Y}_{-i} = (1/q_{ii}) \sum_{j=1}^{n} q_{ij}(Y_j - \hat{\mu}).$$

However, when the mean is estimated, it is no longer true that $E(Y_i - \hat{Y}_{-i})^2 = 1/q_{ii}$. Indeed, if we let $q_i = (q_{ij}, j = 1, \cdots, n)'$, we can write

$$Y_i - \hat{Y}_{-i} = \frac{1}{q_{ii}} q_i' \left( I_n - \frac{11'V^{-1}}{1'V^{-1} 1} \right) Y.\frac{1}{1'V^{-1} 1}.$$ \hspace{1cm} (15)

Simple calculation yields that

$$\sigma^2_{-i} = E(Y_i - \hat{Y}_{-i})^2 = \frac{1}{q_{ii}} \left( q_{ii} - \frac{(q_i'1)^2}{1'V^{-1} 1} \right).$$ \hspace{1cm} (16)
The standardized prediction error in (4) now equals

$$z_i = \left( q_{ii} - \frac{(q'1)^2}{1'V^{-1}1} \right)^{1/2} q_i \left( I_n - \frac{11'V^{-1}}{1'V^{-1}1} \right) Y. \quad (17)$$

We note that kriging is often carried out in terms of the semivariogram. Dubrule (1983) presented an algorithm for calculating the drop-one predictions using the semivariogram, which avoids repeated calculation of $(n-1) \times (n-1)$ matrices. However, for the low rank model (1), it is more convenient to use covariogram than semivariogram.

3 Estimation to optimize predictive scores

In this section, we develop an algorithm for estimation that minimizes some predictive scores. We will estimate the mean parameters by the best linear unbiased estimators, which is identical to the MLE if the joint distribution is normal, and estimate the parameters in the correlation function by minimizing the MSE, and variance by minimizing the logarithmic score. In the next two subsections, we will consider the algorithm for the predictive process model, which is capable of handling very large datasets, and also the local kriging.

3.1 Predictive process models

We assume that the underlying process $Y(s)$ follows the predictive process model as defined in Banerjee et al. (2008): 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 $u_1, \cdots, u_m$ be the $m$ known locations, also called “knots” by Banerjee et al. (2008). The predictive process model is

$$Y(s) = \mu + k(s)' R^{-1} W + \epsilon(s), \quad (18)$$

where $W = (W(u_1), \cdots, W(u_m))'$, $R$ is the correlation matrix of $W$ whose $(i, j)$th element is $\rho(u_i - u_j, \phi)$, and $k(s) = Corr(W, W(s))$ is the correlation of $W$ and $W(s)$ whose $j$th element is $\rho(s - u_j, \phi)$. If $Y = (Y(s_1), \cdots, Y(s_n))'$ represents the observations at $n$ locations, we can write

$$Y = \mu 1 + AW + \epsilon,$$

for $A = KR^{-1}$ and $K = (k(s_1), \cdots, k(s_n))'$ is an $n \times m$ matrix (whose $(i, j)$th element is $\rho(s_i - u_j, \phi)$). The $\Gamma$ matrix defined in (12) is now, because $AR = K$,

$$\Gamma = K(wR + K'K)^{-1} K'. \quad (19)$$

The best linear unbiased estimator for $\mu$ is

$$\hat{\mu} = \frac{1'(I_n - \Gamma)Y}{1'(I_n - \Gamma)1}. \quad (20)$$
Note $\hat{\mu}$ depends on $w$ and $\phi$, which are estimated by minimizing the MSE (14), i.e.,

$$
(\hat{w}, \hat{\phi}) = \text{ArgMin}_{w, \phi} \|(I_n - \gamma)^{-1}(I_n - \Gamma)(Y - \hat{\mu}1)\|^2.
$$

(21)

