Mathematical Geology, Vol. 32, No. 3, 2000

Variogram Model Selection via Nonparametric Derivative Estimation[1]

David J. Gorsich and Marc G. Genton[2]

Before optimal linear prediction can be performed on spatial data sets, the variogram is usually estimated at various lags and a parametric model is fitted to those estimates. Apart from possible a priori *knowledge about the process and the user's subjectivity, there is no standard methodology for choosing among valid variogram models like the spherical or the exponential ones. This paper discusses the nonparametric estimation of the variogram and its derivative, based on the spectral representation of positive definite functions. The use of the estimated derivative to help choose among valid parametric variogram models is presented. Once a model is selected, its parameters can be estimated—for example, by generalized least squares. A small simulation study is performed that demonstrates the usefulness of estimating the derivative to help model selection and illustrates the issue of aliasing. MATLAB software for nonparametric variogram derivative estimation is available at http://www-math.mit.edu/~gorsich/derivative.html. An application to the Walker Lake data set is also presented.*

KEY WORDS: nonparametric, variogram fitting, derivative estimation, generalized least squares, model selection, aliasing.

INTRODUCTION

In optimal linear spatial prediction, or kriging, the choice of weights is completely determined by the choice of the variogram model. Therefore, choosing the variogram model as closely as possible to the underlying variogram that defines the dependence of the data is crucial. Current geostatistical practice in selecting a variogram model is often rather subjective. Sometimes, *a priori* knowledge about the underlying stochastic process can be helpful. For example, topography is fairly smooth and might easily be modeled by a Gaussian variogram, whereas some geologic processes are cyclical in nature and could suggest a hole effect model. However, when such information is not available or not relevant enough, one has to rely on some empirical guidelines (e.g., Journel and Huijbregts, 1978; Clark,

[1]Received July 1998; accepted 9 February 1999.

[2]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307. e-mail: gorsich@math.mit.edu, genton@math.mit.edu

1979), related to the range, sill, and nugget effect of the variogram. These rules are often lacking because the shape of some variogram models is very similar: for instance, compare the exponential, Gaussian, and spherical models. The new tool this paper presents is a nonparametric variogram derivative estimator shown to be helpful in the discrimination of variogram models. Effectively, although variogram models look similar, their derivatives are often quite different. This paper starts by giving an example of the subjectivity involved in the selection of a parametric variogram model, and its effect on kriging. Next the use of the nonparametric estimator of the variogram to determine its derivative is discussed: the nonparametric estimator is introduced so that the estimated variogram is conditionally negative definite. MATLAB software is written that performs the nonparametric estimates for both the variogram and derivative. Results are displayed using a graphical user interface so that the closest parametric model can be selected easily. In the end, a small simulation study is performed to test the new tool on four different variogram models that look similar but have dissimilar derivatives. An application to the Walker Lake data set is also presented.

The importance of the variogram model on the kriging weights can be seen by studying one of the simplest stochastic processes. Consider a spatial stochastic process $\{Z(\mathbf{x}): \mathbf{x} \in D\}$, where D is a fixed subset of $\mathbb{R}^d$, $d \geq 1$. The process is assumed to be isotropic, and intrinsically stationary, i.e., for all $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ in D:

$$\begin{aligned} \mathrm{E}[Z(\mathbf{x})] &= \mu = \text{constant} \\ \mathrm{Var}[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})] &= 2\gamma(\mathbf{h}) \\ \gamma(\mathbf{h}) &= \gamma(\|\mathbf{h}\|) \end{aligned}$$

Consider a realization of Z at a finite number of points $\{Z(\mathbf{x}_i): \mathbf{x}_i \in D\}$ for $i = 1, \ldots, n$. The optimal linear predictor for $Z(\mathbf{x}_0)$, $\mathbf{x}_0 \in D$, in a mean squared error sense, can be found using a linear combination of the known Z values:

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{x}_i) \tag{1}$$

The weights λ_i are constrained so that $\sum_{i=1}^{n} \lambda_i = 1$, which guarantees uniform unbiasedness and are chosen to minimize $\mathrm{E}(Z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0))^2$. This gives the following minimization problem:

$$\mathrm{E}\left(Z(\mathbf{x}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{x}_i)\right)^2 - 2\xi\left(\sum_{i=1}^{n} \lambda_i - 1\right) \tag{2}$$

where ξ is the Lagrange multiplier. Working through the algebra and differentiating

gives the following set of equations to solve:

$$-\sum_{j=1}^{n} \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j) + \gamma(\mathbf{x}_0 - \mathbf{x}_i) - \xi = 0 \tag{3}$$

for $i = 1, \ldots, n$ and with the restriction that $\sum_{i=1}^{n} \lambda_i = 1$. In this way Cressie (1993) arrives at the following formula for the weights:

$$\boldsymbol{\lambda}^T = \left(\boldsymbol{\gamma} + \mathbf{1}_n \frac{1 - \mathbf{1}_n^T \Gamma^{-1} \boldsymbol{\gamma}}{\mathbf{1}_n^T \Gamma^{-1} \mathbf{1}_n}\right)^T \Gamma^{-1} \tag{4}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$, $\boldsymbol{\gamma} = (\gamma(\mathbf{x}_0 - \mathbf{x}_1), \ldots, \gamma(\mathbf{x}_0 - \mathbf{x}_n))^T$, $\mathbf{1}_n = (1, \ldots, 1)^T$, and Γ is a symmetric matrix with $\Gamma(i, j) = \gamma(\mathbf{x}_i - \mathbf{x}_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$, $\Gamma(i, j) = 1$ for $i = n + 1$ and $j = 1, \ldots, n$, $\Gamma(i, j) = 1$ for $j = n + 1$ and $i = 1, \ldots, n$, and $\Gamma(n + 1, n + 1) = 0$. The prediction variance σ^2 of the estimator at point $\mathbf{x}_0$ is given by

$$\sigma^2(\mathbf{x}_0) = \boldsymbol{\gamma}^T \Gamma^{-1} \boldsymbol{\gamma} - \frac{\left(\mathbf{1}_n^T \Gamma^{-1} \boldsymbol{\gamma} - 1\right)^2}{\mathbf{1}_n^T \Gamma^{-1} \mathbf{1}_n} \tag{5}$$

The weights and the prediction variance can not be solved for without estimating the variogram γ from the data. The equations for both the weights and the prediction variance only depend on the variogram function. In situations when the data set is small, incorrectly modeling the variogram can have a significant effect on the new predicted values. For an example of the importance of choosing the correct variogram model for fitting, consider a stochastic process Z of 256 points on a 16 by 16 grid in $\mathbb{R}^2$. Let Z be generated using an exponential variogram with a sill of 1, a nugget of 0 and a range of 9 (where 95% of the sill is reached). Figure 1 is a plot of one realization of such a stochastic process.

