Estimating Variance Function of A Nonstationary Process Using A Difference Filter

E. J. Kim[1], Z. Zhu[2]

[1]Department of Mathematics and Statistics, Amherst College, Box 2239, Amherst, MA 01002
+1 413-542-5422
ekim@amherst.edu

[2]Department of Statistics, Iowa State University, Snedecor Hall 1211, Ames, IA, 50010
zhuz@iastate.edu

Abstract

Many geophysical processes exhibit nonstationary features, and it is important to estimate a non-constant variance function of such process so that an accurate prediction interval be provided. We propose a difference-based approach to estimating the variance function from a single process where the errors are nonstationary and correlated. We assume that the mean function is smooth and that the error process is a product of a smooth standard deviation function and a second-order stationary process. A numerical study shows that the mean squared error depends on the choice of filter and the strength of correlation in the error process. Symmetric-weight filters are preferred for errors with strong correlation, and Hall-Kay-Titterington weight filters are preferred for weakly correlated or independent data.

Keywords: nonstationary process, variance function, filter, symmetric weight

1. Introduction

We develop a method to account for a variance function of a continuous nonstationary process. For example, consider an average daily high temperature map of the U.S. April. In general, it exhibits warm temperature in the south and cool temperature in the north. There is a mesoscale temperature dip in the Great Lakes and the Rockies where the temperature is cooler than the locations in the same latitudes and in the east and west coasts the temperature is higher than the locations in the same latitudes. These trends can be well accounted for by physics-driven models. However, the observed average daily high temperature of April may consistently hit above or below the projected average daily high temperature in patches (relatively small scale to the entire U.S.). The size of the deviations would vary depending on the locations. In order to account for the potential range, a prediction interval, of the average daily highs in the coming April, we need to accurately estimate the variance function of these random deviations.

Estimating the variance by data differencing is a method detailed by von Neumann et al. (1941). When data have natural ordering, a gradual change in mean, and independent and identically distributed errors, simple differencing would efface the effect of estimating a mean structure on the estimation of variance. That is, random errors would directly contribute to the estimation of variance. This idea is extended to estimating a variance function by Gasser & Müller (1984), Buckley et al. (1988), and Hall & Carroll (1989). Simple differencing between neighboring points in sequential data have led to a nonparametric estimation of a one-dimensional variance function. Hall et al. (1991) extended the idea to image processing. The data model assumes independent and identically distributed errors added to a true image. The variance is estimated as a linear combination of squared filtered data. The weights are numerically optimized to achieve minimum variance of the variance estimation. Differencing may create bias due to non-constant mean function. However, as the grid becomes finer and finer (in infill asymptotics), the bias becomes negligible. Our method extends the variance function estimation to two-dimensional data, where a differencing idea is linked to a definition of variogram. Similarly, Zhu & Stein (2002) have introduce generalized variogram and used difference filters to estimate the fractal dimension of fractional Brownian fields.

The paper is organized in the following order. In Section 2, we describe our data model, define a linear filter, and introduce a filter variogram and local variogram. We assume that the error process is isotropic when it is standardized. In Section 3, we introduce a variance function estimator, and in Section 4, different filter shapes and weights are explored. A difference filter is used to remove a local mean structure and to reduce a positive correlation structure in the errors. Via simulation, we compare the effect of applying symmetric weights to Hall-Kay-Titterington weights from Hall et al. (1991). In Section 5, we summarize the results and conclude.

2. Data Model and Variance Function Estimator

2.1 Data Model

Consider a continuous process on a two-dimensional plane. Our data model assumes that the process at location $\boldsymbol{s}$ is centered at smooth mean $\mu(\boldsymbol{s})$ and has an additive non-constant error $\sigma(\boldsymbol{s})\boldsymbol{X}(\boldsymbol{s})$ where $X(\boldsymbol{s})$ is a stationary process mean 0, variance 1, and $cor\,(X(\boldsymbol{s}), X(\boldsymbol{s}')) = c\|\boldsymbol{s} - \boldsymbol{s}'\|^{\alpha}$ where $0 < \alpha < 2$ for $0 < c < 1$. The data model of a random process $\{Z\}_{\boldsymbol{s}\in\mathbb{R}^2}$ is

$$Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \sigma(\boldsymbol{s})X(\boldsymbol{s}). \tag{1}$$

Suppose we have a set of observations on a regular lattice grid.

