Statistica Sinica **25** (2015), 135-149
doi:http://dx.doi.org/10.5705/ss.2013.205w

VARIANCE ESTIMATION AND KRIGING PREDICTION FOR A CLASS OF NON-STATIONARY SPATIAL MODELS

Shu Yang and Zhengyuan Zhu

Iowa State University

Abstract: This paper discusses the estimation and plug-in kriging prediction of a non-stationary spatial process assuming a smoothly varying variance function with an additive independent measurement error. A difference-based kernel smoothing estimator of the variance function and a modified likelihood estimator of the measurement error variance are used for parameter estimation. Asymptotic properties of these estimators and the plug-in kriging predictor are established. A simulation study is presented to test our estimation-prediction procedure. Our kriging predictor is shown to perform better than the spatial adaptive local polynomial regression estimator proposed by Fan and Gijbels (1995) when the measurement error is small.

Key words and phrases: Bandwidth selection, heteroscedasticity, K-fold cross-validation, local polynomial regression, rates of convergence, variance function.

1. Introduction

Stationary spatial models play an important role in such areas as mining, environmental monitoring, meteorology, soil science, economics, and epidemiology. It has long been recognized that the assumption of stationarity is often violated, and the problem is more pronounced when one has high resolution spatial data over large spatial domain. With the influx of such large spatial data in recent years, there has been a substantial amount of research directed at modeling and estimating non-stationarity in spatial data. Examples of non-stationary models include process deformation models (Guttorp, Sampson, and Newman (1992); Bornn, Shaddick, and Zidek (2012)), kernel convolution models (Higdon (1998); Paciorek and Schervish (2006)), spectral approach (Fuentes (2002a,b); Porcu, Gregori, and Mateu (2009)), a wavelet approach (Nychka, Wikle, and Royle (2002); Matsuo et al. (2011)), and many more. Examples of estimation methods include moment-based methods (Nychka and Saltzman (1998); Nychka, Wikle, and Royle (2002); likelihood-based methods (Anderes and Stein (2011)), and Bayesian methods (Higdon, Swall, and Kern (1999); Damian, Sampson, and Guttorp (2001); Schmidt and O'Hagan (2003); Sanso, Schmidt, and Nobre (2005); Schmidt, Schelten, and Roth (2011)). After adopting a non-stationary spatial

model, kriging can be used to make predictions at locations where measurements of the process are not available.

Alternatively, one can model non-stationary spatial data using nonparametric methods, and make spatial predictions using smoothing with spatially adaptive bandwidth. For kernel regression, Fan and Gijbels (1996) developed a method to estimate smoothly varying bandwidth, and discussed local polynomial models with adaptive window widths. For smoothing splines, Cummins, Filloon, and Nychka (2001) developed local generalized cross validation (GCV) to fit spatially adaptive smoothing splines, and Luo and Wahba (1997) proposed a hybrid adaptive spline approach. More recently, Pintore, Speckman, and Holmes (2006) treated spatially adaptive smoothing splines as a function minimization problem.

When the process is stationary in space, it is well known that there is a close connection between kriging and nonparametric regression methods. Wahba (1985) and Stein (1990, 1993) showed kriging under certain simple stationary models is equivalent to smoothing splines, and the restricted maximum likelihood (REML) estimator of the smoothing parameter is more efficient than the GCV estimator if the underlying model is correctly specified. However, a similar connection between kriging under non-stationary models and spatially adaptive nonparametric regression methods has not been established so far.

In this paper, we study this connection under the simple model

$$Z_i = Z(x_i) = \sigma(x_i)W(x_i) + \epsilon_i, \tag{1.1}$$

$i = 1, \ldots, n$, where $x_i = i/n \in [0, 1]$, $\sigma(x)$ is a smoothly varying function, and $W(x)$ is a Brownian motion. Here $\sigma(x)W(x)$ accounts for the heteroscedasticity and spatial correlation in the data. The ϵ_i's are independent normal errors with zero mean and variance σ_ϵ^2, representing measurement error. This model is a generalization of one in Stein (1990) that assumed that $\sigma(x) = \sigma$ is a constant. We consider kriging with estimated parameters under this non-stationary model. One objective is to develop an estimation and prediction method for this non-stationary model, and to derive corresponding asymptotic results, with the goal of comparing them to those from spatially adaptive non-parametric methods.

To estimate the variance function $\sigma^2(x)$, we consider a difference-based kernel smoothing estimator, which is essentially a Method-of-Moment approach. Similar techniques had been investigated by many authors for variance function estimation in heteroscedastic nonparametric regression models. See for example, Von Neumann et al. (1941), Gasser, and Sroka, and Jennen-Steinmetz (1986), Hall, Kay, and Titterington (1990, 1991), Brown and Levine (2007), Klipple and Eubank (2007), Cai and Wang (2008), Cai, Levine, and Wang (2009), Duran, Hardle, and Osipenko (2012). In the context of non-parametric regression, the

motivation for taking the differences is to eliminate the effect of the mean function and turn the problem of estimating the variance function in the model into a conventional regression problem of estimating the mean function. We draw heavily on Brown and Levine (2007) to develop the estimation method and derive asymptotic results. The novelty here is that we assume a model in which the observations are spatially dependent. A kernel smoothing technique is applied to squared differences to obtain the variance function estimator. To estimate σ_ϵ^2, a modified likelihood estimator is proposed, similar to the profile likelihood estimator except that when profiling the variance function $\sigma^2(x)$ we use the difference-based kernel smoothing estimator instead of the maximum likelihood estimator. The estimator of σ_ϵ^2 is then obtained by maximizing the modified likelihood function.