To estimate the variogram points from this realization, Matheron's classical estimator is used (Matheron, 1962). This unbiased estimator is given by

$$2\hat{\gamma}_M(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{N(\mathbf{h})} (Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2$$

where $N(\mathbf{h}) = \{(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i - \mathbf{x}_j = \mathbf{h}\}$ and $N_{\mathbf{h}}$ is the cardinality of $N(\mathbf{h})$. Since the estimated points do not necessarily satisfy the conditionally negative definiteness requirement for the variogram, a valid model is fitted to them by least squares. Figure 2 shows three variogram models, an exponential, a Gaussian, and a spherical, fitted to Matheron's estimates by ordinary least squares. Even though the data were generated from an exponential variogram, it is not clear that the exponential

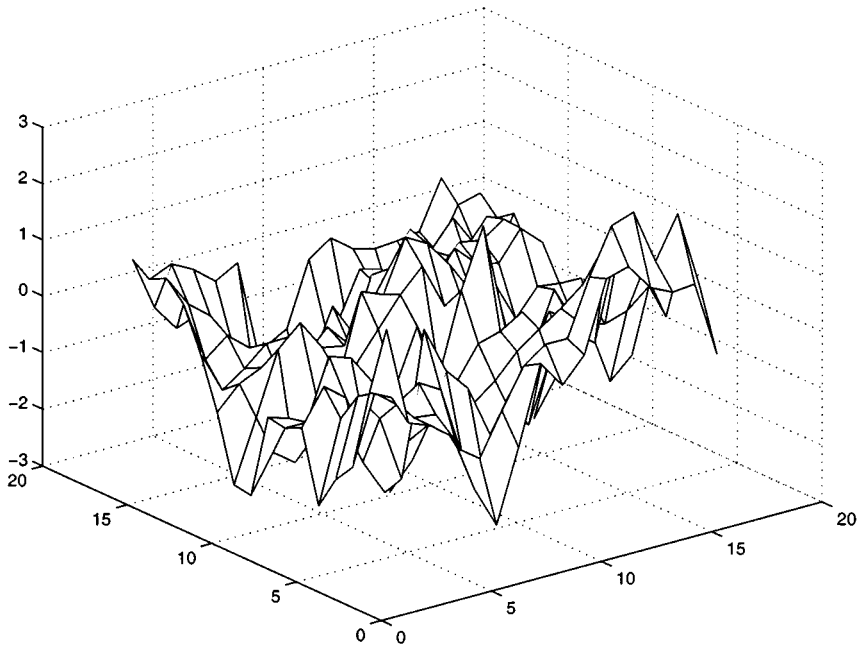

Figure 1. One realization of an isotropic stochastic process Z consisting of 256 regularly spaced points, with a mean of zero. The process is defined by an exponential variogram model with a range of 9, a sill of 1, and a nugget of 0.

variogram should be chosen instead of the Gaussian or the spherical. In fact, based on the residual sum of squared error of the exponential, Gaussian and spherical model least squares fits (0.2397, 0.2213, and 0.2158 respectively), the best model to choose for the variogram is not the exponential variogram which generated the data.

Although the three variograms fit Matheron's points closely, the effects on newly predicted points and on the predictor's variance can be large. Consider 15×15 newly predicted values at the midpoints of the known Z values. Should the exponential model be chosen, the maximum predictor variance is 0.210. In the case of the spherical model, the maximum variance is 0.506 and is 0.537 for the Gaussian model. Although the spherical and Gaussian models fit closer to Matheron's estimates, the predicted points for those models have a much larger variance.

Assume Matheron's variogram estimates are fitted with the Gaussian model. This gives a different set of predicted values than the exponential model would. The absolute value difference between the exponential prediction and the Gaussian prediction is shown in Figure 3. The average absolute difference is 0.290, and the

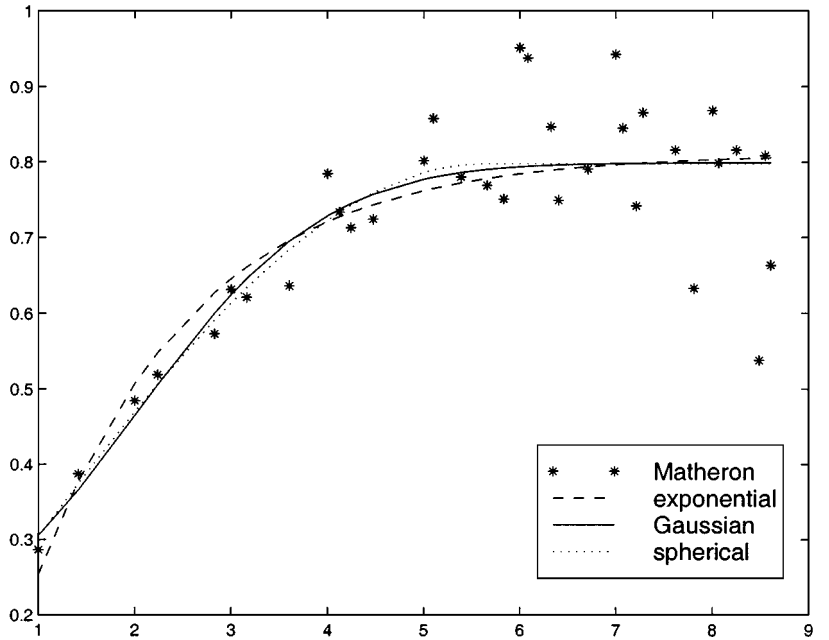

Figure 2. Matheron's variogram estimates are determined from the stochastic process in Figure 1. Exponential, spherical, and Gaussian variogram models are fitted equally well by ordinary least squares to the variogram points.

maximum difference is 1.156. Even though the fits between the Gaussian and exponential variograms are close in Figure 2, the predicted values at the midpoints are quite different. Even worse is the fact that the confidence bands are different. The 95% confidence bands for the Gaussian model are $Z(\mathbf{x}_0) \pm 1.96\sigma(\mathbf{x}_0)$, which are on average $Z(\mathbf{x}_0) \pm 1.436$. The exponential 95% confidence bands are only $Z(\mathbf{x}_0) \pm .897$ on average so that the real confidence intervals are much tighter.

Although the choice of the variogram model is critical for small data sets, misspecification of the model is not as important for larger data sets. Stein (1988, 1990) showed in a series of papers that an incorrect variogram model can be used to still achieve an optimal estimator in the asymptotic case where the number of data points realized becomes dense in the domain D (infill asymptotics). Unfortunately, it is rarely the case that the number of spatial locations is dense in the domain given, and we are interested in small data sets. Notice also that should the geometry of the data's spatial locations be changed, misspecification of the variogram could give even larger errors in confidence bands and predicted values.

The Gaussian, exponential, and spherical variogram models are only three of several commonly used models, depicted in Figure 4. Formulas for these parametric

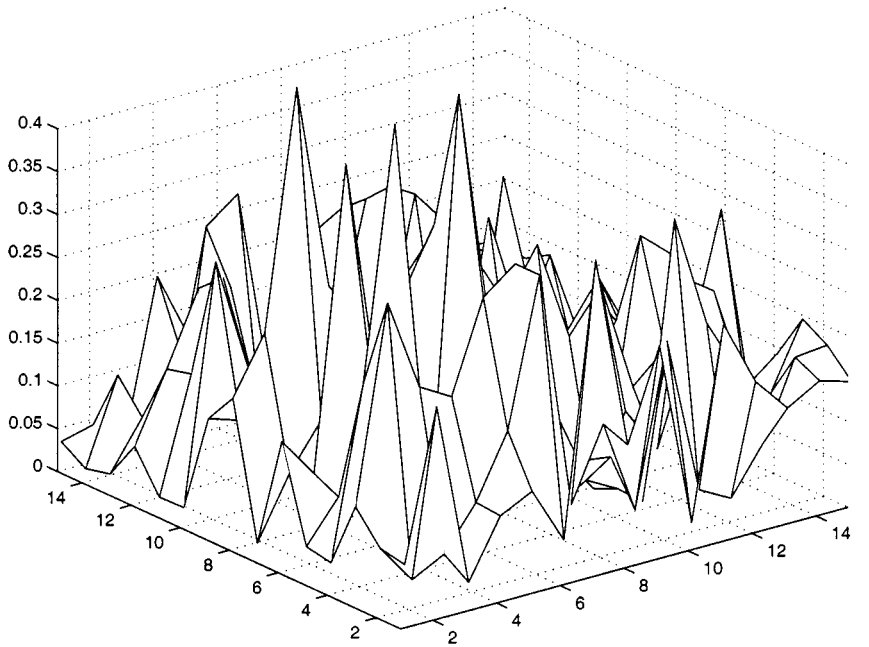

Figure 3. The absolute value of the difference between the predicted surfaces determined with the Gaussian and exponential variograms. The peak difference between the two surfaces is 1.156.

models can be found in Journel and Huijbregts (1978) and Cressie (1993). Their corresponding derivatives are shown in Figure 5. Most of the variogram models appear very similar, but their derivatives with respect to the lags are not. The main idea of this paper is to differentiate the nonparametric variogram estimates to help determine which parametric model is most suitable. Then, a least squares technique can be used to fit the selected parametric model, for example weighted least squares (Cressie, 1985) or generalized least squares (Cressie 1993; Genton 1998b). The parametric model can then give an estimate of the nugget, sill, and range, which are not well defined in the nonparametric fit. The next two sections discuss the nonparametric estimation of the variogram and its derivative.