We estimate the nugget effect $\tau^2$ by minimizing the logarithmic score (5). Let $\gamma_{ij}$ denote the $(i, j)$th element of $\Gamma$ and $\gamma_i = (\gamma_{ij}, j = 1, \cdots, n)'$. Since $V^{-1} = \tau^{-2}(I_n - \Gamma)$, from (16),

$$
\sigma^2_{-i} = \tau^2 \left( \frac{1}{1 - \gamma_{ii}} - \frac{1}{(1 - \gamma_{ii})^2} \frac{(1 - \gamma_i')^2}{n - 1} \frac{1}{1' \Gamma 1} \right) = \alpha_i^2 \tau^2.
$$

(22)

The logarithmic score (5) equals

$$
\frac{1}{2} \log(2\pi) + \frac{1}{2n} \sum_{i=1}^{n} \left( \log(\tau^2 \alpha_i^2) + \frac{(Y_i - \hat{Y}_i)^2}{\alpha_i^2 \tau^2} \right)
$$

Minimizing $\text{LogS}$ leads to the estimator

$$
\hat{\tau}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{\alpha_i^2}.
$$

(23)

This is clearly an unbiased estimator for $\tau^2$. We note that the vector of $\alpha_i^2, i = 1, \cdots$ can be calculated easily in R by the following, where $G$ is the $\Gamma$ matrix:

$$
\frac{1}{1 - \text{diag}(G)} - (1 - \text{rowSums}(G))^2 / ((1 - \text{diag}(G))^2 * \text{n-sum}(G))
$$

One may also estimate $\tau^2$ by the following estimator, which is obtained by maximizing the likelihood function when all other parameters are fixed at their estimates:

$$
\hat{\tau}^2_{mle} = \frac{1}{n} (Y - \hat{\mu}1)'(I_n - \Gamma)(Y - \hat{\mu}1).
$$

(24)

This estimator is biased and leads to higher LogS score than $\hat{\tau}^2$ in (23). Since our primary objective is spatial prediction with lower predictive scores, we will use the estimator (23) in this work.

We note that the maximum likelihood estimators for all parameters are attainable for the predictive process model as will be discussed in the example of next section. However, the likelihood function can be quite flat at the maximum point which can cause some numerical optimization algorithms to stop working. Although the predictive scores can be flat too, the flatness tells us that the prediction is not sensitive to the parameter values and we therefore could use any parameter value in the region where the scores are flat. This is an advantage of the finding estimates through the predictive scores. These estimates require less computation than the maximum likelihood estimates because we do not need to evaluate the determinant.

### 3.2 Local kriging

As in the previous section, we assume the correlation function $\rho$ depends on a parameter $\phi$ which could be a vector. We will get a rough estimate for the parameter and treat the the parameter as being known in
the local kriging. Such a rough estimate can be obtained by the least squares methods using all data or the maximum likelihood method using a subset of data. The parameters to be estimated in each neighborhood are therefore \( \mu, \sigma^2 \) and \( \tau^2 \). Here we consider the case that in each neighborhood \( EY(s) = \mu \) is a constant and the process is stationary. To predict at a location \( s_0 \), we first find the nearest \( m \) locations and let \( Y_0 \) denote the observations at the \( m \) locations. Write the covariance of the \( m \) observed variables as \( V = \sigma^2 R + \tau^2 I \). We estimate \( w = \tau^2 / (\tau^2 + \sigma^2) \) by minimizing the sum of squared errors:

\[
\hat{w} = \text{ArgMin}_w \sum_{s_i \in O_i} (Y(s_i) - \hat{Y}_{-i}(s_i))^2 = \text{ArgMin}_w \| D^{-1}(wI + (1 + w)R)^{-1}(Y_0 - \mu(w)1) \|^2
\]

(25)

where \( O_i \) is the set of \( m \) nearest locations, \( D \) is a diagonal matrix consisting of the diagonal elements of \( (wI + (1 + w)R)^{-1} \), \( 1 \) is the vector of all 1’s, and \( \mu = \mu(w) \) is the best linear unbiased estimator of \( \mu \), i.e.,

\[
\mu(w) = \frac{Y'(wI + (1 + w)R)^{-1}Y}{Y'(wI + (1 + w)R)^{-1}1}.
\]

Once \( \hat{w} \) is obtained, \( \mu \) is estimated by \( \mu(\hat{w}) \) and the sill \( \sigma^2 + \tau^2 \) is estimated by

\[
\frac{1}{m} \sum_{s_i \in O_i} \gamma_{ii}(Y(s_i) - \hat{Y}_{-i}(s_i))^2
\]

where \( \gamma_{ii} \) is the \( i \)th diagonal element of the matrix \( (wI + (1 + w)R)^{-1} \).

