Statistica Sinica **25** (2015), 41-59
doi:http://dx.doi.org/10.5705/ss.2013.227w

A COVARIANCE PARAMETER ESTIMATION METHOD FOR POLAR-ORBITING SATELLITE DATA

Michael T. Horrell and Michael L. Stein

University of Chicago

Abstract: We consider the problem of estimating an unknown covariance function of a Gaussian random field for data collected by a polar-orbiting satellite. The complex and asynoptic nature of such data requires a parameter estimation method that scales well with the number of observations, can accommodate many covariance functions, and uses information throughout the full range of spatio-temporal lags present in the data. Our solution to this problem is to develop new estimating equations using composite likelihood methods as a base. We modify composite likelihood methods through the inclusion of an approximate likelihood of interpolated points in the estimating equation. The new estimating equation is denoted the I-likelihood. We apply the I-likelihood method to 30 days of ozone data occurring in a single degree latitude band collected by a polar orbiting satellite, and we compare I-likelihood methods to competing composite likelihood methods. The I-likelihood is shown capable of producing covariance parameter estimates that are equally or more statistically efficient than competing composite likelihood methods and to be more computationally scalable.

Key words and phrases: Composite likelihood, estimating equations, Gaussian random elds, Godambe information, remote sensing.

1. Introduction

Analysis of spatial and spatio-temporal data sets with a large number of observations has driven many recent advances in statistical modeling and computation (Stein, Chi, and Welty (2004), Kaufman, Schervish, and Nychka (2008), Cressie and Johannesson (2008) Banerjee et al. (2008), Liang et al. (2013), among many others). Spatial and spatio-temporal data are commonly modeled using Gaussian Random Fields (GRFs) with parameterized mean and covariance functions determined by unknown parameter vectors, μ and θ respectively. However, evaluating the likelihood of n observations from a GRF is generally a procedure of $O(n^3)$ complexity and requires $O(n^2)$ RAM storage. Maximum likelihood estimation via numerical optimization can therefore become difficult on desktop computers when n is only moderately large. Though classification of a data set as 'large' depends on the problem and the computational resources available, following Kaufman, Schervish, and Nychka (2008), for systems with 2GB memory,

a data set of size 10,000 may yield an unstructured covariance matrix too large to store in RAM and thus requires frequent and slow reading and writing to hard disk. With the advance of remote sensing techniques and the development of spatio-temporal data sets with billions of observations, a number of covariance parameter estimation methods for large n data sets have been developed to work in different contexts. In this paper, we consider the problem of covariance parameter estimation in the context of data produced by recording instruments aboard polar-orbiting satellites.

Polar-orbiting satellites are often used to collect daily global data on atmospheric conditions and other natural phenomena. Currently, NASA employs a group of polar-orbiting satellites called the A-Train to collect information on total column ozone, temperature, rainfall, aerosols and other data of environmental interest. In 2014, NASA is set to expand this program to include instruments that measure global concentrations of carbon dioxide (NASA (2012)). The daily global coverage provided by these satellites makes them attractive scientific instruments from a data collection perspective, but it also places these data sets solidly in the large n category. For example, the Ozone Monitoring Instrument (OMI) on NASA's Aura polar-orbiting satellite collects over 1 million observations per day.

In addition to the computational difficulties in analyzing these data sets, two more characteristics of polar-orbiting satellite data create problems for covariance parameter estimation. First, the natural processes being monitored by polar-orbiting satellites are complex. This complexity requires modeling flexibility and thus requires an estimation method that can accommodate many different covariance models without loss of computational viability. Second, polar-orbiting satellites move simultaneously through space and time as data is collected. This movement sparsely distributes and confounds spatial and temporal information in the data, making certain effects difficult to estimate. Figure 1(a) roughly depicts the shape of a polar orbit, and subfigures (b) and (c) give plots of the spatio-temporal locations of OMI observations in a single latitudinal band. Figure 1(b) shows how OMI observations in single latitude band centered at 39° N are situated in time, and Figure 1(c) is a close-up view of two consecutive orbits of data collected by the Aura satellite again at 39° N. These plots show how information in data collected by a polar-orbiting satellite can be distributed. At 39° N, consecutive orbits occur at different spatial and temporal locations with a small amount of overlap in space. Due to the shape of a polar orbit, less spatial replication occurs across consecutive orbits near the equator and more occurs near the poles. Spatial replication is needed to untangle the confounding of spatial and temporal effects; however, as Figure 1(b) shows, this replication occurs largely at daily intervals. Therefore, the information needed to satisfactorily estimate certain structures in the data is distributed within orbits, across

