Bayesian Variogram Modeling for an Isotropic Spatial Process

Mark D. Ecker* and Alan E. Gelfand**

June 26, 1997

Abstract

The variogram is a basic tool in geostatistics. In the case of an assumed isotropic process, it is used to compare variability of the difference between pairs of observations as a function of their distance. Customary approaches to variogram modeling create an empirical variogram and then fit a valid parametric or nonparametric variogram model to it.

Here we adopt a Bayesian approach to variogram modeling. In particular, we seek to analyze a recent data set of scallop catches. We have the results of the analysis of an earlier data set from the region to supply useful prior information. In addition, the Bayesian approach enables inference about any aspect of spatial dependence of interest rather than merely providing a fitted variogram. We utilize discrete mixtures of Bessel functions which allow a rich and flexible class of variogram models. To differentiate between models, we introduce a utility based model choice criterion that encourages parsimony. We conclude with a fully Bayesian analysis of the scallop data.

Key Words & Phrases: Bessel functions, correlation functions, importance sampling, mixtures, model determination, stationary process
*Mark D. Ecker is Assistant Professor in the Department of Mathematics, University of Northern Iowa, Cedar Falls, IA 50614-0506 and **Alan E. Gelfand is Professor in the Department of Statistics, University of Connecticut, Storrs 06269-3120. The work of the second author was supported in part by NSF DMS 96-25383. The authors acknowledge the associate editor and three referees for insightful comments on an earlier version of the manuscript which substantially improved the revision.
1 Introduction

Modeling the variogram is at the heart of geostatistics. The variogram expresses the variability of a spatial process as a function of distance and direction. In particular, suppose the data are collected according to a random spatial process in \( \mathbb{R}^n \), i.e. we observe \( Y(s_i), s_i \in \mathbb{R}^n \). It is assumed that \( \text{Var}(Y(s_i) - Y(s_j)) \) is only a function of the separation vector \( s_i - s_j \). If, it depends only upon \( \|s_i - s_j\| \), the distance between sites, the process is said to be *isotropic* and the function \( 2\gamma(\|s_i - s_j\|) = \text{Var}(Y(s_i) - Y(s_j)) \) is called the *variogram*. Typically, the variogram is assumed to increase in \( \|s_i - s_j\| \), the rationale being that the differences between pairs of observations closer in space should tend to exhibit less variability than that for pairs farther apart.

The variogram only specifies a single feature of the spatial process \( Y(s) \). To implement likelihood and/or Bayesian inference requires specification of the joint distribution of an arbitrary set \( Y(s_1), Y(s_2), \ldots, Y(s_N) \). This is typically done by assuming \( Y(s) \) to come from a stationary Gaussian process with constant mean and isotropic covariance function \( \text{Cov}(Y(s_i), Y(s_j)) = C(\|s_i - s_j\|) \). Then \( C \) determines the variogram \( 2\gamma \), i.e., \( 2\gamma(\|s_i - s_j\|) = 2(C(0) - C(\|s_i - s_j\|)) \). In general, \( \gamma \) does not determine \( C \) since stationary increments for \( Y(s_i) - Y(s_j) \) (the so-called *intrinsic hypothesis*) do not imply that \( Y(s) \) is stationary. In fact, \( C(0) \) need not exist. We work in the former context here using a rich class of valid covariance functions yielding a rich class of valid variograms. This contrasts with the usual practice of selecting “safe” parametric variogram models such as the exponential, Gaussian and spherical forms, i.e., those known to be associated with positive definite covariance functions.

Once a parametric model is chosen for \( \gamma \) or \( C \), techniques such as maximum likelihood (Cressie, 1993, section 2.6.1), weighted least squares (Cressie, 1985) or fitting by inspection are usually used to estimate the model parameters. We adopt a Bayesian approach for several reasons. First, for the data set of interest, we want to use the information gained through a previously studied data set from the region. No current variogram modeling technique routinely incorporates prior information. Second, the Bayesian paradigm allows reliable estimates of the variability of the variogram parame-
ters. Maximum likelihood techniques allow only asymptotic estimates. Weighted least squares estimates are dependent upon the arbitrary construction of the variogram (see Cressie, 1993, section 2.6.2). Fitting by inspection is not sound statistical practice and provides no estimation of variability. Third, the Bayesian paradigm provides an entire posterior distribution for each variogram parameter avoiding possibly inappropriate approximate normality assumptions. In addition, a fully Bayesian analysis allows inference (again an entire posterior distribution) for any aspects of the variogram of interest, e.g., the sill, the nugget, and the range. It also allows a confidence band for the variogram itself. In return for a full parametric distributional specification for the spatial process \( Y(s) \), full inference is available. If one is only prepared to specify a parametric form for \( \gamma \) and then fit it to a particular empirical variogram, inference is limited to a point estimate of \( \gamma \).

Our work is motivated by two data sets consisting of scallop counts off the New Jersey and Long Island coastline. Since 1982, the Northeast Fisheries Center of the National Marine Fisheries Service in Woods Hole, Massachusetts has annually sampled on the continental shelf off the Northeastern United States to estimate the abundance of sea scallops and other shellfish. Their methodology is to stratify the region from the Georges Bank to Cape Hatteras based upon water depth and latitude. One geographical region of interest is the New York Bight which encompasses an area in the Atlantic Ocean from the mouth of the Delaware River to the eastern tip of Long Island. A total of 148 sites, sampled in 1990 in the New York Bight, were previously analyzed in Ecker and Heltshle (1994) and are readily available as part of the Splus SpatialStats version 1.0 module. Subsequently in 1993, 147 different locations were sampled in this region. Figure 1 displays the sampled locations for both 1990 and 1993 while Figure 2 depicts smoothed contours of the log-scaled scallop counts for the 1993 data.

The class of all permissible variogram models in \( \mathbb{R}^n \) has been characterized (see Schoenberg, 1938, Theorem 1). In \( \mathbb{R}^2 \), mixtures of Bessel functions of the first kind of order zero define the class of allowable covariance structures. Thus, we propose finite mixtures of such functions to provide an arbitrarily rich covariance specification, hence a very flexible class of variogram models. Similar strategies in the literature include Samp-
son and Guttorp (1992) who use mixtures of Gaussian type correlations and Shapiro and Botha (1991) and Cherry, Banfield and Quimby (1996) who consider discrete mixtures of Bessel functions when \( n = 2 \). However, all of this work uses least squares fitting of these mixture models to an arbitrarily constructed empirical variogram. Instead, we insert these forms into the likelihood, enabling inference conditional on all the observed data, not on a particular summary. We obtain full inference about the fitted model rather than merely an estimated variogram.

One aspect of variogram modeling infrequently addressed in the literature is which of the fitted models best explains the data. Often, it is argued that a particular variogram specification is appropriate for a particular form of spatial data. Such claims are bolstered by practical experience. While empirical feel is invaluable, it would be useful to provide firm support for choosing one variogram model over another. Such comparison should be done through suitable predictive performance of a variogram model. A classical approach uses a goodness of fit criteria perhaps with a penalty for model dimension as in the Akaike Information Criterion (AIC) (1973) or the Schwartz or Bayesian information criterion (BIC) (1978). See Webster and McBratney (1989) for an adaptation of the AIC and BIC methodologies to variogram modeling. Under the Bayesian perspective, we adopt the utility based approach of Gelfand and Ghosh (1997) which minimizes a so-called balanced expected loss with respect to the posterior predictive distribution. Such a loss penalizes actions which depart from what we have observed, but also from what we expect to observe under the model. In the context of variogram estimation, a related point arises. The signal (the true variogram) is often accompanied by enormous noise (variability about the true variogram) making virtually any plausible variogram model fit poorly. We devise a method to sharpen model choice given this difficulty.

