Semiparametric Estimation of Spectral Density with Irregular Observations

Hae Kyung Im, Michael L. Stein, and Zhengyuan Zhu*

Abstract

We propose a semiparametric method to estimate spectral densities of isotropic Gaussian processes with scattered data. The spectral density function (Fourier transform of the covariance function) is modeled as a linear combination of B-splines up to a cutoff frequency and, from this point, a truncated algebraic tail. We calculate an analytic expression for the covariance function and tackle several numerical issues that arise when calculating the likelihood. The parameters are estimated by maximizing the likelihood using the simulated annealing method. Our method directly estimates the tail behavior of the spectral density, which has the biggest impact on interpolation properties. The use of the likelihood in parameter estimation takes fully into account the correlations between observations. We compare our method with a kernel method proposed by Hall et al. (1994) and a parametric method using the Matérn model. Simulation results and an application to a rainfall dataset show that our method out-

*Hae Kyung Im is Research Associate, CISES, the University of Chicago, Chicago, IL 60637. Michael L. Stein is Professor, Department of Statistics, the University of Chicago, Chicago, IL 60637. Zhengyuan Zhu is Assistant Professor, Department of Statistics and Operations Research, the University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (E-mail: zhuz@email.unc.edu).

performs the other two by several criteria.

Key words: covariance function; maximum likelihood; B-spline; kriging;

1 Introduction

Estimation of the covariance structure of physical processes observed at a finite set of locations is fundamental to understand the behavior of such processes and to interpolate to locations where measurements are not available. Kriging, an interpolation method widely used by the geophysical community, is based on the knowledge of the covariances between observed and interpolated locations.

We will concentrate on Gaussian isotropic processes, which are invariant under all rigid motions. Under this assumption, the covariance of the process at two locations only depends on the distance between them, so a covariance function on $\mathbb{R}^+$ fully describes the second order properties of the process. This function has to be positive definite in order to ensure that the variance of any linear combinations of values of the process at various locations is positive. The most common solution to this problem is to restrict the estimation to parametric forms that are proven to be positive definite.

There has been some work in using nonparametric methods or a broad class of positive definite functions based on the spectral representation of covariance functions. Before describing these methods some review of positive definite functions is in order. Bochner's Theorem (Yaglom, 1987) states that a function is continuous and positive definite if and only if it is the Fourier transform of a positive bounded measure on $\mathbb{R}^d$, i.e.,

$$C(\mathbf{x}) = \int_{\mathbb{R}^d} \exp(i\mathbf{w}\mathbf{x}) F(\mathbf{dw}). \tag{1}$$

For isotropic processes (1) can be reduced to a one-dimensional integral

$$C(r) = 2^{(d-2)/2}\Gamma(d/2) \int_0^\infty (ru)^{-(d-2)/2} J_{(d-2)/2}(ru) dG(u), \tag{2}$$

where $G(u) = \int_{|\mathbf{w}|<\mathbf{u}} dF(\mathbf{dw})$ is a bounded positive measure on $\mathbb{R}$, $\Gamma(\cdot)$ is the Gamma function, and $J_\nu(\cdot)$ is the Bessel function of the first kind of order ν (Abramowitz and Stegun, 1965).

Shapiro and Botha (1991) proposed using a finite discrete measure with nodes placed at $t_1, ..., t_n$ so that the integral in (2) is reduced to a finite sum:

$$\tilde{C}(r) = \sum_{j=1}^{m} p_j \Omega_d(t_j h), \tag{3}$$

where the p_j's are positive and $\Omega_d(x) = \left(\frac{2}{x}\right)^{(d-2)/2} \Gamma\left(\frac{d}{2}\right) J_{(d-2)/2}(x)$. For a random field Z on $\mathbb{R}^d$, they use a raw covariogram estimate given by

$$\hat{C}(h) = \frac{1}{N(h)} \sum_{N(h)} (Z(\mathbf{x}_i) - \bar{Z})(Z(\mathbf{x}_j) - \bar{Z}), \tag{4}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are observation sites, $\bar{Z}$ is the average of the observations, the sum in (4) runs over all pairs of observations that are approximately distance h apart, and $N(h)$ is the total number of such pairs. They estimate the values of p_j by minimizing the mean squared difference between the raw covariogram estimate at different lags and their estimator, with positivity constraint.

Genton and Gorsich (2002) follow this idea but propose using the zeros of the Bessel functions as the nodes of the discrete measure and show that their method is computationally simpler, needs fewer nodes, and does not show spurious oscillations. The problem with this choice of nodes is that these numbers are nondimensional. It is not clear what scale should be used to translate these nodes into nodes in the frequency domain (which have dimension 1 / unit of distance). Although not totally explicit, they seem to propose using $1/r_{\max}$ as their scale. This would mean that if we added an additional observation at distance 1.5 $r_{\max}$ from the most distant existing observation, then the nodes would be shifted by a factor of 1.5 in the frequency domain. This behavior seems problematic.

Hall et al. (1994) propose using a kernel estimator for a preliminary covariogram estimate and, in order to ensure positive definiteness, they propose Fourier transforming the kernel estimator, setting the negative values to zero and Fourier transforming back to the spatial domain. For $d = 2$, denoting $\hat{Z}_{ij} = (Z(\mathbf{x}_i) - \bar{Z})(Z(\mathbf{x}_j) - \bar{Z})$, h_{ij} the distance between the observations Z_i and Z_j, K a kernel function (a positive symmetric probability density), and δ the bandwidth, the first step estimator of the covariogram is

$$\tilde{C}(h) = \frac{\sum_{i,j} \hat{Z}_{ij} K(\frac{h-h_{ij}}{\delta})}{\sum_{i,j} K(\frac{h-h_{ij}}{\delta})}. \tag{5}$$

The final estimate of the covariogram is

$$\bar{C}(h) = \int_0^\infty \left(\int_0^\infty \tilde{C}(x) x J_o(wx) dx \right)_+ w J_o(wh) dw, \tag{6}$$

where the subscript + means to take the positive part of the expression. We will call this function the kernel estimator of the covariance function.

All three methods use the raw covariogram as the basis for estimation, which ignores the correlation between the values of the observations at different distances. Furthermore, it is well known that the high frequency properties of the spectral density determine the performance of interpolation procedures (Stein, 1999). None of the above methods give proper consideration to the tail properties.

We propose a flexible family of models for the spectral density that is a linear combination of B-splines of order 4 (cubic splines) up to a cutoff frequency w_t and an algebraically decaying tail from w_t to infinity. We use positive coefficients for the B-splines, which ensure positiveness of the spectral density and, as a consequence, positive definiteness of the covariance function. Assuming the process is well described by a Gaussian random field, we find the parameters that maximize the likelihood. This method estimates the tail property of the spectral density in an explicit way. It does exclude exponential decay of the tail. However, we consider this restriction to be beneficial, since such a fast decay would imply an

unrealistically smooth process (Stein, 1999, 2002). It also excludes oscillatory tails such as $w^{-\gamma}\cos^2 w$, which is generally undesirable (Stein, 1999, 2002). Additionally, through the use of likelihood, our method takes fully into account all the correlations between observations. This is, to the extent of our knowledge, the first work that uses a likelihood approach for scattered spatial data without a parametric model.

In section 2 we present our model and the methodology to estimate the covariance function. Section 3 describes how to deal with the numerical challenges that arise when calculating the likelihood. In section 4 we describe several performance measures and, via a simulation study, compare our method with a parametric method using the Matérn model and the nonparametric kernel method in Hall et al. (1994). Section 5 compares the three methods using a rainfall dataset. Both the simulation study and the real data example show that our method is substantially better than the kernel method. Section 6 summarizes the paper and discusses possible further work, and the proofs are included in the appendix.

2 Methodology

We assume the observations come from realizations of a Gaussian random field whose value at location x is of the form

$$Z(\mathbf{x}) = \mathbf{m}(\mathbf{x})^T\beta + \epsilon(\mathbf{x}), \tag{7}$$

where $\mathbf{m}(\mathbf{x})$ is a known vector valued function, β is a vector of unknown coefficients, ϵ has mean 0 with covariance function $C(\epsilon(\mathbf{x}), \epsilon(\mathbf{y})) = C_\theta(|\mathbf{x}-\mathbf{y}|)$, and θ is the vector of unknown parameters of the covariance function.

2.1 The splines+tail (S+T) model and the Matérn model for the spectral density

Let $f(w)$ be the spectral density of $\epsilon(\mathbf{x})$ in (7), our S+T model can be written as

$$f_\theta(w) = \sigma^2 \sum_{i=-1}^{l+1} b_i B_i(w) \mathbf{1}_{[0,w_t]}(w) + C_f \left(\frac{w_t}{w}\right)^\gamma \mathbf{1}_{(w_t,\infty)}(w), \tag{8}$$

where $\mathbf{1}_A(w)$ is an indicator function of value one if $w \in A$, and zero otherwise. B_is are B-splines of order 4 (See Appendix A for a brief description and references) with node sequence $(w_0, ..., w_l)$ on the interval $[0, w_t]$, with w_t the threshold frequency. The sum goes from -1 to $l+1$ in order to include all B-splines that have support on the interval $[0, w_t]$. We chose order 4 because of the flexibility that cubic splines give to represent a wide range of smooth functions. B-splines of other orders could be used with minor adjustments. We require the spectral density to be continuous and have continuous derivative at w_t. The constant C_f is chosen to achieve continuity at w_t; more explicitly, $C_f = \sigma^2 \sum_{i=-1}^{l+1} b_i B_i(w_t)$. The coefficients of the B-splines are constrained to be positive except for b_{l+1}, which is chosen so that the derivative of $f_\theta(w)$ is continuous at w_t. It is shown in Appendix D that the function is still positive. Restricting the coefficients to be positive is a simple way of ensuring positivity of the function. The B-spline coefficients closely follow the function they represent as the number of nodes increases, so they eventually will become positive for a positive twice differentiable function. This suggests that the positivity condition is not too restrictive.