### 4 An example

In this section, we illustrate the methods by applying them to a set of US precipitation data which consists of the monthly total precipitation record in the conterminous US for April 1948 consisting of 5,906 stations. The standardized square root values, known as anomalies, are analyzed here. The anomaly data are available in the \( \text{R} \) package `spam`. The weather stations are shown in Figure 1 and the empirical semivariogram is plotted in Figure 2. We will use an exponential covariogram with a nugget effect and assume a constant mean.

We first employ the predictive process model with \( m = 760 \) knots. The knots are marked in Figure 1 by “+”. The number of knots is usually kept to be below 1,000 as in Cressie and Johannesson (2008) and Banerjee et al. (2008). The number of knots we use here is larger than theirs (396 in Cressie and Johannesson and 144 in Banerjee et al.) though the sample size in this example is smaller than theirs (173,405 and 9,500 in Cressie and Johannesson and Banerjee et al., respectively). The evenly spaced knots are desirable in practice from the prediction point of view though it may not be the optimal design. Although it is possible in principle to find the best design, one may doubt about the worth of possible gains in prediction precision relative to the computational effort. Furthermore, Banerjee et al. (2008) compared knots on a regular lattice with knots on a regular knots plus some infill knots which are closer together and found little difference between the prediction results given by the two sets of knots. Finally, we note that the choice of knots is not the classical
sampling design problem because the best choice of the knots now depend on the sampling locations. Since the aim of this section is to illustrate different methods and compare their predictive performances, we will not seek optimal knot selection.

We note that the log likelihood function can be calculated on an ordinary computer because we only need to calculate the determinants of $m \times m$ matrices and a quadratic form involving an $m \times m$ matrix. As mentioned earlier, the Woodbury formula implies

$$V^{-1} = \tau^{-2}(I_n - K(wR + K'K)^{-1}K'),$$

where $K_{n \times m}$ is defined in section 3.1. The log likelihood function is

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\tau^2) - \frac{1}{2}|I_n - K(wR + K'K)^{-1}K'|$$

$$-\frac{1}{2}(Y - \mu)^T(\tau^{-2}(I_n - K(wR + K'K)^{-1}K')(Y - \mu)).$$

The profile log likelihood for $\phi$ and $w$ are defined as

$$PLL(\phi, w) = \max_{\mu, \tau^2} \log L.$$ 

For any $\phi$ and $w$, $\log L$ is maximized by

$$\mu(\phi, w) = \frac{1}{Y^T(I_n - K(wR + K'K)^{-1}K')}Y$$

$$\tau^2(\phi, w) = \frac{1}{n}(Y - \mu\mathbf{1})^T(I_n - K(wR + K'K)^{-1}K')(Y - \mu\mathbf{1}) \text{ for } \mu = \mu(\phi, w).$$
Therefore,
\[
P_{LL}(\phi, w) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tau^2) - \frac{1}{2} \log(|I_n - K(wR + K'K)^{-1}K'|) - n/2.
\]
It is well known that 
\[|I_n - UBV| = |B^{-1} - VU||B|\] provided the dimensions of the matrices are compatible with each other (e.g., Harville, 1997). Then
\[
|I_n - K(wR + K'K)^{-1}K'| = |wR||wR + K'K|^{-1},
\]
and
\[
P_{LL}(\phi, w) = -\frac{n}{2} (1 + \log(2\pi)) - \frac{n}{2} \log(\tau^2) - \frac{m}{2} \log(w) - \log(|R|) + \frac{1}{2} \log(|wR + K'K|).
\]
We encountered some convergence problem when maximizing this bivariate function because $P_{LL}(\phi, w)$ is quite flat near the maximum. We therefore worked on the profile log likelihood function for $\phi$—that is $\max_w P_{LL}(\phi, w)$. In Figure 3, we plot the profile log likelihood function for $\phi$. The profile function has a minimum at $\phi = 425$ though it is quite flat when $\phi > 425$. The maximum likelihood estimates are therefore $\hat{\phi}_{mle} = 425, \hat{w}_{mle} = 0.0588, \hat{\tau}^2_{mle} = 0.0865, \hat{\mu}_{mle} = 0.18731$. The corresponding predictive scores $RMSE, LogS$ and $CRPS$ are $0.30888, 0.24227$ and $0.29280$, respectively.