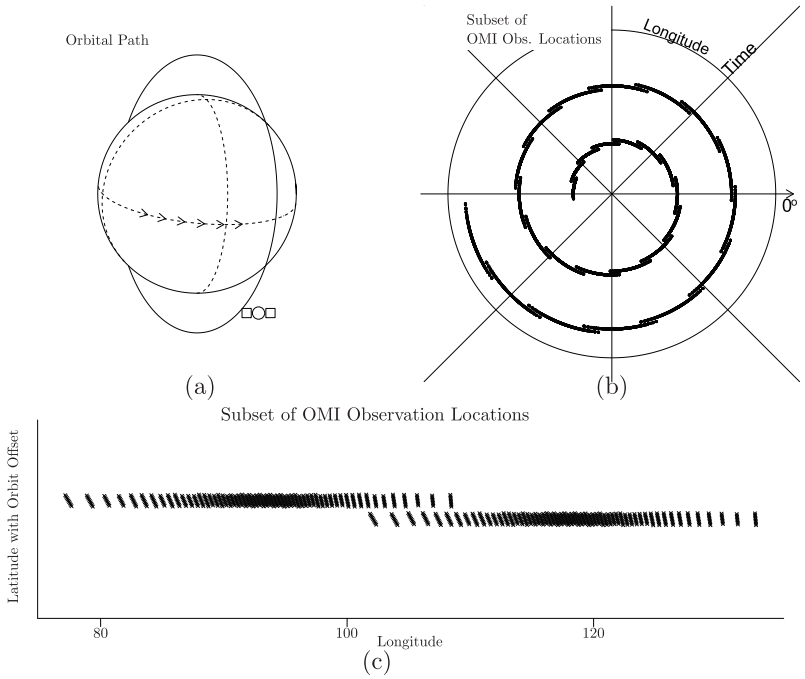

Figure 1. (a) A basic visual of the orbital path of some types of polar orbiting satellites. Arrows indicate rotation of the earth. Note other types of polar orbits do not pass exactly over the poles. (b) OMI Observational locations in a single latitude band centered at 39.5° N on two consecutive days viewed from above the north pole. Time is represented by distance from origin and longitude is the angle with the ray marked 0°. (c) Two consecutive orbits of the OMI data in the same latitude band. An offset is applied by orbit to better show the spatial—but not temporal—overlap of these data.

orbits and across days. A useful covariance parameter estimation method must respect this information structure and thus must be capable of accommodating the relationships between observations over the full range of spatio-temporal lags present in the data.

The problems inherent to polar-orbiting satellite data have been considered previously by Fang and Stein (1998) but not in the context of covariance parameter estimation. Current large n estimation methods are not specifically geared toward addressing the data structure of polar-orbiting satellites and hence may be ill-suited to solve these problems, though several versatile large n techniques are currently available.

One class of estimation techniques that has proven to be quite flexible can be described as composite likelihood methods (Vecchia (1988), Curriero and Lele (1999), Stein, Chi, and Welty (2004), Caragea and Smith (2007),

Bevilacqua et al. (2013), Eidsvik et al. (Accepted). These methods produce estimations through maximization of an objective function that relates the data to the parameters, where the objective function is the product of marginal or conditional likelihoods (Lindsay (1988)). If $Y_1, \ldots, Y_k$ and $C_1, \ldots, C_k$ are subvectors of the vector of observations, Y, a log composite likelihood can be written

$$\log(\mathcal{L}_C) = \sum_{i=1}^{k} \ell(\theta; Y_i | C_i), \tag{1.1}$$

where each ℓ is a marginal or conditional log-likelihood. Note that any C_i may be empty.

The difficulty in using composite likelihood methods is in choosing the subvectors, $Y_1, \ldots, Y_k$ and in choosing the conditioning sets, $C_1, \ldots, C_k$. Many composite likelihood methods choose both sets with a focus only on preserving close spatio-temporal lag relationships, but this strategy is incompatible with the structure of polar-orbiting data. Stein, Chi, and Welty (2004) advocate use of longer range conditioning sets, but their focus is in the spatial setting. Extension of their method to polar-orbiting satellite data or general spatio-temporal settings remains unclear. Caragea and Smith (2007) provide a composite likelihood method that considers longer range relationships, but scaling this method to extremely large data sets remains difficult.

Another class of large n estimation methods focuses on approximating a GRF with a Gaussian markov random field (GMRF) (Lindgren, Rue, and Lindström (2011)). These methods have considerable flexibility but ultimately rely on specific forms of the covariance function to produce computational savings. Additionally, since a GMRF relies on placing observations on a graph, the special observational location structure of polar-orbiting satellite data makes implementation of this procedure problematic.

In this paper, we propose a new method of space and space-time covariance model estimation that matches the needs created by polar-orbiting satellite data. We develop an estimating equation that captures relationships between observations at all spatio-temporal lag scales while simultaneously retaining flexibility in terms of covariance function modeling and maintaining computational feasibility. Our approach builds on composite likelihood methods. The estimating equation we develop uses (1.1) as a base estimating equation; however, this is modified through the addition of an approximate log-likelihood of strategically interpolated points into the estimating equation. In calculating this log-likelihood, the interpolated values are treated as actual observations of the continuous component of the spatial or space-time process being measured. For this reason, we say we use the interpolated points as pseudo-observations, and we name the

estimating equation developed here the Interpolation likelihood or I-likelihood. Using interpolated points as pseudo-observations adds simplicity to this estimation method and provides additional computational savings. The estimation procedure we present in this paper may be usefully applied to data from sources other than polar-orbiting satellites and is thus quite general; any large n data set can be analyzed using this method, and in particular, analysis of any data set produced by a scientific instrument that moves as it collects data may especially benefit from the methods presented here. The example of polar-orbiting satellite data, however, remains the primary motivator for this research.

In the following section of this paper, we more formally state the problem and our method of solution in generality. In Section 3, we apply the I-likelihood method to a large data set gathered by a polar-orbiting satellite, and we explicitly show the computational viability of our method as well as the statistical efficiency gains produced in using it.

2. Methodology

2.1. Problem

We assume observations from a polar-orbiting satellite or some known function of the observations are samples from a GRF, $\mathcal{Y}(s,t)$, indexed by space and time locations s and t. The GRF $\mathcal{Y}(s,t)$ is determined by the sum

$$\mathcal{Y}(s,t) = \mathcal{X}(s,t) + \eta\mathcal{W}(s,t)$$

with $\eta \geq 0$, where $\mathcal{X}(s,t)$ is a continuous GRF and $\mathcal{W}(s,t)$ is a GRF independent of $\mathcal{X}$ that has values distributed as independent standard normal random variables at distinct space-time locations. The distribution of observations $Y = (\mathcal{Y}(s_1,t_1),\ldots,\mathcal{Y}(s_n,t_n))^T = (y_1,\ldots,y_n)^T$ is therefore Gaussian distributed with some mean and covariance. We assume the covariance structure of $\mathcal{Y}$ is determined by a function $K_\theta(\cdot,\cdot)$ with parameter $\theta \in \mathbb{R}^d$. Note η is included in θ. We also assume $\mathbb{E}(\mathcal{Y}(s,t)) = 0$ for simplicity. Extension of this problem to the unknown mean setting is straightforward. If the mean is linear in the unknown parameters, we suggest using restricted likelihood methods (Stein, Chi, and Welty (2004)). We write the distribution of observations: $Y \sim N(0,\Sigma(\theta))$, where $\Sigma(\theta)$ is a positive definite matrix with the ij-th entry, $\Sigma(\theta)_{ij} = K_\theta(y_i,y_j)$. The log-likelihood function of Y is $\ell(\theta;Y)$, the form of which is given by

$$\ell(\theta;Y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\big|\Sigma(\theta)\big| - \frac{1}{2}Y^T\Sigma(\theta)^{-1}Y. \tag{2.1}$$

In this work, we also consider the distribution of the continuous GRF, $\mathcal{X}$. We write the log-likelihood of a sample from $\mathcal{X}$ as $\ell_\mathcal{X}(\theta;\cdot)$. In this model, $\ell_\mathcal{X}(\theta;\cdot) = \ell(\theta;\cdot)$ when $\eta = 0$.

