Spatial Statistics 8 (2014) 1–19

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Limitations on low rank approximations for covariance matrices of spatial data

Michael L. Stein *

Department of Statistics, University of Chicago, Chicago, IL 60637, United States

ARTICLE INFO

Article history:
Received 22 January 2013
Accepted 25 June 2013
Available online 18 July 2013

Keywords:
Fixed-domain asymptotics
Gaussian processes
Kullback–Leibler divergence
Random effects
Subset of regressors
Total column ozone

ABSTRACT

Evaluating the likelihood function for Gaussian models when a spatial process is observed irregularly is problematic for larger datasets due to constraints of memory and calculation. If the covariance structure can be approximated by a diagonal matrix plus a low rank matrix, then both the memory and calculations needed to evaluate the likelihood function are greatly reduced. When neighboring observations are strongly correlated, much of the variation in the observations can be captured by low frequency components, so the low rank approach might be thought to work well in this setting. Through both theory and numerical results, where the diagonal matrix is assumed to be a multiple of the identity, this paper shows that the low rank approximation sometimes performs poorly in this setting. In particular, an approximation in which observations are split into contiguous blocks and independence across blocks is assumed often provides a much better approximation to the likelihood than a low rank approximation requiring similar memory and calculations. An example with satellite-based measurements of total column ozone shows that these results are relevant to real data and that the low rank models also can be highly statistically inefficient for spatial interpolation.

1. Introduction

Fitting models to large spatial datasets can be challenging both conceptually and computationally, especially when the observations are irregularly located. Even under the strong assumption that the process is Gaussian, in which case only the mean and covariance structure need to be modeled, model

* Tel.: +1 7737028326.
E-mail address: stein@galton.uchicago.edu.

2211-6753/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spasta.2013.06.003

selection and estimation can still be difficult. If one is willing to posit parametric models for the mean and covariance structure, then estimation via a likelihood-based approach (maximum likelihood or Bayesian) is natural. For a random vector $\boldsymbol{Z} \sim N(X\boldsymbol{\beta}, K(\boldsymbol{\theta}))$ with X a known $n \times p$ matrix, $\boldsymbol{\beta}$ a p-vector of unknown coefficients and $K(\boldsymbol{\theta})$ a family of $n \times n$ covariance matrices indexed by the unknown q-vector $\boldsymbol{\theta}$, then $-2\times$ loglikelihood is

$$n \log(2\pi) + \log |K(\boldsymbol{\theta})| + (\boldsymbol{Z} - X\boldsymbol{\beta})' K(\boldsymbol{\theta})^{-1} (\boldsymbol{Z} - X\boldsymbol{\beta}). \tag{1}$$

When n is large and $K(\boldsymbol{\theta})$ does not have any exploitable structure, then exact calculation of the loglikelihood is generally done using the Cholesky decomposition of $K(\boldsymbol{\theta})$, which requires $O(n^3)$ flops (floating point operations) and $O(n^2)$ memory, both of which can cause problems when n is much bigger than 10,000.

There are many strategies for reducing these computational and memory requirements (see Sun et al., 2012 for a recent review), but one attractive approach is to assume that $K(\boldsymbol{\theta})$ can be written as a sum of a diagonal matrix (often a multiple of the identity matrix) and a low rank matrix. Specifically, suppose

$$K(\boldsymbol{\theta}) = D(\boldsymbol{\theta}) + WV(\boldsymbol{\theta})W', \tag{2}$$

where $D(\boldsymbol{\theta})$ is diagonal for all $\boldsymbol{\theta}$, W is a known $n \times r$ matrix and $V(\boldsymbol{\theta})$ is a family of $r \times r$ positive semidefinite matrices indexed by $\boldsymbol{\theta}$, so that $WV(\boldsymbol{\theta})W'$ has rank at most r. Then, using the Woodbury formula and the matrix determinant lemma, (1) can be written as

$$\begin{aligned} &n \log(2\pi) + \log |V(\boldsymbol{\theta})^{-1} + W'D(\boldsymbol{\theta})^{-1}W| + \log |V(\boldsymbol{\theta})| + \log |D(\boldsymbol{\theta})| \\ &\quad + (\boldsymbol{Z} - X\boldsymbol{\beta})' \left[D(\boldsymbol{\theta})^{-1} - D(\boldsymbol{\theta})^{-1} W \left\{ V(\boldsymbol{\theta})^{-1} + W'D(\boldsymbol{\theta})^{-1}W \right\}^{-1} W'D(\boldsymbol{\theta})^{-1} \right] (\boldsymbol{Z} - X\boldsymbol{\beta}). \end{aligned} \tag{3}$$

Treating r and n/r as both large, the flop count for calculating (3) is dominated by computing $W'D(\boldsymbol{\theta})^{-1}W$, which can be done by carrying out calculations of the $\binom{r}{2}$ inner products of all pairs of columns of $D(\boldsymbol{\theta})^{-1/2}W$ together with calculations requiring many fewer flops. Since an inner product of two vectors of length n requires n multiplications and $n - 1$ additions, the total flop count for calculating (3) is, to a first approximation, nr^2. Here and throughout this work, I implicitly assume that p/r^2 is small, so that the computational effort for calculating $X\boldsymbol{\beta}$, which requires $O(np)$ flops, is negligible compared to that for calculating $W'D(\boldsymbol{\theta})^{-1}W$. Furthermore, assuming that the elements of W (and $D(\boldsymbol{\theta})$) can be calculated as needed, the memory requirement for evaluating (3) is only $O(r^2)$, since one only needs to store the Cholesky decomposition of $V(\boldsymbol{\theta})^{-1} + W'D(\boldsymbol{\theta})^{-1}W$. Note that if W cannot be stored, the elements of W may need to be calculated repeatedly and these calculations may then be a substantial fraction of the overall computational effort, although if W is taken to be sparse as in Cressie and Johannesson (2008), then this effort can be reduced substantially. Even if W is not sparse, assuming $K(\boldsymbol{\theta})$ to be of the form (2) with r much smaller than n greatly reduces both the flops and memory requirements for exact calculation of the loglikelihood. It is common to assume that $D(\boldsymbol{\theta})$ is a multiple of the identity matrix when using low rank approximations and I will assume $D(\boldsymbol{\theta})$ to be of this form hereafter, although more general diagonal and nondiagonal matrices are considered briefly in the discussion.