In section 2, we review standard variogram models and describe the relationship between the variogram and the correlation structure of the process. In section 3, we examine the class of all valid variogram models in \( \mathbb{R}^2 \). The computationally intensive Bayesian model-fitting procedure, outlined in section 4, is used to estimate the model parameters. In section 5, the aforementioned model choice criterion is developed. We also clarify the signal to noise problem and suggest a remedy. Section 6 analyzes the
scallops data mentioned above while section 7 provides conclusions.

2 Isotropic Variogram Models

For sites $s_1, s_2, \ldots, s_N$ in $R^n$, let $y = (Y(s_1), Y(s_2), \ldots, Y(s_N))^\top$ be the response vector. Under the intrinsic hypothesis of Matheron (1963), the increments $Y(s_i) - Y(s_j)$ are stationary, i.e.,

$$E(Y(s_i) - Y(s_j)) = 0$$

$$\text{Var}(Y(s_i) - Y(s_j)) = 2\gamma(s_i - s_j).$$

(1)

When $\gamma$ is invariant to rotations of the spatial coordinates (i.e., the spatial variability is the same for any direction), the process is said to be isotropic in which case $\gamma$ depends only upon $d_{ij} = \sqrt{\sum_{\ell=1}^n h_{ij}^2}$, the Euclidean distance between sites $s_i$ and $s_j$. Here, $h_{ij}$ denotes the distance between sites $s_i$ and $s_j$ in the $\ell$th dimension and $2\gamma(d_{ij})$ is called the variogram of the process while $\gamma(d_{ij})$ is the semivariogram. If we further assume that $y$ is second-order stationary, then $E(Y(s_i)) = \mu$ and $\text{Cov}(Y(s_i), Y(s_j)) = C(d_{ij}) < \infty$ is also a function of the Euclidean distance between sites.

The fitting of models for $\gamma(d_{ij})$ takes two routes. The first is a parametric version that makes a distributional assumption on the process such as normality (or transformable to normality). If the multivariate normal covariance structure of the data is only a function of $d_{ij}$, then second-order stationarity immediately follows and modeling the variogram is equivalent to modeling the covariance. Likelihood based techniques can be used to estimate model parameters and to compare models.

The second or nonparametric methodology imposes no distribution on $y$, hence, second-order stationarity need not follow. This approach views variogram modeling as a curve fitting exercise. For each of the $\frac{N(N-1)}{2}$ pairs of data points in $R^n$, a plot of $\frac{1}{2}(Y(s_i) - Y(s_j))^2$ versus $d_{ij}$ is termed the semivariogram cloud. The semivariogram cloud for the 1993 scallop data appears in Figure 3a. The Matheron (1963) estimator of the semivariogram,

$$\gamma^*(d_{ij}) = \frac{1}{2N_Br} \sum_{\{(i,j) : d_{ij} \in B_r\}} (Y(s_i) - Y(s_j))^2$$

(2)
smoothes the semivariogram cloud by aggregating distances into $R$ sets $B_1, B_2, \ldots, B_R$. $B_r = \{d_{ij} : b_{r-1} \leq d_{ij} < b_r\}$ for constants $b_{r-1}$ and $b_r, r = 1, \ldots, R$ with $b_0 = 0$ and $N_{B_r}$ equals the number of $d_{ij}$ in $B_r$ with $\sum_r N_{B_r} = \frac{N(N-1)}{2}$. Usually, the constants $b_r$ are chosen to be $r\delta$ where $\delta$ is a specified lag. If the data are systematically sampled, then the choice of $\delta$ is clear. A plot of $\gamma^*$ versus the midpoint of the interval $(b_{r-1}, b_r)$ for $r = 1, \ldots, R$ is called the Matheron empirical semivariogram. It need not be a valid semivariogram; that is, $\gamma^*$ need not be conditionally non-negative definite (see e.g. Armstrong and Diamond (1984) and Christakos (1984)). Other variogram estimators exist (see David, 1988, pp 41-42) but the Matheron estimator is the most frequently used in practice. The Matheron empirical semivariogram for the 1993 scallop data appears in Figure 4.

A parametric specification for $y$ customarily assumes that it is second-order stationary and normally distributed. Then we can write as in, e.g., Diggle, Liang and Zeger (1994, p 87),

$$y \sim N(\mu, \Sigma(\alpha))$$

(3)

where $\alpha = [\tau^2 \sigma^2 \phi]'$ and $\Sigma(\alpha) = \tau^2 I + \sigma^2 H(\phi)$ with $(H(\phi))_{ij} = \rho(d_{ij}, \phi)$ a valid parametric correlation function depending upon the distance between sites $s_i$ and $s_j$. Examples of standard parametric forms of $\rho(d, \phi)$ are given in Table 1 where $\phi$ becomes a scalar capturing the rate of correlation decay.

For the scallop data described in the introduction, we take the response $Y(s_i)$ is $\log(\text{total catch at } s_i + 1)$ where the constant one is added to address the observed zero catches. (We assume that zero catches do not arise because the site is unsuitable for scallop habitation. Under the latter assumption a different modeling approach incorporating a point mass at zero might be appropriate.) Both Ecker and Heltshe (1994) and Kaluzny et al (1996) model the 1990 log transformed scallop catches as approximately normally distributed with justification by exploratory data analysis. With non-geographical covariates, one can generalize the mean structure in (3) to a linear form, $E(Y(s_i)) = \sum_{j=0}^p X(s_i)j_0 \beta_j$ with $X(s_i)0 = 1$. With only geographical covariates (as in the case of the scallop data), the mean is often modeled as a trend surface, a polynomial in the spatial locations. Tension arises when capturing spatial effects through both the
Table 1: Common Parametric Correlation Forms

<table>
<thead>
<tr>
<th>Name</th>
<th>$\rho(d, \phi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$\exp(-\phi d)$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$\exp(-\phi d^2)$</td>
</tr>
<tr>
<td>Cauchy</td>
<td>$\frac{1}{1 + \phi d^2}$</td>
</tr>
<tr>
<td>Bessel</td>
<td>$J_0(\phi d)$</td>
</tr>
</tbody>
</table>
| Spherical | $\frac{1}{2}(\phi^3 d^3 - 3\phi d + 2)$ if $d \leq \frac{1}{\phi}$
              | $0$ if $d > \frac{1}{\phi}$

mean and the variance structure (see, e.g., Journel and Rossi, 1989). Additionally, a nonconstant mean complicates the notion of a variogram, both theoretically and empirically. Hence, we adopt a constant mean in the sequel; a general linear mean structure in (3) can be handled similarly.

Because $y$ is second-order stationary, the variogram is bounded and the relationship between the variogram and the covariance is given by

$$2\gamma(d, \alpha) = 2\gamma(d, \tau^2, \sigma^2, \phi) = 2(\tau^2 + \sigma^2(1 - \rho(d, \phi))).$$  \hfill (4)