NONPARAMETRIC VARIOGRAM ESTIMATION

In order to estimate the derivative without assuming a prior model, a nonparametric estimator is needed for the variogram that guarantees its conditional negative definiteness. Standard derivative estimators cannot be used directly since they are not guaranteed to be derivatives of a conditionally negative definite function. Nonparametric approaches to variogram estimation first appeared in Shapiro

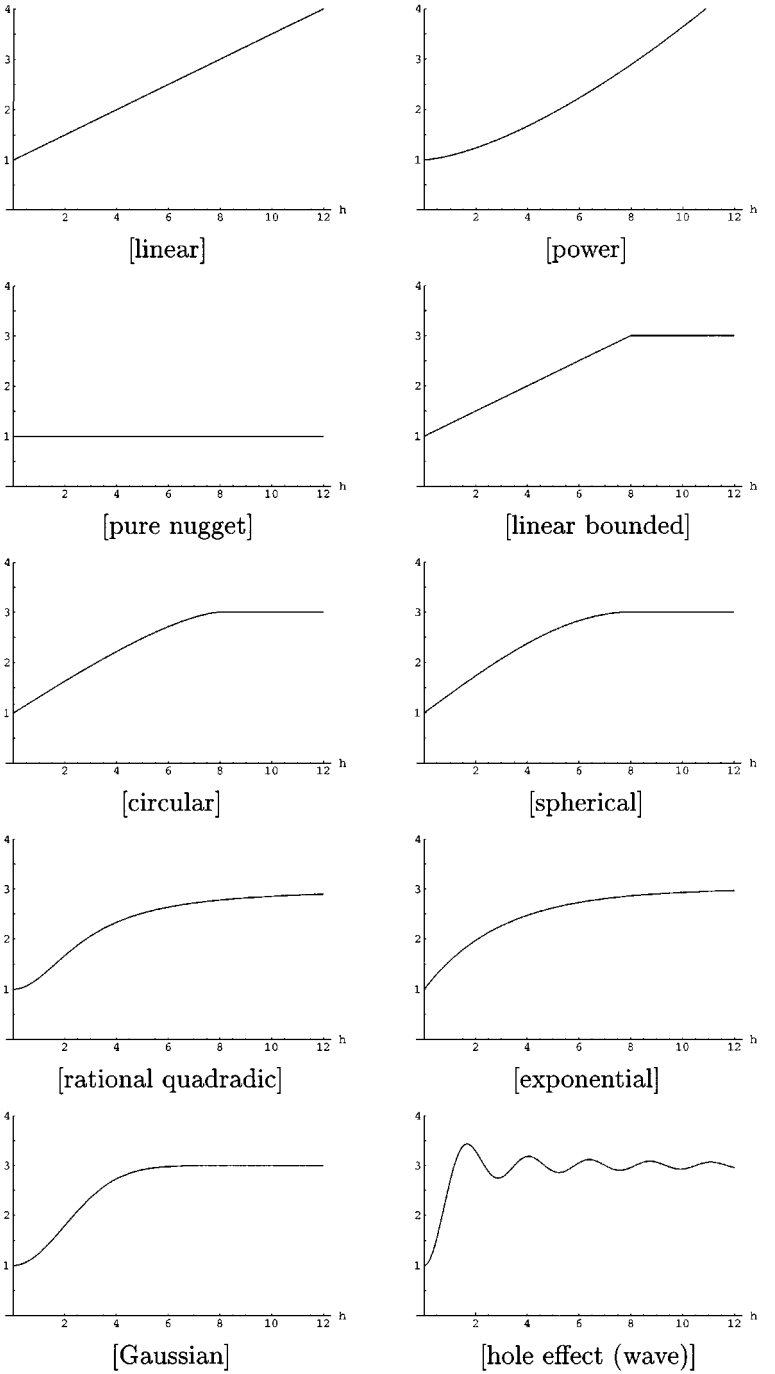

Figure 4. Some variogram models.

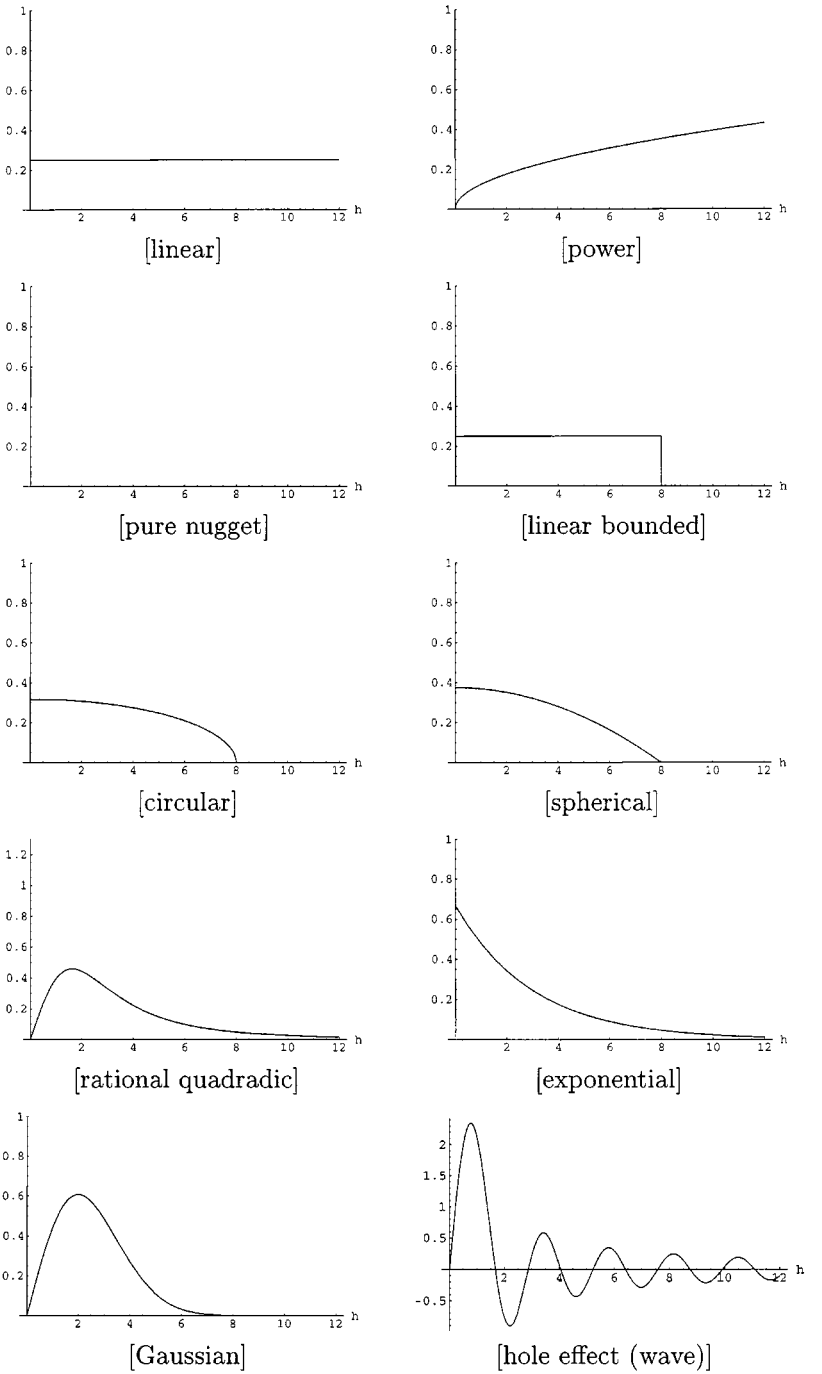

Figure 5. Derivatives of some variogram models.