To specify a S+T model, we first need to determine the number and location of the nodes for the B-splines. In this paper, we restrict the B-splines to have uniformly distributed nodes between 0 and w_t, i.e., given the number of nodes $(l+1)$ and the cutoff frequency (w_t), we place the nodes at locations iw_t/l for $i = -1, 0, ..., l, l+1$.

Conditional on knowing the number of nodes, the parameter θ of our S+T model includes smoothness parameter γ, sill σ^2, cutoff frequency w_t, and the coefficients of B-splines (b_i for $i = 0, ..., l$). The coefficients b_{-1} and b_{l+1} determine the derivatives of the function at the end points. We chose b_{-1} to equal b_1 in order to make $f'(0) = 0$ and b_{l+1} to be such that the derivative of the function is continuous at w_t (See Appendix C).

Let us briefly describe the Matérn model in order to compare it to our model. This class is considered to be a sensible model for a wide range of processes arising in environmental problems (Stein, 1999; Handcock and Wallis, 1994). With only three easily interpretable parameters (σ^2, ϕ and ν), the Matérn class allows considerable flexibility in the type of processes it can represent. σ^2 is simply the variance of the process at a given location, ϕ is the inverse range parameter, and ν is a measure of the smoothness of the process. A process with smoothness parameter ν is $\lceil \nu \rceil$ - 1 times (mean square) differentiable, where $\lceil \cdot \rceil$ represents the smallest integer greater than or equal to the number. The spectral density of the Matérn class has the form

$$f(w) = \frac{\sigma^2 \lambda(\phi, \nu)}{(\phi^2 + w^2)^{\nu + d/2}} \tag{9}$$

with $\lambda(\phi, \nu) = \frac{\Gamma(\nu + d/2)}{\pi^{d/2} \Gamma(\nu)} \phi^{2\nu}$ such that the variance, $C(0)$, is σ^2. At high frequencies both the Matérn and our S+T model approach zero at the rate $1/w^\gamma$, with $\gamma = 2\nu + d$. In the simulation studies, we will use ν as the parameter for the S+T model.

2.2 Compute covariance function using the Hankel transform

The covariance function can be calculated from the spectral density $f_\theta(w)$ by applying (2) for $d = 2$:

$$C_\theta(r) = 2\pi \int_0^\infty w J_o(rw) f_\theta(w) dw. \tag{10}$$

The transform in (10) is called the Hankel transform of order 0. $f_\theta(w)$ in the S+T family is a linear combination of B-splines and an algebraic tail, the Hankel transform of both can be calculated analytically. The Hankel transform of B-splines requires calculating two Bessel functions of the first kind of orders 1 and 2 and two Struve functions of orders 1 and 2 (Abramowitz and Stegun, 1965) for each node. The computation of this part is straightforward albeit moderately time consuming. See Appendix B for details. The Hankel transform of the truncated algebraic tail is

$$\begin{aligned}
&\int_{w_t}^{\infty} w^{1-\gamma} J_o(wr) dw \\
=& r^{\gamma-2} \int_{w_t r}^{\infty} u^{1-\gamma} J_o(u) du \\
=& r^{\gamma-2} \left(\frac{\gamma \Gamma(-\gamma/2)}{2^\gamma \Gamma(\gamma/2)} + \frac{(w_t r)^{2-\gamma} {}_1F_2(1-\gamma/2; 1, 2-\gamma/2; -(r w_t)^2/4)}{\gamma - 2} \right),
\end{aligned} \tag{11}$$

where ${}_1F_2(a; b, c; z)$ is a generalized hypergeometric function with series representation $\sum_{k=0}^{\infty} \frac{(a)_k}{(b)_k (c)_k} \frac{z^k}{k!}$. Here $(\cdot)_k$ represents the Pochhammer's symbol (Abramowitz and Stegun, 1965), which is defined by $(z)_0 = 1$ and $(z)_k = z(z+1)(z+2)...(z+n-1) = \frac{\Gamma(z+n)}{\Gamma(z)}$.

Several features make the Hankel transform of the tail (11) numerically hard to compute. First, there is no easy way of evaluating this hypergeometric function accurately without resorting to summing a large number of terms of its series expansion, which can lead to severe numerical errors. Second, the Γ function is infinite when the argument is a negative integer, and we have no reason to exclude negative integer values for $-\gamma/2$. Third, the first term ($\frac{\gamma\Gamma(-\gamma/2)}{2^\gamma\Gamma(\gamma/2)}$) is the limit of the second term ($\frac{(w_t r)^{2-\gamma} {}_1F_2(1-\gamma/2;1,2-\gamma/2;-(rw_t)^2/4)}{\gamma-2}$) as $rw_t \to \infty$, so in the case where rw_t is large we need to take the difference of two very similar numbers.

The first problem is addressed by using arbitrary precision arithmetic libraries (code downloaded from http://www.mpfr.org/). For the second problem, we note that the divergence of the Γ function is compensated by the divergence of one of the terms of the series

expansion of the hypergeometric function. When $\gamma/2$ is an integer, we use an asymptotic expansion of the Γ function when the argument is close to $-\gamma/2$, and subtract it from the series expansion of the hypergeometric function. Only one term in each series expansion diverges as the arguments approaches $-\gamma/2$, and we get a modified series expansion for the difference between the two terms, which can be computed in the same fashion as the hypergeometric function, i.e., by adding the series until convergence is achieved and using arbitrary precision libraries to avoid numerical errors. The final expression for the tail integral when $\gamma/2 = n+1$ is

$$\begin{aligned}&\int_{w_t}^{\infty} w^{1-\gamma} J_o(wr)dw \\ &= r^{2n}\frac{\log(2) - \log(rw_t) + \psi(n+1)}{(-4)^n n!^2} + \frac{w_t^{2n}}{2n}\sum_{k=0,k\neq n}^{\infty}\frac{-n}{-n+k}\frac{(-(rw_t)^2/4)^k}{k!^2},\end{aligned} \tag{12}$$

where $\psi(n)$ is the digamma function (Abramowitz and Stegun, 1965), which for positive integer arguments can be evaluated as $\sum_{k=1}^{n-1}\frac{1}{k} - \gamma_{eg}$,where $\gamma_{eq} = 0.577216...$ is the Euler's constant. The details are given in Appendix E.

To solve the third problem of subtracting two very similar numbers, we again use an asymptotic expansion of the hypergeometric function, whose leading term is $\frac{\gamma\Gamma(-\gamma/2)(\gamma-2)}{2^\gamma\Gamma(\gamma/2)(rw_t)^{2-\gamma}}$. As a result, we are left with an expression that directly calculates the difference. See details in Appendix F. For large values of rw_t the Hankel transform of the truncated tail is approximated by

$$\begin{aligned}&\int_{w_t}^{\infty} w^{1-\gamma} J_o(wr)dw \\ &\approx \left(\frac{\cos(rw_t) - \sin(rw_t)}{\sqrt{\pi}(rw_t)^{\gamma-\frac{1}{2}}} - \frac{\left(-15 + 16\,\gamma + 128\,\gamma^2\right)\,(\cos(rw_t) - \sin(rw_t))}{128\,\sqrt{\pi}(rw_t)^{\gamma+\frac{3}{2}}}\right. \\ &\quad \left. + \frac{(-3 + 8\,\gamma)\,(\cos(rw_t) + \sin(rw_t))}{8\,\sqrt{\pi}(rw_t)^{\gamma+\frac{1}{2}}} + \dots\right) r^{\gamma-2}.\end{aligned} \tag{13}$$

The approximation (13) also apply when $\gamma/2$ is an integer.

2.3 Likelihood based parameter estimation

Let $\mathbf{Z} = (Z(\mathbf{x}_1), ..., Z(\mathbf{x}_n))^T$ be our observation and $\mathbf{M} = (\mathbf{m}(\mathbf{x}_1)...\mathbf{m}(\mathbf{x}_n))^T$. Under model (7), the log-likelihood has the form

$$l(\theta, \beta; \mathbf{Z}) = -\frac{1}{2}\log|\det\mathbf{\Sigma}_\theta| - \frac{1}{2}(\mathbf{Z} - \mathbf{M}\beta)^T\mathbf{\Sigma}_\theta^{-1}(\mathbf{Z} - \mathbf{M}\beta) - \frac{n}{2}\log(2\pi), \tag{14}$$

where $\mathbf{\Sigma}_{ij} = C_\theta(|\mathbf{x}_i - \mathbf{x}_j|)$ can be computed using (10) and (8). If the mean parameter β can be assumed known, one can estimate the parameter θ of the S+T model by maximizing (14). When β is unknown, θ can be estimated by maximizing the following restricted likelihood (Stein, 1999; McCullagh and Nelder, 1989) (ignoring an additive constant):

$$rl(\theta; \mathbf{Z}) = -\frac{1}{2}\log|\det\mathbf{\Sigma}_\theta| - \frac{1}{2}\log|\det\mathbf{W}| - \frac{1}{2}\mathbf{Z}^T(\mathbf{\Sigma}_\theta^{-1} - \mathbf{\Sigma}_\theta^{-1}\mathbf{M}\mathbf{W}^{-1}\mathbf{M}^T\mathbf{\Sigma}_\theta^{-1})\mathbf{Z} \tag{15}$$

where $\mathbf{W} = \mathbf{M}^T\mathbf{\Sigma}_\theta^{-1}\mathbf{M}$.