The problem we address here is to develop an estimator of θ that has three specific properties. Property A, the estimation procedure must make use of the relationships between observations at all spatio-temporal lag scales. Property B, the number of floating point operations and the amount of RAM storage needed to carry out the estimation must scale well with n. And Property C, the estimation procedure must not rely on computational shortcuts that require a specific form of the covariance matrix, $\Sigma(\theta)$, or the covariance function, K_θ.

2.2. Estimation method

Our covariance parameter estimation method builds off composite likelihood methods in a way that observational relationships occurring over the full range of spatio-temporal lags are explicitly considered in the estimation. We construct a new function, $g(\theta; Y)$, based on a composite likelihood that we maximize with respect to θ. Our primary departure from composite likelihood methods is that g contains one or more likelihoods based on interpolated points. We denote g as the log of the I-likelihood. A more in-depth comparison of the I-likelihood to more familiar composite likelihoods is in Section 2.3. Here we discuss the general construction of g.

The I-likelihood is constructed in at least two stages that we call tiers. The first tier, g_1, is an ordinary composite log-likelihood. An appropriate composite log-likelihood for the first tier will preserve the relationships between points that are close in space and/or time. For this paper, we use a Vecchia (1988) type formulation of (1.1) that involves partitioning the data into blocks of computable size and constructing the first tier composite likelihood as the product of the likelihoods of these blocks conditioned on the data in other blocks. We also use a first tier composite likelihood that treats each block of data as independent. In practice, conditioning and blocking considerations must follow the application of interest while simultaneously respecting the computational resources at hand, but loosely, the closer g_1 is to the true log-likelihood, ℓ, the better the composite likelihood will approximate the true likelihood, likely improving overall estimation.

The second tier function, g_2, is built using interpolated points. The method and placement of these interpolated points may vary, but generally the interpolated points should represent information at a coarser spatial or temporal scale than is captured by g_1. We use the following block interpolation method. First, the vector of observations, Y, is partitioned into k blocks, $Y_1, \ldots, Y_k$, of computable size. This partitioning may follow that used in g_1, but it is not necessary. Second, using only data from the i-th block, m_i new points are interpolated, $\hat{X}_i(Y_i) \in \mathbb{R}^{m_i}$. Denote the set of interpolated points as $\hat{X} =$

$(\hat{X}_1^T(Y_1), \ldots, \hat{X}_k^T(Y_k))^T$. With these interpolated points, the second tier function is calculated $g_2 = \ell_{\mathcal{X}}(\theta, \hat{X})$. For this reason, we say we use $\hat{X}$ as pseudo-observations, where pseudo-observations are defined to be the values of $\mathcal{X}$ at the interpolation locations. The log I-likelihood for two tiers, g, is subsequently given by $g = g_1 + g_2$. Our estimator is $\hat{\theta} = \arg\max_\theta \ g(\theta; Y)$.

In the case where the number of interpolated observations is large, it may be that the second tier function g_2 is still too difficult to calculate. In this scenario, a third tier would be generated by calculating a composite likelihood for the second tier and using a second set of interpolated points to generate the third tier log-likelihood. The function g would therefore be the sum of g_1, g_2 and now g_3. Further tiers may be included in g to ensure this estimation method scales to larger and larger data sets. An outline of the calculation of the I-likelihood is below.

ALGORITHM FOR CALCULATION OF $g(\theta; Y)$, THE I-LIKELIHOOD

1. Calculate a composite log-likelihood of Y that preserves small scale spatio-temporal lag relationships. Denote this by $g_1(\theta; Y)$.
2. Break Y into k blocks, $Y \Rightarrow Y_1^T, \ldots, Y_k^T$ of computable size.
3. Interpolate m_i new points within each block, $\hat{X}_1(Y_1), \ldots, \hat{X}_k(Y_k)$, where $\hat{X}_i(Y_i) \in \mathbb{R}^{m_i}$, and calculate the log-likelihood of these points as pseudo-observations. Denote this by $g_2(\theta; Y) = \ell_{\mathcal{X}}(\theta; (\hat{X}_1^T(Y_1), \ldots, \hat{X}_k^T(Y_k))^T)$.
4. The I-Likelihood is given by $g(\theta; Y) = g_1(\theta; Y) + g_2(\theta; Y)$.
5. If g_2 is too difficult to calculate, use a composite likelihood for g_2 and repeat Steps 2-4 using a second set of interpolated points to form g_3. The locations of this second set of interpolated points should be more sparsely distributed than those used for g_2. Form $g = g_1 + g_2 + g_3$.

One interpretation of this formulation is the following: the first calculation, g_1, captures the small scale information, while the second calculation g_2 captures the larger scale information that may get omitted in the first tier calculation. In this way, relationships at all spatio-temporal lags are accounted for in the formation of g; thus, this estimation method has Property A outlined in Section 2.1. Note further that nowhere is the covariance function assumed to take any specific form; hence, this estimator has Property B.