and Botha (1991), and then later in Cherry and others (1996), and Cherry (1997). Barry and Ver Hoef (1996) fit a mixture of piecewise-linear variogram models to the empirical variogram. They prove that in $\mathbb{R}^1$, any continuous variogram with a sill can be approximated arbitrarily closely by piecewise-linear variograms. Ecker and Gelfand (1997) discuss variogram modeling in a Bayesian framework using an expected utility function (to be maximized) as a model choice criterion. The key idea behind a nonparametric estimator for the variogram is Bochner's theorem (Bochner, 1955). His theorem gives the spectral representation for any positive definite function. In particular, a covariance function $C(\mathbf{h})$ is positive definite if and only if it has the following form:

$$C(\mathbf{h}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \cos(\mathbf{u}^T\mathbf{h}) F(d\mathbf{u}) \tag{6}$$

where $F(d\mathbf{u})$ is a positive bounded symmetric measure. If $C(\mathbf{h})$ is isotropic, Bochner's theorem can be written as (Cressie, 1993)

$$C(h) = \int_0^{\infty} \Omega_r(ht) F(dt) \tag{7}$$

where Ω_r is a basis for functions in $\mathbb{R}^r$ given by

$$\Omega_r(x) = (2/x)^{(r-2)/2} \Gamma(r/2) J_{(r-2)/2}(x) \tag{8}$$

where $F(dt)$ is a nondecreasing bounded function, $\Gamma(r/2)$ is the gamma function, and J_v is the Bessel function of the first kind of order v. Some familiar examples of Ω_r are $\Omega_1(x) = \cos(x)$, $\Omega_2(x) = J_0(x)$, and $\Omega_3(x) = \sin(x)/x$. There is a considerable amount of freedom in the choice of r. The only requirement to maintain positive definiteness is that $r \geq d$, where d is the dimension of the spatial domain D. The effects of choosing different r on the estimator will be discussed later.

A spectral representation of the variogram is derived from Equation (7) by means of the relation:

$$\gamma(h) = C(0) - C(h) \tag{9}$$

To solve for $\gamma(h)$, we choose a vector $\mathbf{t}$, which represents the locations of the jump points in a discretization of $F(t)$. Let the length of $\mathbf{t}$, i.e., the number of jump points, be m and the size of each jump point p_j for $j = 1, \ldots, m$. For the simulations that follow we choose 260 jump points as $\mathbf{t} = [\pi/600 : \pi/130 : 2\pi]^T$, where $[a : b : c]^T$ is a column vector starting at a, ending at c, and with increments of b. The vector $\mathbf{t}$ should be chosen very carefully. The smallest value of $\mathbf{t}$ is critical, and so is the

largest one. The best jump points will depend on the problem, and the lags h that accompany it. More details on how $\mathbf{t}$ is chosen follows in the next section.

In order to find $\hat{\gamma}(h)$ from $\hat{\gamma}_M$, let $F(t) = \sum_{j=1}^{m} p_j \Delta(t - t_j)$, where Δ is the step function:

$$\Delta(t - t_j) = \begin{cases} 1, & \text{if } t \geq t_j, \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Given $\mathbf{t} = (t_1, \ldots, t_m)^T$ and $F(t)$, the nonparametric estimator has the form

$$\hat{\gamma}(h_i) = \sum_{j=1}^{m} p_j (1 - \Omega_r(h_i t_j)) \tag{11}$$

where $i = 1, \ldots, l$ is the lag number and the estimate of $C(0)$, the sill, is $\sum_{j=1}^{m} p_j$. This is a discrete version of the variogram giving its values only at h_i, the ith lag value found using Matheron's estimator. Assuming Equation (11) can be used to estimate the variogram at any lag, we have

$$\hat{\gamma}(h) = \sum_{j=1}^{m} p_j (1 - \Omega_r(h t_j)) \tag{12}$$

To find the jumps p_j we estimate the empirical variogram points from a realization of Z using Matheron's estimator. Therefore, using the points $\hat{\gamma}_M(h_i)$, p_j is estimated by minimizing S given by

$$S[\mathbf{p}] = \sum_{i=1}^{l} w_i \left(\hat{\gamma}_M(h_i) - \sum_{j=1}^{m} p_j (1 - \Omega_r(h_i t_j)) \right)^2 \tag{13}$$

or equivalently in matrix notation,

$$S[\mathbf{p}] = (\hat{\gamma}_M - M\mathbf{p})^T W (\hat{\gamma}_M - M\mathbf{p}) \tag{14}$$

where $M_{ij} = 1 - \Omega_r(h_i t_j)$, $\hat{\gamma}_M = (\hat{\gamma}_M(h_1), \ldots, \hat{\gamma}_M(h_l))^T$, $\mathbf{p} = (p_1, \ldots, p_m)^T$, and $W = \text{diag}(w_1, \ldots, w_l)$ is a weighting matrix that we assume to be the identity for simplicity. The estimator $\hat{\gamma}(h)$ can now be differentiated as is done in nonparametric kernel estimation (Härdle, 1989; Wand and Jones, 1995), or estimated with some other standard derivative estimators. In the next section, the problem of aliasing is shown to be a serious issue for the differentiation of the kernel $\Omega_r(ht)$. Finite differences of the nonparametric estimator avoid this problem.

NONPARAMETRIC VARIOGRAM DERIVATIVE ESTIMATION

A classical approach to estimate the derivative is to use the same jumps **p** and differentiate the kernel $\Omega_r(x)$ (Härdle, 1989; Wand and Jones, 1995). A standard derivative estimator would therefore be:

$$\hat{\gamma}' = M'\mathbf{p} \tag{15}$$

where M', the derivative of the matrix M with respect to h is defined by $M'_{ij} = -\Omega'_r(h_i t_j)$ and **p** is the same vector as before. The derivative estimate $\hat{\gamma}'$ is a vector giving derivative estimates only at discrete lags h_i, $i = 1, \ldots, l$. The way in which Ω_r is differentiated is very important near the origin where we are dividing $0/0$. The most stable way to perform the differentiation is to use the following relation:

$$\frac{\partial}{\partial x}\frac{J_v(x)}{x^v} = -\frac{J_{v+1}(x)}{x^v} \tag{16}$$

Therefore,

$$\frac{\partial}{\partial x}\Omega_r(x) = -\left(\frac{x}{r}\right)\Omega_{r+2}(x) \tag{17}$$

For example, for $\Omega_2(x)$ the derivative is $-J_1(x)$. If standard chain rules are used instead, the estimate of Ω'_r does not go to zero at the origin due to numerical roundoff errors.