As explained in Section 3, the need for high computational speed forced us to discretize some of the parameters (smoothness and cutoff frequencies), so the usual continuous optimization routines are not applicable, and we use the simulated annealing method (Section 3.1) to maximize the likelihood.

The number of nodes can be determined using some model selection criteria. We compared the performance of Akaike Information Criterion (AIC, Akaike (1974)) and Bayes information criterion (BIC, Schwarz (1978)). Our simulation results indicate that AIC is more likely to select the correct number of nodes when the data are simulated using S+T model, and the prediction performance using AIC is also slightly better than BIC for all the models we used for simulation. We use AIC to select nodes in all the simulation studies presented in section 4. With the AIC, one additional node in the model should be compensated by an increase of at least one unit of log-likelihood in order to accept the larger model. Other criteria for selecting the number of nodes, as well as data driven methods for determining the locations of the nodes, could be considered, which we do not address in this paper.

3 Numerical Implementation

3.1 Simulated Annealing

We implemented the maximization using the simulated annealing method (Givens and Hoeting, 2005). This method is based on the way a physical system finds its minimum energy state when it is first heated to high temperature and then cooled down slowly to zero temperature. In our problem, the energy to be minimized is the negative of the log-likelihood function. One starts with an initial value of the parameters and calculates the energy. New values of the parameters are drawn from a proposal distribution and the new energy is calculated. If the new energy is lower than the previous one, the parameters get updated with the new values. If the new energy is higher than the initial energy, the new parameters are accepted with probability $\exp(-(E_f - E_i)/T)/(1 + \exp(-(E_f - E_i)/T))$, where E_i and E_f are the initial and final energies and T is the temperature. This helps keep the system from being trapped in local minima. These steps are repeated several times, after which the temperature is lowered and the same procedure is followed until the temperature is close to 0.

We have noticed that the convergence depends on the starting values of the parameters, most notably on the threshold frequency w_t. We use three different starting values of w_t ($1/r_{\min}$, $1/r_{\max}$, and the average of the two), and choose the estimated parameters that have the largest likelihood. We start by fitting a Matérn model and using the estimated sill (σ) and smoothness (ν) as starting values for the optimization. The initial values for the coefficients of the B-Splines are set to be constant 1 at all nodes. Since we normalize the spectral density so that it yields variance σ^2, the overall scale of the coefficients is not relevant.

The proposal distribution for the coefficients of the B-splines is a mixture of two lognormal

distributions, one with mean parameter zero and variance parameter one, and the other one with mean centered at the initial value and variance 0.1. The proposal distribution for the sill is also a mixture of lognormals, one with mean given by the sample variance of the observations and variance 1, and the other one centered at the initial value with variance 0.1. These numbers were chosen so that the convergence was satisfactory. We let w_t take 100 discrete values between $1/r_{\max}$ to $1/r_{\min}$. The proposal was a mixture of two uniform distributions, one that ranges over all 100 values, and the other one centered at the previous value with a range that is 10% of the whole range. Likewise, we let the smoothness parameter take 100 discrete values between 0.05 and 5. Larger values of smoothness give rise to almost singular covariance matrices. The proposal was also a mixture of uniforms, one centered at the previous value with a range that is 10% of the total range, and the other one over the whole range.

Several cooling schedules were tested. The one that gave slightly better convergence was one that updated the temperature in each step according to $T_i = \frac{T_{i-1}}{1+aT_{i-1}}$ with $a = 30$ and $T_0 = 1000$. We stopped after 10000-20000 iterations, after which no changes in parameters occurred. Each optimization took around 2-5 minutes on a Linux machine with dual AMD Opteron™ processors and 2Gb of memory.

3.2 Tabulation of Hankel transforms

To speed up the computation of the Hankel transforms of the truncated tail, we resort to some further approximations and shortcuts. We calculated the covariance function for $n_r = 100$ equispaced values between $r_{\min}$ and $r_{\max}$ and interpolated using cubic spline interpolation for distances between these points. Also, we restricted the values of the threshold frequency w_t and the power of the tail γ to $n_w = 100$ and $n_\gamma = 100$ discrete values. Namely, $w_t(j) = \frac{1}{r_{\max}} + \frac{j}{n_w}\left(\frac{1}{r_{\min}} - \frac{1}{r_{\max}}\right)$ and $\gamma(j) = 2 + 2\left(0.05 + \frac{j}{n_\gamma}(5 - 0.05)\right)$.

The Hankel transform of the truncated tail $\left(t(i,j,k) = \int_{w_t(k)}^{\infty} w^{1-\gamma(j)} J_o(wr(i))dw\right)$ was tabulated into an array of dimensions $n_r \times n_\gamma \times n_w = 100 \times 100 \times 100$. The Hankel transform of piecewise polynomials of the form $\mathbf{1}_{[w_i,w_{i+1})}(w - w_i)^m$ for $m = 0, 1, 2$, and 3 were tabulated into an array of dimensions $l \times 4 \times n_r \times n_w = l \times 4 \times 100 \times 100$, where l is the number of polynomial pieces used in the representation of the spectral density. In order to take advantage of this tabulation at the time of calculating the transform, we converted the splines $S(w)$ (linear combination of B-splines) into a piecewise polynomial form

$$S(w) = \sum_{i=0}^{l} \sum_{j=0}^{3} a_{ij} \mathbf{1}_{[w_i,w_{i+1})}(w - w_i)^j. \tag{16}$$

Hence the Hankel transform was reduced to multiplying the tabulated values by the corresponding coefficients.

4 Simulations

We first simulated Gaussian random fields with mean zero and various covariance functions and estimated the spectral density using S+T family of functions. For comparison purposes, we also estimated the covariance function using the Matérn model as well as the kernel method proposed by Hall et al. (1994). The locations were chosen to be where the National Acid Deposition Program sites are situated. We used a total of 63 sites that are shown in Figure 1. The smallest distance between sites is 14 km, the largest distance is 2000 km, and the median distance is 802 km. For simplicity, we use chordal distance and ignore the fact that the surface is spherical.

The covariance models we used to simulate the data are Matérn, polynomial Matérn, S+T, and spectral exponential. Polynomial Matérn is a family of spectral density which is the product of Matérn spectral density and a positive polynomial (f(w)=$\frac{((w-u)^2+v^2)((w+u)^2+v^2)}{(a^2+w^2)^{\nu+d/2}}$).

This function is positive on $\mathbb{R}^+$, thus a valid spectral density. The spectral density of the spectral exponential model has the form $f(w) = \exp(-w/\phi)$. In the simulation study we set $\phi = w_t$. Matérn, S+T and polynomial Matérn share the same high frequency behavior, namely, $1/\omega^\gamma$. The spectral exponential model has a much faster decay and has analytic realizations of the process. Though we do not consider this type of behavior to be reasonable for modeling natural physical processes, it is included here to test the method.

In order to assess the performance of each method, we look at the following quantities.

- Parameter values: When the true model and the model used to estimate the covariance functions have common parameters (for example, σ^2 is common to all models) the difference between the true and the estimated parameters is an obvious measure of performance.

- Likelihood values: The value of the likelihood also gives us an indication of how good the models are fitting the data. Although it is a bit unfair to compare methods that seek maximizing the likelihood with methods that seek to optimize other criteria, large deviations from the true likelihood should give us an idea of how good the estimated function is. We compute the log likelihood ratio between alternative methods and the S+T method. Positive value indicates that the alternative model is a better fit, and negative value indicates worse fit.

- Covariance function: Comparing the distance between the true and estimated covariance function or spectral densities seems to be an obvious and esthetically pleasing way of assessing the performance of methods. We use the $\mathcal{L}^2$-norm between the true and estimated covariance function as a measure of performance.

If we are interested in estimating the covariance function for interpolation purposes, this last method could be misleading. The following example from Stein (1999) illustrates this

point. Suppose the true model is $\exp(-r)$. The function $\exp(-r^2/2)$ is closer in mean squared sense to the true covariance function than $\exp(-2r)/2$ is. However, interpolating with the latter function can give smaller prediction errors and dramatically better estimates of uncertainties than the squared exponential covariance function $\exp(-r^2/2)$. We will see this effect in some of our simulation results. In most applications, the ultimate goal of estimating the covariance functions is the prediction of the random field at unobserved locations. To that end we use the following two quantities.

- Prediction error: It is most common to use the mean squared prediction error (MSPE) as an indicator of the goodness of fit. Following Stein (1999), we use $E_0 e_i^2 / E_0 e_0^2$, where E_i indicates expectation using the i^{th} covariance function, and e_i is the prediction error using the i^{th} covariance function, i.e., the difference between the true value of the random field and the predicted value using covariance function i. The true covariance function is labeled using subscript 0. It is easy to show that $E_0 e_i^2 / E_0 e_0^2 = 1 + E_0(\hat{Z}_i(x) - \hat{Z}_0(x))^2 / E_0 e_0^2$. We estimate the numerator in the second term by taking the sample mean (out of 100 simulations) of the squared difference between the interpolated values with the misspecified covariance function ($\hat{Z}_i$) and the interpolated values with the true covariance function ($\hat{Z}_0$).

- Error in estimating prediction variance: A common approach to estimating the variance of the predictions is to calculate it by plugging in the estimated variance parameters ($E_1 e_1^2$). Therefore, we would like the ratio between $E_1 e_1^2$ and the actual prediction variance ($E_0 e_1^2$) to be as close to one as possible. To treat underestimate and overestimate of prediction variance equally, we use $|\log(E_1 e_1^2 / E_0 e_1^2)|$ as the performance measure.