Low rank approximations to covariance matrices have been used widely in the statistics (Cressie and Johannesson, 2006, 2008; Banerjee et al., 2008; Eidsvik et al., 2012; Finley et al., 2009; Katzfuss and Cressie, 2012) and machine learning (Smola and Bartlett, 2001; Quiñonero-Candela and Rasmussen, 2005; Quiñonero-Candela et al., 2007) literatures for prediction and fitting of random processes. In machine learning, the approach is often called the "subset of regressors" approach. Problems with these low rank approximations have been noted previously (Quiñonero-Candela and Rasmussen, 2005; Stein, 2007b; Banerjee et al., 2010; Sang et al., 2011), but there is little in the way of mathematical results delineating when exactly one might expect the approximation to be poor. Chalupka et al. (2013) provides a more thorough account of the efficacy of low rank and other approximations to large covariance matrices with a focus on prediction. Here, the focus is on model approximation as measured by the Kullback–Leibler (KL) divergence. Using the KL divergence, the theoretical results in Section 2 indicate that there is a class of problems for which even the best possible low rank approximations perform terribly.

An alternative approach to reducing the memory and computational burden of computing (1) is to assume that $K(\boldsymbol{\theta})$ is block diagonal, which, under the Gaussian assumption, corresponds to assuming observations in the different blocks to be independent. If, to a first approximation, there are n/r blocks of size r, then the dominant component of the flop count in computing (1) is the n/r Cholesky decompositions of matrices of size $r \times r$. Since a Cholesky decomposition requires $\frac{1}{3}r^3$ flops (Golub and van Loan, 1996, p. 145), the leading term in the total flop count is $\frac{1}{3}nr^2$, which is $\frac{1}{3}$ the value needed for (3). One never needs to store more than the Cholesky decomposition of a single $r \times r$ matrix, so the storage requirements of this approach are essentially identical to that for the low rank approximation (assuming r is the same in each case and that W does not need to be stored for the low rank approximation). However, the low rank approximation does have an important advantage over the independent block approach when W is a fixed matrix and D is a multiple of some fixed matrix, even when r is the same in each case. Specifically, thinking of r as fixed and letting n grow, all of the computations requiring $O(n)$ flops, such as $W'W$, $W'\boldsymbol{Z}$ and $W'X$, can be done once and for all, whereas for the independent block approach, the matrix decompositions of the covariance matrices for the blocks generally need to be redone for each $\boldsymbol{\theta}$. Note, though, that in the approaches of Banerjee et al. (2008) and Sang and Huang (2012), the matrix W depends on the unknown parameters, so these calculations cannot be done just once in their approaches to likelihood approximation.

The assumption of independence between blocks is obviously unrealistic for most spatial data, so one might expect this approximation to work poorly in many circumstances. Indeed, if the purpose for using the independent block approximation is purely to reduce the computation relative to that needed to compute (1), there are more sophisticated likelihood approximations that may work considerably better in some circumstances (Vecchia, 1988; Stein et al., 2004; Caragea and Smith, 2007). Nevertheless, results in Section 2 show that there is a class of problems for which the independent block approach yields much larger expected loglikelihoods than the low rank approach for comparable values of r.

Section 2 gives theoretical results on the expected decrease in loglikelihood due to using low rank and independent block approximations for evenly spaced observations from a periodic process that behaves like a Matérn model. The results are asymptotic as the number of observations on a fixed interval is increased, and hence are an example of fixed-domain asymptotics (Stein, 1999). By taking limits this way, we get that an increasing fraction of the variation in the observations is captured by low frequency variation, so a low rank approximation might be thought to be particularly appropriate as the sample size increases. Unfortunately, the results show that when the error/nugget variance is sufficiently small and observations are sufficiently dense, the low rank approximation performs disastrously and the independent block model much better. Section 3 shows that unless one is willing to make sufficiently strong assumptions about $V(\boldsymbol{\theta})$ in (2), one might as well use the matrix W as a covariate matrix in the mean of the model and fit the coefficients by least squares methods. Section 4 describes some numerical results that support the theory in Section 2. Additional numerical results allow us to examine how the two approximations perform as one varies the range parameter of the covariance function and show that, when there is a substantial nugget effect, the independent block approximation tends to dominate the low rank approximation at shorter ranges, whereas for sufficiently large range parameters, the results depend on the rank of the approximation. A final example shows that the low rank approximation does not necessarily provide a good solution to the modeling of nonstationary processes. Section 5 compares statistical analyses of 83,305 total column ozone observations over the Pacific Ocean based on low rank and independent block approaches. Section 6 considers the appropriateness of KL divergence as a measure of model quality and discusses some other approaches to reducing computations when fitting Gaussian process models to large datasets.

2. Theoretical results for KL divergences

For probability measures P and Q, the KL divergence of Q from P, written as $\mathrm{KL}(P, Q)$, equals $E_P \log(dP/dQ)$, where E_P indicates expectation under P and dP/dQ is the Radon–Nikodym derivative of P with respect to Q. Since the KL divergence is just the expected difference in loglikelihood when using model Q rather than P when P is true, it provides a natural way of assessing the accuracy

of a model when using likelihood methods. For two nonsingular multivariate normal distributions $P = N(\boldsymbol{\mu}_0, \Sigma_0)$ and $Q = N(\boldsymbol{\mu}_1, \Sigma_1)$ on $\mathbb{R}^n$ with Σ_0 and Σ_1 positive definite,

$$2\,\mathrm{KL}(P, Q) = \mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\Sigma_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - \log|\Sigma_0| + \log|\Sigma_1| - n. \tag{4}$$

Banerjee et al. (2008) and Sang and Huang (2012) also use KL divergence to measure the quality of low rank and other approximations to the covariance structure of spatial data. The focus here is on the influence of the covariance structure on the divergence, so I will assume $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$, in which case, the second term on the right-hand side of (4) drops out.