We derive the asymptotic mean squared error bound of the variance function estimator and establish its asymptotic normality. The asymptotic bias of the plug-in kriging predictor is also obtained. Our theoretical results indicate that both the kernel smoothing estimator of the variance function $\sigma^2(x)$ and the modified likelihood function of σ_ϵ^2 are consistent with small measurement error. The convergence rate deteriorates as the variance of measurement error increases, and when measurement error variance is too large, variance estimation is no longer consistent. This is seen in our simulation results, where we compare the kriging prediction with estimated parameters with a spatially adaptive local polynomial regression estimator (Fan and Gijbels (1995)). The kirging predictor out-performs the local polynomial estimator when measurement error is small, and under-performs it when the measurement error is large.

The rest of the paper is organized as follows. Section 2 describes the difference-based kernel estimator of the variance function, the modified likelihood estimator of the measurement error variance, and the plug-in kriging predictor with the unknown parameters replaced by their estimates. A bandwidth selection procedure is also included. Section 3 presents the asymptotic mean squared error bound of the variance function estimator and the asymptotic bias of the plug-in kriging predictor. In Section 4 we provide a limited-scope simulation study to show the finite sample performance of our estimation procedure. Discussion is in Section 5, and proofs can be found in the supplementary document online.

2. Methodology

2.1. Difference-based kernel estimator

Our estimation method for the variance function is similar to that of Brown and Levine (2007) for estimating the variance function in a nonparametric regression model. They use the difference squares of observations, transforming variance function estimation into mean function estimation, which is easier to

handle. The estimation procedure has two steps: take the square of the first-order differences of the observations, apply local polynomial regression estimation with squared differences to obtain a smoothed estimator of $\sigma^2(x)$.

Let $D_h(Z(x)) = Z(x+h) - Z(x)$ and $D_{h,i} = Z(x_i+h) - Z(x_i)$. For a Brownian motion W, $Cov(W(x_i), W(x_i+h)) = x_i$ for $h \geq 0$. Under some regularity conditions

$$E(D_{h,i}^2) = \sigma^2(x_i)h + \{\sigma^{(1)2}(x_i)x_i + \sigma^{2(1)}(x_i)\}h^2 + 2\sigma_\epsilon^2 + o(h^2),$$

where the notation $f^{(k)}(\cdot)$ denotes the k-th derivative of $f(\cdot)$. We can write

$$E(D_{h,i}^2) = \sigma^2(x_i)h + 2\sigma_\epsilon^2 + o(h).$$

$\sigma^2(x_i)$ is what we wish to estimate, σ_ϵ^2 is the measurement error variance, and $o(h)$ is a higher order bias term caused by heteroscedasticity. If variances at different locations are constant, the higher order bias term is zero. The correlation of the differences is negligible. Here except for successive differences which share a observation at the same location. Thus $\sigma(x_{i+1})$ at x_i,

$$\begin{aligned} Z(x_{i+1}) - Z(x_i) = {} & \sigma(x_i)(W(x_{i+1}) - W(x_i)) \\ & + \{\sigma^{(1)}(x_i)h + o(h)\}W(x_{i+1}) + \epsilon_{i+1} - \epsilon_i. \end{aligned}$$

And, due to independent increments, for $j - i > 1$,

$$Cov(D_{h,i}, D_{h,j}) = \sigma^{(1)}(x_i)\sigma^{(1)}(x_j)x_{i+1}h^2 + o(h^2). \tag{2.1}$$

A number of nonparametric regression procedures for estimating the mean function can be applied to estimate the variance function. Here we consider a local polynomial regression estimator. That automatically adjusts boundary effects, preserving the asymptotic order of the bias (Fan and Gijbels (1996)). The local polynomial regression estimator $\hat{D}_{h,\lambda}^2(x)$ of $D^2(x) = \sigma^2(x)h + 2\sigma_\epsilon^2$ based on $D_{h,i}^2$ is

$$\hat{D}_{h,\lambda}^2(x) = \hat{a}_0, \text{ where}$$

$$\begin{aligned} & (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p) \\ & = \arg\min_{a_0, a_1, \ldots, a_p} \sum_{i=1}^{n-1} \left[D_{h,i}^2 - a_0 - a_1(x - x_i) - \cdots - a_p(x - x_i)^p\right]^2 K(\frac{x - x_i}{\lambda}), \end{aligned}$$

with $K(\cdot)$ the kernel function and λ the bandwidth.

Definition 1. $K(\cdot)$ is a kernel function of order $p+1$ if $K(x) \geq 0$ with support $[-1, 1]$ satisfies $\int_{-1}^1 K(x)dx = 1$, and $\int_{-1}^1 K^2(x)dx < \infty$; $\int_{-1}^1 K(x)x^j dx = 0$, for $j = 1, 2, \ldots, p$; $\int_{-1}^1 K(x)x^{p+1}dx > 0$.