2.2 Notations and Definitions

A linear filter function L is defined by a set of neighboring points $\mathcal{J}$ about $\boldsymbol{p}_0$, the center of a configuration, such that

$$\mathcal{J} = \left\{ \boldsymbol{p}_j = (p_{1j}, p_{2j}) \in \mathbb{Z}^2 : \sum_j (\boldsymbol{p}_j - \boldsymbol{p}_0) = \mathbf{0} \right\}. \tag{2}$$

A set of non-zero weights $A = \{a_j : j \in \mathcal{J}\}$ is assigned to each point $\mathcal{J}$. Then,

$$L\left(Z(\boldsymbol{s})\right) = \sum_{j \in \mathcal{J}} a_j Z(\boldsymbol{s} + \boldsymbol{p}_j) \tag{3}$$

represents a filter L applied to a process $\boldsymbol{Z}$ about $\boldsymbol{s}$. Throughout this paper, we use these shorthands: $Z(\boldsymbol{s}+\boldsymbol{p}_j) = Z_{\boldsymbol{s}+j},\ Z(\boldsymbol{s}+h\boldsymbol{p}_j) = Z_{\boldsymbol{s}+jh},\ \rho(\|\boldsymbol{s_i} - \boldsymbol{s_j}\|) = \rho_{\|\boldsymbol{i}-\boldsymbol{j}\|}$, and use $j \in \mathcal{J}$ for $\boldsymbol{p}_j \in \mathcal{J}$.

Definition 2.1 *Define an L-filter variogram at scale h as*

$$\varrho_L(h) = 1 - 2 \sum_{j \in \mathcal{J}} \sum_{\substack{k \neq j \\ k \in \mathcal{J}}} a_j a_{j+k} \rho_{h\|k\|}. \tag{4}$$

Filter weights A should satisfy these basic **conditions**:

1. $\sum_{j \in \mathcal{J}} a_j = 0$ which implies that $E(\sum_{j \in \mathcal{J}} a_j X_{i+j}) = 0$.
2. $\sum_{j \in \mathcal{J}} a_j^2 = 1$ which implies that $E\left(\sum_{j \in \mathcal{J}} a_j X_{i+j}\right)^2 = \varrho_L(h)$.
3. $\sum_{j \in \mathcal{J}} a_j \boldsymbol{p}_j = (0,0)$ which implies that $L(Z_{\boldsymbol{s}})$ gives a pseudo-residual at $\boldsymbol{s}$.

These are not sufficient conditions to determine the weights uniquely. The number of conditions should match the number of nodes in each filter to determine a unique set of weights. We impose a **symmetry condition** below:

4. The weights are symmetrically distributed about (0,0) or $\boldsymbol{p_0}$.

Definition 2.2 *A filter is called a symmetric-weight filter when the set of weights on each node of a filter satisfies Conditions 1 - 4.*

Some filter configuration will not satisfy the fourth condition explicitly. In that case, we rotate the filter and achieve the symmetric weighting. Instead of Condition 4, we could choose weights such that the variance of the estimator is minimized. Hall et al (1991) have developed an array of difference filters to estimate the variance of additive i.i.d. errors, and we call their proposed weight *Hall-Kay-Titterington* weight (HKT). Symmetric-weight filters have the weight centered at the center of each filter $\boldsymbol{p_0}$, whereas the center of HKT weight is on one extremity of a filter configuration as shown in Section 4.1 (marked with '$\times$').