In almost all cases, the differentiation of the kernel does not work well for estimating the variogram's derivative because of the basis Ω_r. For most data sets, D will contain irregularly spaced locations and there will typically not be enough samples per period to avoid aliasing. The basis functions $\Omega_r(x)$ are oscillatory in nature, and there will be an aliasing problem since the Nyquist sampling rate (Oppenheim and Schafer, 1989) will not normally be achieved. Assume the variogram γ is a bandlimited function that is sampled every T lags or with sampling frequency $1/T$ in samples per unit distance. This means $\gamma(h_i) = \gamma_c(iT)$, $i = 1, \ldots, l$ where γ_c is the continuous version of the variogram defined by Equation (12). Since γ is bandlimited, its does not contain frequencies above ω. To avoid aliasing, the function must be sampled with a sampling frequency of ω_s given by the Nyquist sampling theorem (Oppenheim and Schafer, 1989):

$$\omega_s > 2\omega \tag{18}$$

where $\omega_s = 2\pi/T$. If this inequality does not hold, high frequency artifacts known as aliasing are introduced into the estimation. As an example, consider the function

$\cos(\omega x)$. It does not contain frequencies above ω. In fact, it only contains the frequency ω. Let the sampling period be T so that sampling is regular. If $\omega = 1$, then there must be at least two samples within lags of π. The Fourier transform of $\cos(\omega x)$ is just two impulses at $\pm\omega$. When $\cos(\omega x)$ is sampled at discrete lags, the Fourier transform of the sampled function becomes shifted repeated copies of the original transform of $\cos(\omega x)$, i.e., shifted repeated copies of two impulses. If $\omega_s < 2\omega$ then repeated copies overlap causing aliasing and values of $\gamma_c(h)$ at locations other than h_i cannot be found. With aliasing, the reconstructed function from its samples is not $\cos(\omega x)$, but instead $\cos((\omega_s - \omega)x)$. The original function cannot be recovered. This means the function of higher frequency, $\cos(\omega x)$, is now the same as the lower frequency function of $\cos((\omega_s - \omega)x)$. In general, since the lags are not regularly spaced, the Fourier transform of the sampled variogram function will not be periodic, but aliasing will still occur.

The basis functions for the nonparametric estimator of the variogram are in fact asymptotically equivalent to a weighted $\cos(x)$ with some phase amount. The frequencies of the basis functions in Equation (8) can be seen through the asymptotics of the Bessel functions of order v, where $v = (r - 2)/2$. For large x, $x \gg (\frac{1}{2}v^2 - \frac{1}{8})$ (Arfken, 1985):

$$J_v(x) \approx \sqrt{\frac{2}{\pi x}} \cos\left(x - \left(v + \frac{1}{2}\right)\frac{\pi}{2}\right) \tag{19}$$

Therefore, for any r and all i, we require samples within lags of length $(h_{i+1} - h_i)t_j$ to be less than π. Of course, there is no way to force lags to always be under π. The values of h_i are fixed from Matheron's estimator. To prevent aliasing for the derivative, it is required for all i and j that $h_{i+1}t_j - h_i t_j < \pi$. Unfortunately, $\mathbf{t}$ can rarely be chosen to satisfy this. The largest value of $\mathbf{t}$ must be at least π in order that the higher frequency variations in γ can be fitted.

The aliasing issue does not appear in the nonparametric fitting, because the vector $\mathbf{p}$ is found using the aliased vectors, which still provide a valid basis for the variogram. However, this vector does not correspond to the aliased columns in the matrix M', and aliasing now becomes important.

As an example, consider bounded linear variograms of ranges 5, 10, 15, and lags $[0.5{:}1{:}39.5]^T$. Now to avoid aliasing, $t_j < \pi$ for all j must be satisfied. Assume that Matheron's estimator has given the variogram values perfectly, with sill $= 1$ and no nugget. This variogram is valid in $\mathbb{R}^1$ so the basis $\Omega_1(x) = \cos(x)$ can be used. Now the discussion of aliasing immediately applies with regularly spaced lags. The jump points $\mathbf{t}$ must be chosen sufficiently small to capture the low frequency content of the variogram. Variograms with larger ranges require smaller values of $\mathbf{t}$. With a range of 15, the linear bounded variogram requires at least some values of $\mathbf{t}$ smaller than $\pi/26$, whereas for ranges 10 and 5 the values are $\pi/23$ and $\pi/20$, respectively. With t_1 larger than those values, the sum squared

error in the nonparametric fit jumps by a factor of 10, and the error in the derivative will become much worse. Smaller values of t_1 only improve slightly on the error as long as the last value of $\mathbf{t}$ is held fixed. Note that t_1 must be larger than zero to avoid singular matrices. The fit works well for jumps at $\mathbf{t} = [\pi/40{:}\pi/320{:}\pi/2]^T$, but there are still small variations in the fit which are amplified by the derivative. To improve the fit further, higher frequencies are needed and $t_{\max}$ must be greater than $\pi/2$, which is not a problem for the nonparametric variogram at fixed h_i. For variograms like pure nugget, the maximum node value must be around 2π. For the bounded linear variogram example, the derivative is

$$\frac{\partial}{\partial h}\gamma(h) = \begin{cases} \text{sill/range}, & \text{if } h < \text{range}, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

To approximate the jump from sill/range to 0, high frequencies above $\pi/2$ are needed. The key thing to notice is that t_j plays the same role as ω, and in general, there will be aliasing in M. The two things that needs to be taken care of are choosing small t_1 and $t_{\max}$ in order that low frequencies are fitted and aliasing is reduced, and to have $t_{\max}$ large enough to provide a good fit to Matheron's points. Cherry and others (1996) choose $t_{\max} = 20$ for their simulations.

When the derivative is taken, high frequencies are amplified, and the need for larger $t_{\max}$ becomes more urgent. The magnitude of the frequency response of an ideal differentiator H is (Oppenheim and Schafer, 1989):

$$|H(\omega)| = |\omega| \tag{21}$$

for $|\omega| < \pi/T$ and zero otherwise. Convolution in time is the same as multiplication in the frequency domain. The higher frequencies that get amplified are exactly the columns with the worst aliasing in M and in general $M'\mathbf{p}$ will not provide a good fit to the derivative.

To avoid the aliasing problem, a new matrix $\tilde{M}'$ can be used that reflects the fact that only samples of γ have been estimated. Now, finite differences along every column of M are used instead. This is another approximation, but it reflects the true basis that was used to calculate $\mathbf{p}$ given the irregularly sampled points, and is minor in comparison to the aliasing. $\tilde{M}'$ is calculated by taking centered differences along every column of M and forward and backward differences for the first and last two rows of M. The matrix $\tilde{M}'$ more closely resembles the derivative of the basis that has been chosen and in general:

$$\tilde{M}' \neq M' \tag{22}$$

except in the first few columns where $h_{i+1}t_j - h_i t_j \ll \pi$. Figure 6 shows the sum squared error between the columns of $\tilde{M}'$ and M'. The error begins to jump significantly due to aliasing around $\pi/2$.

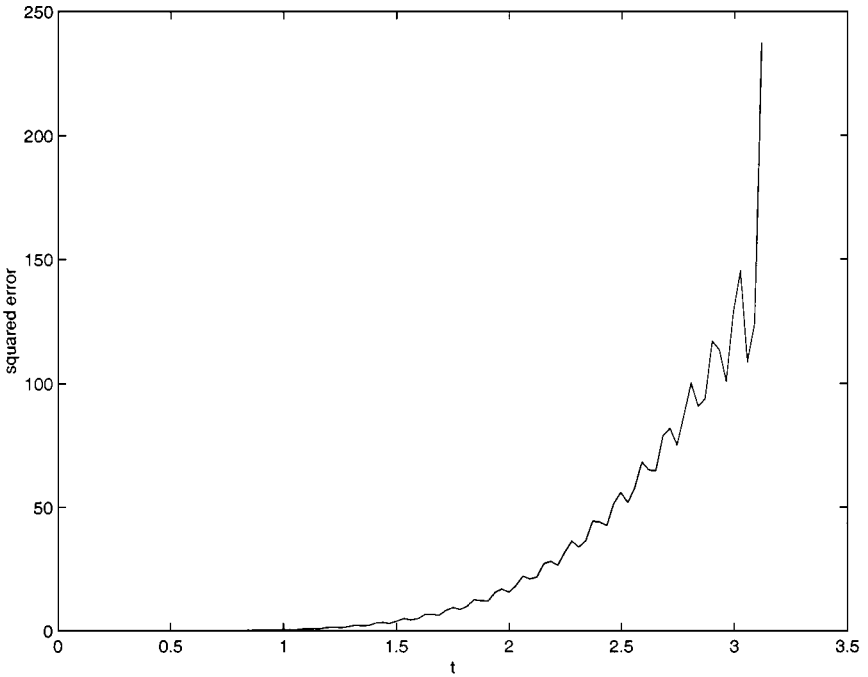

Figure 6. The squared error between columns of M' and $\tilde{M}'$ is displayed. Error comes from two sources: aliasing and the finite difference approximation. Error comes mainly from the finite differences.