The local polynomial regression estimator $\hat{D}^2_{h,\lambda}(x)$ can be expressed as a weighted average of $D^2_{h,i}$'s,

$$\hat{D}^2_{h,\lambda}(x) = \sum_{i=1}^{n-1} K_n(\frac{x - x_i}{\lambda}) D^2_{h,i},$$

where $K_n((x - x_i)/\lambda)$ are the kernel weights, satisfying the discrete moment conditions $\sum_{i=1}^{n-1} K_n((x - x_i)/\lambda) = 1$; $\sum_{i=1}^{n-1}(x - x_i)^j K_n((x - x_i)/\lambda) = 0$, for any $j = 1, \ldots, p$; $K_n((x - x_i)/\lambda) = 0$ for all $|x - x_i| > \lambda$.

The local polynomial regression estimator of $\sigma^2(x)$ is given by

$$\begin{aligned}\hat{\sigma}^2_\lambda(x; \sigma^2_\epsilon) &= \frac{(\hat{D}^2_{h,\lambda}(x) - 2\sigma^2_\epsilon)}{h} \\ &= \sum_{i=1}^{n-1} K_n(\frac{x - x_i}{\lambda})\Delta_i, \end{aligned} \tag{2.2}$$

where $\Delta_i = (D^2_{h,i} - 2\sigma^2_\epsilon)/h$.

2.2. Modified likelihood estimator of σ^2_ϵ

Note that $\hat{\sigma}^2_\lambda(x; \sigma^2_\epsilon)$ depends on σ^2_ϵ, which in general is unknown and needs to be estimated from the data. We consider a modified likelihood approach to estimate σ^2_ϵ, similar to profile likelihood estimation except that when profiling $\sigma^2(x)$ we use the kernel smoothing estimator instead of the maximum likelihood estimator. Take

$$P(\sigma^2_\epsilon) = l(\hat{\sigma}^2_\lambda(x; \sigma^2_\epsilon), \sigma^2_\epsilon; \mathbf{d}), \tag{2.3}$$

where $\mathbf{d} = (Z(x_2) - Z(x_1), Z(x_3) - Z(x_2), \ldots, Z(x_n) - Z(x_{n-1}))$ is the difference vector, and $l(\sigma^2(x), \sigma^2_\epsilon; \mathbf{d})$ is the log likelihood function of $\sigma^2(x)$ and σ^2_ϵ based on $\mathbf{d}$. Since the correlation of non-overlapping differences is negligible, the joint distribution of $\mathbf{d}$ can be approximated by a multivariate normal distribution with mean $\mathbf{0}$ and variance

$$\Sigma = \begin{pmatrix} \sigma^2(x_1)h + 2\sigma^2_\epsilon & -\sigma^2_\epsilon & \cdots & 0 \\ -\sigma^2_\epsilon & \sigma^2(x_2)h + 2\sigma^2_\epsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \sigma^2(x_{n-2})h + 2\sigma^2_\epsilon & -\sigma^2_\epsilon \\ 0 & \cdots & -\sigma^2_\epsilon & \sigma^2(x_{n-1})h + 2\sigma^2_\epsilon \end{pmatrix}.$$

As a result, we have

$$P(\sigma^2_\epsilon) = -\frac{1}{2}\log|\hat{\Sigma}(\sigma^2_\epsilon)| - \frac{1}{2}\mathbf{d}^T\{\hat{\Sigma}(\sigma^2_\epsilon)\}^{-1}\mathbf{d}, \tag{2.4}$$

where

$$\hat{\Sigma}(\sigma_\epsilon^2) = \begin{pmatrix} \hat{D}^2_{h,\lambda}(x_1) & -\sigma_\epsilon^2 & \cdots & 0 \\ -\sigma_\epsilon^2 & \hat{D}^2_{h,\lambda}(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \hat{D}^2_{h,\lambda}(x_{n-2}) & -\sigma_\epsilon^2 \\ 0 & \cdots & -\sigma_\epsilon^2 & \hat{D}^2_{h,\lambda}(x_{n-1}) \end{pmatrix},$$

with the diagonal elements $\sigma^2(x_i)h + 2\sigma_\epsilon^2$ in Σ replaced by the kernel smoothing estimator $\hat{D}^2_{h,\lambda}(x_i)$. The modified likelihood estimator of σ_ϵ^2 is obtained by maximizing (2.4).

Replacing σ_ϵ^2 in (2.2) by $\hat{\sigma}_\epsilon^2$, the kernel smoothing estimator $\hat{\sigma}^2_\lambda(x)$ is

$$\begin{aligned} \hat{\sigma}^2_\lambda(x) &= \hat{\sigma}^2_\lambda(x; \hat{\sigma}_\epsilon^2) \\ &= \sum_{i=1}^{n-1} K_n(\frac{x - x_i}{\lambda})\hat{\Delta}_i, \end{aligned} \tag{2.5}$$

where $\hat{\Delta}_i \equiv (1/|h|)(D^2_{h,i} - 2\hat{\sigma}_\epsilon^2)$. The impact of using $\hat{\sigma}_\epsilon$ rather than σ_ϵ^2 on the asymptotic behavior of $\hat{\sigma}^2_\lambda(x)$ will be discussed in Section 3.

2.3. Bandwidth selection

A kernel smoothing estimator requires a choice of bandwidth. Two popular methods here are the plug-in-type procedure such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) methods, and the data-driven procedure based on minimizing an estimator of the mean squared error (MSE) such as the cross validation (CV) method. We use the K-fold cross-validation approach suggested by Levine (2006). Since the sequence $\{D^2_{h,i}\}$ has a relatively small correlations, we expect the K-fold cross-validation to perform well.