Table 1 shows the average results of running 100 simulations for each of the following

models: Matérn ($\nu = 3$, $\phi = 9.4$), polynomial Matérn ($\nu = 3$, $\phi = 9.4$, $u = 4.7$, $v = 0.94$), S+T ($\nu = 3$, $w_t = 9.4$, $l = 4$, $\mathbf{b} = (1, 0.2, 2, 0.6, 0.4)$), and spectral exponential ($\phi = 9.4$). For all the models $\sigma^2 = 1$, and the unit for ϕ and w_t is 1/1000 km.

Each simulation includes 200 independent realizations of a Gaussian random field at 63 locations, totaling 12,600 observations with the given covariance functions. The covariance matrix corresponding to this type of datasets is block diagonal, which allows us to have a large number of observations (so that the parameters can be estimated well) while keeping the computational load at a manageable level. For each simulation, the ML parameters for S+T and Matérn models were estimated. Also the kernel estimate of the covariance function was calculated.

The first three rows show the smoothness parameters, true and estimated, using S+T and Matérn models. When the true model is Matérn or polynomial Matérn, the estimated smoothness for S+T model is around 2.2, a bit smaller than the truth, which was 3. This is expected for the Matérn model, since the rate of decay of the estimating tail function $1/w^{2\nu+2}$ is faster than the rate of decay of the true tail function $1/(a^2 + w^2)^{\nu+1}$. We can see this more clearly by comparing the derivatives of the logs of the two tail functions

$$-\frac{2\nu+2}{w} \quad \text{vs.} \quad -\frac{2(\nu+1)w}{a^2+w^2} = -\frac{2\nu+2}{w} + \frac{2a^2(\nu+1)}{w(w^2+a^2)},$$

or more clearly

$$\nu \quad \text{vs.} \quad \nu - \frac{a^2(\nu+1)}{w^2+a^2}. \tag{17}$$

Hence the estimated power should be smaller in absolute value than the true power. The estimation process will try to match the left and right side of (17) in an average sense. A similar argument works for the polynomial Matérn function.

When simulating under the S+T model, there is no approximation in the tail, and the S+T model gives an estimated value that is very close to the truth. In the case of the

spectral exponential model, we do not have a true parameter with which to compare. We notice that S+T and Matérn methods give similar estimates of the smoothness when the true model is spectral exponential.

The estimated values of the sill (σ^2) are very close to the true value of 1 for all three methods except for the Matérn method when the true model is S+T, which yields a mean of 1.15 with standard deviation 0.03, and the kernel method when the truth is spectral exponential, which yields a mean of 0.97 with standard deviation 0.01. The S+T method gives 1.00 with error 0.02 when the true model is Matérn, polynomial Matérn and S+T. When the true model is spectral exponential, we get 1.00 with standard deviation 0.01.

The next two blocks of rows show the cutoff frequency w_t and the inverse range. For the S+T model, the average estimated value of w_t is 10.6 (in 1/1000km units) with standard deviation 1.4. The true parameter is 9.4, which is within one standard deviation of the estimated value. For the Matérn model both the true and estimated inverse range parameters are 9.4. The standard deviation of the estimated value is 0.3.

The $\mathcal{L}^2$-norm between the true and estimated covariance functions are shown next. The S+T method is better by factors of 2.3, 2.8, 2.8 and 1.8 compared to the kernel method when the true models are Matérn, polynomial Matérn, S+T, and spectral exponential, respectively. The S+T method also gives smaller $\mathcal{L}^2$-norm compared to Matérn method except when the Matérn model is the truth. The $\mathcal{L}^2$-norms were calculated by averaging the squared differences between the true and estimated covariance function values at 100 equispaced points in the range 0 to $r_{\max}$.

The last block shows the log likelihood ratio between alternative models and the S+T model. The S+T model gives larger likelihood in all cases except when we use the Matérn model to estimate and the data were simulated from the Matérn model. Even then, the advantage of using the correct parametric model is modest.

Table 2 shows the prediction performance of each method. The first three rows compare the prediction error using the median of MSPE scaled by the error of the best linear predictor minus one, $E_0e_i^2/E_0e_0^2 - 1$, for 100 sites located on a lattice that covers the observed region. The interquartile range (IQR), the difference between the third and the first quartiles, is shown between the parenthesis as a measure of variability. The MSPE of the predictors using the S+T method are less than 0.2% larger than the best linear predictor for all four simulated models. In absolute terms, the kernel method gives prediction errors 17.5%, 3.8%, 68.5% and 0.3% higher than the best linear predictor while the Matérn method gives errors 0.01%, 1.8%, 4.2% and 0.4% higher, and for S+T method it is only 0.16%, 0.12%, 0.05%, and 0.12% higher.

The next three rows compare the error in estimating the prediction variance using the square root of the median of $|\log(\frac{E_ie_i^2}{E_0e_i^2})|$ for the 100 sites described in the previous paragraph, where the $E_0e_0^2$ is evaluated using sample variance over the 100 simulations. One could also look at the median of $\frac{E_ie_i^2}{E_0e_i^2} - 1$, which gives similar results for this simulation.

Our method gives the estimated MSPE that differ about half a percent from the actual MSPE. As a comparison, the kernel method gives MSPE that differed from the actual MSPE by 59.1%, 27.0%, 100.8% and 5.8% for the four true models respectively, and the Matérn method's differences were 0.3%, 0.9%, 8.3%, and 1.2%, both substantially worse than the S+T method except for the Matérn method when the truth is Matérn.

We have done simulation studies for other parameter values with qualitatively similar results. In practice, we often need to estimate the mean. When the mean is unknown, we can use restricted likelihood to estimate the parameters, and the simulation results are very similar to the case when the mean is assumed known. These numerical results can be found in Im (2005) and are not presented here.

Figures 2-5 show the covariance function (left) and the spectral density (right) for the

Matérn, polynomial Matérn, S+T models, and spectral exponential models. These figures show one typical simulation from each model. As suggested by the $\mathcal{L}^2$-norm values from Table 1, the covariance functions estimated using S+T method (dots) are closer to the true covariance functions (solid line) than the kernel estimators (cross). The kernel estimator of the covariance function becomes wiggly for large distances, mainly because there are fewer pairs of observations that contribute to this region. The right side of Figures 2-5 show that our method yields estimates of the spectral densities that are quite close to the truth except near the origin. The Matérn method is not able to reproduce the structure of the spectral density at low and mid frequencies, but it captures reasonably well the tail behavior. The kernel estimates of the spectral density follow the overall shape of the function, but they are very wiggly. We also see several intervals of frequencies where the kernel estimator take value 0, which are due to the truncation of the function necessary to ensure positive definiteness of the corresponding covariance function. When we calculated the spectral density, we followed the ad hoc solution proposed by Hall et al. (1994): from some point T_1 use a straight line that goes from the value of the estimator at T_1 to zero at some other distance T_2. We chose $T_1 = 1500$km and $T_2 = 2000$km. The actual transformation was done using the fast Hankel transform method proposed by Siegman (1977).

5 Application to rainfall data

In this section we applied the three aforementioned methods to an annual rainfall dataset in the eastern US to compare their performance in prediction. We chose the study region to be between latitude 27.1 to 49.0 and longitude -100.5 to -80.2, which includes 2742 stations (dots on the right of Figure 1). The time period of the data is between 1960 and 1999. Detailed documentation of this dataset can be found at

http://www1.ncdc.noaa.gov/pub/data/documentlibrary/tddoc/td9651.pdf. We used 2182 stations that had no missing data. For each station data we subtracted the mean over the 40 years of data and modeled the difference as a Gaussian random process. We looked at the normal quantile plots of the difference for all the stations at each year, and in most of the years the data are consistent with the Gaussian assumption. Some of those normal quantile plots appear to have a few large outliers. The number of possible outliers are never larger than 0.5% of the stations. No attempts were made to identify those outliers. Instead we used robust measures to compare the three methods. An examination of the autocorrelation and cross correlation of the observations at different stations reveals no significant time dependence, thus we model the observations from each year as independent realizations of the same process with a different mean. We randomly chose 200 stations among 2182 stations to estimate the covariance function using all three methods (S+T, Matérn, and kernel). The empirical variogram suggested the need of including a nugget term in the covariance model. This is straightforward for the two likelihood methods (S+T and Matérn). The kernel method needs some modification in order to handle the nugget effect. One possible way is to estimate the covariance function using only distinct pairs, i.e., $(Z_i - \bar{Z})(Z_j - \bar{Z})$ with $i \neq j$. This would leave out the nugget term. One could estimate the nugget by subtracting the estimated covariance function at zero from the sample variance. With the estimated covariance functions, we kriged the data at the remaining 1982 stations, and calculated the median squared prediction error (SPE, squared difference between the predicted values and the actual data) as a measure of prediction performance, as well as the median of the log ratio of the SPE and the estimated prediction variance (EPV) as a measure of performance in estimating prediction variance. The average results for 10 different samples of the 200 stations are given in Table 3, and the location of the first sample of the 200 stations are show as "+" on the right of Figure 1. For both prediction and

estimating the prediction variance, our method slightly outperforms the Matérn method, and is substantially better than the kernel method. For prediction, the average median SPE of our method is 43% smaller than that of the kernel method, and for estimating the prediction variance, our method has 25% smaller average median log(SPE/EPV). We also note that the Matérn method consistently outperforms the kernel method. The results shown for the kernel method did not take into account the nugget effect. When the method was modified to include the nugget term, the performance was much worse. We have tried using 100 stations and the results were similar.

6 Summary and discussion

In this paper we propose a new method to estimate spectral densities of isotropic Gaussian processes with scattered data using a flexible semiparametric S+T model, whose parameters can be estimated using ML or REML methods. We have calculated explicit expressions of the Hankel transform of the spectral density and tackled several numerical issues arising during the computation of the covariance function. Simulated annealing method is used to maximize either the likelihood (when the mean is known) or the restricted likelihood.