Suppose Σ_0 has eigenvalue decomposition $S\Lambda S'$ for orthogonal S and diagonal Λ with positive diagonal elements $\lambda_1 \geq \cdots \geq \lambda_n$. Write S_r for the first r columns of S, Λ_r the $r \times r$ upper left submatrix of Λ and $\bar{\lambda}_r = \frac{1}{n-r}\sum_{j=r+1}^{n}\lambda_r$. Consider Σ_1 of the form $\tau I_n + R$, where I_n is the $n \times n$ identity matrix, $\tau \geq 0$ and R is an $n \times n$ positive semidefinite matrix of rank at most r. Among such Σ_1, the matrix that minimizes $\mathrm{KL}(P, Q)$ is $S_r(\Lambda_r - \bar{\lambda}_r I_r)S_r' + \bar{\lambda}_r I_n$ and then

$$2\,\mathrm{KL}(P, Q) = -\sum_{j=r+1}^{n} \log\frac{\lambda_j}{\bar{\lambda}_r}. \tag{5}$$

This result is essentially an "expected value" version of a standard result on maximum likelihood estimation when Σ_0 is the sample covariance matrix based on independent replicates from a multivariate normal distribution (Stoica and Jansson, 2009). Williams and Seeger (2001) use eigendecompositions to obtain low rank approximations to covariance matrices, but the focus there is on prediction for a fixed covariance structure, not covariance estimation.

We see that (5) measures the relative variation of the eigenvalues beyond the first r. Thus, even if the sum of the first r eigenvalues is much larger than the sum of the remaining eigenvalues, (5) may still be large. This circumstance can occur when neighboring observations are strongly correlated and the eigenvectors associated with the largest eigenvalues capture the large scale variation of the process. Nevertheless, (5) shows that the KL divergence will still be large if the fine scale variations are not well described by white noise.

Stationary periodic processes on the line provide a convenient setting for theoretical development. These processes have a discrete spectrum f_j for integers j with f even, nonnegative and summable. If the period of the process is 1, the corresponding autocovariance function is

$$K(x) = \sum_{j=-\infty}^{\infty} f_j e^{2\pi i j x}. \tag{6}$$

When such a process Z is observed at j/n for $j = 1, \ldots, n$, the corresponding Σ_0 is circulant; plugging (6) into (Brockwell and Davis, 1991, Proposition 4.5.1) yields the eigenvalues

$$f_{j,n} = n\sum_{q=-\infty}^{\infty} f_{j+nq} \tag{7}$$

for $j = 0, \ldots, n-1$. If Z is observed with white noise with variance $Bn^{-\beta}$, then the eigenvalues of the covariance matrix of the observations are $g_{j,n} = f_{j,n} + Bn^{-\beta}$ for $j = 0, \ldots, n-1$. Letting the noise variance depend on n makes it possible to explore how the level of noise affects the quality of the best possible low rank approximation. For many environmental datasets, the noise variance is quite small and it will be reasonable to think of taking $\beta > 0$ to obtain useful asymptotic results. For deterministic computer model output, which is often modeled using Gaussian processes (see, e.g., Kennedy and O'Hagan, 2001), it might be reasonable to treat the noise variance as 0 or at least extremely small.

To compute the minimum possible KL divergence, we need to know the ranking of the $f_{j,n}$'s. From (7), $f_{j,n} = f_{n-j,n}$ for $j = 1, \ldots, n-1$. Rearranging terms and using f even, we can write

$$f_{j,n} = n\sum_{q=-\infty}^{\infty}(f_{j+nq} + f_{n-j+nq}).$$

If f_j is convex for $j \geq 0$, then for all $q \geq 0$, $f_{j+nq} + f_{n-j+nq}$ is nonincreasing in j for $0 \leq j \leq \frac{1}{2}n$, so $f_{j,n}$ is nonincreasing in j for $0 \leq j \leq \frac{1}{2}n$. Writing $\lambda_{1,n} \geq \cdots \geq \lambda_{n,n}$ for the ordered eigenvalues of Σ_0, it follows that $\lambda_{1,n} = g_{0,n}$, $\lambda_{2j,n} = g_{j,n}$ for $1 \leq j \leq \frac{1}{2}n$ and $\lambda_{2j+1,n} = g_{j,n}$ for $1 \leq j \leq \frac{1}{2}(n-1)$.

So let us consider a class of models for f_j such that f_j is convex for $j \geq 0$ and that includes processes with a range of degrees of local smoothness similar to the popular Matérn model (Stein, 1999, p. 48). Specifically, let us assume that for $j \neq 0$, $f_j = C|j|^{-\alpha-1}$ with C and α positive (so f is summable) and $f_0 = 2C$. The larger α is, the smoother the corresponding process is. In particular, Z has m mean square derivatives if and only $\alpha > 2m$, so $\frac{1}{2}\alpha$ plays the role of the smoothness parameter ν in the Matérn model (Stein, 1999, p. 48). It is easy to show that f_j is convex for $j \geq 0$. The propositions below surely hold under weaker conditions on the f_j's – perhaps just $f_j = Cj^{-\alpha-1} + O\left(j^{-\alpha-1-\epsilon}\right)$ as $j \to \infty$ for some $\epsilon > 0$ is sufficient – but the proofs would get much more difficult without any meaningful improvement in understanding of the issues.

Now let the rank of the low dimensional approximation, r, depend on n. To avoid some minor technical difficulties, from now on take n and r_n odd, although, except for minor modifications of (10) and (22), all of the asymptotic results given in this section still apply when n and/or r_n are even. Define $n' = \frac{1}{2}(n-1)$, $r_n' = \frac{1}{2}(r_n - 1)$ and $\bar{g}_{r_n} = \frac{1}{n'-r_n'}\sum_{j=r_n'+1}^{n'} g_{j,n}$. Then, writing P_n and Q_n to emphasize the dependence of the probability models on n,

$$2\,\mathrm{KL}(P_n, Q_n) = -2\sum_{j=r_n'+1}^{n'} \log\frac{g_{j,n}}{\bar{g}_{r_n}}.$$

It is worthwhile to consider three cases for r_n, one in which r_n is approximately a fixed fraction of n, one in which r_n grows with n but in such a way that $r_n/n \to 0$ as $n \to \infty$ and one in which r_n does not grow with n. To be specific, let us consider the following three cases:

Case 1: $r_n = Dn + O(1)$ as $n \to \infty$ for $0 < D < 1$.
Case 2: $r_n = Dn^{\delta} + O(1)$ as $n \to \infty$ for $D > 0$ and $0 < \delta < 1$.
Case 3: $r_n = r_0$ for all n for some positive integer r_0.