Randomly divide $\{D^2_{h,i},\ i = 1, \ldots, n\}$ into K subsets; leave out one fold, say K_s, estimate the parameters using the remaining data K_{-s}; predict the omitted points in the leave-out fold. A good summary criterion is the mean of the squared prediction errors. Here we use the discrete mean and refer to it as cross-validated discrete mean squared error (CDMSE),

$$CDMSE(\lambda) = n^{-1} \sum_{i=1}^{n} (D^2_{h,i} - \hat{D}^2_{h,-s,i})^2,$$

where $\hat{D}^2_{h,-s,i} = \hat{D}^2_{h,-s}(x_i)$ for $i \in K_s$, $\hat{D}^2_{h,-s}(x)$ is the difference-based kernel smoothing estimator of $D^2(x)$ fitted to the remaining data K_{-s}, $\hat{D}^2_{h,-s}(x) = n_{-s}^{-1} \sum_{k \in K_{-s}} K_{n_{-s}}((x - x_k)/\lambda) D^2_{h,k}$, with n_{-s} being the sample size of K_{-s}. The cross-validation bandwidth is

$$\lambda_{CV} = \arg\min CDMSE(\lambda).$$

2.4. Kriging prediction

Consider the kriging prediction of the underlying process $f(x_0) = \sigma(x_0)W(x_0)$ for $x_0 \in [0,1]$ based on the observations $\mathbf{z} = (Z(x_1), \cdots, Z(x_n))$. For simplicity we suppress the dependence of $\hat{\sigma}(x)$ on λ. When the parameters are known, the best linear unbiased predictor of $f(x_0)$ is the conditional expectation of $f(x_0)$ conditional on $\mathbf{z}$,

$$\begin{aligned} p(x_0) &= Cov(f(x_0), \mathbf{z})^T \Sigma_{\mathbf{z}}^{-1} \mathbf{z} \\ &= \sigma(x_0) Cov(W(x_0), \mathbf{z})^T \Sigma_{\mathbf{z}}^{-1} \mathbf{z}, \end{aligned}$$

where $\Sigma_{\mathbf{z}}$ is the covariance matrix of $\mathbf{z}$. The plug-in kriging predictor replaces the unknown parameters $\sigma(x)$ and σ_ϵ^2 in $p(x_0)$ with the kernel smoothing estimator of $\sigma(x)$ and the modified likelihood estimator of σ_ϵ^2.

3. Theoretical Results

In this section, we establish the asymptotic properties of the variance function estimators and the plug-in kriging predictor. Proofs can be found in the on-line supplementary document.

We need some smoothness condition on $\sigma^2(x)$. We make the standard assumption (see for example, Brown and Levine (2007)) that $\sigma^2(x)$ belongs to Lipschitz classes $C_\beta^+(M)$ for $\beta > 0$ and $M > 0$.

Definition 2. The Lipschitz class $C_\beta(M) = \big\{g : \text{ for all } 0 \le x, y \le 1, k = 0, \ldots, \llcorner\beta\lrcorner - 1, |g^{(k)}(x)| \le M \text{ and } |g^{(\llcorner\beta\lrcorner)}(x) - g^{(\llcorner\beta\lrcorner)}(y)| \le M|x-y|^{\beta'}\big\}$, where $\llcorner\beta\lrcorner$ is the largest integer less than β and $\beta' = \beta - \llcorner\beta\lrcorner$.

Definition 3. $C_\beta^+(M) = \big\{g : \ g \in C_\beta(M) \text{ and } \exists\delta > 0, \text{ s.t. for all } 0 \le x \le 1, \ g(x) > \delta\big\}$.

Theorem 1. *In model* (1.1)*, assume* $\sigma^2(x)$ *belongs to the functional classes* C_β^+ *for* $\beta > 0$ *and the variance of measurement error* σ_ϵ^2 *is* $O(n^{-\alpha})$ *with* $\alpha > 1/2$*. The estimator* $\hat{\sigma}_\lambda^2(x; \sigma_\epsilon^2)$ *at* (2.2) *is consistent for* $\sigma^2(x)$ *for any* $x \in [0,1]$*, with bias* $O(\max(n^{-1}, \lambda^\beta))$ *and variance* $O((n\lambda)^{-1}\max(1, n^{2-2\alpha}))$*.*

When $\alpha \ge 1$*, the optimal bandwidth is* $\lambda = O(n^{-1/(1+2\beta)})$*, and the mean squared error is* $O(n^{-2\beta/(1+2\beta)})$ *. When* $1/2 < \alpha < 1$*, the optimal bandwidth is* $\lambda = O(n^{-(2\alpha-1)/(1+2\beta)})$*, and the mean squared error is* $O(n^{-(2\alpha-1)2\beta/(1+2\beta)})$*.*