Next, we examine the predictive scores for different estimates of $\phi$. For each $\phi = 50, 100, 150, 200, 250, 300, 350, 400, 700, 1000, 1500, 2000$, we find the $w$ that minimizes $RMSE$, and calculate the estimates $\hat{\tau}^2$ and $\hat{\mu}$ given by (23) and (20), respectively. We provide in Table 1 the predictive scores corresponding each $\phi$ and plot the predictive scores against $\phi$ in Figure 4. We see that the predictive scores change very little from $\phi = 450$ to $\phi = 2000$ though the predictive scores decrease slightly as $\phi$ increases. Therefore, from the
Figure 3: The profile log likelihood for $\phi$.

Figure 4: Predictive scores against $\phi$: RMSE (solid line), LogS (dotted line) and CRPS (dashed line)
Table 1: Estimates and predictive scores for each fixed $\phi$ for the anomaly data. The last column $p$ is the percentage of the drop-one 95% prediction intervals containing the observations.

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>$\hat{w}$</th>
<th>$\hat{\tau}^2$</th>
<th>RMSE</th>
<th>LogS</th>
<th>CRPS</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.16179</td>
<td>0.10209</td>
<td>0.33803</td>
<td>0.33621</td>
<td>0.31834</td>
<td>0.05689</td>
</tr>
<tr>
<td>2</td>
<td>0.15442</td>
<td>0.08791</td>
<td>0.31323</td>
<td>0.26115</td>
<td>0.29672</td>
<td>0.05621</td>
</tr>
<tr>
<td>3</td>
<td>0.11937</td>
<td>0.08637</td>
<td>0.31057</td>
<td>0.25226</td>
<td>0.29498</td>
<td>0.05571</td>
</tr>
<tr>
<td>4</td>
<td>0.09427</td>
<td>0.08584</td>
<td>0.30968</td>
<td>0.24926</td>
<td>0.29444</td>
<td>0.05604</td>
</tr>
<tr>
<td>5</td>
<td>0.07723</td>
<td>0.08559</td>
<td>0.30926</td>
<td>0.24784</td>
<td>0.29418</td>
<td>0.05621</td>
</tr>
<tr>
<td>6</td>
<td>0.06521</td>
<td>0.08546</td>
<td>0.30902</td>
<td>0.24705</td>
<td>0.29402</td>
<td>0.05604</td>
</tr>
<tr>
<td>7</td>
<td>0.05634</td>
<td>0.08537</td>
<td>0.30888</td>
<td>0.24657</td>
<td>0.29391</td>
<td>0.05588</td>
</tr>
<tr>
<td>8</td>
<td>0.04954</td>
<td>0.08531</td>
<td>0.30878</td>
<td>0.24624</td>
<td>0.29384</td>
<td>0.05588</td>
</tr>
<tr>
<td>9</td>
<td>0.02863</td>
<td>0.08519</td>
<td>0.30856</td>
<td>0.24552</td>
<td>0.29366</td>
<td>0.05604</td>
</tr>
<tr>
<td>10</td>
<td>0.02010</td>
<td>0.08515</td>
<td>0.30851</td>
<td>0.24533</td>
<td>0.29361</td>
<td>0.05604</td>
</tr>
<tr>
<td>11</td>
<td>0.01343</td>
<td>0.08514</td>
<td>0.30848</td>
<td>0.24523</td>
<td>0.29358</td>
<td>0.05604</td>
</tr>
<tr>
<td>12</td>
<td>0.01005</td>
<td>0.08513</td>
<td>0.30847</td>
<td>0.24520</td>
<td>0.29357</td>
<td>0.05604</td>
</tr>
</tbody>
</table>