To compare our method with other existing method for estimating spectral density, we simulated observations with Matérn, polynomial Matérn, S+T, and spectral exponential spectral densities. Our simulation results showed that our method (S+T) outperforms the non-parametric kernel method in terms of estimated sill, $\mathcal{L}^2$-norms of the covariance functions, the likelihood values (this is expected since we are maximizing the likelihood, but the large differences seen may be indicative of poor performance), MSPE, and errors in the estimated variances of the predictions. Our method also outperforms the parametric method using Matérn covariance model when the true model is not Matérn by all these per-

formance criteria. All three methods are applied to a rainfall data to compare the prediction performance, with our method doing better than the other two.

The MSPE and the errors of the estimated variances of the predictions are the most relevant measures of performance when our ultimate goal is interpolation to locations where there are no observations. With this criteria, the Matérn method outperforms the kernel method, although it generally has larger $\mathcal{L}^2$-norm values of the covariance function than the kernel method when the truth is not Matérn. The reason for the better prediction properties of the Matérn method is that the tail properties of the spectral function play a fundamental role in the prediction. Our method directly estimates the tail property just like the Matérn method does, while it also offers more flexibility for modeling the lower frequencies, which improves the predictions. Our method outperforms the kernel method even when the true model is spectral exponential, which has an exponential tail, whereas our method assumes an algebraic tail.

We have performed simulations with smaller number of replicates (20 and 1 instead of 200) per simulation. Using the prediction performance criteria, we found that our method outperforms the Matérn method when the number of replicates is 20, but not so when we only have one replicate of the spatial process. So our method should be applied with caution when one does not have a large amount of data.

We applied all three methods to annual rainfall data and showed that our method slightly outperforms Matérn method and substantially outperforms the kernel method in terms of prediction error and estimated uncertainty. Including a nugget term in both our method and Matérn method is straightforward. The kernel method can be modified in order to include the nugget effect, but its performance for this dataset was worse than the original method.

In this work we used equally spaced nodes for the B-splines and did not attempt to estimate the optimal spacing of the nodes. Our method can be modified to allow the spacing

of the nodes to be estimated from the data. One possible way of doing this efficiently is to utilize the information contained in the kernel estimator. Even though the kernel method fails to capture the tail behavior of the spectral density, it does seem to have useful information about the mid frequency shape of the function. We plan to address this possibility in a separate paper.

Acknowledgements

The research described herein has been funded wholly or in part by the United States Environmental Protection Agency through STAR Cooperative Agreement #R-82940201-0 to The University of Chicago. It has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred. Ken Wilder provided us with C code necessary for calculating hypergeometric functions accurately, for which we are very grateful. Thanks also goes to Pavel (Pasha) Groisman from the National Climatic Data Center for providing the rainfall dataset.

References

Abramowitz, M. and Stegun, I. (1965), *Handbook of Mathematical Functions*, New York: Dover, 9th ed.

Akaike, H. (1974), "A new look at the statistical identification model," *IEEE Transactions on Automatic Control*, 19, 716–723.

de Boor, C. (2001), *A Practical Guide to Splines*, Springer.

Genton, M. G. and Gorsich, D. J. (2002), "Nonparametric variogram and covariogram estimation with Fourier-Bessel matrices," *Computational statistics & Data Analysis*, 41, 47–57.

Givens, G. H. and Hoeting, J. A. (2005), *Computational Statistics*, Wiley.

Hall, P., Fisher, N., and Hoffmann, B. (1994), "On the nonparametric estimation of covariance functions," *Annals of Statistics*, 22, 2115–2134.

Handcock, M. S. and Wallis, J. R. (1994), "An approach to statistical spatial temporal modeling of meteorological fields," *Journal of the American Statistical Association*, 89, 368–378.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman and Hall, 2nd ed.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Shapiro, A. and Botha, J. D. (1991), "Variogram fitting with a genral class of conditionally nonnegative definite functions," *Computational Statistics & Data Analysis*, 11, 87–96.

Siegman, A. E. (1977), "Quasi fast Hankel transform," *Optics Letters*, 1, 13–15.

Stein, M. L. (1999), *Statistical Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer.

— (2002), "The screening effect in kriging," *Annals of Statistics*, 30, 298–323.

Wolfram Research,Inc. (2001a), "Gamma function," `http://functions.wolfram.com/Gamma`.

— (2001b), "Gamma function," `http://functions.wolfram.com/GammaBetaErf/PolyGamma`.

— (2001c), "Generalized Hypergeometric Function ${}_1F_2$," http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric1F2/.

Yaglom, A. M. (1987), *Correlation Theory of Stationary and Related Random Functions*, vol. 1, New York: Springer-Verlag.

Appendix

A B-splines

The following is a summary of the properties of B-splines relevant for this work. For a more extensive description see de Boor (2001).

A piecewise polynomial of order k with strictly increasing break (node) sequence ξ is a function of the form

$$\sum_j \mathbf{1}_{[\xi_j,\xi_{j+1})} p^{(k)}(x) \tag{18}$$

where $p^{(k)}(x)$ is a polynomial of degree $k-1$ or smaller. The set of all piecewise polynomial functions of order k with break sequence ξ is denoted $\Pi_{<k,\xi}$.

B-splines are defined in terms of a non-decreasing knot sequence $\mathbf{t} = (t_j)$. The j^{th} B-spline of order 1 for knot sequence $\mathbf{t}$ is the indicator function of the interval $[t_j, t_{j+1})$:

$$B_{j,1} := B_{j,1,\mathbf{t}} = \mathbf{1}_{[t_j,t_{j+1})}. \tag{19}$$

If $t_j = t_{j+1}$, $B_j = 0$. The j^{th} B-spline of order $k > 1$ is defined by the following recurrence relation

$$B_{jk} := B_{j,k,\mathbf{t}} := w_{jk} B_{j,k-1} + (1 - w_{j+1,k}) B_{j+1,k-1} \tag{20}$$

with $w_{jk} := w_{j,k,\mathbf{t}} := \frac{x-t_j}{t_{j+1}-t_j}$. B_{jk} is a piecewise polynomial function with break sequence $t_k, ..., t_{j+k}$. It is positive on $[t_j, t_{j+k}]$ and 0 outside this interval. B-splines of order k with

knot sequence $\mathbf{t}$ span the space of piecewise polynomial functions of order k with break sequence ξ and continuity conditions on the breaks given by the multiplicity of the knots. More specifically, the sum of the number of continuity conditions at a break ξ_j and the number of repeated knots at ξ_j equals the order k.

For the uniform knot sequence $t = (..., -\Delta, 0, \Delta, 2\Delta, ...)$ the corresponding B-splines are

$$B^{\Delta}_{j,k,\mathbf{t}}(x) = \sum_{r=0}^{k} \frac{(-1)^{k-r}}{(k-1)!} \binom{k}{r} (r - x/\Delta + j)_+^{k-1} \tag{21}$$

In particular, for $k = 4$, $j = 0$, and $\mathbf{t} = \mathbb{Z}$ and $\Delta = 1$

$$B(x) := B^1_{0,4,\mathbb{Z}}(x) = \begin{cases} x^3/6 & \text{if } 0 \le x < 1; \\ (-3x^3 + 12x^2 - 12x + 4)/6 & \text{if } 1 \le x < 2; \\ (3x^3 - 24x^2 + 60x - 44)/6 & \text{if } 2 \le x < 3; \\ (-x^3 + 12x^2 - 48x + 64)/6 & \text{if } 3 \le x < 4; \end{cases} \tag{22}$$

For arbitrary Δ and $j \neq 0$ the B-splines are obtained from (22) by translating the argument by $j\Delta$ and scaling it by $1/\Delta$, i.e., $B^{\Delta}_{j,4,\mathbb{Z}}(x) = B^1_{0,4,\mathbb{Z}}(\frac{x - j\Delta}{\Delta})$.

B Hankel transform of polynomials

The Hankel Transforms of piecewise polynomials of the form $(w - c)^m$ are given by

$$\int_{a'}^{b'} (w - k')^m w J_o(wr) dw = r^{-m-2} \int_{a'r}^{b'r} u(u - k'r)^m J_o(u) du, \tag{23}$$

which for $a = a'r$, $b = b'r$, $k = k'r$, and $m = 0$, 1 ,2, and 3 are

$$\int_a^b u J_o(u) du = -a J_1(a) + b J_1(b),$$

$$\begin{aligned} \int_a^b (u - k) u J_o(u) du &= J_1(a)(-a^2 + a(k + (\pi H_o(a))/2)) + J_1(b)(b^2 + b(-k - (\pi H_o(b))/2)) \\ &\quad -(a\pi J_o(a) H_1(a))/2 + (b\pi J_o(b) H_1(b))/2, \end{aligned}$$

$$\begin{aligned}\int_a^b (u-k)^2 u^2 J_o(u)du &= J_1(a)(-a^3+2a^2k+a(4-k^2-k\pi H_o(a))) + J_o(a)(-2a^2+ak\pi H_1(a)) \\ &\quad + J_1(b)(b^3-2b^2k+b(-4+k^2+k\pi H_o(b))) + J_o(b)(2b^2-bk\pi H_1(b)),\end{aligned}$$

and

$$\begin{aligned}\int_a^b (u-k)^3 u J_o(u)du &= J_1(a)(-a^4+3a^3k+a^2(9-3k^2)+a(-12k+k^3+(-9/2+(3k^2)/2)\pi H_o(a))) \\ &\quad + J_1(b)(b^4-3b^3k+b^2(-9+3k^2)+b(12k-k^3+(9/2-(3k^2)/2)\pi H_o(b))) \\ &\quad + J_o(a)(-3a^3+6a^2k+a(9/2-(3k^2)/2)\pi H_1(a)) \\ &\quad + J_o(b)(3b^3-6b^2k+b(-9/2+(3k^2)/2)\pi H_1(b)),\end{aligned}$$

where $J_\nu(\cdot)$ are Bessel functions of the first kind of order ν and $H_\nu(\cdot)$ are Struve functions of order ν (Abramowitz and Stegun, 1965).