There are many results that one could prove under these conditions, but I will focus on distinguishing between when $2\,\mathrm{KL}(P, Q) = o(n)$ and when it does not, since if the KL divergence is not much smaller than the sample size, then it might generally be agreed upon that one measure is not a good approximant of the other. Define

$$S_\alpha(x) = \sum_{q=-\infty}^{\infty}\left(|q+x|^{-\alpha-1} + |q+1-x|^{-\alpha-1}\right),$$

so that $g_{j,n} = Bn^{-\beta} + Cn^{-\alpha}S_\alpha\left(\frac{j}{n}\right)$ and

$$\bar{g}_{r_n} = Bn^{-\beta} + \frac{Cn^{-\alpha}}{n'-r_n'}\sum_{j=r_n'+1}^{n'} S_\alpha\left(\frac{j}{n}\right).$$

Proposition 1. *In Case* 1,

$$\bar{g}_{r_n} = Bn^{-\beta} + \frac{2Cn^{-\alpha}}{1-D}\int_{D/2}^{1/2} S_\alpha(x)\,dx + O\left(n^{-\alpha-1}\right). \tag{8}$$

In Case 2,

$$\bar{g}_{r_n} = Bn^{-\beta} + \frac{2^{\alpha+2}C}{\alpha D^\alpha}n^{-\delta\alpha} + O\left(n^{-\delta(\alpha+1)} + n^{-\delta\alpha+\delta-1} + n^{\delta-\alpha-1}\right). \tag{9}$$

In Case 3,

$$\bar{g}_{r_n} = Bn^{-\beta} + 4C\sum_{j=\frac{1}{2}(r_0+1)}^{\infty} j^{-\alpha-1} + O\left(n^{-1}\right). \tag{10}$$

A proof is given in the Appendix. In every case, there exists $\epsilon > 0$ such that the relative error in the remainder term is $O(n^{-\epsilon})$.

Next we need approximations for $\sum_{j=r_n'+1}^{n'} \log g_{j,n}$. We have

$$\sum_{j=r_n'+1}^{n'} \log g_{j,n} = -\alpha(n' - r_n') \log n + \sum_{j=r_n'+1}^{n'} \log \left\{ Bn^{\alpha-\beta} + CS_\alpha\left(\frac{j}{n}\right)\right\}. \tag{11}$$

Write $a_n \approx_n b_n$ if $a_n = b_n + o(n)$.

Proposition 2. *For Case* 1, *if* $\beta < \alpha$,

$$\sum_{j=r_n'+1}^{n'} \log g_{j,n} \approx_n -\frac{1}{2}\beta(1-D)n \log n + \frac{1}{2}n(1-D)\log B, \tag{12}$$

if $\beta = \alpha$,

$$\sum_{j=r_n'+1}^{n'} \log g_{j,n} \approx_n -\frac{1}{2}\beta(1-D)n \log n + n\int_{D/2}^{1/2} \log\{B + CS_\alpha(x)\}\,dx \tag{13}$$

and if $\beta > \alpha$,

$$\sum_{j=r_n'+1}^{n'} \log g_{j,n} \approx_n -\frac{1}{2}\beta(1-D)n \log n + n\int_{D/2}^{1/2} \log\{CS_\alpha(x)\}\,dx. \tag{14}$$

For Cases 2 *and* 3, *if* $\beta < \alpha$,

$$\sum_{j=r_n'+1}^{n'} \log g_{j,n} \approx_n -\frac{1}{2}\beta n \log n + \frac{1}{2}n \log B, \tag{15}$$

if $\beta = \alpha$,

$$\sum_{j=r_n'+1}^{n'} \log g_{j,n} \approx_n -\frac{1}{2}\alpha n \log n + n\int_{0}^{1/2} \log\{B + CS_\alpha(x)\}\,dx \tag{16}$$

and if $\beta > \alpha$,

$$\sum_{j=r_n'+1}^{n'} \log g_{j,n} \approx_n -\frac{1}{2}\alpha n \log n + n\int_{0}^{1/2} \log\{CS_\alpha(x)\}\,dx. \tag{17}$$

A proof in a typical case is given in the Appendix.

Now we just have to combine the results of Propositions 1 and 2 to obtain asymptotic results on $2\,\mathrm{KL}(P_n, Q_n)$. It is apparent that there are two critical cutoffs for β: $\beta = \alpha$ (from Proposition 2) and $\beta = \delta\alpha$ (from (9)). In the following proposition, take $\delta = 0$ for Case 3. These results follow readily from the previous propositions:

Proposition 3. *If* $\beta < \delta\alpha$,

$$2\,\mathrm{KL}(P_n, Q_n) \approx_n 0. \tag{18}$$

For Case 1, *if* $\beta > \alpha$,

$$2\,\mathrm{KL}(P_n, Q_n) \approx_n n(1-D)\log\left\{\frac{2C}{B(1-D)}\int_{D/2}^{1/2} S_\alpha(x)\,dx\right\} \tag{19}$$

and if $\beta = \alpha$,

$$2\,\mathrm{KL}(P_n, Q_n) \approx_n n(1-D)\left[\log\left\{\frac{2}{1-D}\int_{D/2}^{1/2}(B + CS_\alpha(x))\,dx\right\} - \frac{2}{1-D}\int_{D/2}^{1/2}\log\{B + CS_\alpha(x)\}\,dx\right]. \tag{20}$$