The parameters of the variogram $\tau^2, \sigma^2$ and $\phi$ have a geostatistical interpretation. If $\rho(d, \phi) \to 0$ as $d \to \infty$, then the sill of the variogram is defined to be $\lim_{d \to \infty} 2\gamma(d) = 2(\tau^2 + \sigma^2)$ and represents the stationary variance of the process. Some parametric forms, such as the spherical, achieve their sill for finite $d$, while others, e.g., the exponential, Gaussian and Cauchy (rational quadratic), reach their sills asymptotically. For monotonically increasing variograms that reach their sill exactly, the range is defined to be the distance at which the process reaches its sill or equivalently the distance at which $\rho(d, \phi)$ becomes zero. Intuitively, points separated by distances greater than the range are spatially uncorrelated. For asymptotically silled variograms, two points will only be spatially uncorrelated in the limit as $d \to \infty$. Here, we speak of the effective range, a notion which is not uniquely defined. We note two possible definitions. One facilitates interpretation in the variogram space of the process and the other in the correlation or
covariance space. The first definition (McBratney and Webster, 1986, p 623) defines the range, \( r_C \), as the distance \( d \) where the variogram reaches \( 2(\tau^2 + 0.95\sigma^2) \). Thus, the range \( r_C \) is achieved when the serial correlation of the process is 5\%. Mathematically, \( r_C \) solves \( \rho(r_C, \phi) = 0.05 \). A second definition (Cressie, 1993, pp 67-68) defines the range, \( r_V \) to be the distance where the variogram reaches 95\% of its sill. Then, \( r_V \) solves \( \rho(r_V, \phi) = 0.05 \left( \frac{\tau^2 + \sigma^2}{\sigma^2} \right) \). Note that \( r_V \) would not exist if \( \frac{\tau^2 + \sigma^2}{\sigma^2} > 20 \), but this would be unlikely in practice. For the asymptotically silled variograms given in Table 1, the relationship between the scalar correlation decay parameter \( \phi \) and the ranges \( r_C \) and \( r_V \) are presented in Table 2. It is obvious that \( r_V \leq r_C \) with equality if \( \tau^2 = 0 \). Many authors adopt \( r_C \) and parameterize the correlation decay \( \phi \) to be the range \( r_C \). For nonmonotonic variograms such as the Bessel, the range is not defined. The nugget of the variogram is \( \lim_{d \to 0} 2\gamma(d) = 2\tau^2 \) which need not be zero due to measurement error and/or a microscale effect resulting from extrapolating the variogram from the minimum sampled distance, \( \min(d_{ij}) \), to the origin.

### Table 2: Range Definitions for Asymptotic Correlation Forms

<table>
<thead>
<tr>
<th>Name</th>
<th>( r_C )</th>
<th>( r_V )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>( \frac{3}{\phi} )</td>
<td>( \frac{3}{\phi} ) ( \frac{\log\left( \frac{\phi}{\phi} \right)}{\phi} )</td>
</tr>
<tr>
<td>Gaussian</td>
<td>( \frac{3}{\phi} )</td>
<td>( \frac{3}{\phi} ) ( \frac{\log\left( \frac{\phi}{\phi} \right)}{\phi} )</td>
</tr>
<tr>
<td>Cauchy</td>
<td>( \frac{19}{\phi} )</td>
<td>( \frac{19}{\phi} ) ( \frac{\log\left( \frac{\phi}{\phi} \right)}{\phi} )</td>
</tr>
</tbody>
</table>

3 Extensions of Parametric Variogram Forms

A correlation function \( \rho(d, \phi) \) is permissible (valid) only if it is positive definite in \( d \), \( \rho(0, \phi) = 1 \) and \( |\rho(d, \phi)| \leq 1, \forall d \). From Bochner’s theorem (Cressie, 1993, p. 84), the characteristic function of a symmetric distribution in \( R^n \) satisfies these constraints. For example, since \( \exp(-\phi d^2) \) is the characteristic function of a \( N_n(0, 2\phi I) \) random variable where we emphasize \( d = \sqrt{\sum_{i=1}^{n} h_i^2} \) for \( h = [h_1, h_2 \ldots h_n] \), the Gaussian variogram is valid
in any dimension. Feller (1966, p 476) shows that the exponential and Cauchy among
others are permissible in $\mathbb{R}^1$. For validity of the spherical and Bessel, see Montoglou and
477) states that a convex combination of characteristic functions is itself a characteristic
function, extending $\rho(d, \phi)$ to mixture forms. For transect data in $\mathbb{R}^1$, the Pólya criterion
(Chung, 1974, pp 182-183) provides sufficient conditions for recognizing permissible forms
of $\rho(d, \phi)$. Christakos (1984) and Armstrong and Diamond (1984) provide rigorous
criteria for testing the validity of an arbitrary form $\rho(d, \phi)$ in $\mathbb{R}^n$.

The role of dimensionality deserves mention. If $\rho(d, \phi)$ is valid in $\mathbb{R}^n$, then for any
$n_2 < n_1$, $\rho(d, \phi)$ is valid in $\mathbb{R}^{n_2}$. For example, Feller (1966, p 497) shows that $\exp(-\phi d)$
where $d = \sqrt{h_1^2 + h_2^2}$ is the characteristic function of the bivariate Cauchy distribution.
Hence, $\exp(-\phi d)$ is immediately valid in $\mathbb{R}^1$ where $d = \sqrt{h_1^2}$. The converse in not
necessarily true. A counterexample (Armstrong and Jabin, 1981) is the tent correlation
function, $\rho(d, \phi) = \begin{cases} 
\phi d & 0 \leq d \leq \phi^{-1} \\
0 & d > \phi^{-1} 
\end{cases}$ which is valid in $\mathbb{R}^1$, but not in $\mathbb{R}^2$.

From Khinchin’s Theorem (Yaglom, 1987, p 106), the class of all valid functions
$\rho(d, \phi)$ in $\mathbb{R}^n$ can be expressed as (also see Schoenberg, 1938)

$$
\rho(d, \phi) = \int_0^\infty \Omega_n(zd) \, dG_\phi(z)
$$

(5)

where $G_\phi$ is nondecreasing integrable and $\Omega_n(x) = (2^{2v})^{n_2-1} \Gamma\left(\frac{n}{2}\right)J_v(\frac{\pi x}{2})$. Here, $J_v(\cdot)$
is the Bessel function of the first kind of order $v$. For $n = 1, \Omega_1(x) = \cos(x)$; for $n = 2, \Omega_2(x) = J_0(x)$; for $n = 3, \Omega_3(x) = \frac{\sin(x)}{x}$; for $n = 4, \Omega_4(x) = \frac{2}{\pi}J_1(x)$ and
for $n = \infty, \Omega_\infty(x) = \exp(-x^2)$. Specifically, $J_0(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} (\frac{x}{2})^{2k}$ and $\rho(d, \phi) = 
\int_0^\infty J_0(zd) \, dG_\phi(z)$ provides the class of all permissible variogram models in $\mathbb{R}^2$.

In practice, a convenient simple choice for $G_\phi(z)$ is a step function that assigns
positive mass (jumps or weights) $w_\ell$ at points (nodes) $\phi_\ell$, $\ell = 1, \ldots, p$ yielding, with
$w = (w_1, w_2, \ldots, w_p)$,

$$
\rho(d, \phi, w) = \sum_{\ell=1}^{p} w_\ell \Omega_n(\phi_\ell d).
$$

(6)

The forms in (6) are referred to as nonparametric variogram models in the literature
because a standard or parametric form for $\rho(d, \phi)$ i.e., a form given in table 1, is not
specified. Thus, the functional form for $\rho$ (hence $\gamma$) can be chosen either parametrically (from table 1) or nonparametrically (of the form (6)). This is a separate issue from selecting a parametric or nonparametric methodology for parameter estimation as discussed in section 2. Both Sampson and Guttorp (SG) (1992) and Shapiro and Botha (SB) (1991) use a step function for $G_{\phi}$. Cherry, Banfield and Quimby (1996) further evaluate the SB method. Barry and Ver Hoef (1996) employ a mixture of piecewise linear variograms in $\mathbb{R}^1$ and piecewise-planar models for sites in $\mathbb{R}^2$. Hall, Fisher and Hoffmann (1994) transform the problem from choosing $\phi_\ell$’s and $w_\ell$’s in (6) to determining a kernel function and its associated bandwidth. Also, Lele (1995) proposes an iterative nonparametric variogram estimation technique based upon spline smoothing of the variogram and approximating these results by a positive definite covariance matrix. His resulting covariance structure is not easily seen to be of the form (5). Most of these nonparametric techniques fit their respective models to an arbitrarily lagged empirical variogram such as (2).