We now consider Property C, the computational viability of this method. Composite likelihood methods at the first tier can be made quite scalable. For example, if we consider the block independent composite likelihood and use the

blocks formed in Step 2 of the algorithm, the calculation complexity at the first tier is $O(k(n/k)^3)$ with storage $O((n/k)^2)$. Writing c for n/k, the complexity is $O(c^2 n)$ with storage $O(c^2)$. If we can think of c as bounded independent of n, these orders are $O(n)$ and $O(1)$, respectively, but we retain the factors of c^2 here because c can be quite large. In this example, c is essentially a constant representing the maximum number of observations ideally considered at any given time. In practice c can be chosen based on time or memory constraints.

At the second tier (and if necessary the third and beyond tiers), if we bound the number of observations in any single block in any tier by the constant c, the storage burden at any given time is $O(c^2)$. Under the assumption that a fixed fraction of the blocks needed in one tier are needed in the subsequent tier (implying $O(\log n)$ tiers), the order of the floating point operations is $O(c^2 n \log n)$. However, use of more tiers than absolutely needed should be avoided since each tier adds further approximation into the estimation.

Beyond the complexity and storage calculations, this procedure can be quite friendly to parallelization. Each tier can be calculated independently, and if block structure at any level is utilized, with sufficient memory, a naïve parallelization technique can produce estimations at a fraction of the time a serial algorithm would require.

2.3. Comparison to composite likelihood and additional computational benefits of the I-Likelihood

The similarity between the I-Likelihood and composite likelihood estimation methods warrants a more thorough comparison. The difference between an I-Likelihood and a general composite likelihood is the inclusion of a likelihood of interpolated points, using the assumption that the interpolated points are pseudo-observations. This is a critical assumption, and it creates tremendous computational savings by itself. At first glance, this assumption may seem unnecessary. Interpolated points are often calculated as linear combinations of the actual observations, denoted $\hat{X} = AY$; therefore $\mathbb{E}\hat{X}\hat{X}^T = A\Sigma(\theta)A^T$, and thus the exact log-likelihood of $\hat{X}$ can be calculated for any given θ. By contrast, we approximate $\mathbb{E}\hat{X}\hat{X}^T$ by simply using the locations of $\hat{X}$ to generate $\Sigma_{\hat{X}}(\theta)$ under the assumption $\hat{X}$ are pseudo-observations. Using $A\Sigma(\theta)A^T$ instead of the approximation, one could cast our method as a purely composite likelihood method similar to that proposed by Caragea and Smith (2007), but doing so does not produce a procedure that is computationally scalable. Specifically, $A\Sigma(\theta)A^T$ involves the $n \times n$ matrix, $\Sigma(\theta)$. Calculation of this matrix can be carried out piece by piece to avoid memory limits, however, it is an $O(n^2)$ calculation. Therefore, any estimation procedure that scales well with n must avoid use of all entries of $\Sigma(\theta)$. As a brief example, the calculation savings in the application in Section

3 can be shown to be tremendous. Use of $\Sigma_{\hat{X}}(\theta)$ instead of $A\Sigma(\theta)A^T$ amounts to replacing a 220,426 $\times$ 220,426 matrix with a 4,360 $\times$ 4,360 matrix; therefore, for every calculation of g, using $A\Sigma(\theta)A^T$ instead of its approximation lengthens the calculation by a minimum of over 24 billion covariance function evaluations.

In practice, the closeness of the approximation of $A\Sigma(\theta)A^T$ with $\Sigma_{\hat{X}}(\theta)$ must be evaluated. Consideration of the Kullback-Leibler divergence provides one method of comparison. Comparison of the Kullback-Leibler divergence can be made considering models with different parameter values, and an example included in the supplemental material shows these divergences can be quite small in some circumstances. As expected, the divergence is closest to zero when interpolated locations are in regions near many actual observations.

Finally, a comparison of the I-likelihood to a similarly multi-tiered composite likelihood is informative. If a likelihood of some subset of the observations was used in place of g_2, g would simply be a composite likelihood with irregularly shaped blocks (we numerically compare this approach to the I-likelihood approach in Section 3.4). The advantages to using interpolated points instead of actual observations at the second tier are two-fold. First, interpolated points may provide a better summary of local information than any single actual observation due to the removal of a nugget effect. Second, the interpolation locations can be chosen by the statistician. For certain models and data, interpolation to a grid can produce $\Sigma_{\hat{X}}(\theta)$ with Toeplitz or Block-Toeplitz structure. Exploiting this structure can speed up computations and save substantially on memory requirements (Akaike (1973)). As an added benefit, computational savings at the second level allows more interpolated observations to be considered and reduces the need to use a third tier in the I-likelihood. This likely improves the overall performance of the estimator.

2.4. Estimating efficiency and inference

A measure of estimating efficiency can be obtained for $\hat{\theta}$ via estimating functions theory. If we assume g is twice differentiable with respect to the set of parameters, θ, and denote the gradient of g as G and the hessian as $\dot{G}$, we form the inverse of the Godambe information matrix (Godambe (1960));

$$\mathcal{G}^{-1} = \mathbb{E}_\theta(\dot{G})^{-1}\mathbb{E}_\theta\left(GG^T\right)\mathbb{E}_\theta(\dot{G})^{-1}. \tag{2.2}$$

Extending the geometric argument from Godambe (1960), for an unbiased estimating equation, a desirable g has a gradient at the true value θ that is close to zero over repeated simulations, and g is concentrated: it should take large values near the true parameter value and very small values away from the true parameter value. These properties are encoded in the components $\mathbb{E}_\theta\left(GG^T\right)$ and $\mathbb{E}_\theta(\dot{G})^{-1}$, implying that $\mathcal{G}^{-1}$ should be small in the set of positive definite

matrices. More specifically, if $\mathcal{G}_a^{-1}$ and $\mathcal{G}_b^{-1}$ correspond to separate equations g_a and g_b, the estimator from g_a would be preferred to that of g_b if $\mathcal{G}_b^{-1} - \mathcal{G}_a^{-1}$ is a positive definite matrix. Less precise but still informative comparisons can be made considering the diagonal elements of $\mathcal{G}^{-1}$, smaller loosely indicating a parameter more efficiently estimated.