For Case 2, *if* $\beta = \delta\alpha$,

$$2\,\mathrm{KL}(P_n, Q_n) \approx_n n\log\left(1 + \frac{2^{\alpha+2}C}{B\alpha D^\alpha}\right). \tag{21}$$

For Case 3, *if* $\beta = 0$,

$$2\,\mathrm{KL}(P_n, Q_n) \approx_n n\log\left(1 + \frac{4C}{B}\sum_{j=\frac{1}{2}(r_0+1)}^{\infty} j^{-\alpha-1}\right). \tag{22}$$

For Cases 2 *and* 3, *if* $\delta\alpha < \beta < \alpha$,

$$2\,\mathrm{KL}(P_n, Q_n) \sim (\beta - \delta\alpha)n\log n \tag{23}$$

and if $\beta \geq \alpha$,

$$2\,\mathrm{KL}(P_n, Q_n) \sim \alpha(1-\delta)n\log n. \tag{24}$$

Note that the term in square brackets on the right-hand side of (20) is positive by Jensen's inequality.

To make sense of these results, it helps to consider how α affects the behavior near the origin of the autocovariance function K given by (6). Define $\langle x\rangle^\alpha$ to mean $|x|^\alpha$ when α is not an even integer and $|x|^\alpha \log|x|$ when it is an even integer. Then there exist constants $C_0, \ldots, C_q$ with q the largest integer not greater than $\frac{1}{2}\alpha$, such that

$$K(x) - \sum_{j=0}^{q-1} C_j x^{2j} \sim C_q\langle x\rangle^\alpha,$$

which follows from Pitman (1968); see Stein (1999, p. 34). The term $C_q\langle x\rangle^\alpha$ is called the principal irregular term (Stein, 1999, p. 28) of K; it controls the local behavior of the corresponding process. We see that at least for α not an even integer, $C_q\langle\frac{1}{n}\rangle^\alpha$ and the nugget variance $Bn^{-\beta}$ are of the same order of magnitude when $\alpha = \beta$. Thus, the noise is arguably negligible when $\beta > \alpha$ and it is not surprising that in Proposition 3, whenever $\beta > \alpha$, the value of β does not affect the asymptotic result and the results are the same as in the case of no nugget.

For Cases 2 or 3 (Case 1 is not really "low rank"), when $\beta > \alpha$, the coefficient multiplying $n\log n$ in the asymptotic KL divergence is proportional to α. Thus, in this negligible nugget case, the smoother the process, the worse the low rank approximation does. However, (18) indicates that the size of the nugget necessary for a low rank approximation to have KL divergence small relative to the sample size is smaller for larger α. Furthermore, (23) shows that for β in the intermediate range $\delta\alpha < \beta < \alpha$, the effect of increasing δ (and hence the rank of the low rank approximation) is greater for larger α.

Next consider approximating Σ_0 by a block diagonal matrix with specified block diagonal structure. It is not hard to show that the block diagonal Σ_1 that minimizes the KL divergence is the matrix whose entries equal those of Σ_0 in the block diagonals and is 0 elsewhere. Writing Σ_{0j} for the jth diagonal block of Σ_0, we get $2\,\mathrm{KL}(P_n, Q_n) = \sum_{j=1}^{b}\log|\Sigma_{0j}| - \log|\Sigma_0|$. For the circulant setting considered in Propositions 1–3, write Σ_n^r for the $r \times r$ covariance matrix of r evenly spaced observations with spacing $\frac{1}{n}$.

Proposition 4. *If* $0 < \alpha < 2$ *and* $r > 1$, *then*

$$\log |\Sigma_n^r| \sim -\min(\beta, \alpha)(r-1)\log n. \tag{25}$$

This result is proven in the Appendix. I would expect this result to hold when $\alpha \geq 2$ as long as $r > \frac{1}{2}\alpha$. From (15) to (17), $\log |\Sigma_0| \sim -\min(\alpha, \beta)n \log n$. For any fixed r and any n that is a multiple of r, we then have, at least for $\alpha < 2$,

$$2\,\mathrm{KL}(P_n, Q_n) \sim \frac{1}{r}\min(\alpha, \beta)n \log n \tag{26}$$

for Q_n the independent block model with contiguous blocks of size r. (26) still holds when n is not a multiple of r if we take all of the blocks besides one to have size r and the last block to have size less than r. Thus, the KL divergence is nearly proportional to the reciprocal of the block size. In contrast, for any fixed r, using the low rank approximation of rank r, if $\beta > 0$, (23) and (24) imply $2\,\mathrm{KL}(P_n, Q_n) \sim \min(\alpha, \beta)n \log n$ independent of r. Thus, whenever the noise is small, independent blocks of a fixed size $r > 1$ yield a much better approximation to the true measure than the best low rank approximation with rank r.

3. Likelihoods for fixed versus random effects

One way to think of the low rank approach is as a random effects model. Specifically, writing $\mathbf{0}$ for a vector of zeros, suppose $\boldsymbol{Z} = X\boldsymbol{\beta} + \boldsymbol{e}$, where X is fixed, $\boldsymbol{\beta}$ and $\boldsymbol{e}$ are independent, $\boldsymbol{\beta} \sim N(\mathbf{0}, V)$ and $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2 I)$. Then the marginal distribution of $\boldsymbol{Z}$ is $N(\mathbf{0}, \sigma^2 I + XVX')$. In a random effects model, it would be common to treat at least the intercept term as a fixed effect, but from a Bayesian perspective, all unknown coefficients are treated as random, although there is no need to take their prior means to be 0. Generally when one uses a random effects model, some kind of structure is assumed for V. With additional assumptions, there can sometimes be substantial benefits in treating $\boldsymbol{\beta}$ as random rather than fixed. Thus, for example, for space–time data, if one is willing to assume stationarity in time but not space, a large number of replications across time may provide a basis for allowing estimation of the covariance structure of spatial random effects without strong additional assumptions on the spatial dependence. For stationary spatial processes observed on a grid, the (possibly tapered) spatial periodogram can provide another basis for estimating covariance structures by adopting the usual approximation that the values of the spatial periodogram at different Fourier frequencies are independent and the expected values of the periodogram at nearby frequencies are similar (Whittle, 1954; Dahlhaus and Künsch, 1987).