Sampson and Guttorp (1992) fit their model, using $\Omega_{\infty}(x)$ in (6), to the semivariogram cloud rather than to the smoothed Matheron version given by (2). Their example involves a data set with twelve sites yielding only 66 points in the semivariogram cloud making this feasible. Application of their method to a much larger, hence “noisier” data set would be expected to produce a variogram mixing hundreds perhaps thousands of Gaussian forms. The resulting variogram will follow the semivariogram cloud too closely to be plausible.

The SB method fixes the nodes $\phi_\ell$ and estimates the jumps $w_\ell$ by an iterative weighted least squares fit of $2(1 - \rho(d, \phi, w))$ from (6) to the empirical variogram. Depending on the desired smoothness of the resulting variogram, this method offers three different sets of constraints. One set forces the estimated variogram to be concave; another imposes only monotonicity. A third set of constraints forces only smoothness whence the resulting variogram may again follow the empirical variogram values too closely. Cherry, Banfield and Quimby (1996) evaluate the SB method and advocate $p = 200$ in (6) for modeling the spatial correlation.

Barry and Ver Hoef (BV) (1996) fit linear-piecewise variograms to the empirical variogram and prove that their model can approximate any continuous variogram in $\mathbb{R}^1$.
with enough components. Instead of mixing $\Omega_1(x) = \cos(x)$ curves in $\mathbb{R}^1$, each defined over the entire support of the variogram as with the SG or SB methods, BV use individual linear components defined on disjoint subsets of the support. Hence, components can be focused on distance regions of higher interest such as those near the origin.

Hall, Fisher and Hoffman (1994) confine themselves to $\mathbb{R}^1$. For the model in (5) assuming $G_{\phi}(z)$ is differentiable, they let $dG_{\phi}(z) = g(z)dz$ and employ a kernel estimator $\hat{g}$ of $g$. Their goal is to provide a valid variogram using $\Omega_1(x) = \cos(x)$ involving a continuous rather than discrete mixture. Their method thus requires choosing a kernel function for $\hat{g}$ and an associated bandwidth. Additionally, $\hat{g}$ needs to be “smoothed” to insure a positive-definite covariance matrix.

We work in $\mathbb{R}^2$ where again $\Omega_2(x) = J_0(x)$. Within the Bayesian paradigm, we introduce (6) directly into the likelihood but keep $p$ small (at most 5) allowing random $w_\ell$ or random $\phi_\ell$. We offer a compromise between the standard parametric forms (Table 1) that specify three parameters for the covariance structure and the SB or SG method, with the practically implausible mixture of hundreds of components. Moreover, by working with the likelihood, inference is conditioned upon the observed $y$, rather than on a summary such as a smoothed version of the semivariogram cloud. Because the semivariogram cloud is often very noisy (see section 5), mixing only a few Bessels is more parsimonious and more satisfying than mixing possibly hundreds.

Returning to (5) when $n = 2$, we obtain

$$\rho(d, \phi) = \int_0^\infty \sum_{k=0}^\infty \frac{(-1)^k}{k!^2} \left( \frac{zd}{2} \right)^{2k} dG_{\phi}(z).$$

(7)

Only if $z$ is bounded, i.e., $G_{\phi}$ places no mass on say $z > \phi_{\max}$ can we interchange summation and integration to obtain

$$\rho(d, \phi) = \sum_{k=0}^\infty \frac{(-1)^k}{k!^2} \left( \frac{d}{2} \right)^{2k} \delta_{2k}$$

(8)

where $\delta_{2k} = \int_0^{\phi_{\max}} z^{2k} dG_{\phi}(z)$. The simplest such choice for $G_{\phi}$ puts discrete mass $w_\ell$ at a finite set of values $\phi_\ell \in (0, \phi_{\max}), \ell = 1, \ldots, p$ resulting in a finite mixture of Bessels model for $\rho(d, \phi)$ which yields as in (4)

$$\gamma(d_{ij}) = \tau^2 + \sigma^2(1 - \sum_{\ell=1}^p w_\ell J_0(\phi_\ell d_{ij})).$$

(9)
Under a Bayesian framework for a given $p$, if the $w_\ell$’s are each fixed to be $\frac{1}{p}$, with $\phi_\ell$’s unknown hence random, they are constrained by $0 < \phi_1 < \phi_2 < \ldots < \phi_p < \phi_{\max}$ for identifiability. The result is an equally weighted mixture of random curves. If a random mixture of fixed curves is desired, then the $w_\ell$’s are random and the $\phi_\ell$’s are systematically chosen to be $\phi_\ell = \left(\frac{\ell}{p+1}\right)\phi_{\max}$. We examine $p = 2, 3, 4, 5$ for fixed nodes and $p = 1, 2, 3, 4, 5$ for fixed weights. Mixture models using random $w_\ell$’s and random $\phi_\ell$’s might be considered but, in our limited experience, the posteriors have exhibited weak identifiability in the parameters and thus are not recommended.

In choosing $\phi_{\max}$, we essentially determine the maximum number of sign changes we allow for the dampened sinusoidal Bessel correlation function over the range of $d$’s of interest. For say $0 \leq d \leq d_{ij}^{\max}$, the larger $\phi$ is the more sign changes $J_0(\phi d)$ will have over this range. This suggests making $\phi_{\max}$ very large. However, as noted earlier in this section, we seek to avoid practically implausible $\rho$ and $\gamma$ which can arise from implausible $J_0(\phi d)$. For illustration, we chose to allow at most eleven sign changes (5 periods from the initial zero of $J_0$). Figure 5 graphs the correlation function $J_0(x)$ with five periods. The first eleven zeros of $J_0$ occur at 2.4048, 5.5201, 8.6797, 11.7915, 14.9309, 18.0711, 21.2116, 24.3525, 27.4935, 30.6346 and 33.7758. Letting $\kappa$ be the value of $x$ where $J_0(x) = 0$ attains its $k$th sign change (completes its $\frac{k-1}{2}$ period) we would set $\kappa = \phi_{\max}d_{ij}^{\max}$, thus determining $\phi_{\max}$. We reduce the choice of $\phi_{\max}$ to choosing the maximum number of Bessel periods allowable. For a given $p$ when the $\phi$’s are random, the posterior distribution for $\phi_p$ will reveal how close to $\phi_{\max}$ the data encourages $\phi_p$ to be.

4 Bayesian Model Fitting

A handful of recent papers have dealt with modeling spatial data from a Bayesian perspective. These include Handcock and Stein (1993) and Handcock and Wallis (1994) who model with the Matern class of correlation functions, Gaudard et. al (1995) who use a mixture of exponential and Gaussian forms and DeOliveira, Kedem and Short (1997) who employ the general exponential correlation function. All focus on prediction, while
our emphasis is on explanation of the spatial correlation structure. While the two issues
are often related, the model which best fits the data (our focus) need not be the model
which best predicts responses at unsampled sites.

The likelihood for the data is given by (3) using (6). Hence, to complete the Bayesian
model, specification of prior distributions for \( \mu \) and \( \alpha \) is required. For the parametric
models of Table 1, we assume the prior \( \pi(\mu, \alpha) \) takes the form

\[
\pi(\mu, \alpha) = \pi_1(\mu)\pi_2(\tau^2)\pi_3(\sigma^2)\pi_4(\phi).
\]

Although the parameters \( \mu, \tau^2, \sigma^2 \) and \( \phi \) are not truly thought to be independent, the
alternative, specifying a joint prior incorporating dependence, is arbitrary and difficult
to justify. We prefer to let the data modify our independence assumption through the posterior. In fact, we prefer to let the data drive our inference so for the covariance
parameters, we assume rather uninformative inverse gamma distributions by setting the
shape parameter equal to two (implying an infinite variance) i.e., \( \tau^2 \sim IG(2, b_{\tau^2}), \sigma^2 \sim
IG(2, b_{\sigma^2}) \) and \( \phi \sim IG(2, b_{\phi}) \). We use the 1990 scallop data to provide a prior mean for
each covariance parameter thus determining, e.g., \( b_{\sigma^2} = \frac{1}{E(\sigma^2)} \). The prior mean guesses
are point estimates obtained by fitting (4) to the Matheron semivariogram of the 1990
data using any convenient algorithm described in Cressie (1993). Finally, \( \mu \sim N(a_{\mu}, b_{\mu}) \)
where \( a_{\mu} \) is the 1990 log scaled mean of 3.5 and \( b_{\mu} = 1 \). For the Bessel mixtures, we
must add a prior for either \( w \) or \( \phi \). We use constant priors in both cases assuming
\( w \) to be Dirichlet(\( \alpha_1 = 1, \alpha_2 = 1, \ldots, \alpha_p = 1 \)) and \( \phi \) to be \( \phi_{max} \) times an ordered
Dirichlet(\( \alpha_1 = 1, \alpha_2 = 1, \ldots, \alpha_p = 1 \)).