Remark 1. Wang et al. (2008) derived the minimax rate of convergence for variance function estimation in a heterogeneous nonparametric regression model. They characterized explicitly how the smoothness of the unknown mean function

influences the rate of convergence of the variance estimator and showed that the minimax rate of convergence under both pointwise MSE and global MISE is $\max\{n^{-4\alpha}, n^{-2\beta/(2\beta+1)}\}$ if the mean function has α derivatives and the variance function has β derivatives. One goal is to establish asymptotic bounds of the bias and variance of the variance function estimator for non-stationary spatial processes and study how the magnitude of measurement error influences the variance function estimation. Here α differs from that of α in Wang et al. (2008). The optimal bandwidth and mean squared error can be obtained accordingly. For $\alpha \geq 1$ the rate of convergence of the variance function estimator is $n^{-2\beta/(2\beta+1)}$, which coincides with the minimax rate of the convergence of variance function estimator in heterogeneous nonparametric regression. For $1/2 < \alpha < 1$, the rate of the convergence of the variance function estimator depends on the variability of the measurement error, and deteriorates as $\alpha \to 1/2$. This is consistent with the intuition that, when the variability of the measurement error increases, the differences of observations are dominated by measurement error and therefore carry little information about the variance function under estimation. For $\alpha < 1/2$, the asymptotic theory for the difference-based kernel smoothing method breaks down and it is no longer possible to have consistent estimates of the variance function.

Theorem 2. *In model* (1.1)*, assume* $\sigma^2(x)$ *belongs to the functional classes* C_β^+ *for* $\beta > 0$ *and the variance of measurement error* σ_ϵ^2 *is* $O(n^{-\alpha})$ *with* $\alpha > 1/2$*. For* $\hat{\sigma}_\lambda^2(x; \sigma_\epsilon^2)$ *at* (2.2)*,* $\alpha \geq 1$*, and* $\lambda = O(n^{-1/(1+2\beta)})$ *(the optimal bandwidth),*

$$n^{\beta/(1+2\beta)}(\hat{\sigma}_\lambda^2(x; \sigma_\epsilon^2) - \sigma^2(x) - O(\lambda^\beta)) \to^d Z_1,$$

as $\lambda \to 0$*,* $n \to \infty$*, and* $n\lambda \to \infty$*. For* $1/2 < \alpha < 1$*, and* $\lambda = O(n^{-(2\alpha-1)/(1+2\beta)})$ *(the optimal bandwidth),*

$$n^{(2\alpha-1)\beta/(1+2\beta)}(\hat{\sigma}_\lambda^2(x; \sigma_\epsilon^2) - \sigma^2(x) - O(\lambda^\beta)) \to^d Z_2,$$

as $\lambda \to 0$*,* $n \to \infty$ *and* $n\lambda \to \infty$*, where* Z_1 *and* Z_2 *are normal distributions with mean zero and variance* σ_1^2 *and* σ_2^2*, respectively,* $0 < \sigma_1^2, \sigma_2^2 < \infty$*.*

Remark 2. Brown and Levine (2007) proposed difference-based estimators for nonparametric regression model and established their asymptotic normality. The asymptotic normality of the variance function estimator in our model can be established by using similar arguments. The proof of Theorem 2 relies on Theorem 2.2 in Peligrad and Utev (1997).

Theorems 1 and 2 assumes σ_ϵ^2 is known, while in most applications, σ_ϵ^2 is unknown. We first estimate σ_ϵ^2 using the modified likelihood estimator, then plug in $\hat{\sigma}_\epsilon^2$ to obtain the variance function estimator. In Theorem 3 and 4, we establish asymptotic properties of the modified likelihood estimator of σ_ϵ^2, and the plug-in variance function estimator $\hat{\sigma}_\lambda^2(x, \hat{\sigma}_\epsilon^2)$.

Theorem 3. *In model* (1.1)*, assume* $\sigma^2(x)$ *belongs to the functional classes* C_β^+ *for* $\beta > 0$ *and the variance of measurement error* σ_ϵ^2 *is* $O(n^{-\alpha})$ *with* $\alpha > 1/2$*. If* $\hat{\sigma}_\epsilon^2$ *is the modified likelihood estimator of* σ_ϵ^2*,* $\lim_{n\to\infty} \hat{\sigma}_\epsilon^2 = \sigma_\epsilon^2$ *in probability. For* $\alpha \geqslant 1$*,* $\hat{\sigma}_\epsilon^2 = \sigma_\epsilon^2 + O_p(n^{-3/2})$*. For* $1/2 < \alpha < 1$*,* $\hat{\sigma}_\epsilon^2 = \sigma_\epsilon^2 + O_p(n^{-(1+2\alpha)/2})$*.*

Remark 3. Theorem 3 shows that $\hat{\sigma}_\epsilon^2$ converges to σ_ϵ^2 at rate $n^{-3/2}$ when the measurement error is of order $n^{-\alpha}$ with $\alpha > 1/2$. If $\alpha \geq 3/2$, the convergent rate of $\hat{\sigma}_\epsilon^2$ is slower than the rate of the measurement error going to zero. In such cases the measurement error is too small to have any impact on the estimation of $\sigma^2(x)$. Conversely, if $1/2 < \alpha < 1$, then the convergence rate of $\hat{\sigma}_\epsilon^2$ depends on α, with larger α corresponds to slower convergence rate.