When g is unbiased, and under sufficient regularity conditions, $\hat{\theta}$ is asymptotically normally distributed centered at θ with covariance matrix $\mathcal{G}^{-1}$ (Heyde (1997)); Varin, Reid, and Firth (2011)). This result is an extension of the asymptotic normality of the MLE. However, since the gradient of the I-likelihood does not have an exactly zero expectation at the true parameter value, this result does not specifically hold for our estimator. Further, the complex structure of polar-orbiting satellite data makes establishing regularity conditions problematic; hence, asymptotic normality would be difficult to prove even for exact likelihood methods. Nonetheless, since the bias of the I-likelihood estimator will be small in many cases, we believe $\mathcal{G}^{-1}$ to be a reasonable measure of efficiency in the non-asymptotic case, and we expect $\mathcal{G}^{-1}$ to be close to the covariance matrix of $\hat{\theta}$ when using larger sample sizes.

To calculate an approximation of (2.2), we plug $\hat{\theta}$ in for θ, but the calculation remains quite difficult. Insertion of the observed hessian, $\dot{G}(\hat{\theta})$, in place of the expected hessian is often used to shorten this calculation. For the I-likelihood, we propose an additional shortcut to approximate the middle term, the variance of the gradient, $\mathbb{E}_{\hat{\theta}}\left(GG^T\right)$. The shortcut is to again use $\hat{X}$ as pseudo-observations. Briefly, for any θ, computation of $\mathbb{E}_{\theta}\left(GG^T\right)$ involves covariance calculations of quadratic forms of normally distributed random variables. This calculation therefore uses interpolated points covariances and cross covariances, $A\Sigma(\theta)A^T$ and $\Sigma(\theta)A^T$. The shortcut we propose is to replace these matrices with the corresponding approximations, $\Sigma_{\hat{X}}(\theta)$ and $\Sigma_{Y\hat{X}}(\theta)$, created using $\hat{X}$ as pseudo-observations. Further details of this calculation can be found in the supplemental materials. A problem with this shortcut is that the approximate $\mathcal{G}^{-1}$ is calculated assuming $\hat{X}$ are additional observations, whereas without this shortcut, the covariance matrix treats $\hat{X}$ more accurately as linear combinations of Y. This assumption has the potential to produce measures of efficiency that are too optimistic, but this appears not to happen in the settings we consider in this work. We address this issue more rigorously in the simulation study included in the supplemental materials. Results there show (2.2) calculated with our shortcut produces valid standard errors in several reasonable scenarios.

The proposed substitution still requires use of the full covariance matrix, $\Sigma(\theta)$. This inflates computation time, but in many applications, the variance need only be calculated a single time at the end of an estimation. The estimation time itself is not affected by the estimator's covariance calculations. For

faster inference, the variance of the gradient may be approximated using stochastic trace approximations (Hutchinson (1990)) as described in Stein, Chen, and Anitescu (2013).

3. Application to Total Column Ozone Data

We applied the I-likelihood estimation technique to data from the Ozone Monitoring Instrument (OMI) from NASA's Aura satellite. To assess the performance of the I-likelihood methods, we compared two I-likelihood estimators to four separate composite likelihoods that may be seen as competing estimation methods for fitting this model to these data.

3.1. The data

The Aura satellite collects data at approximately local noon over nearly the entire globe each day. To do this, the satellite takes measurements as it moves over the sunlit side of the globe in a south to north direction. No data is collected as the satellite passes from north to south on the unlit portion of the globe. The satellite orbits the earth 14-15 times each day, collecting data on each orbit and generating around 1.2 million observations per day. Several observations on air quality are derived from spectrometer readings at different wavelength ranges, total column ozone being one of these (Levelt et al. (2006)). More information on both air quality measurement programs and associated satellites can be found at `http://ozoneaq.gsfc.nasa.gov/` and `http://aura.gsfc.nasa.gov/`.

The data are a subset of the OMI OMDOAO3G product. We considered data recorded over a 30 day period from December 27, 2006 through January 25, 2007. These dates were chosen to avoid missing orbits of data. This completeness allows us to illustrate gridding interpolated points at the second tier, but generally, the I-likelihood can be used even when there are missing orbits. We consider a one degree latitude band centered at 39° North latitude; hence our observations lie within the latitude band 38.5° N – 39.5° N. The data set contains 436 orbits with a total of 220,426 irregularly spaced ozone observations, the associated longitude, latitude positions of these observations, and the times (measured in seconds since January 1, 1993) and orbit numbers indicating when the observations were taken. We used hours as the unit of time because the time gap between orbits is more naturally measured in hours. Figure 1(c) again shows the observational locations of a subset of these data. Additional plots of the data are in the supplemental material.

We considered data for a single month to avoid seasonal non-stationarities. Since ozone levels are different over land and sea, a mean function was estimated by fitting a spline depending only on longitude for three years worth of January data. Observations were then centered using this mean function. In fitting the

spline, we placed extra knots at land and sea borders to account for possible sharp changes in mean ozone levels at these points. We do not consider mean variation in latitude, since the one degree latitude window we consider is small.

To perform the blocking and interpolation steps in the I-likelihood, we set single orbits as blocks, and we interpolated 10 points evenly within an orbit at exactly 39° N. Specific details of these steps are given in the supplement.

3.2. The covariance model

The covariance model we considered has several components. We included spatial and temporal range and standard deviation parameters. We considered anisotropy in space oriented on the sphere, and we considered a drift in ozone through time. We also included a parameter indicating the smoothness or degree of differentiability of the process and a nugget standard deviation term as well. Specific details of the model construction can be found in the supplemental materials. Briefly, if Z_1 and Z_2 are independent isotropic (after rescaling spatial and temporal distances by the range parameters, α and β respectively) Matérn processes in $\mathbb{R}^4$ restricted to the sphere $\times$ time and indexed by latitude, L, longitude, l and time, t, with smoothnesses $\nu + 1$ and ν respectively, the model for the detrended ozone is

$$Y(L,l,t|\alpha,\beta,\phi,\lambda,\omega,\sigma,\eta,\nu) = \phi\left(\sin(\lambda)\frac{\partial}{\partial L} + \cos(\lambda)\frac{\partial}{\partial l}\right) Z_1(L, l_\omega, t) + \sigma Z_2(L, l_\omega, t) + \eta W(L,l,t), \tag{3.1}$$

where $l_\omega = l - \omega/24 \cdot t$ and $W(L,l,t)$ is the nugget effect.