The resulting posterior distribution of \( \theta = (\mu, \alpha) \) is not a standard form; we do
not even know the integrating constant to make it a density. Instead, to investigate
features of the posterior, the duality between a density and samples from the density
is utilized. Any desired attribute of the distribution (ex. mean, quantiles, skewness,
etc.) can be obtained to arbitrary accuracy by sampling from the density. In general,
define \( f(\theta) = L(Y|\theta) \cdot \pi(\theta) \) with the goal being to obtain a sample from the posterior
\( f(\theta)/ \int f(\theta) d\theta \). Since the dimensionality of \( \theta \) is low (at most nine in our examples),
we use a noniterative Monte Carlo method employing an importance sampling density
(ISD), say \( g(\theta) \).

Once \( g(\theta) \) is chosen, draw \( \theta_1, \theta_2, \ldots, \theta_V \) from \( g(\theta) \) and form weights \( \nu_i = \frac{f(\theta_i)}{g(\theta_i)} \).

Let \( \nu = \sum_{i=1}^{V} \nu_i \) and \( q_i = \frac{\nu_i}{\nu}, i = 1, \ldots, V \). Monte Carlo integration for any posterior expectation say \( E(b(\theta)|y) \) takes the form \( \sum q_i b(\theta_i) \) while resampling the \( \theta_i \) using the probabilities \( q_i \) provides an approximate sample from the posterior (Smith and Gelfand, 1992).

West’s adaptive mixture method (1993) is used to iteratively construct a mixture distribution to use as a \( g(\theta) \). Since West’s procedure employs mixtures of multivariate normal distributions, all parameters should be transformed to the real line to improve the resulting \( g(\theta) \). Starting from some \( g^0(\theta) = N(\theta_0, \Sigma_0) \), draw \( \theta_1, \theta_2, \ldots, \theta_V \) from \( g^0(\theta) \) and form the weights \( \nu^0_i = \frac{f(\theta_i)}{g(\theta_i)} \) to calculate

\[
\tilde{\theta}_1 = \frac{\sum_{i=1}^{V} \nu^0_i \theta_i}{\sum_{i=1}^{V} \nu^0_i} \quad \text{and} \quad \Sigma_1 = \frac{\sum_{i=1}^{V} \nu^0_i (\theta_i - \tilde{\theta}_1)(\theta_i - \tilde{\theta}_1)^T}{\sum_{i=1}^{V} \nu^0_i}
\]

and let \( g^1(\theta) = \sum_{i=1}^{V} \nu^0_i N(\theta_i; \Sigma_1) \). Now repeat the process using \( g^1(\theta) \) to form \( g^2(\theta) \).

When small variability in the weights \( \nu_i \) is achieved, we stop. In practice, a few iterations usually provides a reasonable ISD for \( f(\theta) \). Notice that in developing \( g \) and in evaluating the \( \nu_i \), we require many evaluations of \( f(\theta) \), i.e., many calculations of \( |\Sigma(\alpha)| \) and \( \Sigma(\alpha)^{-1} \).

## 5 Model Choice

General discussion of the model determination issue with regard to variogram specification appears to be virtually ignored in the literature. Penalized likelihood techniques such as the Akaike Information Criterion (Akaike, 1973) and the Schwartz or Bayes Information Criterion (Schwartz, 1978) can be used if \( y \) is parametrically specified. Otherwise, Webster and McBratney (1989) propose minimizing a penalized residual sum of squares criterion within a class of variogram models such as the Gaussian to an arbitrarily lagged empirical variogram such as (2). However, for most variogram modeling applications, geoscientists seem content to use empirical wisdom to propose a suitable class of variogram models. Then, adopting a goodness-of-fit criterion, they obtain the best member of the class. There is no comparison across variogram classes. Is an exponential form bet-
ter than a Gaussian? Is the flexible mixture of Bessel functions class significantly better than a Cauchy? In addressing this problem three issues emerge. What is an appropriate goodness-of-fit criterion? How do we penalize a variogram model for complexity? How do we adjust our model adequacy notion in the context of fitting variograms to data with enormous noise and weak signal? We clarify and answer these questions below.

Indexing models by \( m \), with model \( m \) having parameters \( \theta^{(m)} = (\mu, \tau^2, \sigma^2, \phi^{(m)}) \), the formal Bayesian approach is to compute the marginal density of the data evaluated at the observed \( y \). Unfortunately, this quantity can be difficult to compute and moreover, it is only interpretable if the prior on \( \theta^{(m)} \) is proper. Additionally, even under a proper prior, this criterion arises from a hypothesis testing form of utility (Kadane and Dickey, 1980) which does not, in a practical sense, reflect our utility for a variogram model.

Recall that our focus is explanation of spatial dependence, not spatial prediction. There is no elaboration of the mean structure; for all proposed models it is constant as in (3). Under the intrinsic hypothesis, dependence is captured in the data by the set of \( z_{ij} = (Y(s_i) - Y(s_j))^2 \). That is under model \( m \), \( E(z_{ij} | \theta^{(m)}) = \eta^{(m)}_{ij} = 2\gamma(d_{ij}, \theta^{(m)}) \).

Hence, model performance is reflected in the fit of the variogram to the set of \( z_{ij} \). Since under (3), \( z_{ij} \sim \eta_{ij} \chi^2_1 = Gamma(\frac{1}{2}, \frac{1}{2\eta_{ij}}) \), variogram models elaborate mean structures for the \( z_{ij} \)'s.

Then, to compare variograms we must work in the predictive space of the \( z_{ij} \)'s. (In fact, comparison of \( \theta^{(m)} \)'s makes no sense!) Hence, we need to specify a utility, equivalently a loss (negative utility), for taking action \( a \in \mathbb{R}^+ \) when \( z \) is realized. A natural loss to use is the deviance associated with the Gamma distribution, which from McCullagh and Nelder (1989, p 34) is

\[
L(z, a) = \frac{z - a}{a} - \log \left( \frac{z}{a} \right) \tag{10}
\]

Note that \( L(z, a) \) takes its minimum at \( a = z \) and increases strictly as \( a \) moves away from \( z \) in either direction.

Following Gelfand and Ghosh (1997), we extend (10) to a so-called balanced loss function (Zellner, 1994)

\[
L(z_{rep}, a; z_{obs}) = L(z_{rep}, a) + kL(z_{obs}, a) \tag{11}
\]
In (11), $z_{rep}$ is viewed as an independent replication of $z_{obs}$, i.e. if $z_{ij,obs}|\theta^{(m)} \sim \eta_{ij}^{(m)} \chi_1$, then so is $z_{rep}$. Then (11) views $a$ as a compromise action which rewards closeness to $z_{rep}$ but also penalizes for being too far from $z_{obs}$. The weight $k$ indicates relative regret for departure from $z_{obs}$ compared with departure from $z_{rep}$. We would set $k = 0$ if $z_{rep}$ were in fact a new $z$ at a new distance $d$ where there was no associated $z_{obs}$. Model choice is usually insensitive to specification of $k$ and we use $k = 1, 3, 9$ for illustration in section 6.