Theorem 4. *In model* (1.1)*, for the kernel smoothing estimator at* (2.5)*,* $\alpha \geq 1$*, and* $\lambda = O(n^{-1/(1+2\beta)})$ *(the optimal bandwidth),*

$$n^{\beta/(1+2\beta)}(\hat{\sigma}_\lambda^2(x, \hat{\sigma}_\epsilon^2) - \sigma^2(x) - O(\lambda^\beta)) \to^d Z_1,$$

as $\lambda \to 0$*,* $n \to \infty$*, and* $n\lambda \to \infty$*. For* $1/2 < \alpha < 1$*, and* $\lambda = O(n^{-(2\alpha-1)/(1+2\beta)})$ *(the optimal bandwidth),*

$$n^{(2\alpha-1)\beta/(1+2\beta)}(\hat{\sigma}_\lambda^2(x, \hat{\sigma}_\epsilon^2) - \sigma^2(x) - O(\lambda^\beta)) \to^d Z_2,$$

as $\lambda \to 0$*,* $n \to \infty$*, and* $n\lambda \to \infty$*, where* Z_1 *and* Z_2 *are normal with mean zero and variance* σ_1^2 *and* σ_2^2*, respectively,* $0 < \sigma_1^2, \sigma_2^2 < \infty$*.*

To prove Theorem 4, we have $\hat{\Delta}_i = \Delta_i + O_p(\max(n^{-1/2}, n^{-(2\alpha-1)/2}))$ from Theorem 3, and thus

$$\begin{aligned}
\hat{\sigma}_\lambda^2(x; \hat{\sigma}_\epsilon^2) &= \sum_{i=1}^{n-1} K_n(\frac{x - x_i}{\lambda})\hat{\Delta}_i \\
&= \sum_{i=1}^{n-1} K_n(\frac{x - x_i}{\lambda})\{\Delta_i + O_p(\max(n^{-1/2}, n^{-(2\alpha-1)/2}))\} \\
&= \hat{\sigma}_\lambda^2(x; \sigma_\epsilon^2) + O_p(\max(n^{-1/2}, n^{-(2\alpha-1)/2})) \\
&= \sigma^2(x) + O_p(\lambda^\beta) + O_p(\max(n^{-1/2}, n^{-(2\alpha-1)/2})). \qquad (3.1)
\end{aligned}$$

When $\alpha \geq 1$, the optimal bandwidth is $\lambda = O(n^{-1/(1+2\beta)})$, under which the third therm is negligible. When $1/2 < \alpha < 1$, the optimal bandwidth is $\lambda = O(n^{-(2\alpha-1)/(1+2\beta)})$, under which $O(\lambda^\beta) = O(n^{-(2\alpha-1)\beta/(1+2\beta)})$. Since $(2\alpha - 1)/2 > (2\alpha - 1)\beta/(1 + 2\beta)$ always holds, the third term is again negligible compared to the second term in (3.1). By Theorem 2 and Slutsky's theorem, the results in Theorem 4 follow.

Remark 4. According to Theorem 4, substituting σ_ϵ^2 with $\hat{\sigma}_\epsilon^2$ in estimating $\sigma^2(x)$ has negligible effect, and the asymptotic property of $\hat{\sigma}_\lambda^2(x, \hat{\sigma}_\epsilon^2)$ is the same as $\hat{\sigma}_\lambda^2(x; \sigma_\epsilon^2)$.

Theorem 5. *The plug-in kriging predictor $\hat{p}(x_0)$ is asymptotically unbiased for $\sigma(x_0)W(x_0)$. When $\alpha \geqslant 1$,*

$$E\{\hat{p}(x_0)\} = \sigma(x_0)W(x_0) + O(n^{-\beta/(1+2\beta)}),$$

and when $1/2 < \alpha < 1$,

$$E\{\hat{p}(x_0)\} = \sigma(x_0)W(x_0) + O(n^{-(2\alpha-1)/(1+2\beta)}).$$

Remark 5. Theorem 5 shows that the bias of the plug-in kriging predictor is small when $\alpha \geq 1$, and it is dependent on α when $\alpha < 1$. The bias term becomes non-negligible when α is close to $1/2$, due to the deterioration of the variance function estimator as shown in Theorem 1. The performance of the kriging prediction using the estimated variance function deteriorates as the variability of measurement error increases, and it becomes harder to recover the underlying variance function in the estimation stage.

4. Simulation Studies

We report the results of two simulation studies, one on variance estimation and the other on prediction.

4.1. Simulation one - variance estimation

In Simulation One, we tested the performance of our proposed method for recovering the underlying variance function. $B = 100$ Monte Carlo samples of sizes n were generated from $z_i = z(x_i) = \sigma(x_i)W(x_i) + \epsilon_i$ on a regular grid $x_i = i/n$ on $[0, 1]$, where $W(x)$ is the Brownian motion on $[0, 1]$, and ϵ_i i.i.d. $\sim N(0, \sigma_\epsilon^2)$. Consider the variance of the measurement error to be $\sigma_\epsilon^2 = 0.1/n$. We considered the following parameter values $n = 200,\ 500$ and 1,000, and used the variance functions $\sigma^2(x) = 16(x - 1/2)^2 + 1/2$ and $\sigma^2(x) = 0.2\sin(x/0.15) + 1.0$. We chose the optimal bandwidth by K-fold cross validation with $K = 10$. The performance of the difference-based kernel smoothing estimator was measured using discrete mean squared error

$$\text{DMSE} = n^{-1}\sum_{i=1}^{n}\{\hat{\sigma}_{\lambda_{CV}}^2(x_i) - \sigma(x_i)\}^2,$$

where $\lambda_{CV}$is the K-fold cross-validation bandwidth.

Table 1 and Table 2 show the median DMSE for the difference-based kernel smoothing estimator, the median bandwidth, and the mean of $\hat{\sigma}_\epsilon^2$ over 100 Monte

Table 1. Performance of variance function estimator and $\hat{\sigma}_\epsilon^2$ with a quadratic variance function.