Another good reason to use finite differences on M is that the basis Ω_r' goes to zero very strongly near small lags. Because of Equation (7), the value of the estimators at lag zero must be zero. At lag zero, $\Omega_r(0) = 1$ and $\Omega_r'(0) = 0$. Finite differences avoid this problem by only relying on Ω_r. For most data sets the nugget will not be zero. However, the derivative estimates close to the origin will not be affected because finite differences rely only on the first few lags, not on the nugget.

Since there is a great amount of freedom in the choice of r, and therefore the smoothness of the solution, there is no need to smooth the data by binning in Matheron's estimator. It is better to have more points to estimate $\hat{\gamma}$ so the bins are taken to be small. As long as $r \geq d$, the nonparametric variogram estimate will be conditionally negative definite, and will be the best fit to the points in a least squares sense. Choosing larger r gives smoother fits to the data. The reason for this was first proved by Schoenberg (1938) by demonstrating that Equation (7) is $(r-1)/2$ times differentiable. This implies that in higher dimensional spaces, there is a smoothing effect caused by the positive definiteness condition. When $r = \infty$ then $\Omega_\infty(x) = e^{-x^2}$. An example of the behavior of the estimator for r between 2

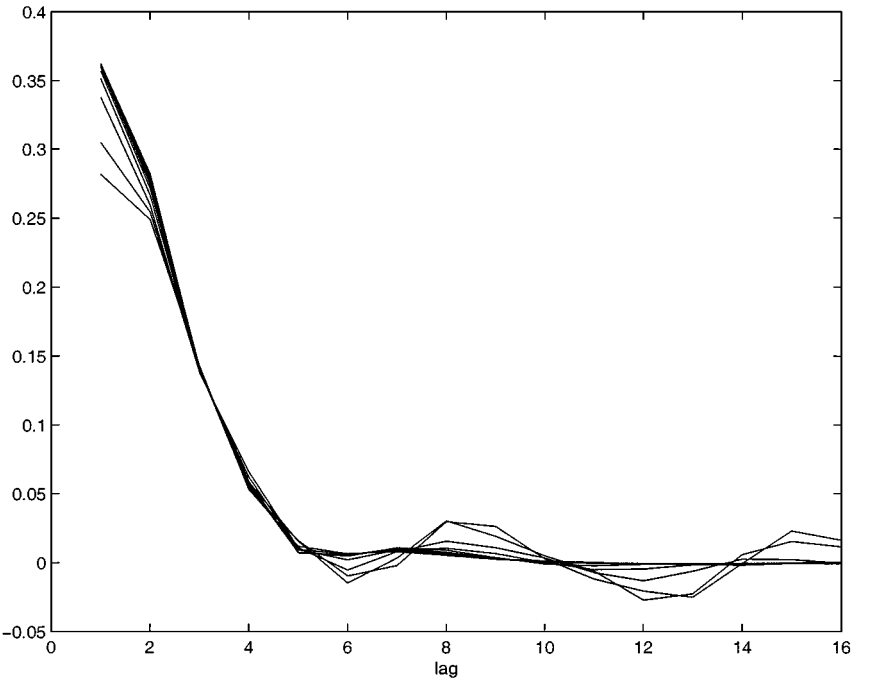

Figure 7. The nonparametric derivative estimation of a spherical variogram with r between 2 and 8. As r increases, the fit becomes smoother and the value of the derivative near zero decreases.

and 8 is shown in Figure 7. The nonparametric derivative fits quickly approach the smoothest fit, given by using Ω_∞ as a basis. Here the original data set was simulated with a spherical variogram, and the derivative in fact is very close to the true spherical derivative. Notice, as r increases from 2–8, the small scale variation near zero lag is lost and the derivative begins to look like it could be an exponential one. The derivative near small lags begins to increase with increasing r, but even at $r = 8$ the derivative falls off quickly at larger lags making it still closer to a spherical variogram model. It was found that values ranging from $d + 1$ to $d + 4$ were optimal for the simulations that were performed in this paper.

Another advantage to taking $r \geq d$ is that aliasing is reduced. As r increases, the basis becomes less like a periodic function and more like a Gaussian. The asymptotic approximation in Equation (19) holds for only larger and larger lags as r increases. Therefore r can be chosen to take the basis out of the range of the lags in a given problem. Although choosing r somewhat larger than d helps smooth out the estimates, taking r too large causes the matrices M and M' to become more and more singular, and also wipes out smaller variations in the data as seen in Figure 7. Oversmoothing can easily blur the distinction between variogram derivatives.

THE MATLAB SOFTWARE

For a tool to be useful, it should also be easy to use as well as an aid in the selection of a model. MATLAB has built in graphical user interface (GUI) tools that were used to build pop-up windows. One window for the estimates of the variogram, and the other for the estimates of the derivative. These estimates are displayed above some parametric models and their derivatives. Figures 8 and 9 display the GUI for the data presented in the introduction (Figures 1 and 2), simulated with an exponential variogram. Two windows appear, one with the

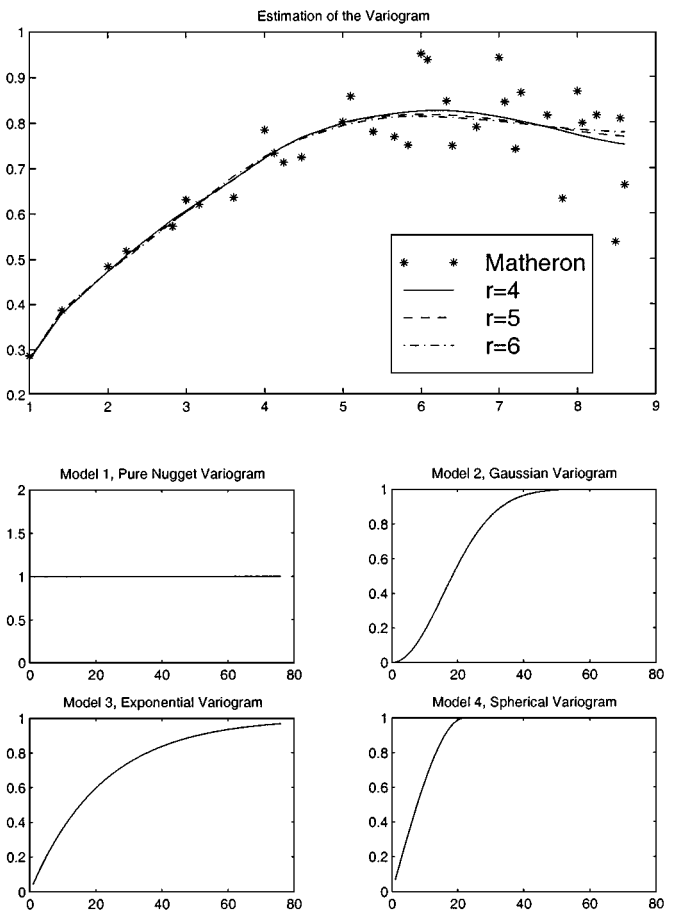

Figure 8. The graphical user interface displays the nonparametric fit for three values of r in the larger plot. The four smaller plots show the shapes or four parametric models of variograms.