We convert (11) to a model choice criterion by computing the sum over the $\frac{N(N-1)}{2}$ $z_{ij,rep}$'s,

$$D_{k,m} = \sum_{i,j} \min_{a_{ij}} \left\{ E_{z_{ij,rep}|y,m} L(z_{ij,rep}, a_{ij}) + k L(z_{ij,obs}, a_{ij}) \right\}. \quad (12)$$

(One might criticize the additive form in (12) since the $z_{ij}$ are not independent. However, the joint dependence structure is complicated and an additive utility is often natural.) That is, under model $m$, for each $(i,j)$ pair we choose $a_{ij}$ to minimize the expected posterior predictive loss under the distribution $f(z_{ij,rep}|y,m)$. We then choose the model which makes $D_{k,m}$ the smallest thus selecting the model with the largest expected utility.

Fortunately, the minimization in (12) can be done explicitly under (10). After a bit of manipulation, we obtain

$$D_{k,m} = (k + 1) \sum_{i,j} \left\{ \log \left( \frac{\lambda_{ij}^{(m)} + kz_{ij,obs}}{k + 1} \right) - \log \left( \frac{\lambda_{ij}^{(m)}}{k + 1} \right) + k \log(z_{ij,obs}) \right\}$$

$$+ \sum_{i,j} \left( \log(\lambda_{ij}^{(m)}) - E(\log(z_{ij,rep}|y,m)) \right) \quad (13)$$

where $\lambda_{ij}^{(m)} = E(z_{ij,rep}|y,m)$. The concavity of the log function ensures that both summations on the right hand side of (13) are positive. (As an aside, in theory $z_{ij,obs} > 0$ a.s., but in practice we may observe some $z_{ij} = 0$ as, for example, using log counts in the scallop data example. A correction is needed and can be achieved by adding $\epsilon$ to $z_{ij,obs}$ where say $\epsilon$ is one-half of the smallest possible positive $z_{ij,obs}$.) The first term in (13), henceforth $G_{k,m}$, can be viewed as a goodness-of-fit piece. It would be 0 if each prediction for $z_{ij,rep}$, $\lambda_{ij}^{(m)}$, equalled $z_{ij,obs}$. The second term can be viewed as a penalty for model complexity, henceforth $P_m$. That is, expanding $\log(z_{ij,rep})$ around $\log(\lambda_{ij}^{(m)})$ to second order and taking expectations we find $\log(\lambda_{ij}^{(m)}) - E(\log(z_{ij,rep}|y,m)) \approx Var(z_{ij,rep}|y,m)/(\lambda_{ij}^{(m)})^2$. For
underfitted models, predictive variances will tend to be large, hence so will $P_m$. But also, for overfitted models, we expect inflated predictive variances, again making $P_m$ large. Models which are too simple will do poorly in both $G_{k,m}$ and $P_m$. As the variogram model becomes increasingly complex, we anticipate a trade-off; $G_{k,m}$ will decrease but eventually $P_m$ will increase. As a result, complexity is discouraged and parsimony is encouraged. Here we find ourselves in the same spirit as familiar penalized likelihood approaches, e.g., Akaike (1973) and Schwartz (1978) but, for us the penalty term falls out as a result of the utility maximization.

The criterion in (13) is readily computed. If $z_{rep}$ is the vector of $z_{ij,rep}$’s then

$$f(z_{rep}|y, m) = \int f(z_{rep}|\theta^{(m)}) f(\theta^{(m)}|y, m) \, d\theta^{(m)}.$$  \hspace{1cm} (14)

The expression in (14) appears foreboding. Apart from the multidimensional integration, we know neither density function under the integral explicitly. The power of the simulation-based model fitting approach is that we do not need to. Suppose $\theta^{(m)}_\ell$, $\ell = 1, \ldots, L$ is a sample from the posterior $f(\theta^{(m)}|y, m)$ and for each $\theta^{(m)}_\ell$, we draw $y^*_\ell$ according to (3). Transforming $y^*_\ell$ to $z^*_\ell$ immediately yields a random draw from $f(z_{rep}|y, m)$. But then the set of $z^*_\ell$ is a sample from which we can compute a Monte Carlo integration for $\lambda^{(m)}_{ij}$, i.e., $\sum z^*_\ell / L$ and for $E(\log(z_{ij,rep}|y, m))$, i.e. $\sum \log(z^*_\ell) / L$.

Finally, we note that any plausible variogram model will provide an inadequate fit to the $z_{ij,obs}$. In the simplest sense, we are fitting a low dimensional parametric variogram model (perhaps ten parameters at most) to $\frac{N(N-1)}{2}$ data points. (For the 1993 scallop data in section 6, $N = 147$ so $\frac{N(N-1)}{2} = 10,731$). This is evident from the semivariogram cloud of the $(d_{ij}, \frac{1}{2}z_{ij,obs})$ pairs in Figure 3a.

The situation is analogous to the design problem in regression on a single explanatory variable. There, the design points, say $x_1, x_2, \ldots, x_R$ are selected and independent replications $y_{ij}, j = 1, \ldots, n_i$ are taken at $x_i$. Letting $\hat{\mu}^{(m)}_i$ be the fitted value for $E(y_{ij})$ under model $m$, the error sum of squares is factored as

$$\sum_{i,j} (y_{ij} - \hat{\mu}^{(m)}_i)^2 = \sum_i n_i(\bar{y}_i - \hat{\mu}^{(m)}_i)^2 + \sum_{i,j} (y_{ij} - \bar{y}_i)^2.$$  \hspace{1cm} (15)

The first term on the right hand side of (15) measures the goodness-of-fit of model $m$ by comparison with $\bar{y}_i$, the sample guess for $E(y_{ij})$, i.e., the signal. The second term is
a pure error expression capturing the variability of the $y_{ij}$ about $E(y_{ij})$, i.e., the noise. In such a setting the second term typically dominates. The $R^2$ will be very small and in fact, it is often noted that $R^2$ is an inappropriate measure of model performance in this case. The first term is more informative in this regard.

By analogy, when looking at the set of $z_{ij,\text{obs}}$, we customarily create a lagged empirical semivariogram such as the Matheron estimator in (2). The $\gamma_c$'s which define this estimator play the role of the $\tilde{y}_i$'s above. Goodness-of-fit of the semivariogram model is considered with respect to the empirical semivariogram. The semivariogram cloud of $\frac{N(N-1)}{2}$ points is reduced to $R$ points where $R$ is the number of sets $B_r$.

Three important remarks are needed at this point. First, unlike the design problem, here the choice of $R$ is arbitrary. Which empirical variogram should we reduce to? Second, again distinct from the design problem, the set of $z_{ij,\text{obs}}$ with $d_{ij} \in B_r$ are not independent and more importantly are not symmetrically distributed about their mean. In fact up to a scale parameter, they follow a $\chi^2_1$ distribution rather than a symmetric normal. Figure 3b shows a boxplot of the $\frac{z_{ij,\text{obs}}}{s^2}$ within each $B_r$ for the scallop data and reveals their substantial skewness. Why should $\gamma_c$, the sample mean of the $z_{ij,\text{obs}}$, be a suitable center against which fit is judged? Third, Bayesian (or likelihood) model fitting conditions on $y$ not $z$. There is no attempt to find $\theta^{(m)}$ to provide a good fit to the $z_{ij,\text{obs}}$'s. We should not expect a Bayesian analysis to provide a “best-fitting” variogram to an arbitrarily selected empirical variogram. We reiterate that if the sole objective is to achieve goodness-of-fit to a specified empirical variogram, we would recommend the approaches of Sampson and Guttorp (1992) and Shapiro and Botha (1991) described in section 3. The Bayesian approach is appropriate if we seek full inference, i.e., inference about any aspect of the dependence structure. Thus, the role of $D_{k,m}$ is clarified. It is a criterion for introducing our utility for a model into the model selection process which is apart from employing the Bayesian framework to fit the models.