Weight	n	θ						
Symmetric	40	$\theta = 0.1$	0.16	0.18	0.16	0.14	0.16	0.22
		$\theta = 0.01$	0.89	0.91	0.87	0.86	0.88	0.96
	100	$\theta = 0.1$	0.07	0.07	0.06	0.06	0.06	0.09
		$\theta = 0.01$	0.56	0.59	0.53	0.51	0.54	0.68
HKT	40	$\theta = 0.1$	0.31	N/A	0.28	0.37	0.25	0.40
		$\theta = 0.01$	0.96		0.96	0.99	0.94	0.97
	100	$\theta = 0.1$	0.14	N/A	0.12	0.17	0.11	0.20
		$\theta = 0.01$	0.75		0.73	0.85	0.67	0.82

Table 1: L-filter variogram values for different filter configurations, weighting options, and the strengths of correlation, θ, assuming an exponential correlation function for the error process.

Definition 2.3 *Define an L-filter local variogram for a two-dimensional nonstationary process as the leading term of* $E\left[L(Z(\boldsymbol{s},h))^2\right]$.

$$\Gamma_\Lambda\left(\boldsymbol{s}; L(Z(\boldsymbol{s},h))\right) = \sigma^2(\boldsymbol{s})(1 - \sum_{\substack{j\neq k \\ j,k\in\mathcal{J}_L}} a_j a_k \rho_{h\|j-k\|}) = \sigma^2(\boldsymbol{s})\varrho_L(h). \tag{5}$$

The L-filter variogram describes the dispersion in correlated data as a function of lag size. Table 1 focuses on displaying the dispersion measure by filter configuration and weighting option. With symmetric weight filters L-filter variogram matches closely with a regular variogram. With HKT weights, however, L-filter variogram exaggerates the dispersion because the weight is heavily loaded on one node, and this reduces the size of cross terms.

3. Variance Function Estimator

L-filter local variogram is, by definition 2.3, a variance function embedded in an observed process scaled by a filter-specific variogram. We use this property to derive an estimator for a variance function. First, we apply a linear filter L to a set of observations that rise from smoothly varying mean and standard deviation functions. The filtered observations serve as pseudo-residuals. Then, we take a local average of the squared filtered process. This is a surface estimate of the L-filter local variogram, $\hat{\Gamma}_\Lambda\left(\boldsymbol{s}; L(Z(\boldsymbol{s},h))\right)$. When taking a local average, the boundary of 'local' and the form of the smoothing kernel need to be determined. We defined a two-dimensional kernel $\mathbf{K}_\Lambda(\cdot)$ as the Kronecker product of Gasser-Müller kernels (Gasser & Müller (1984)) where Λ represents a two-tuple bandwidth vector. We have L-filter local variogram estimator at location $\boldsymbol{s_0}$ as

$$\hat{\Gamma}_{\Lambda}\left(\boldsymbol{s_0};\, L(h)\right) = \sum_{\boldsymbol{i}\in\mathcal{R}} K_{\Lambda}(\boldsymbol{i},\boldsymbol{0}) L(Z(\boldsymbol{s}_i,\, h))^2. \tag{6}$$

To estimate $\hat{\varrho}_L(h;\hat{\theta})$, it seems that we need an insight on the correlation function $\varrho(\cdot)$ of the process. From our one-dimensional study, however, we see that the full description of the correlation function is not needed because the lag size h is fixed when filtering the data. While h is relatively small, we assume a simple parametric form of the correlation structure, estimate the correlation function parameters $\hat{\theta}$, and then estimate the variance at location $\boldsymbol{s_0}$ as

$$\hat{\sigma}_{\Lambda}(\boldsymbol{s_0}) = \frac{\hat{\Gamma}_{\Lambda}(\boldsymbol{s_0}; L(h))}{\varrho_L(h;\hat{\theta}).} \tag{7}$$

4. Exploring Filter Options

We are interested in identifying a set filter configurations and weighting options that would provide a statistically consistent and efficient estimation of variance function. The bias and variance of the variance estimator in equation (7) can be examined by long and complex analytical derivations for the several filters we chose to investigate, but instead we performed a simulation study to understand the statistical performance of the proposed estimator of variance.

4.1 Configuration and Weights

We have proposed to use a difference filter to recast the data as a filtered error process since it leads to less bias in the estimation of a continuous and smooth variance function given that the mean function changes slowly (the degree differentiability is small) in comparison to the variance function. The shape of the filter we consider is all in the span of a 3×3 grid and is symmetric about an axis that goes through the areal center of the filter.