C Continuity of derivative

For uniform knots with spacing Δ, the values of the spectral density on (w_{n-1}, w_n) where $w_n = \Delta n = w_t$ is

$$f(w) = f_n B(\frac{w-w_{n-2}}{\Delta}) + f_{n-1} B(\frac{w-w_{n-3}}{\Delta}) + f_{n-2} B(\frac{w-w_{n-4}}{\Delta}) + f_{n+1} B(\frac{w-w_{n-1}}{\Delta}). \tag{24}$$

It follows that

$$f'(w) = f_n B'(\frac{w-w_{n-2}}{\Delta}) + f_{n-1} B'(\frac{w-w_{n-3}}{\Delta}) + f_{n-2} B'(\frac{w-w_{n-4}}{\Delta}) + f_{n+1} B'(\frac{w-w_{n-1}}{\Delta}).$$

At $w = w_t$, $f'(w_t) = -\frac{nf_{n-1}}{2w_n} + \frac{nf_{n+1}}{2w_t}$. If we want $f'(w)$ to be continuous at w_t, we need $f'(w_t) = -\frac{-\gamma f_t}{w_t}$, where $f_t = \frac{1}{6}f_{n-1} + \frac{2}{3}f_n + \frac{1}{6}f_{n+1}$. Thus we have

$$f_{n+1} = \frac{3n-\gamma}{3n+\gamma} f_{n-1} - \frac{4\gamma}{3n+\gamma} f_n. \tag{25}$$

D Positivity of the spectral density

Our model allows the values of the $(n+1)^{\text{th}}$ coefficient to be negative, and we need to check whether the spectral density is still positive on the interval $(0, w_t)$. Because of the linearity of the derivative, it is enough to consider separately the cases where $f_n = 0$ and $f_{n-1} = 0$. If we get positive spectral density in each case the sum of the two will also result in a positive function. Also, since the support of the B-spline with coefficient f_{n+1} (B) is (w_{n-1}, w_{n+3}), we only need to worry about the interval (w_{n-1}, w_n).

Case $f_{n-1} = 0$

Since each of the terms in (24) is nonnegative except possibly for the term corresponding to f_{n+1}, we have

$$f(w) \geq f_n B(x) + f_{n+1} B(x-1) := g(x), \tag{26}$$

where $x = \frac{w - w_{n-2}}{\Delta}$. To show $f(w) \geq 0$ for $w \in (w_{n-1}, w_n)$, it is enough to show $g(x) \geq 0$ for $x \in (1, 2)$. Substituting the corresponding piecewise polynomial in each case and using (25), we get

$$\begin{aligned} g(x) &= \frac{f_n}{6}\left(-3x^3 + 12x^2 - 12x + 4 - \frac{4\gamma}{3n+\gamma}(x-1)^3\right) \\ &= \frac{f_n}{6}\left(\frac{2\gamma + 3n}{2(3n+\gamma)}(2-x)(4 - 10x + 7x^2) + \frac{3n}{2(3n+\gamma)}x^3\right). \end{aligned} \tag{27}$$

Since $4 - 10x + 7x^2 > 0$, it follows that $g(x) \geq 0$ for $x \in (1, 2)$.

Case $f_n = 0$

We have

$$f(w) \geq f_{n-1} B(x) + f_{n+1} B(x-2) := g(x), \tag{28}$$

where $x = \frac{w - w_{n-3}}{\Delta}$, and $x \in (2, 3)$ for $w \in (w_{n-1}, w_n)$. To show $g(x) \geq 0$ for $x \in (2, 3)$, we next show that the function $g(x)$ is monotone decreasing and that the value at $x = 3$ is

positive. Since

$$g(x) = \frac{f_{n-1}}{6}\left(3x^3 - 24x^2 + 60x - 44 + \frac{3n-\gamma}{3n+\gamma}(x-2)^3\right), \tag{29}$$

we have $g(3) = f_{n-1}\frac{n}{3n+\gamma} > 0$, and

$$g'(x) = 6f_{n-1}\frac{(x-2)(\gamma(x-4)+6n(x-3))}{3n+\gamma} \leq 0 \tag{30}$$

for $x \in (2,3)$. Thus $g(x) > 0$ for $x \in (2,3)$

E Tail integral with integer smoothness

Using the series representation of the hypergeometric function, the integral of the tail given in (11) can be written as

$$\int_{s_t}^{\infty} u^{1-\gamma} J_o(u)du = \frac{(-\gamma/2)\Gamma(-\gamma/2)}{2^{\gamma-1}\Gamma(-\gamma/2)} + \frac{s_t^{2-\gamma}}{\gamma-2}\sum_{k=0}^{\infty}\frac{-\gamma/2+1}{-\gamma/2+1+k}\frac{(-st^2/4)^k}{k!^2}. \tag{31}$$

Let $-\gamma/2+1 = -n+\delta$ with $n \in \mathbb{N}$. As δ goes to zero, the first term in (31) and the n^{th} term in the second term of (31) go to infinity but the total contribution of the diverging terms is finite. This can be shown by using the asymptotic expansion of the gamma function when the argument is close to a negative integer. We reorder the terms in (31) to make explicit the two terms that diverge when $\gamma/2$ is integer valued:

$$\begin{aligned}\int_{s_t}^{\infty} u^{1-\gamma} J_o(u)du =& \frac{\Gamma(-n+\delta)}{2^{\gamma-1}\Gamma(n+1-\delta)} + \frac{s_t^{2-\gamma}}{\gamma-2}\frac{-n+\delta}{\delta}\frac{(-st^2/4)^n}{n!^2} + \\ &+ \frac{s_t^{2-\gamma}}{\gamma-2}\sum_{k=0,k\neq n}^{\infty}\frac{-\gamma/2+1}{-\gamma/2+1+k}\frac{(-st^2/4)^k}{k!^2}.\end{aligned} \tag{32}$$

Using $\Gamma(-n+\delta) = \frac{(-1)^n}{n!\delta} + \frac{(-1)^n\psi(n+1)}{n!} + O(\delta)$ (Wolfram Research,Inc., 2001a), where $\psi(n+1)$ is the digamma function (Wolfram Research,Inc., 2001b), we get

$$\int_{s_t}^{\infty} u^{1-\gamma} J_o(u)du = -\frac{2^{-1-2\,n}\,{s_t}^{2\,(\delta-n)}\left(-{s_t}^2\right)^n}{\delta\,n!^2} - \frac{(-1)^n\,2^{-1+2\,\delta-2\,n}}{\delta\,n!\,\Gamma(1-\delta+n)} + \frac{(-1)^n\,2^{-1+2\,\delta-2\,n}\,\psi(0,1+n)}{n!\,\Gamma(1-\delta+n)} + \sum_{k=0,k\neq n}^{\infty} ... + O(\delta). \tag{33}$$

By letting δ go to zero, we have

$$\int_{s_t}^{\infty} u^{1-\gamma} J_o(u)du = \frac{\log(2) - \log(s_t) + \psi(n+1)}{(-4)^n n!^2} + \frac{s_t^{2-\gamma}}{\gamma-2} \sum_{k=0,k\neq n}^{\infty} \frac{-\gamma/2+1}{-\gamma/2+1+k} \frac{(-s_t^2/4)^k}{k!^2}. \tag{34}$$

F Asymptotic expansion of tail

We use the following asymptotic expansion of ${}_1F_2$ for large z to find an approximate expression for the truncated tail integral.

$$\begin{aligned}
{}_1F_2(a_1;b_1,b_2;z) \approx{}& \frac{\Gamma(b_1)\,\Gamma(b_2)\,(-z)^{a_1}}{\Gamma(-a_1+b_1)\,\Gamma(-a_1+b_2)} \left(1 + \frac{a_1\;(1+a_1-b_1)\;(1+a_1-b_2)}{z}\right. \\
&\left. + \frac{a_1\;(1+a_1)\;(1+a_1-b_1)\;(2+a_1-b_1)\;(1+a_1-b_2)\;(2+a_1-b_2)}{2\,z^2} + \dots\right) \\
&+ \frac{(-z)^{\chi}\,\Gamma(b_1)\,\Gamma(b_2)}{2\,\sqrt{\pi}\,\Gamma(a_1)} \left(\cos\left(2\,\sqrt{-z} + \pi\chi\right)\,\left(1 + \frac{d_2}{z} + \dots\right) + \sin\left(2\,\sqrt{-z} + \pi\chi\right)\left(\frac{d_1}{\sqrt{-z}} + \dots\right)\right)
\end{aligned} \tag{35}$$

for large z (Wolfram Research,Inc., 2001c), where

$$\begin{aligned}
\chi =& \frac{1}{2}\left(\frac{1}{2} + a_1 - b_1 - b_2\right), \\
d_1 =& \frac{1}{16}\left(-3 + 12\,{a_1}^2 - 4\,{b_1}^2 + 8\,b_2 - 4\,{b_2}^2 + 8\,b_1\;(1+b_2) - 8\,a_1\;(1+b_1+b_2)\right), \\
d_2 =& \frac{1}{512}\left(-15 + 144\,{a_1}^4 + 16\,{b_1}^4 + 16\,b_2 + 56\,{b_2}^2 - 64\,{b_2}^3 + 16\,{b_2}^4\right. \\
& -64\,{b_1}^3\;(1+b_2) - 64\,{a_1}^3\;(7+3\,b_1+3\,b_2) + 8\,{b_1}^2\;\left(7+8\,b_2+12\,{b_2}^2\right) \\
& +16\,b_1\;\left(1+25\,b_2+4\,{b_2}^2-4\,{b_2}^3\right) - 8\,{a_1}^2\;\left(-43+4\,{b_1}^2-72\,b_2+4\,{b_2}^2-8\,b_1\;(9+5\,b_2)\right) \\
& \left.+16\,a_1\;\left(-1+4\,{b_1}^3-25\,b_2-4\,{b_2}^2+4\,{b_2}^3-4\,{b_1}^2\;(1+b_2)-b_1\;\left(25+40\,b_2+4\,{b_2}^2\right)\right)\right).
\end{aligned}$$

Substituting this expression for ${}_1F_2$ in (11) and we have (13) for $wr \to \infty$.