In this regard, working with $D_{k,m}$, how can we separate signal from noise analogous to (15) in the design problem? Returning to (13), suppose for a given set $B_1, B_2, \ldots, B_R$
we set
\[
G_k(c_1, c_2, \ldots, c_R) = \sum_{r=1}^{R} \sum_{\{(i,j) : d_{ij} \in B_r\}} \left\{ \frac{1}{k+1} \log \left( \frac{c_r + k z_{ij, \text{obs}}}{k+1} \right) - \frac{\log(c_r) + k \log(z_{ij, \text{obs}})}{k+1} \right\}
\]
(16)

For a given \( r \), consider minimizing the inner summation in (16), which we define by \( \zeta_k(c_r) \). Clearly, \( \zeta_k(c_r) \to \infty \) as \( c_r \to 0 \) and as \( c_r \to \infty \). Also, since \( \frac{\sum_{\{(i,j) : d_{ij} \in B_r\}} c_r}{N_{B_r}} = \frac{1}{N_{B_r}} \sum_{\{(i,j) : d_{ij} \in B_r\}} c_r + k z_{ij, \text{obs}} \frac{1}{k+1} \in \left( -\frac{1}{k+1}, \frac{k}{k+1} \right) \) and increases in \( c_r \), the minimizing value is unique. We denote it by \( \hat{c}_r \) and obtain it by any convenient root-finding algorithm.

Then, the set \( \hat{c}_1, \hat{c}_2, \ldots, \hat{c}_R \) provides the smallest possible value of \( G_{k,m} \) when using a constant \( c_r \) over \( B_r \) and thus provides an empirical variogram arising from (13). We may argue that, since the \( z_{ij} \) follow a scaled \( \chi^2_1 \) distribution, the \( \hat{c}_r \)'s provide a more natural variogram estimator than say the \( 2\gamma^* \)'s in (2). We define this new empirical variogram to be the deviance variogram. It is data based, motivated only by the presumed normality of \( y \), not by any model for \( \gamma \). Moreover, since \( \frac{n_r \zeta_1(c_r)}{N_{B_r}} \approx E\left( \frac{c_r}{c_r + \sqrt{z}} \right) z \sim \eta_r \chi^2_1 \) \(- \frac{1}{k+1} \), if, for instance \( k = 1 \), \( \frac{n_r \zeta_1(n_r)}{N_{B_r}} \approx E((1 + \omega)^{-1}| \omega \sim \chi^2_1) - \frac{1}{2} = 0.152 \)

while \( \frac{0.385 n_r \zeta_1(0.385 n_r)}{N_{B_r}} \approx E(0.385(0.385 + \omega)^{-1}| \omega \sim \chi^2_1) - \frac{1}{2} = 0.000 \). Hence, we expect \( \hat{c}_r (\approx 0.385 \eta_r) < 2\gamma^* (\approx \eta_r) \), i.e., that the deviance empirical semivariogram will lie below the Matheron estimator. In Figure 4, we overlay the deviance empirical semivariogram for \( k = 1 \) on the Matheron estimate for the 1993 scallop data using lag \( \delta = 0.05 \) and observe this to be the case. Lastly, analogous to (15), we can subtract \( G_k(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_R) \) from \( G_{k,m} \) to obtain an adjusted goodness-of-fit value. This value reflects the lack of fit of the variogram model \( m \) relative to the best fitting empirical variogram using \( B_1, B_2, \ldots, B_R \).

6 Example

The dataset of scallop catches in the Atlantic Ocean mentioned in section 1 is examined for spatial correlation. After some preliminary exploratory discussion, we examine the results of parametric and nonparametric variogram model fitting to these data. After choosing the best fitting model of each type, we summarize resultant inference for each.
6.1 Data and Exploratory Techniques

For the 1993 scallop data, Figure 3 shows the semivariogram cloud (3a) together with boxplots (3b) formed from the cloud using lag $\delta = 0.05$. The 10,731 pairs of points which produce the semivariogram cloud do not reveal any distinct pattern. However, the boxplots and the Matheron and deviance empirical semivariograms (see section 5) each based on lag $\delta = 0.05$ (Figure 4) clearly exhibit spatial dependence in the sense that when separation distances are small, the spatial variability tends to be less. The arbitrary lag choice of $\delta = 0.05$ is chosen for presentation of results, not for any formal variogram modeling. Figure 3 is routinely produced by the SpatialStats version 1.0 module of Splus.

For the choice of $\phi_{\text{max}}$ in the nonparametric setup, we selected seven sign changes or three Bessel periods. With $d_{ij}^{\text{max}} = 2.83$ degrees, $\phi_{\text{max}}$ becomes 7.5. A sensitivity analysis with two Bessel mixtures ($p = 2$) having a fixed weight $w_1$ and random nodes was undertaken. Two, four and five Bessel periods revealed little difference in results as compared with three. However, when one Bessel period was examined ($\phi_{\text{max}} = 3$), the model fit poorly and in fact $\phi_p$ was just smaller than 3. This is an indication that more flexibility (a larger value of $\phi_{\text{max}}$) is required.

6.2 Fitted Semivariogram Models

All of the parametric models of Table 1 and nonparametric Bessel mixtures with different combinations of fixed and random parameters were fit to the 1993 scallop data. Figure 5 shows the posterior mean of each respective semivariogram while Table 3 provides the model choice criteria for each model along with the independence model ($\Sigma(\alpha) = (\tau^2 + \sigma^2)I$). The simulation error associated with the entries in table 3 can most easily be assessed by repeated fitting of the models. In doing so, we conclude that a 3$\sigma$ range is at most 30. All variogram models fit better than the independence model which again supports the presence of spatial correlation. Of the parametric models, the Cauchy and Gaussian fit best using the $D_{k,m}$ and $G_k$ criteria. Taking $k = 1$, the value of $G_1(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_R)$ associated with lag $\delta = 0.05$ is 8961.6. For other lags, $\delta = 0.1$ and
\( \delta = 0.025 \), the values are 8971.1 and 8942.8 respectively, confirming the level of noise about the deviance variogram. The \( G_{1,m,\text{adj}} \) (\( = G_{1,m} - G_1(\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_R) \)) value in table 3 is the sharpened goodness of fit criterion. The variogram model with the smallest \( G_{1,m,\text{adj}} \) value (the nonparametric mixture of five Bessel functions with random weights and fixed \( \phi \)'s) is the best fitting variogram model; however, the model with the smallest penalty component, \( P_m \), is the Cauchy. The wide range of estimated (posterior mean) variograms in Figure 6 indicates considerable sensitivity to the choice of the parametric form.

Of the Bessel mixtures, the five component model with random weights and fixed \( \phi \)'s fit best where, given \( \phi_{\text{max}} = 7.5 \), the nodes were fixed to be \( \phi_1 = 1.25, \phi_2 = 2.5, \phi_3 = 3.75, \phi_4 = 5 \) and \( \phi_5 = 6.25 \). One would expect that the fit measured by the \( G_{k,m} \) criterion should improve with increasing \( p \). However, the models are not nested by \( p \) except, for instance, the \( p = 2 \) model is a special case of the \( p = 5 \) model. Thus, it can occur that the four component fixed \( \phi \) model performs worse than the three component model. The random \( \phi \) Bessel mixture models were all very close and, as a class, these models fit as well as or better than the best parametric model. Hence, modeling mixtures of Bessel functions appears more sensitive to the choice of fixed \( \phi \)'s than to fixed weights.