Setting the smoothness parameter for Z_1 to $\nu + 1$ implies that $\left(\sin(\lambda)\frac{\partial}{\partial L} + \cos(\lambda)\frac{\partial}{\partial l}\right) Z_1(L, l_\omega, t)$ has the same degree of mean square differentiability in any direction as Z_2. A visual of this covariance model is given in Figure 2.

This model was picked after favorable model fit comparisons to other models containing the same general components, but this model does not exhaustively account for all structure in the data. There remain several possible areas of improvement, including use of different range parameters for Z_1 and Z_2 and consideration of temporal non-stationarities as well as allowing for different degrees of smoothness in space and time. We avoided these complications in this application to simplify fitting procedures.

The components of the model can be loosely broken into parameters reflecting information at different time and distance scales. The parameters that can be reasonably estimated using only within orbit information (α, σ, ϕ, λ, η, ν) are the spatial range α, the standard deviation of the isotropic component σ, the standard deviation of the anisotropic component ϕ, the anisotropy angle λ, the

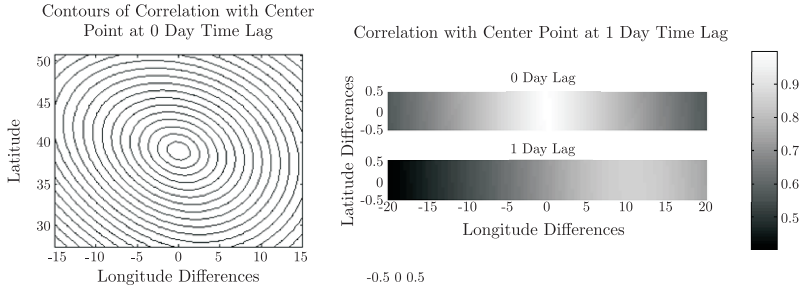

Figure 2. Correlations given by model (3.1) at $\hat{\theta}$ given by estimating equation (B, I) in Table 1. The first figure gives a contemporaneous view of the covariance function with a much wider latitude band than in the data to help visualize the spatial anisotropy. The second shows the estimated drift of the process through time.

nugget standard deviation η and the smoothness parameter ν. These parameters can be estimated using a composite likelihood that preserves only short-range spatio-temporal lags or within-orbit relationships. Since each orbit has a relatively short spatial domain, however, better estimations result if longer spatio-temporal lags are taken into account. The temporal range parameter β cannot be estimated using only within orbit information, but at 39° N, results indicate consecutive orbits contain enough temporal variability and spatial replication to make this parameter estimable. The parameter ω is the number of degrees longitude drift or rotation observed in one day. Positive ω indicates westerly drift. We expect the speed of the drift of the process on the globe to be slow relative to the time difference observed for observations within a single orbit (less than 100 seconds); hence, this parameter effectively cannot be estimated using a composite likelihood that only uses within-orbit information. Use of consecutive orbits may make this effect estimable, but our results show that it is best estimated using data across multiple days.

3.3. Notation

We denote different estimating equations using an ordered pair. The estimating equations we compare have two components, a first tier likelihood, g_1 and a second tier likelihood, g_2. Some of the estimating equations we use are standard composite likelihoods. In these cases $g_2 = 0$.

The set of six estimating equations we consider are denoted by the ordered pairs: (B, N), (B, I), (B, S), (C, N), (C, I), (C, S). The first letters characterize the composite likelihood g_1. Estimating equations under B calculate g_1 assuming blocks of observations are independent. This is a block independent composite likelihood. Estimating equations under C calculate g_1 assuming observations in

blocks are conditionally independent of prior blocks given the previous block. From the form in (1.1), C_i is the set of observations from orbit $i-1$, with C_1 empty.

The second letters I, N, and S characterize g_2. The letter I indicates g_2 is calculated based on the interpolated points. These estimating equations follow exactly the I-likelihood recipe given in Section 2.2. Letter N indicates there is no second tier, $g_2 = 0$. These are previously considered composite likelihoods of the form in (1.1) where either the conditioning set is empty or contains observations from the previous orbit. Letter S indicates g_2 is formed using a subset of actual observations that are omitted from the calculation of g_1 for this estimating equation. Since g_2 is formed using actual observations, we include the nugget effect in this second tier calculation. The observations included in g_2 from each block in this case are the ten closest observations to the interpolation locations. Note that (B, S) and (C, S) are composite likelihoods but they are not, as far as we are aware, explicitly considered previously in the literature. Among the methods considered here, (B, S) and (C, S) are the closest composite likelihoods to the I-likelihoods (B, I) and (C, I) respectively.

3.4. Results

Using these six estimating equations, we fit the model to 30 days worth of data or 220,426 observations in total. The estimations were carried out in MATLAB on a single 2.67Ghz thread. Details of the optimizations are given in the supplemental material. Estimation results along with the associated computational resources used for each computation are in Table 1. To capture estimating efficiency information, we calculated the inverse of the Godambe information matrix in (2.2) using the computational shortcuts presented in Section 2.4. A section of the supplement is devoted to evaluating the accuracy of this calculation with these data. The results of this evaluation showed the square root of the diagonal elements of this matrix can serve as reasonable standard error estimates for each of the parameter estimators; hence we call these values standard errors here. The set of standard errors for $\hat{\beta}$ and $\hat{\omega}$ for (B, N) are empty because these effects are not estimable without cross block comparisons. Under preliminary model fitting, convergence issues led to use of (B, N) fixing $\hat{\beta} = \infty$ and $\hat{\omega} = 0$.