Variance function: $16(x-1/2)^2+1/2$

n	Median DMSE	Median Bandwidth	Mean $\hat{\sigma}_\epsilon^2$
200	0.201	1.00	0.00050
500	0.095	1.00	0.00017
1,000	0.053	1.00	9.9e-05

Table 2. Performance of variance function estimator and $\hat{\sigma}_\epsilon^2$ with a sine variance function.

Variance function: $2\sin(x/0.15)+2.8$

n	Median DMSE	Median Bandwidth	Mean $\hat{\sigma}_\epsilon^2$
200	0.694	0.26	0.00070
500	0.429	0.22	0.00023
1,000	0.274	0.21	0.00012

Carlo samples for $\sigma^2(x)=16(x-1/2)^2+1/2$ and $\sigma^2(x)=0.2\sin(x/0.15)+1.0$, respectively. The performance of the variance estimator improves as n increases, which is consistent with Theorem 4. Similarly from the column "Mean $\hat{\sigma}_\epsilon^2$", one can see that the bias of $\hat{\sigma}_\epsilon^2$ gets smaller as n increases, as predicted by Theorem 3.

4.2. Simulation two - kriging versus spatially adaptive local polynomial fitting

In Simulation Two, we compared the performance of our proposed method of plug-in kriging to non-parametric methods. $B=100$ Monte Carlo samples of sizes $n=200$ were generated from $z_i=z(x_i)=\sigma(x_i)W(x_i)+\epsilon_i$ on a regular grid $x_i=i/n$ on $[0,1]$, with $\sigma^2(x)=1.6(x-0.5)^2+0.8$, $W(x)$ the Brownian motion on $[0,1]$, and ϵ_i i.i.d. $\sim N(0,\sigma_\epsilon^2)$. We took σ_ϵ^2 to be $0.1/n$, $1/n$, and $10/n$. The plug-in kriging predictor was compared with the spatially adaptive local polynomial regression estimator (ALPRE), and the local polynomial regression estimator (LPRE) with a global bandwidth. In ALPRE, the adaptive bandwidth was obtained by a procedure similar to the one proposed by Fan and Gijbels (1995). The interval$[0,1]$ was split into $[1.5n/(10\log(n))]$ sub intervals, and a leave-one-out cross validation method is used in each interval to obtain a local bandwidth. These bandwidths are then smoothed to obtain the bandwidth for each point.The performance of the prediction was measured using the discrete mean squared error (DMSE).

Table 3 shows the median of DMSE over 100 Monte Carlo samples for the plug-in kriging predictor(Kriging), adaptive local polynomial regression estima-

Table 3. Performance of plug-in kriging, adaptive local polynomial regression estimator(ALPRE) and local polynomial estimator(LPRE).

σ_ϵ^2	Methods	Median DMSE
$0.1/n$	Kriging	0.00048
	ALPRE	0.00220
	LPRE	0.00380
$1/n$	Kriging	0.00320
	ALPRE	0.00390
	LPRE	0.00450
$10/n$	Kriging	0.03200
	ALPRE	0.01400
	LPRE	0.01200

tor(ALPRE) and local polynomial estimator(LPRE) with a global bandwidth. When $\sigma_\epsilon^2 = 0.1/n$, Kriging outperformed ALPRE. When $\sigma_\epsilon^2 = 1/n$, the performance of Kriging and that of ALPRE are comparable. When $\sigma_\epsilon^2 = 10/n$, Kriging underperformed ALPRE. When the measurement error is small, the realized process is very close to the underlying true process, and all three methods did well in predicting the underlying process. Nevertheless, kriging outperforms the other two methods, with its median DMSE less than 1/4 of ALPRE. As the measurement error increases, the realized process is subject to more noises, and at some point, the measurement error is too large for our method to estimate reliably the underlying variance function. Kriging did poorly in recovering the underlying true process compared with ALPRE and LPRE. (See Figures in the supplementary material for the support of the above argument). It is also interesting to note that in this case ALPRE is no better than LPRE. From Table 3, when σ_ϵ^2 is large, the median DMSE of LPRE with a global bandwidth is 14% better than ALPRE. This suggests that a global bandwidth is enough.

5. Discussion

In this paper we developed a difference-based estimation method to estimate the variance function of a non-stationary spatial process based on one realization, whereas, the non-stationary model is usually fit to spatial temporal data where there are time replications of spatial process or spatial replications of time series, see Fonseca and Steel (2011), Bornn, Shaddick, and Zidek (2012), among others. spatial process is an advantage of our method.

The estimation procedure we developed can be applied to more flexible non-stationary spatial processes. For example, Brownian motion can be replaced by a Gaussian process with Matern covariance structure that allows for a fairly flexible class of non-stationary covariance structure. The variance function estimation

under such models can be done similarly, though it would be more difficult to derive asymptotic results.

We have limited our attention to non-stationary spatial processes on $\mathbb{R}^1$. In principle, our methodology can be applied to the estimation of variance function of non-stationary spatial process in higher dimensions. For example, Hall, Kay, and Titterington (1991) discussed estimation of noise variance in two-dimensional signal processing using a difference-based approach. A similar approach can be used to estimate the variance function of a two-dimensional non-stationary spatial process. We also restricted our difference-based estimator to the first-order difference to limit the technical derivations. Properties of the estimator based on higher order differences will be addressed in a future work.