Here is a 2×2 square filter, which I name Square2, with symmetric weight on the left and HKT weight on the right:

symmetric: $\begin{matrix} -a & a \\ & \times \\ a & -a \end{matrix}$ $a = \pm\frac{1}{2}$ HKT: $\begin{matrix} -3a & a \\ & \times \\ a & a \end{matrix}$ $a = \pm\frac{1}{\sqrt{12}}$

I call a 3×3 square filter, which incorporates eight nodes about $\boldsymbol{p_0}$ and excluding the weight center, Square3. I do not present the exact picture here to reserve space. Refer to the Appendix of Kim (2013). Next is a + -shape filter, with symmetric weight on the left and HKT weight on the right:

a 0.231

symmetric: a -4a a $a = \pm\frac{1}{\sqrt{20}}$ HKT:×0.263 0.167 -0.892

a 0.231

Lastly, a Y-shaped filter is presented only with a symmetric weight.

a

Symmetric: a -3a $a = \pm\frac{1}{\sqrt{12}}$

a

4.2 Simulation Set-up

We simulated $n = 100$ zero-mean, stationary Gaussian processes with varying levels of dependent structure on a unit square. Each innovation was read from both $N = 40 \times 40$ and 100×100 equally-spaced grid points. We used an exponential correlation function at two levels of range parameter $\theta = 0.01$ (weak correlation) and $\theta = 0.1$ (strong correlation) to simulate innovations, and we also generated innovations with independent errors. We, then, scaled the stationary innovations by the standard deviation functions shown in figure 1.

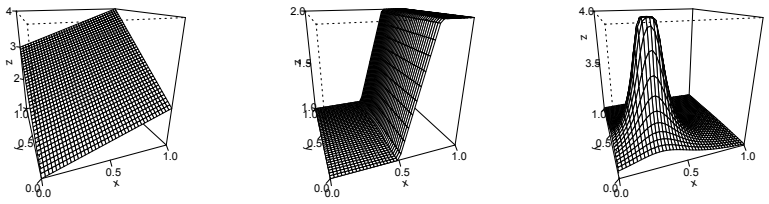


Figure 1: Standard deviation functions $\sigma(s_x, s_y)$ in 3-dimensional perspective drawing.

Since the variance is in quadratic units of the actual observations, overestimation is easily pronounced. So, we scale the estimate by the true value:

$$\hat{\epsilon}_\Lambda(\boldsymbol{s}) = \frac{\hat{\sigma}^2_\Lambda(\boldsymbol{s}) - \sigma^2(\boldsymbol{s})}{\sigma^2(\boldsymbol{s})}, \tag{8}$$

and use it in the measures of discretely-integrated mean-squared-error ($DMSE$), median-absolute-deviation (MAD), and maximum deviation ($rMAX$) defined with

the relative size of the error:

$$rDMSE_{\Lambda}(L_{\nu}) = \frac{1}{N_l}\sum \hat{\epsilon}_{\Lambda}(\boldsymbol{s_i})^2 \quad (9)$$

$$rMAD_{\Lambda}(L_{\nu}) = \frac{1}{N_l}\sum |\hat{\epsilon}_{\Lambda}(\boldsymbol{s_i}) - \text{median}_i\hat{\epsilon}_{\Lambda}(\boldsymbol{s_i})| \quad (10)$$

$$rMAX_{\Lambda}(L_{\nu}) = \max |\hat{\epsilon}_{\Lambda}(\boldsymbol{s_i})| . \quad (11)$$