		Truth			
		Matérn	polMatérn	S+T	spect. exp'l
ν	true	3	3	3	–
	S+T	2.15 (0.10)	2.21 (0.26)	3.00 (0.06)	1.82 (0.70)
	Mat	2.98 (0.09)	10.00 (2.39)	7.47 (0.31)	1.63 (0.18)
σ^2	true	1	1	1	1
	S+T	1.00 (0.02)	1.00 (0.02)	1.00 (0.02)	1.00 (0.01)
	Mat	1.00 (0.02)	1.00 (0.02)	1.15 (0.03)	1.00 (0.01)
	Ker	0.99 (0.03)	0.98 (0.02)	0.99 (0.02)	0.97 (0.01)
$w_t \times 1000$	true	–	–	9.4	–
	S+T	12.3 (1.2)	18.7 (2.1)	10.6 (1.4)	32.7 (5.1)
inv.range	true	9.4	–	–	–
$\times$1000	Mat	9.4 (0.3)	35.4 (4.6)	15.3 (0.5)	41.0 (195.7)
$\mathcal{L}^2$-norm	S+T	8.7 (4.7)	7.9 (3.9)	7.4 (2.7)	5.2 (1.7)
Cov	Mat	7.4 (5.0)	53.6 (0.3)	123.9 (7.8)	10.4 (16.0)
$\times$1000	Ker	20.6 (6.7)	21.9 (4.8)	20.9 (6.2)	9.2 (1.6)
loglik	Mat	4 (4)	-181 (22)	-296 (22)	-20 (96)
ratio	Ker	-1093 (167)	-254 (34)	-3173 (246)	-28 (9)

Table 1: Summary of simulation results for $\nu = 3.00$, $\sigma^2 = 1.00$, and inverse range (or w_t) = 9.4(1/1000 km). Maximum likelihood estimation. Average estimates from 100 simulations are shown. Each simulation consisted of 12,600 observations (200 replications of 63 spatially correlated observations). Standard deviations are shown in parentheses. Columns correspond to true models and rows correspond to estimating methods: ML using S+T, ML using Matérn, and kernel. Log-likelihoods are differences from S+T.

		Matérn	polMatérn	S+T	spect. exp'l
$E_0e_1^2/E_0e_0^2 - 1$	S+T	0.16 (0.10)	0.12 (0.05)	0.05 (0.07)	0.00.12 (0.00.08)
$E_0e_2^2/E_0e_0^2 - 1$	Mat	0.01 (0.00)	1.77 (3.17)	4.24 (3.17)	0.00.40 (0.00.63)
$E_0e_3^2/E_0e_0^2 - 1$	Ker	17.51 (32.30)	4.10 (9.10)	68.54 (118.61)	0.00.26 (0.00.45)
$\lvert\log(E_1e_1^2/E_0e_1^2)\rvert$	S+T	0.93 (2.83)	0.39 (0.68)	0.33 (0.52)	0.31 (1.29)
$\lvert\log(E_2e_2^2/E_0e_2^2)\rvert$	Mat	0.30 (0.76)	0.86 (2.28)	8.57 (17.24)	1.21 (2.16)
$\lvert\log(E_3e_3^2/E_0e_3^2)\rvert$	Ker	59.08 (94.75)	27.06 (43.31)	100.78 (157.13)	5.79 (16.41)

Table 2: Simulations as in Table 1. $\frac{E_0e_i^2}{E_0e_0^2} - 1$ represents the additional prediction error made by using the misspecified covariance in percentual unit. We show the median over 100 prediction locations and the IQR (interquartile range) between parentheses. $\frac{E_ie_i^2}{E_0e_i^2}$ represents the prediction error variance calculated with the misspecified covariance scaled by the actual prediction error variance. We show the median over 100 prediction locations of $\lvert\log(\frac{E_ie_i^2}{E_0e_i^2})\rvert$ and the IQR between parentheses.

	S+T	Matérn	Kernel
median SPE	3.98	4.15	6.96
median log(SPE/EPV)	1.30	1.41	1.74

Table 3: Prediction performance comparison for annual rainfall data in the eastern US.Each entry are average results of ten samples, and each sample represents one incidence where 200 sites are randomly chosen as observations to estimate the model, and the median SPE (squared differences between the predicted values and the actual data) as well as the median of log(SPE/EPV) (log ratio of the SPE and the estimated prediction variance) at the rest of the sites are computed to compare the three methods.

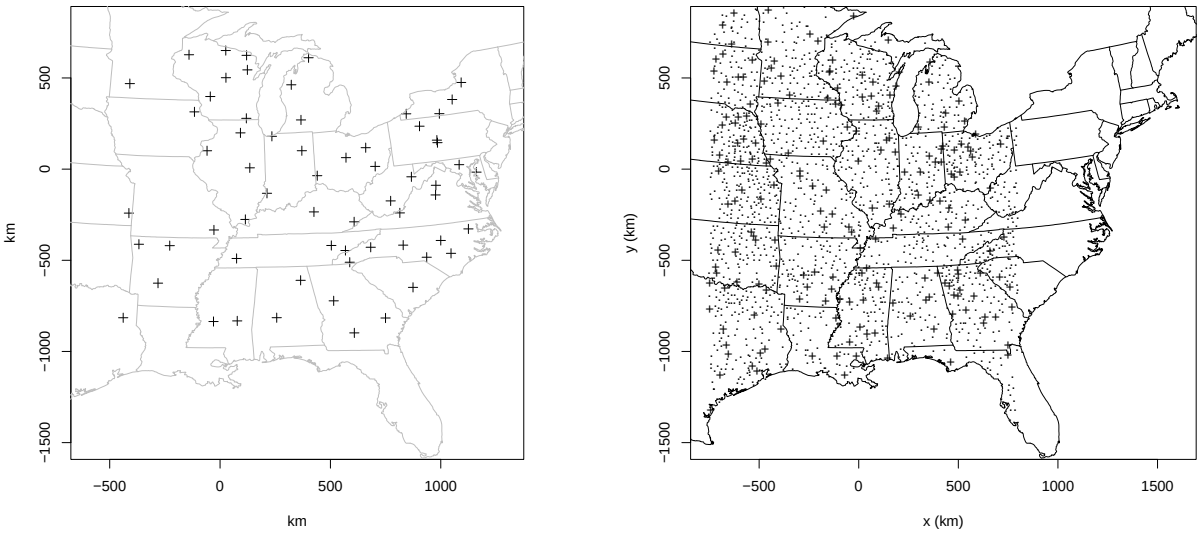

Figure 1: Plot of NADP monitoring sites used for simulations (left), and the monitoring sites for the rainfall dataset (right).

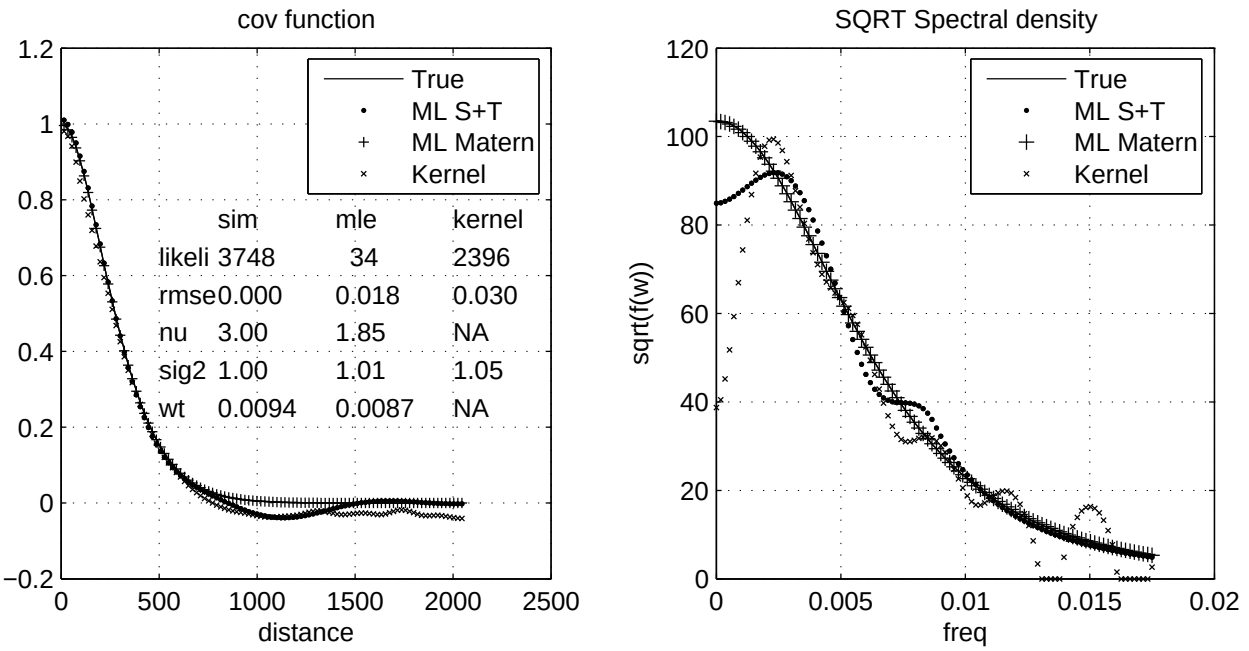

Figure 2: True and estimated a) covariance function and b) spectral density. The true model is Matérn with $\nu = 3$, $\sigma^2 = 1.00$, and inverse range $= 9.4$ (1/1000km)