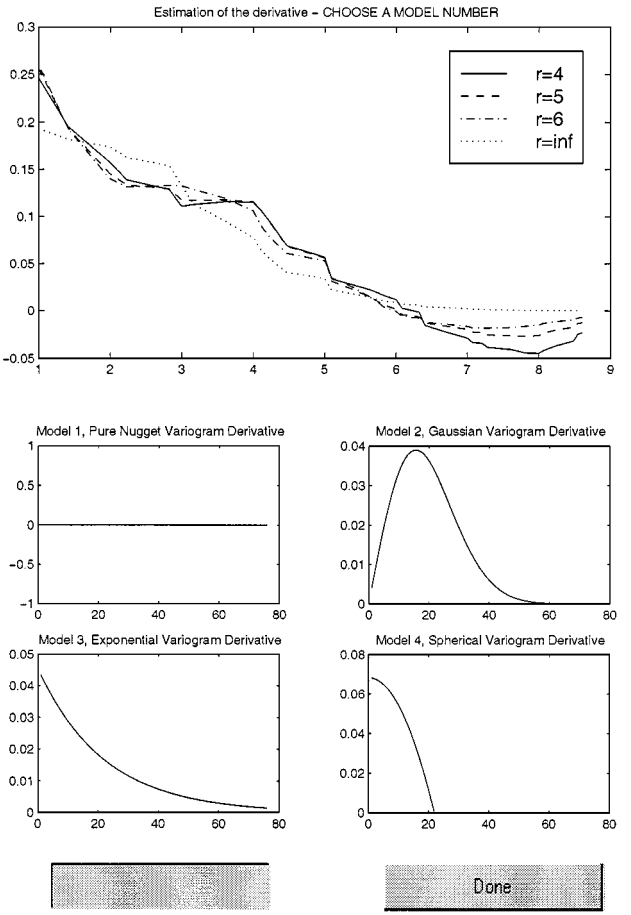

Figure 9. This graphic displays the nonparametric derivative estimates in the larger plot and shows the derivatives of four variogram models. The derivative estimates most closely follow the characteristics of an exponential variogram derivative.

nonparametric fits for $r = 4$, 5, and 6, and the other with the derivatives of those nonparametric fits. The derivative window also displays the fit for $r = \infty$. Four smaller windows below give a reference to the user of what the derivatives should look like for particular models. Recall that the selection of the variogram model was quite difficult in this case, and that one would not choose the exponential model. It can be seen in Figure 9 that the estimated derivative of the variogram suggests the exponential model and rules out the other ones. This is an example of the help that the variogram derivative can provide for model selection.

SIMULATIONS

In order to test our method, a small simulation study was performed in $\mathbb{R}^2$ with Gaussian, exponential, pure nugget and spherical variograms at various ranges, and using different values of r. A linear bounded variogram was not compared since it is only valid in $\mathbb{R}^1$. The Cholesky method was used to simulate the data on 20×20 grids in $\mathbb{R}^2$. The grids were regular with a total of 400 grid points. The covariance matrix Σ is factorized into two matrices: $\Sigma = LL^T$. Using this factorization, $\mathbf{z} = (Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n))^T$ is given by

$$\mathbf{z} = \boldsymbol{\mu} + L\boldsymbol{\epsilon} \tag{23}$$

where $\boldsymbol{\epsilon}$ is a vector of independent and identically distributed (i.i.d.) Gaussian random variables and $\boldsymbol{\mu} = \mu \mathbf{1}_n$ with $\mathbf{1}_n$ the column vector of ones.

Whether or not the estimated derivative could discriminate the variogram model correctly is determined by visual inspection. To determine if there is a benefit of having a derivative estimate, the computer first displays Matherons's points along with the nonparametric fits and asks the user what variogram was used to generate that particular stochastic process. The user responds without knowing what the underlying variogram is. After recording the results for 200 stochastic processes, the user again looks at the nonparametric fits of Matheron's points, but also has the derivative estimate displayed for $r = 3$ and 4. The computer again asks what the user thinks the underlying variogram is and the results are recorded. There are three key ways used to classify the differences between the models from the derivative. The first way is to observe whether the derivative is increasing or decreasing at small lags. This characteristic clearly separates the Gaussian model from the rest. The next characteristic to classify variograms is to observe how quickly the derivative falls to zero. The derivative of the spherical model falls off quickly, whereas the one of the exponential falls off very slowly. The derivative of the spherical model tends to be flat at small lags, whereas the one of the exponential model is rising towards lag zero. Note also that the derivative of the spherical model is concave, whereas the one of the exponential is convex. Finally for a pure nugget variogram, the derivative is close to zero and flat, which is easily distinguished from the others. These characteristics can be seen in Figure 5.

For the simulations, $\mathbf{t}$ is given by $\mathbf{t} = [\pi/600 : \pi/130 : 2\pi]^T$ for the 260 jump points. Choosing more than 300 jump points slows down simulations considerably, and only slightly improves the results. The simulation of 200 random processes is conducted with the variogram randomly chosen using a uniform distribution. There were a total of 53 spherical, 61 exponential, 46 Gaussian, and 40 nugget variograms used. The mean of $\mathbf{z}$ was fixed to zero, and the variograms all had a sill of 1 and a nugget of 0 except for the pure nugget variogram.

Table 1. Percentage of Correct Selections of Variogram Models[a]

Variogram type	Without derivative	With derivative
Exponential	$16/28 = 57\%$	$24/28 = 86\%$
Spherical	$14/24 = 58\%$	$19/24 = 79\%$
Gaussian	$7/17 = 41\%$	$15/17 = 88\%$
Pure nugget	$31/31 = 100\%$	$31/31 = 100\%$

[a] Variograms other than pure nugget have a sill of 1, a nugget of 0, and a range (or equivalent range) of 9. The pure nugget is 1. Each percentage is based on simulations.

Table 1 shows the percentage of correct classifications determined from the key characteristics of the four variograms with and without the derivative available. The range for the spherical variogram and the equivalent range for the exponential and Gaussian to reach 95% of the sill were chosen to be 9. In all cases except for the nugget, the derivative estimate aided in the selection of the underlying variogram model. Without the derivative the correct model was chosen only 54% of the time. When the derivative was used, the classification rate went to 84%, an increase of 30%. The nugget model is not included in these numbers since it is easy to distinguish from either the derivative or the nonparametric fits. For other ranges, the improvement in classification with the derivative is similar. The derivative does very well at helping the user's ability to classify the three variogram models correctly. If the model is not classified correctly with the derivative, it was rarely classified correctly without the derivative.

Although $r = 2$ could be used, there was too much variation in the estimates. Since no binning was performed on Matheron's estimates, a $r = 3$ or higher was used to smooth the estimates. Raising r greater than 6 only makes the small lag variation deteriorate. In a different simulation study it was found that varying r from 3 to 5 did not change the classification results much. The method can work better if there are data points between lags of length 1, but then matrices larger than 900×900 would be required that would take much more simulation time. The use of a highly robust variogram estimator (Genton, 1998a) can also improve the results. The derivative estimate using M' fails almost 75% of the time for all cases, but improves slightly on increasing r. Using large values of r, around 30 or 40, with M or M' was problematic due to numerical errors.