In Table 1, we see the I-likelihood method improves overall estimation of within-orbit, consecutive-orbit, and across-day parameter types compared to B and C type composite likelihood methods. Comparing (B, N) to (B, I), we see a remarkable reduction in the standard errors for all estimates. Each standard error under (B, N) is over 3 times larger than the corresponding standard error from (B, I). Moreover, the drift parameter ω and the temporal range, β, can be estimated with (B, I) where using (B, N) they cannot. Between (B, N) and

Table 1. (*) Note optimization of estimating equation (C, S) was started at the (C, I) point estimate, leading to fewer total function evaluations. Additionally, the algorithm used to calculate g_2 was different between (C, S) and (B, S); hence, memory usage is not comparable between these two estimating equations. Details are in the supplemental material.

Estimating Equation Comparisons on 30 Day Estimations

		Estimates and (Standard Errors) Estimating Equation					
Parameter		(B, N)	(B, I)	(B, S)	(C, N)	(C, I)	(C, S)
Spatial Range	α	4.81	4.24	4.43	3.58	3.51	3.46
(1000s of Km)		(5.21)	(1.35)	(3.03)	(0.28)	(0.28)	(0.26)
Temporal Range	β	∞	116.77	119.84	60.64	66.57	64.79
(Hours)			(37.79)	(78.71)	(5.00)	(5.19)	(4.91)
Isotropic SD	σ	36.23	33.72	34.62	26.54	26.65	26.69
(Dobson Units)		(25.39)	(6.92)	(10.26)	(1.10)	(1.01)	(1.35)
Anisotropic SD	ϕ	17.95	14.48	15.59	11.46	11.05	10.93
(Dobson Units)		(32.51)	(7.54)	(15.41)	(1.41)	(1.36)	(1.47)
Anisotropy Angle	λ	0.551	0.583	0.585	0.574	0.572	0.569
(Radians)		(0.075)	(0.018)	(0.017)	(0.014)	(0.014)	(0.014)
Nugget SD	η	2.404	2.424	2.428	2.373	2.394	2.400
(Dobson Units)		(0.056)	(0.017)	(0.047)	(0.013)	(0.014)	(0.013)
Drift term	ω	0	10.74	10.64	16.16	11.60	11.67
(Degrees/Day)			(0.28)	(0.77)	(2.26)	(0.46)	(0.50)
Smoothness	ν	0.579	0.582	0.581	0.560	0.564	0.566
		(0.024)	(0.006)	(0.024)	(0.004)	(0.005)	(0.004)
Memory Used		1.3G	1.5G	4.1G	1.7G	1.7G	1.6G*
Comp. Time		26h	33h	41h	57h	82h	21h
Func. Evaluations		110	72	152	82	90	27*
Observations	n	220,426					
Thread Speed		2.67Ghz					

(B, I), there is good agreement among the estimates of the other parameters. No estimate pair is more than 0.5 standard errors away from one another when using the larger of the two standard errors.

Comparison of (C, N) to (C, I) tells a similar story. The parameter estimates are consistent with one another (with ω being the most incompatible), and the I-likelihood estimating equation is shown to estimate all parameters with similar or better efficiencies. Most notably, the drift parameter, ω is much more efficiently estimated in (C, I) as compared to (C, N).

Looking at (B, I) and (B, S), we see a high degree of consistency despite the first and second tiers being slightly different. No pair of estimates is more than 0.14 standard errors away from one another. Furthermore, comparison of standard errors shows that, except for λ, the model parameters appear to be much

better estimated in (B, I) than they are in (B, S). In particular, the standard error on the smoothness parameter is almost 4 times larger under (B, S) than it is under (B, I). We attribute these efficiency gains to the removal of the nugget effect at the second tier in (B, I), a procedure that is not possible in (B, S); thus, the interpolated observations serve as a better summary of local information and lead to more efficient estimation.

The estimating equations (C, I) and (C, S) also exhibit consistency, but unlike the (B, I) and (B, S) comparison, the performances of the (C, I) and (C, S) estimators are roughly equal. Unlike B, using C for g_1 provides information on consecutive-orbit relationships; therefore, any efficiencies added by g_2 in the $(C, \cdot)$ case would be primarily due to the added use of cross-day relationships. However, cross-day relationships are much less local in time than consecutive-orbit relationships; hence, precisely measuring local behavior by using $\hat{X}$ in the $(C, \cdot)$ cases might not be as relevant as it is proven to be in the $(B, \cdot)$ cases.

For larger data sets, (C, I) will have substantial computational advantages over (C, S). Specifically, the locations of the interpolated points give Block-Toeplitz $\Sigma_{\hat{X}}(\theta)$. This structure can be exploited via the methods in Akaike (1973) leading to more efficient computation of g_2 and lower memory usage. By comparison, the (C, S) estimator uses an unstructured Cholesky decomposition to calculate g_2; thus, fewer second tier points may be handled using (C, S).

Slight discrepancies arise in comparison of the $(B, \cdot)$ estimators to the $(C, \cdot)$ estimators. In any set of estimates of a single parameter, no two pairs are more than 3.4 standard errors away from one another. The largest difference relative to the standard errors between the $(B, \cdot)$ and $(C, \cdot)$ estimators is in the estimation of the smoothness and drift parameters. The $(C, \cdot)$ estimates of ν are each slightly smaller than all $\hat{\nu}$ given by the $(B, \cdot)$ estimations, and each $\hat{\omega}$ under $(C, \cdot)$ estimations are larger than those using the $(B, \cdot)$ estimators. These discrepancies are likely a sign of modest model misfit.

A surprising outcome in these comparisons is the smaller standard error for $\hat{\omega}$ under (B, I) than under (C, I) and (C, S). With the more complex first tier composite likelihood in (C, I) and (C, S), a more efficient estimator of all parameters is expected. This pattern is observed for all but the $\hat{\omega}$ estimates. The reversed pattern for $\hat{\omega}$ is again likely a sign of model misfit or an insufficiently flexible model. Conflicting information at different scales may have driven the resulting standard error for $\hat{\omega}$ under (C, I) and (C, S) higher. Evidence of this can be seen through comparison of the (C, N) estimate of ω to those from (C, I), (C, S) and (B, I).