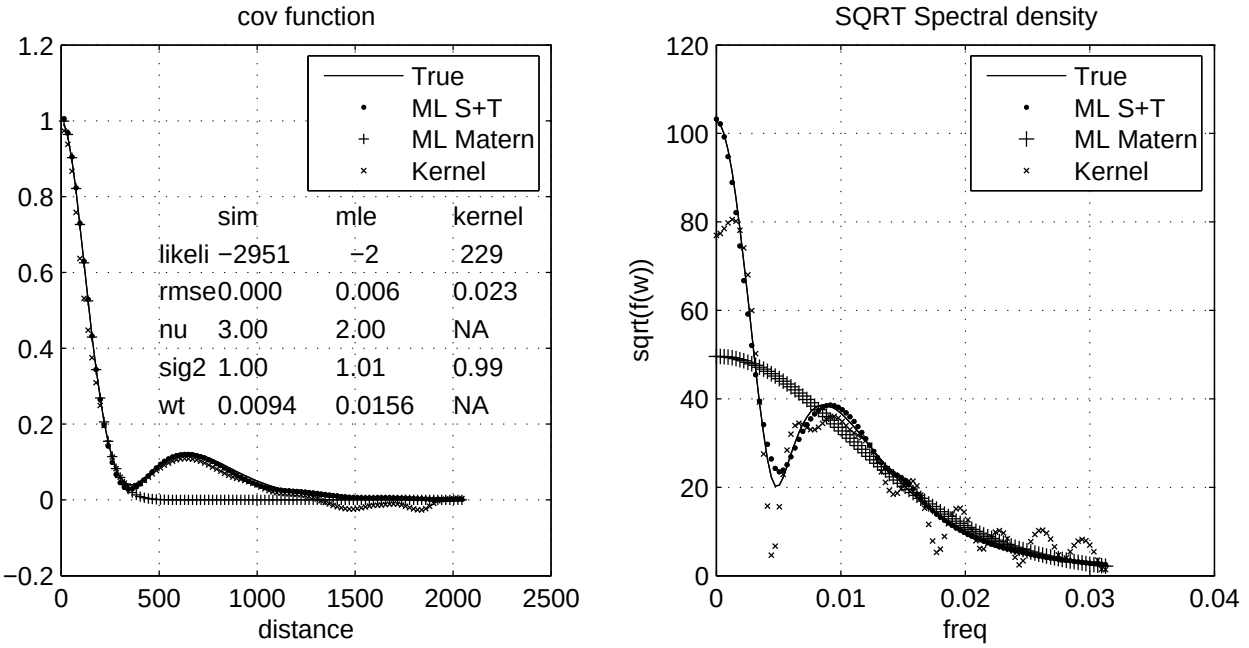

Figure 3: True and estimated a) covariance function and b) spectral density. The true model is polynomial Matérn with $\nu = 3$, $\sigma^2 = 1.00$, inverse range (or w_t) $= 9.4$ (1/1000km), $u = 0.5w_t$, and $v = 0.1w_t$.

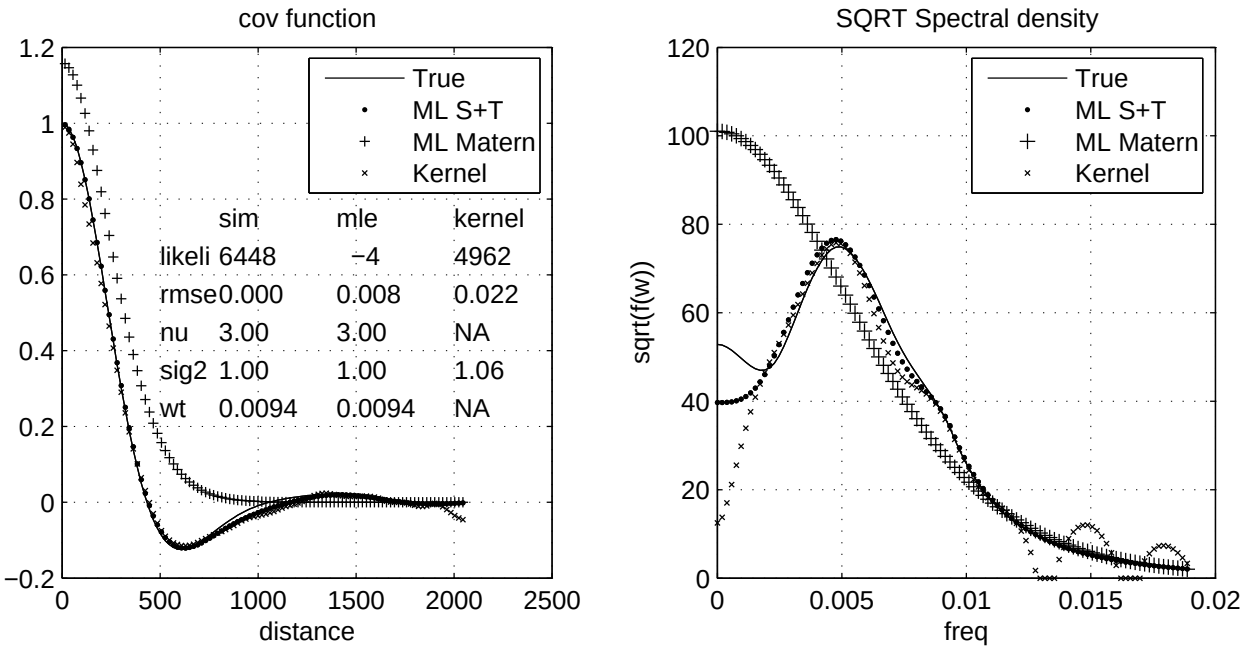

Figure 4: True and estimated a) covariance function and b) spectral density. The true model is S+T with $\nu = 3$, $\sigma^2 = 1.00$, $w_t = 9.4$ (1/1000km), and coefficients $\mathbf{b} = (1, 0.2, 2, 0.6, 0.4)$

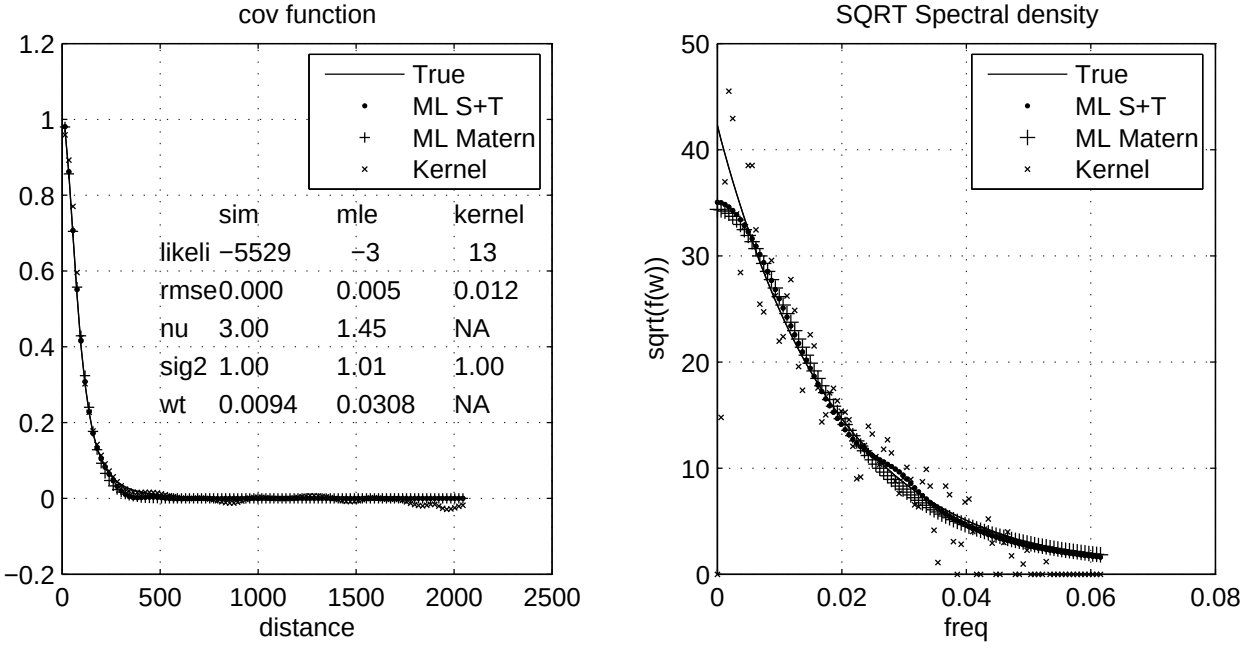

Figure 5: True and estimated a) covariance function and b) spectral density. The true model is spectral exponential with $\sigma^2 = 1.00$ and inverse range $= 9.40$ (1/1000km)