APPLICATION TO THE WALKER LAKE DATA SET

Our technique was also tested on a variogram from the Walker Lake data set. The Walker Lake area is in Nevada, in the western United States. The data set consists of elevation data over 260×300 grid. A complete analysis on a subset of

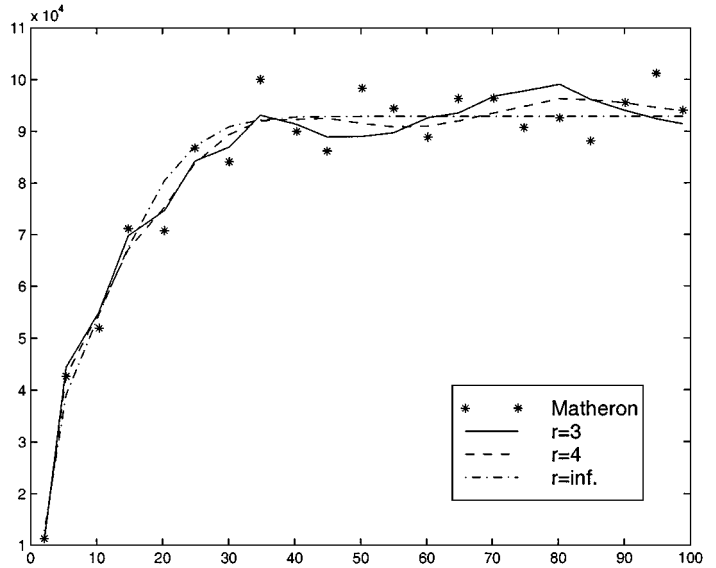

Figure 10. Matheron's variogram estimates on a subset of the Walker data set given by Isaaks and Srivastava (1989). The nonparametric variogram fits the points for both $r = 3$, $r = 4$, and $r = \infty$.

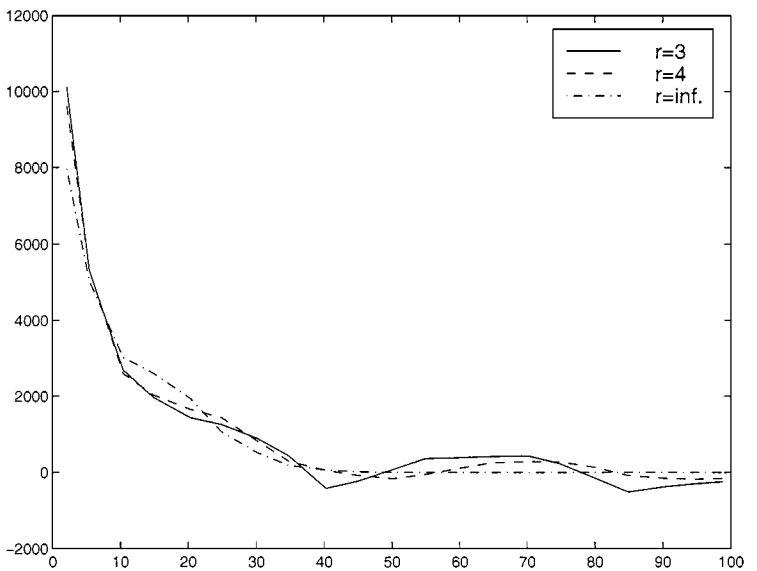

Figure 11. The nonparametric variogram derivative estimates for the Walker Lake data set for both $r = 3$, $r = 4$, and $r = \infty$. The derivative estimates most closely follow the characteristics of an exponential variogram derivative.

this data set can be found in Isaaks and Srivastava (1989). They randomly sampled the data set in the east-west direction at about 5 m apart and with an angular tolerance of 20°. Matheron's estimator was then used to generate 11 variogram estimates and are given in a table in the book. These points are plotted in Figure 10 along with the nonparametric estimates at $r = 3, r = 4$, and $r = \infty$. Figure 11 then shows the resulting derivative estimates that are shaped as an exponential derivative. The exponential model coincides with the model judged by the authors to be the best model to use for kriging.

CONCLUSIONS

In this paper, it has been shown that the derivative of the nonparametric variogram estimator can be used as a tool to aid in variogram model selection. In estimating the derivative, key issues of nonparametric estimation and aliasing were addressed. Simulations were performed with Gaussian, exponential, pure nugget, and spherical variograms, on two-dimensional grids. The Gaussian, exponential, and spherical variograms can be hard to distinguish unless the derivative is estimated, which makes the choice of the variogram model easier. The variogram model was selected correctly approximately 30% more often with the help of the derivative than without. A graphical user interface was developed as a tool to be used by practitioners who need additional insight for interpreting the estimated variogram points. MATLAB software for the GUI is available at http://www-math.mit.edu/~gorsich/derivative.html. Finally, an application to the Walker Lake data set was presented.

REFERENCES

Arfken, G., 1985, Mathematical methods for physicists: Academic Press, New York, 985 p.

Barry, R. P., and Ver Hoef, J. M., 1996, Blackbox kriging: spatial prediction without specifying variogram models: Journal of Agricultural, Biological and Environmental Statistics, v. 1, no. 3, p. 297–322.

Bochner, S., 1955, Harmonic analysis and the theory of probability: University of California Press, Berkley and Los Angeles, 176 p.

Cherry, S., 1997, Non-parametric estimation of the sill in geostatistics: Environmetrics, v. 8, p. 13–27.

Cherry, S., Banfield, J., and Quimby, W. F., 1996, An evaluation of a non-parametric method of estimating semi-variograms of isotropic spatial processes: Journal of Applied Statistics, v. 23, no. 4, p. 435–449.

Clark, I., 1979, Practical geostatistics: Applied Science Publishers, Essex, England, 129 p.

Cressie, N., 1985, Fitting variogram models by weighted least squares: Math. Geology, v. 17, p. 563–586.

Cressie, N., 1993, Statistics for spatial data: John Wiley & Sons, New York, 900 p.

Ecker, M. D., and Gelfand, A. E., 1997, Bayesian variogram modeling for an isotropic spatial process: Journal of Agricultural, Biological and Environmental Statistics, v. 2, no. 4, p. 347–369.

Genton, M. G., 1998a, Highly robust variogram estimation: Math. Geology, v. 30, no. 2, p. 213–221.

Genton, M. G., 1998b, Variogram fitting by generalized least squares using an explicit formula for the covariance structure: Math. Geology, v. 30, no. 4, p. 323–345.
Härdle, W., 1989, Applied nonparametric regression: Cambridge University Press, 333 p.
Isaaks, E. H., and Srivastava, R. M., 1989, An introduction to applied geostatistics: Oxford University Press, 561 p.
Journel, A. G., and Huijbregts, Ch. J., 1978, Mining geostatistics: Academic Press, London, 600 p.
Matheron, G., 1962, Traité de géostatistique appliquée, Tome I: Mémoires du Bureau de Recherches Géologiques et Minières, no. 14, Editions Technip, Paris, 333 p.
Oppenheim, A. V., and Schafer, R. W., 1989. Discrete-time signal processing: Prentice-Hall Signal Processing Series, Englewood Cliffs, New Jersey, 879 p.
Schoenberg, I. J., 1938, Metric spaces and completely monotone functions: Annals of Mathematics, v. 39, no. 4, p. 811–841.
Shapiro, A., and Botha, J. D., 1991. Variogram fitting with a general class of conditionally nonnegative definite functions: Computational Statistics and Data Analysis, v. 11, p. 87–96.
Stein, M., 1988, Asymptotically efficient prediction of a random field with a misspecified covariance function: Annals of Statistics, v. 16, no. 1, p. 55–63.
Stein, M., 1990, Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure: Annals of Statistics, v. 18, no. 2, p. 850–872.
Wand, M. P., and Jones, M. C., 1995, Kernel smoothing: Chapman & Hall, 212 p.