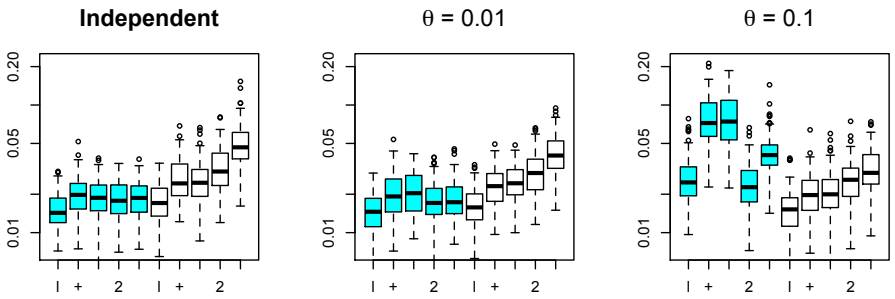
4.3 Results

For the following generalization of the study results, assume that the mean function $\mu(\boldsymbol{s})$ varies more slowly, or has a lower order differentiability, than the standard deviation function $\sigma(\boldsymbol{s})$. There are three takeaways from the simulation study.

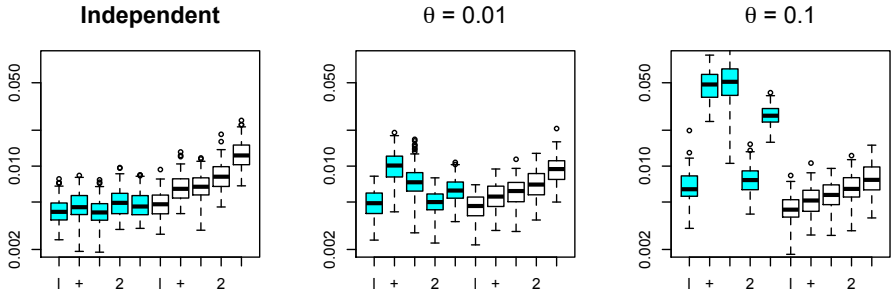
First, the filter weighting option should be chosen depending on the correlation structure of the data. When the errors are independent or when the correlation is weak, HKT weights by construction are the most efficient estimator of variance function across all filter configurations. When the errors are moderately correlated, symmetric weights should be used because the estimator portions the correlation more accurately. In figure 3, the plots on the left have independent error scenario and display lower $rDMSE$ by HKT (in blue) than symmetric weights (in white) for all five filter configurations. The plots on the right reverses this description as the correlation is quite strong with $\theta = 0.1$. The boxplots in the middle column seem to tell a conflicting story, but upon close examination the reason HKT works well on the top is the correlation is not detected on a coarse $\frac{1}{40} \times \frac{1}{40}$ scale versus on a fine $\frac{1}{100} \times \frac{1}{100}$ scale grid over a fixed region.

Secondly, when there is a dominant direction in the filter configuration, such as a line or a Y-shape filter, it is important to achieve a symmetric weighting about the major and minor axes of the filter. Depending on the type of grid and the shape of the filter, it would take four or six directional rotations to filter an observed process. The steps of the directional-averaging should be, first, filtering the process in all directions, and then, taking the average of a set of four or six filtered processes, and lastly, using it as the filtered process, $L(Z)$. The averaging across four or six directions of a filter in a span of 3×3 is, in fact, providing multiple symmetric weighting options. Depending on the ways the weight is distributed on each node, we could increase the statistical efficiency of the variance estiatmion. Table 2 provides a five-number-summary of $rMAD$ where the simulated processes are on a 100×100 grid. By directional-averaging, as shown in the right-most column, the relative median absolute deviation decreases significantly. Comparing five filter configurations by $rDMSE$ in figure 2, we see that the directionally-averaged line filter (8-point star) performs the best among all, and then, follows the + with the scale of 1 and $\sqrt{2}$, Square2, and Square3 filters.

Lastly, it is important to have the configuration of a filter be compact and encompassing of all directions to capture local characteristics. As noted by directional-averaging, filtered data should take information from all directions assuming isotropy of the errors. From table 2, we see that a single Y filter performs, on average, better than a single line filter since a Y configuration extends out more than a line. We also



(a) $N = 40 \times 40$



(b) $N = 100 \times 100$

Figure 2: $rDMSE$ of five filters with HKT weight (in light blue) and symmetric weight (in white). The filter configurations are 8-point star, +, ×, Square2, and Square3. The top row used data on a coarse grid of $N = 40 \times 40$, and the bottom on a fine grid of $N = 100 \times 100$.

investigated the effect of scale h on the estimation but did not include another table here for brevity. To summarize, when the errors are independent, either a scale of $h = 1$ or 2 results in a similar estimation; but when the errors are strongly correlated, the smaller scale of $h = 1$ results in smaller $rMAD$ and $rDMSE$ than $h = 2$. In other words, a filtered process is better de-correlated when the applied filter uses the smallest possible span rather than a larger span.