Overall, the I-likelihood appears to increase the estimating efficiency of parameters that require use of data in multiple blocks compared to standard composite likelihood methods. It does this while simultaneously preserving (and also

improving) estimations of parameters that may be estimated using only within block information. Through comparison of (B, I) and (B, S), the I-likelihood method can also be said to outperform at least some purely composite likelihood methods both in terms of computational resources needed and estimating efficiency in this example. We also find potential in estimating equations of type $(\cdot, S)$, though further study of these estimators may be necessary. Finally, we find differences in estimates using information at different spatio-temporal scales may serve as a useful way of detecting model misfit.

4. Conclusion

In this paper, we present a new method of spatial or spatio-temporal covariance parameter estimation for large data sets that explicitly considers the relationships between observations at all ranges of spatio-temporal lags. This method is strongly motivated by consideration of data collected by polar orbiting satellites. These types of data sets are massive in size, complex, and the data collection structure implies that sensible estimation of traditional covariance parameters as well as parameters of physical interest must be able to take into account the relationships between observations over the full range of spatio-temporal lags.

The method we propose can be seen as a computationally efficient extension of composite likelihood methods. A new estimating function, the I-likelihood, is developed in multiple tiers. A first tier composite likelihood captures close spatio-temporal relationships, while a second tier likelihood based on interpolated points captures the medium to large scale spatio-temporal relationships. The innovation in this technique is in the use of multiple tiers to explicitly capture the full range of spatio-temporal lag information and in the use of interpolated points on top of composite likelihood methods. Using interpolated data as pseudo-observations leads to tremendous computational savings. Through consideration of total column ozone data gathered from a polar orbiting satellite, the I-likelihood estimation method is shown to be capable of providing fast and efficient estimation.

The I-likelihood relies on approximation of the interpolated process with a process based only on the locations of the interpolated points. This approximation is shown to be inconsequential in our application. We believe the I-likelihood to be broadly applicable, but a more general consideration of this approximation remains necessary.

Application of the I-likelihood in a Bayesian setting may also be an area of exploration. Papers by Ribatet, Cooley, and Davison (2012) and Shaby (Accepted) consider Bayesian estimation using composite likelihoods and more generally using consistent estimating equations, respectively. A Bayesian estimation employing I-likelihood methods may follow the methods of these papers closely.

Acknowledgement

We are grateful to Joe Guinness, Professor Montserrat Fuentes, and Professor Alan Gelfand for useful conversations and suggestions of focus. We also thank Eric Janofsky for his pre-processing work with the data. We thank NASA and the GES DISC team for providing the data. The first author was supported by Department of Education award P200A100171 and STATMOS, an NSF-funded Network (NSF-DMS awards 1106862, 1106974 and 1107046). The second author was supported by US Department of Energy grant no. DE-SC0002557.

References

Akaike, H. (1973). Block Toeplitz matrix inversion. *SIAM J. Appl. Math.* **24**, 234-241.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *J. Roy. Statist. Soc. Ser. B* **70**, 825-848.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2013). Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *J. Amer. Statist. Assoc.* **107**, 268-280.

Caragea, P. C. and Smith, R. L. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J. Multivariate Anal.* **98**, 1417-1440.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. Roy. Statist. Soc. Ser. B* **70**, 209-226.

Curriero, F. C. and Lele, S. (1999). A composite likelihood approach to semivariogram estimation. *J. Agric. Biol. Environ. Stat.* **4**, 9-28.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M. and Niemi, J. (Accepted). Estimation and prediction in spatial models with block composite likelihoods. *J. Comput. Graph. Statist.*

Fang, D. and Stein, M. L. (1998). Some statistical methods for analyzing the TOMS data. *J. Geophys. Res.* **103**, 26,165-26,182.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208-1211.

Heyde, C. (1997). *Quasi-likelihood and Its Application.* Springer Series in Statistics, New York.

Hutchinson, M. F. (1990). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simulation Comput.* **19**, 433-450.

Kaufman, C. G., Schervish, M. J. and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Amer. Statist. Assoc.* **103**, 1545-1555.

Levelt, P. F., van den Oord, G. H. J., Dobber, M. R., Mälkki, A., Visser, H., de Vries, J., Stammes, P., Lundell, J. O. V. and Saari, H. (2006). The ozone monitoring instrument. *IEEE Trans. Geoscience Remote Sensing* **44**, 1093-1101.

Liang, F., Cheng, Y., Song, Q., Park, J. and Yang P. (2013). A resampling-based stochastic approximation method for analysis of large geostatistical data. *J. Amer. Statist. Assoc.* **108**, 325-339.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. Roy. Statist. Soc. Ser. B* **73**, 423-498.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 220-239.

NASA. (2012). The afternoon constellation – A-train. `http://atrain.nasa.gov/intro.php`.

Ribatet, M., Cooley, D. and Davison, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statist. Sinica* **22**, 813-846.

Shaby, B. A. (Accepted). The open-faced sandwich adjustment for MCMC using estimating functions. *J. Comput. Graph. Statist.*

Stein, M. L., Chi, Z. and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *J. Roy. Statist. Soc. Ser. B* **66**, 275-296.

Stein, M. L., Chen, J. and Anitescu, M. (2013). Stochastic approximation of score functions for Gaussian processes *Ann. Appl. Statist.* **7**, 1162-1191.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21**, 5-42.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50**, 297-312.

Department of Statistics, The University of Chicago - Chicago, IL 60637, U.S.A.

E-mail: horrell@galton.uchicago.edu

Department of Statistics, The University of Chicago - Chicago, IL 60637, U.S.A

E-mail: stein@galton.uchicago.edu

(Received August 2013; accepted March 2014)