For the low rank approximations considered in this work, it may not be clear what kind of structure to assume for V, so one might consider letting V be any diagonal matrix with nonnegative diagonal elements (e.g., Stein, 2007b) or even any positive semidefinite matrix (Cressie and Johannesson, 2008). However, I now think that leaving V so unconstrained is at best a waste of time. First, consider the fact that conditionally on $\boldsymbol{\beta}$, the distribution of $\boldsymbol{Z}$ does not depend on V. Thus, $\boldsymbol{Z}$ cannot possibly contain more information about V than $\boldsymbol{\beta}$ does. If we now assume $\boldsymbol{\beta} \sim N(\mathbf{0}, D)$, where D is a diagonal matrix with arbitrary nonnegative diagonal entries, we see that $\boldsymbol{\beta}$ provides one observation per variance parameter to estimate, which leaves an essentially hopeless estimation problem without further assumptions. Allowing a nonzero prior mean for $\boldsymbol{\beta}$ or an arbitrary covariance matrix then results in more parameters than observations, which is completely hopeless. One may as well treat $\boldsymbol{\beta}$ as a fixed effect and estimate it by using least squares.

In terms of likelihoods, the maximized likelihood obtained by treating $\boldsymbol{\beta}$ as a fixed effect is at least as large as the likelihood obtained by maximizing over σ^2 and any possible positive semidefinite V. To prove this inequality, consider maximizing the likelihood based on $\boldsymbol{Z} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$ over $\boldsymbol{\beta}$ and σ^2 versus maximizing the likelihood based on $\boldsymbol{Z} \sim N\big(\mathbf{0}, \sigma^2(I + X\tilde{V}X')\big)$ over σ^2 and all positive semidefinite $\tilde{V}$, where $\tilde{V} = \sigma^{-2}V$. For the fixed-effects model, assuming X is of full rank, the maximized loglikelihood is easily shown to be (here and below, leaving out the term $-\frac{n}{2}\log(2\pi)$)

$$-\frac{n}{2} - \frac{n}{2}\log\left[\frac{1}{n}\boldsymbol{Z}'\big\{I - X(X'X)^{-1}X'\big\}\boldsymbol{Z}\right]. \tag{27}$$

For the random effects model, if we fix $\tilde{V}$ and maximize over σ^2, the maximized loglikelihood is

$$-\frac{n}{2}-\frac{n}{2}\log\left[\frac{1}{n}\boldsymbol{Z}'\left\{I-X\left(\tilde{V}^{-1}+X'X\right)^{-1}X'\right\}\boldsymbol{Z}\right]-\frac{1}{2}\log|I+X\tilde{V}X'|. \tag{28}$$

Now, $\log|I+X\tilde{V}X'|\geq 0$, so the difference between (27) and (28) is at least

$$\frac{n}{2}\log\left[\frac{\boldsymbol{Z}'\left\{I-X\left(\tilde{V}^{-1}+X'X\right)^{-1}X'\right\}\boldsymbol{Z}}{\boldsymbol{Z}'\left\{I-X(X'X)^{-1}X'\right\}\boldsymbol{Z}}\right]. \tag{29}$$

Eq. (29) is nonnegative because

$$\begin{aligned}&\boldsymbol{Z}'\left\{I-X\left(\tilde{V}^{-1}+X'X\right)^{-1}X'\right\}\boldsymbol{Z}-\boldsymbol{Z}'\left\{I-X(X'X)^{-1}X'\right\}\boldsymbol{Z}\\&\quad=\boldsymbol{Z}'X\left\{(X'X)^{-1}-\left(\tilde{V}^{-1}+X'X\right)^{-1}\right\}X'\boldsymbol{Z}\end{aligned}$$

is nonnegative since $(X'X)^{-1}-\left(\tilde{V}^{-1}+X'X\right)^{-1}$ is positive semidefinite. Thus, (29) is nonnegative for any $\tilde{V}$ and hence for the $\tilde{V}$ that maximizes (28). Since maximizing over σ^2 and V is the same as maximizing over σ^2 and $\tilde{V}$, we have proven the following result:

Proposition 5. *For X full rank, the maximized likelihood for $\boldsymbol{Z}\sim N(X\boldsymbol{\beta},\sigma^2 I)$ over $\boldsymbol{\beta}$ and σ^2 is at least as large as the maximized likelihood for $\boldsymbol{Z}\sim N\left(\boldsymbol{0},\sigma^2 I+XVX'\right)$ over σ^2 and all positive semidefinite V.*

Further linear algebra shows Proposition 5 also holds if the mean of $\boldsymbol{\beta}$ in the random effects model is not $\boldsymbol{0}$. It is plausible that such an elementary result as Proposition 5 is known in some context, but I have been unable to find a published result to this effect, so have included the elementary proof given here. This result quantifies the notion that there is no point in treating $\boldsymbol{\beta}$ as random unless one is willing to make some meaningful assumptions about V. It also provides a convenient upper bound for the likelihood that can be obtained under any possible assumption about V.