prediction point of view, we can fix $\phi$ at any value greater than 450 and the prediction performance will be very similar according to the prediction scores. This suggests that the prediction results are not much affected if we overestimate $\phi$. The empirical semivariogram provides a rough estimate for an upper bound of the range parameter $\phi$. For example, from Figure 2, it is reasonable to believe that $\phi$ is less than 600. Therefore we can fix $\phi$ at 600 or larger and estimate the other parameters. This results in faster algorithms and avoids possible convergence problems in numerical optimization.

The estimate of the nugget effect $\tau^2$ is hardly affected by the overestimates of $\phi$, but the estimate of $w$ is significantly affected by the choice of $\phi$. This is because the ratio $\sigma^2/\phi$ is a microergodic parameter that can be estimated well (Zhang and Zimmerman, 2005). One can verify that the estimates of $\sigma^2/\phi(= \hat{\tau}^2/(\hat{\phi}\hat{w}))$, vary very little when $\phi$ increases from 450 to 2000. From the predictive point of view, we only need to estimate the microergodic parameters well. The estimates $\tau^2$ and $\sigma^2/\phi$ support our earlier finding that there is very little loss in predictive performance when $\phi$ is overestimated.

To further substantiate this finding, we simulate from the predictive process model (18) at the same 5906 locations in Figure 1 and the same 760 knots. The covariogram is the isotropic exponential with variance $\sigma^2 = 1$. Other model parameters are $\mu = 0$, $\tau^2 = 0.08$ and $\phi = 425$. These parameters mimic the estimates we just obtained. For each $\phi$, we obtain estimates of $w$, $\tau^2$ by minimizing the predictive scores. The estimates and predictive scores are provided in Table 2. We find that even when $\phi$ is severely overestimated,
Next, we simulate the predictive process model (18) with a spherical covariogram having a range $\phi = 425$. The sampling locations and knots are the same and the model parameters are $\sigma^2 = 1$, $\mu = 0$, and $\tau^2 = 0.08$. The estimates and predictive scores are provided in Table 3. Similar conclusions hold although there is no infill asymptotic theoretical results established for the spherical covariogram.

We now revisit the anomaly data and apply local kriging with the exponential covariogram with a nugget effect. The range parameter $\phi$ is fixed at 425 in every neighborhood and the size of neighborhood is fixed at $m = 500$. Within each neighborhood, we minimize the RMSE and the LogS score to estimate the nugget effect and the partial sill. These estimates vary from neighborhood to neighborhood and will not be reported. Instead, we will report the predictive scores where the drop-one prediction $\hat{Y}(s_i)$ is calculated using only the nearest $m$ locations. The RMSE, LogS and CRPS are 0.26702, 0.06453 and $-1.40194$, respectively. These predictive scores are lower than those in the predictive process model, which means that this local kriging resulted in better prediction results. This could be due to the fact that the predictive process model is an approximation to the underlying process or due to non-stationarity.

We increase the number of neighborhood from $m = 500$ to $m = 800$ and obtain RMSE, LogS and CRPS
Table 3: Estimates and predictive scores for each fixed $\phi$ for the simulated data with the spherical co-
variogram. The last column $p$ is the percentage of the drop-one 95% prediction intervals containing the 
observations.

to be 0.26736, 0.06479 and −1.38466. These predictive scores are similar to the previous ones when $m = 500$. 
Therefore, for this dataset, 500 neighboring locations are sufficiently large for local kriging. This example 
indicates that local kriging with hybrid estimation can lead to even superior prediction results although it 
avoids the difficulties of handing a large covariance matrix.

Acknowledgement

The research is supported by the US National Science Foundation grants DMS-0405782 and DMS-
0706835. The authors thank the Editor and a referee for constructive comments.

References

large spatial data sets. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 70(4), 
825–848.

1–3.


*Positive definite functions: From Schoenberg to space-time challenges*, pages 181–196.


*Biometrika*, 92, 921–936.