		$rMAD$ (%)						
Shape	θ	Min.	Q1	Median	Q3	Max.	Mean (Stdev.)	Dir. Avg.
	0	5.10	7.10	8.20	9.20	12.20	8.20 (1.40)	6.52 (1.02)
	0.01	5.00	7.10	7.80	8.60	11.80	7.80 (1.20)	6.13 (0.87)
	0.1	5.70	7.20	7.90	8.60	12.10	8.00 (1.20)	6.11 (1.10)
	0	4.80	6.10	7.00	7.80	9.30	6.90 (1.10)	6.69 (1.08)
	0.01	3.80	6.10	6.60	7.20	9.50	6.70 (1.00)	6.26 (0.95)
	0.1	4.50	5.80	6.60	7.30	10.10	6.60 (1.20)	6.17 (1.06)

Table 2: $rMAD$ comparison between line versus Y-shape filters, and with and without directional rotation and averaging.

5. Discussion

We frequently encounter nonstationary processes in our geography, and those processes not only contain varying levels of mean but also a varying size of scale by location. In this paper, we proposed a nonparametric method for a variance function estimation using a single observed process. First, a difference filter is applied to regularly dispersed data over the area of interest; then we take a local average of the squared filtered data and scale it down by an L-filter local variogram. We have not discussed the practical issues of bandwidth selection. The size of a bandwidth controls the range of averaging and affects the quality of estimation greatly. We recommend taking the cross-sections of the data in $x-$ and $y-$ directions and to perform separate bandwidth selections as discussed in Chapter 3 of Kim (2013).

We have assumed that the data is recorded on a grid. In practice, many georeferenced data are not observable on an exact grid. Still, we have applied a difference filter to a map of annual precipitation of the Midwest. As long as the observations are distributed uniformly across the region, the filter configuration can adapt to the locations of data collection. We have explored different filter configurations and two weighting schemes. Based on the simulation study, we recommend using a compact and directionally extended configuration and placing symmetric weights on the nodes of a filter. When the data contain weak correlation in the error, the HKT weight filter estimates the variance function with small $DMSE$. For many nonstationary processes, it is difficult to include local features in the mean function, therefore, a symmetric weight filter should help consistently estimate a variance function. Also,

we advise employing directional-averaging if there is a dominant direction in a filter configuration, so that the local feature is well accounted for in all directions.

6. References

Buckley, M. J., Eagleson, G. & Silverman, B. W. (1988), 'The estimation of residual variance in nonparametric regression', *Biometrika* **75**(2), 189–199.

Gasser, T. & Müller, H.-G. (1984), 'Estimating regression functions and their derivatives by the kernel method', *Scandinavian Journal of Statistics* **11**(2), 171–185.

Hall, P. & Carroll, R. J. (1989), 'Variance function estimation in regression: the effect of estimating the mean', *Journal of Royal Statistical Society. Series B* **51**(1), 3–14.
URL: *http://www.jstor.org/stable/2345837*

Hall, P., Kay, J. W. & Titterington, D. M. (1991), 'On estimation of noise variance in two-dimensional signal processing', *Advanced Applied Probability* **23**, 476–495.
URL: *http://www.jstor.org/stable/1427618*

Kim, E. J. (2013), 'Hotspot detection and nonstationary process variance function estimation', *Dissertation* .

von Neumann, J., Kent, R. H., Bellinson, H. R. & Hart, B. I. (1941), 'The mean square successive difference', *The Annals of Mathematical Statistics* **12**(2), 153–162.
URL: *http://www.jstor.org/stable/2235765*

Zhu, Z. & Stein, M. L. (2002), 'Parameter estimation for fractional brownian surfaces', *Statistica Sinica* **12**, 863–883.