4. Numerical results

This section considers some numerical results on KL divergences for low rank and independent block approximations. Write $\mathcal{M}_\nu$ for the Matérn autocorrelation function with smoothness parameter ν and range 1: $\mathcal{M}_\nu(d)=\frac{2^{\nu-1}}{\Gamma(\nu)}d^\nu\mathcal{K}_\nu(d)$ for $d\geq 0$, where $\mathcal{K}_\nu$ is a modified Bessel function of order ν. For $\nu=\frac{1}{2}$, the model is just the exponential autocorrelation function and for $\nu=1$, the Whittle model. Consider 1200 observations on a line with spacing $\frac{1}{1200}$ and autocovariance function $K(x)=\mathcal{M}_\nu(\phi|x|)$ plus a nugget. First consider fixing ϕ and varying ν and the nugget; later in the section I consider varying ϕ. The inverse range parameter ϕ is set to 12 in Figs. 1 and 2 and the nugget variances to 0, 0.01 and 0.1. For the exponential model, $K(0)-K(1/1200)\approx 0.01$, so the nugget variance of 0.01 makes the noise comparable to the local fluctuations of the process. Fig. 1 shows KL divergences from the true model for the best possible low rank approximations and for independent blocks with block sizes that are factors of 1200. The white noise approximation corresponds to a rank of 0 or to a block size of 1, so Figs. 1 and 2 are plotted with this appropriate offset on the horizontal axis to highlight that the results from the two approximations are the same in this case. When there is no nugget, the independent block approximation is always better than the comparable low rank approximation. When the nugget is 0.01, the low rank approximation is only better when the rank is 599 and there are two blocks. When the nugget is 0.1, the low rank approximation is better when there are eight or fewer blocks, or a rank of at least 149. Results for the Whittle model are shown in Fig. 2. As expected from Proposition 3 in Section 2, the low rank approximations are now even worse when there is no nugget but improve quickly with increasing nugget, such that for nugget of 0.1, the low rank approximation now outperforms independent blocks when the rank is at least 19 (60 or fewer blocks). In this case, the nugget effect strongly dominates the local behavior of the process ($K(0)-K(1/1200)\approx 0.00036$) and low rank approximations work quite well. We can make the comparisons of KL divergences even

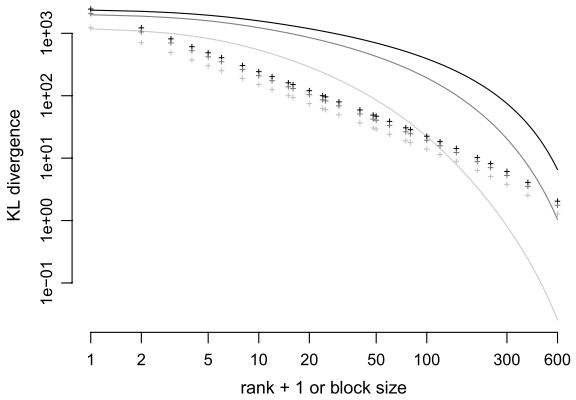

Fig. 1. KL divergences of low rank (curves) and independent block (+symbols) approximations to 1200 observations on a line with spacing $\frac{1}{1200}$ and autocovariance function $K(x) = e^{-12|x|}$ plus a nugget. Black (curves/+symbols) corresponds to a nugget of 0, medium gray to 0.01 and light gray to 0.1. Low rank approximations are best possible.

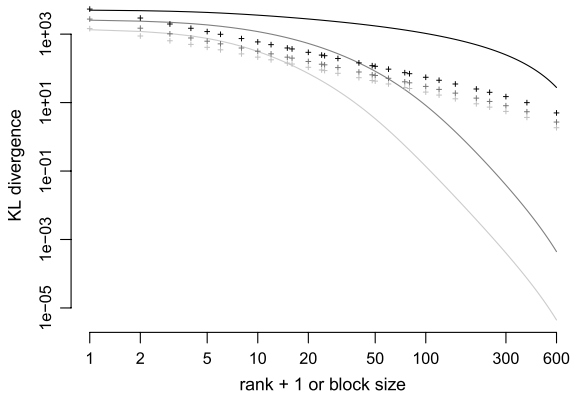

Fig. 2. Same as Fig. 1 except that the exponential autocovariance function is replaced by the Whittle autocovariance function $12|x|\mathcal{K}_1(12|x|)$.

more favorable to the low rank approximation by considering positive nuggets and values for the smoothness parameter ν greater than 1. However, in my experience, values of ν much larger than 1 are not common for environmental processes.

Next consider varying the range of an exponential covariance function. One way to look at the effect of the range parameter is to fix the nugget and consider $K(x) = e^{-\phi|x|}$ for different values of ϕ. The spectral density corresponding to $K(x) = e^{-\phi|x|}$ is $f(\omega) = \frac{\phi}{\pi(\phi^2+\omega^2)}$, so increasing range (decreasing ϕ) corresponds to increasing variation at low frequencies but decreasing variation at sufficiently high frequencies. For evenly spaced observations, the relationship between the eigenvalues of the covariance matrix and the spectrum is similar to the periodic case considered in Section 2, with the decay of the eigenvalues being closely tied to the decay of the spectrum. Thus, when there is a nugget effect, we should expect the low rank approximation to perform increasingly well for sufficiently large ranges, especially for larger values of the rank, which is exactly what we see in the left panel of Fig. 3. Specifically, with the nugget fixed at 0.1 (the largest value in Fig. 1), the independent block approximation dominates for short ranges for a wide range of block sizes, but as the range increases, the KL divergence monotonically increases for the independent block approximation but eventually rapidly decreases for the low rank approximation and outperforms independent blocks, sometimes by a large margin. Indeed, since, as $\phi \to \infty$, $K(x) \to 1$ for any fixed x, the covariance matrix of the

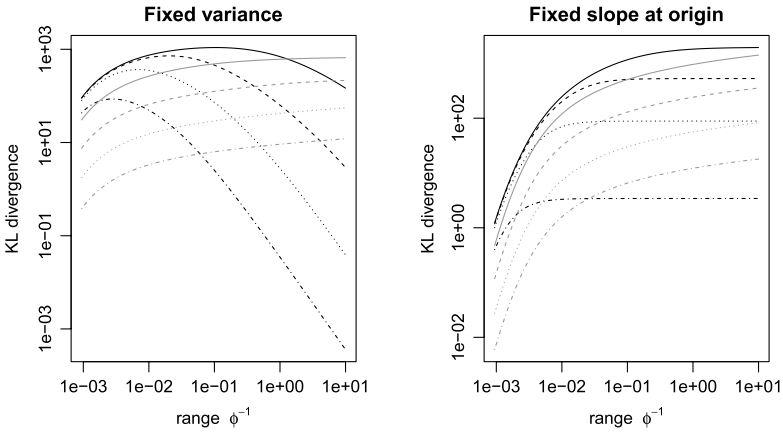

Fig. 3. KL divergences of low rank (black) and independent block (gray) approximations to 1200 observations on a line with spacing $\frac{1}{1200}$ and autocovariance function $K(x) = e^{-\phi|x|}$ plus a nugget of 0.1 (left panel) and $K(x) = 12\phi^{-1}e^{-\phi|x|}$ plus a nugget of 0.1 (right panel). Black curves correspond to low rank approximations (best possible) and gray to independent block approximations. Line types indicate B, the block size, with the rank of the corresponding low rank approximation equal to $B-1$; the solid, dashed, dotted and dash–dot curves correspond to $B = 3$, 12, 50 and 200 respectively.

observations converges to a diagonal matrix plus a rank 1 matrix (the matrix with all elements equal to 1), so the KL divergence of the low rank approximation will tend to 0 as $\phi \to \infty$ for any $B \geq 2$. In contrast, the KL divergence for the independent block approximation will not tend to 0 even for blocks of size 600 (and so two blocks) because it ignores the nontrivial cross-block dependence.