### 6.3 The Best Fitting Parametric and Nonparametric Models

In this section, the best fitting parametric model (Cauchy) and nonparametric model (mixture of five Bessel functions with random weights and fixed \( \phi \)'s) are further examined. Figure 7 displays the posterior means for each with 95\% interval estimates. The Bessel mixture model appears to align with the Matheron estimator better than the Cauchy except near the origin. Point and interval estimates for all parameters are presented in Table 4. As seen in Figure 7 and also from the interval estimates for their respective sills, the Cauchy model has a tighter upper confidence bound for its variogram. Also, negative covariances among the \( \alpha \) parameters of the Cauchy model arise in agreement with intuition. That is, if \( \phi \) were held constant in the Cauchy model and \( \sigma^2 \) were increased, then \( \tau^2 \) must decrease to accommodate the data. Likewise, if \( \tau^2 \) were held constant and \( \sigma^2 \) increased, then the range must also increase forcing \( \phi \) to decrease.
Table 3: Model Choice for Fitted Variogram Models

<table>
<thead>
<tr>
<th>Name</th>
<th>$G_{1,m}$</th>
<th>$G_{1,m,adj}$</th>
<th>$P_m$</th>
<th>$D_{1,m}$</th>
<th>$D_{3,m}$</th>
<th>$D_{9,m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PARAMETRIC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td>10961.6</td>
<td>2000.0</td>
<td>13897.7</td>
<td>24859.3</td>
<td>40905.7</td>
<td>75202.7</td>
</tr>
<tr>
<td>Gaussian</td>
<td>10860.5</td>
<td>1898.9</td>
<td>13843.2</td>
<td>24703.7</td>
<td>40519.6</td>
<td>74317.2</td>
</tr>
<tr>
<td>Cauchy</td>
<td>10695.6</td>
<td>1734.0</td>
<td>13808.5</td>
<td>24504.1</td>
<td>39950.1</td>
<td>72940.2</td>
</tr>
<tr>
<td>Spherical</td>
<td>11450.0</td>
<td>2488.4</td>
<td>13965.7</td>
<td>25415.7</td>
<td>42381.7</td>
<td>78665.1</td>
</tr>
<tr>
<td>Bessel</td>
<td>11042.2</td>
<td>2080.6</td>
<td>14032.5</td>
<td>25074.7</td>
<td>41291.0</td>
<td>75972.3</td>
</tr>
<tr>
<td>Independent</td>
<td>11577.2</td>
<td>2615.6</td>
<td>16159.3</td>
<td>27736.5</td>
<td>44824.7</td>
<td>81426.4</td>
</tr>
<tr>
<td><strong>SEMIPARAMETRIC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed $\phi_{t}$'s/Random $w_{t}$'s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two</td>
<td>11072.8</td>
<td>2111.2</td>
<td>13967.5</td>
<td>25040.3</td>
<td>41320.5</td>
<td>76133.8</td>
</tr>
<tr>
<td>Three</td>
<td>10593.7</td>
<td>1632.1</td>
<td>13818.6</td>
<td>24412.3</td>
<td>39725.0</td>
<td>72436.5</td>
</tr>
<tr>
<td>Four</td>
<td>10934.4</td>
<td>1972.8</td>
<td>13870.2</td>
<td>24804.6</td>
<td>40809.8</td>
<td>75026.9</td>
</tr>
<tr>
<td>Five</td>
<td>10564.5</td>
<td>1602.9</td>
<td>13817.5</td>
<td>24382.0</td>
<td>39627.6</td>
<td>72192.5</td>
</tr>
<tr>
<td>Random $\phi_{t}$’s/Fixed $w_{t}$’s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two</td>
<td>10677.0</td>
<td>1715.4</td>
<td>13906.3</td>
<td>24583.3</td>
<td>40091.9</td>
<td>73236.0</td>
</tr>
<tr>
<td>Three</td>
<td>10685.0</td>
<td>1723.4</td>
<td>13968.6</td>
<td>24653.6</td>
<td>40181.3</td>
<td>73364.9</td>
</tr>
<tr>
<td>Four</td>
<td>10645.8</td>
<td>1684.2</td>
<td>13915.7</td>
<td>24561.5</td>
<td>40000.3</td>
<td>72989.3</td>
</tr>
<tr>
<td>Five</td>
<td>10614.3</td>
<td>1652.7</td>
<td>13902.7</td>
<td>24517.0</td>
<td>39887.3</td>
<td>72725.1</td>
</tr>
</tbody>
</table>
Table 4: Summary Statistics from the Posterior by Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>95% Interval Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAUCHY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget($\tau^2$)</td>
<td>0.84</td>
<td>(0.44 , 1.28)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>2.17</td>
<td>(1.18 , 3.70)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>29.41</td>
<td>(6.71 , 91.09)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>3.02</td>
<td>(2.25 , 3.99)</td>
</tr>
<tr>
<td>Sill</td>
<td>3.01</td>
<td>(2.03 , 4.52)</td>
</tr>
<tr>
<td>$r_C$</td>
<td>0.93</td>
<td>(0.46 , 1.68)</td>
</tr>
<tr>
<td>$r_V$</td>
<td>0.78</td>
<td>(0.41 , 1.39)</td>
</tr>
<tr>
<td>BESSEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget($\tau^2$)</td>
<td>1.47</td>
<td>(1.18 , 1.81)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1.59</td>
<td>(0.69 , 3.68)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>2.86</td>
<td>(1.92 , 3.78)</td>
</tr>
<tr>
<td>Sill</td>
<td>3.05</td>
<td>(2.08 , 4.78)</td>
</tr>
</tbody>
</table>

For the Cauchy model, posterior distributions for the nugget ($\tau^2$), $\sigma^2$, $\mu$, and the sill are each represented by the solid line in Figures 8a-d, respectively. Figure 8e shows $\phi$ for the Cauchy correlation structure, while the ranges $r_C$ (solid line) and $r_V$ (dashed line) for the Cauchy semivariogram are presented in Figure 8f. The posterior means for the ranges $r_C$ and $r_V$ are 0.93 and 0.78 degrees. The nugget ($\tau^2$), $\sigma^2$, $\mu$ and the sill for the five component random weight Bessel model are given by the dotted line in Figures 8a-d. The posterior medians for the five weights are 0.203, 0.271, 0.337, 0.016 and 0.173, respectively. The two models have very different nuggets ($\tau^2$), but their sills are nearly identical with the Bessel mixture being slightly more right skewed. Again, since the Bessel function is not monotone, the range is not defined.
7 Conclusions

We have proposed a fully Bayesian approach to fitting variogram models, arguing its advantages over customary methods. We have examined a broad range of models including an arbitrarily flexible class obtained through mixtures. We have shown how model determination can be carried out in this framework to investigate adequacy and comparison. We also have demonstrated the scope of possible inference.

References


Figure 1: Sites sampled in the Atlantic Ocean for 1990 and 1993 scallop catch data.
Figure 2: Smoothed log scaled scallop counts in 1993.
Figure 3: Semivariogram cloud (3a) for 1993 scallop data and boxplot produced from 0.05 lag (3b).
Figure 4: Matheron and deviance empirical semivariograms for lag $\delta = 0.05$. Matheron $= \bullet$; Deviance $= \circ$. 
Figure 5: A plot of $J_0$ out to eleven sign changes (five periods).
Figure 6: Posterior means for various semivariogram models.
Figure 7: Interval estimates for Cauchy and the five component Bessel mixture with fixed \( \phi \)'s. Matheron empirical semivariogram (●) at lag \( \delta = 0.05 \) shown for reference.
Figure 8: Posterior distributions for $\tau^2$ (a), $\sigma^2$ (b), $\mu$ (c) and Sill (d) under Cauchy model (−) and under the mixture of five Bessels model (⋯). Cauchy model $\phi$ (e) and ranges $r_C$ (−) and $r_V$ (−−) (f).