References

Anderes, E. B. and Stein, M. L. (2011). Local likelihood estimation for non-stationary random fields. *J. Multivariate Anal.* **102**, 506-520.

Bornn, L., Shaddick, G. and Zidek, J. V. (2012). Modeling non-stationary processes through dimension expansion. *J. Amer. Statist. Assoc.* **107**, 281-289.

Brown, L. D. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Ann. Statist.* **35**, 2219-2232.

Cai, T. and Wang, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *Ann. Statist.* **36**, 2025-2054.

Cai, T., Levine, M. and Wang, L. (2009). Variance function in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.* **100**, 126-136.

Cummins, D. J., Filloon, T. G. and Nychka, D. (2001). Confidence intervals for nonparametric curve estimates: Toward more uniform pointwise coverage. *J. Amer. Statist. Assoc.* **96**, 233-246.

Damian, D., Sampson, P. and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics* **12**, 161-178.

Duran, E. A., Hardle, W. K. and Osipenko, M. (2012). Difference based ridge and Liu type estimators in semi-parametric regression models. *J. Multivariate Anal.* **105**, 164-175.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57**, 371-394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

Fonseca, T. O. and Steel, M. F. (2011). Non-Gaussian spatiotemporal modeling through scale mixing. *Biometrika* **98**, 761-774.

Fuentes, M. (2002a). Interpolation of non-stationary air pollution processes: a spatial spectral approach. *Statist. Model.* **2**, 281-298.

Fuentes, M. (2002b). Spectral methods for non-stationary spatial processes. *Biometrika* **89**, 197-210.

Gasser, T. and Sroka, L. and Jennen-Steinmetz, C. (1986), Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625-633.

Guttorp, P., Sampson, P. D. and Newman, K. (1992). Nonparametric estimation of spatial covariance with application to monitoring network evaluation. *Statist. Environmental & Earth Sci.*, 39-51.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521-528.

Hall, P., Kay, J. W. and Titterington, D. M. (1991). On estimation of noise variance in two-dimensional signal processing. *Adv. Appl. Probab.* **23**, 476-495.

Higdon, D. (1998). A process-convolution approach to modeling temperatures in the North Atlantic Ocean. *J. Environmental and Ecolog. Statist.* **5**, 173-190.

Higdon, D., Swall, J. and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian Statist.* **6**, 761-768.

Klipple, K., and Eubank, R. (2007). Difference-based variance estimators for partially linear models. *Festchrift in Honor of Distinguished Professor Mir Masoom Ali on the Occasion of His Retirement*, 313-323.

Levine, M. (2006). Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach. *Comput. Statist. Data Anal.* **50**, 3405-3431.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *J. Amer. Statist. Assoc.* **92**, 107-116.

Matsuo, Y., Misu, T., Sakaida, S. and Shishikui, Y. (2011). Video coding with wavelet image size reduction and wavelet super resolution reconstrunction. *IEEE, Internat. Conf. Image Processing*, 1157-1160.

Nychka, D. and Saltzman, N. (1998). Design of air quality monitoring networks. *Case Studies in Environmental Statist.* **132**, 51-76.

Nychka, D., Wikle, C. K. and Royle, J. K. (2002). Multiresolution models for non-stationary spatial covariance functions, *Statist. Model.* **2**, 315-331.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modeling using a new class of non-stationary covariance functions. *Environmetrics* **17**, 483-506.

Peligrad, M. and Utev, S. (1997). Central limit theorem for linear processes. *Ann. Probab.* **25**, 443-456.

Pintore, A., Speckman, P. L. and Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika* **93**, 113-125.

Porcu, E., Gregori, P. and Mateu, J. (2009). Archimedean spectral densities for non-stationary space-time Geostatistics. *Statist. Sinica* **19**, 273-286.

Sanso, B., Schmidt, A. and Nobre, A. A. (2005). Bayesian spatio-temporal models based on discrete convolutions. Technical Report, Departamento de Metodos Estatsticos, Universidade Federal do Rio de Janeiro, Brazil.

Schmidt, A. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *J. Roy. Statist. Soc. Ser. B* **65**, 743-758.

Schmidt, A., Schelten, K. and Roth, S. (2011). Bayesian deblurring with integrated noise estimation. *IEEE, Computer Vision and Pattern Recognition*, 2625-2632.

Stein, M. L. (1990). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Ann. Statist.* **18**, 1139-1157.

Stein, M. L. (1993). Spline smoothing with an estimated order parameter. *Ann. Statist.* **21**, 1522-1544.

Von Neumann, J., Kent, R. H., Bellinson, H. R. and Hart, B. I. (1941). The mean square successive difference. *Ann. Math. Statist.* **12**, 153-162.

Wahba, G. (1985). Partial and interaction splines for the semiparametric estimation of functions of several variables. University of Wisconsin, Department of Statistics

Wang, L., Brown, L. D., Cai, T. and Levine, M. (2008). Effect of mean on variance function estimation on nonparametric regression. *Ann. Math. Statist.* **36**, 646-664.

Department of Statistics, Iowa State University, Ames, IA, 50010, USA.

E-mail: shuyang@iastate.edu

Department of Statistics, Iowa State University, Ames, IA, 50010, USA.

E-mail: zhuz@iastate.edu

(Received July 2013; accepted February 2014)