Instead of fixing the variance of the process $K(0)$, consider $f(\omega) = \frac{12}{\pi(\phi^2+\omega^2)}$, so that the scale of high frequency variation is fixed, which I would argue is the more natural way to vary the range. The corresponding covariance function is $K(x) = 12\phi^{-1}e^{-\phi|x|}$, so when $\phi = 12$, $K(\cdot)$ is the same as in Fig. 1. Note that $K'(0^+) = -12$ independently of ϕ; see Stein (1999) for more on the relationship of a spectral density at high frequencies and the behavior of the corresponding covariance function at the origin. Again fixing the nugget at 0.1, the right panel of Fig. 3 shows that the KL divergence grows for both approximations as the range increases. At the smaller ranges, the independent block approximation is clearly superior but the curves eventually flatten out for the low rank approximation whereas they continue to increase for the independent block approximation. Therefore, for sufficiently large ranges, the low rank approximation will have lower KL divergence, although Fig. 1 indicates that for moderate block sizes, the range will have to be more than 10 times the length of the observation interval for this to occur. Furthermore, for ranges where the low rank approximation has lower KL divergence, neither approximation is doing very well.

One might argue that, despite using the best possible low rank approximations, these comparisons are somewhat unfair because they assume the true model to be stationary. Undoubtedly, fitting a stationary model with independent blocks will perform less well when the truth is nonstationary, but I am unaware of any argument that suggests the low rank approximations to work better when the truth is nonstationary than when it is stationary. To give one example of what can happen when the truth is nonstationary, suppose Y is a stationary process on the real line with autocovariance function $K(x) = e^{-12|x|}$ and that, for $0 \leq x \leq 1$, $Z(x) = (1+\sin \pi x)Y(x)$. Observe Z at $\frac{j}{1200}$ for $j = 1, \ldots, 1200$ with an independent error term of variance 0.1. Now if we use a stationary exponential plus nugget model with 120 blocks of size 10 and choose the covariance parameters to minimize the KL divergence from the true model, the minimized KL divergence is 763, which is very large considering that there are only 1200 observations. However, if we use the best low rank approximation of rank 10, the KL divergence is 2959, nearly 4 times as large. For the low rank approximation to beat the independent block approximation with blocks of size 10, the rank has to be at least 105. Thus, even when the truth is nonstationary, we may be much better off using a stationary model with the independent block approximation rather than a low rank approximation. It is interesting to note that despite the severe

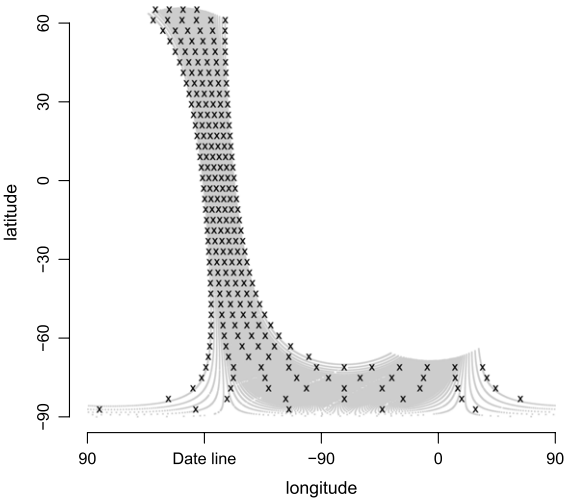

Fig. 4. Locations of total column ozone measurements (gray dots) and knots of basis functions in the subset of regressors model (black $\times$ symbols).

nonstationarity of this process, if we use a stationary exponential plus nugget model without blocking (or, one could say, a single block of size 1200), then the KL divergence is reduced to 312, still quite large, but the best low rank approximation must have rank at least 238 to obtain a smaller KL divergence.

5. Application to total column ozone

The Ozone Monitoring Instrument (OMI) is a satellite-based instrument that measures total column ozone levels (in Dobson units) globally at over a million locations a day. The instrument is aboard a polar-orbiting satellite making about 14.6 orbits each day and we consider the 83,305 Level-2 (ungridded) observations from a single orbit on January 2, 2007, near the international date line. The instrument observes backscattered light, so does not provide data during polar nights. Therefore, on January 2, the data only go up to about 66°N and cover a longitudinal swath of about 3000 km over most latitudes. The observation locations are shown in Fig. 4. Let us first consider a regression model with independent and identically distributed Gaussian errors for the data. Given that the Earth is nearly a sphere, it is natural to use regressors that are spherical harmonics in latitude (L) and longitude (ℓ). On the basis of some numerical experimenting, the following model for the mean function appears to provide a good fit to the observations relative to the number of parameters (214):

$$\sum_{n=0}^{25} \beta_{n0} c_{n0} P_n^0(\sin L) + \sum_{m=1}^{4} \sum_{n=m}^{25} c_{nm} P_n^m(\sin L)(\beta_{nm} \cos m\ell + \gamma_{nm} \sin m\ell) \tag{30}$$

where the P_n^m's are Legendre functions and the c_{nm}'s are known normalizing constants. The limit on n being much larger than that on m reflects the fact that the data cover a wide range of latitudes but a limited range of longitudes (except near the South Pole). The multiple R^2 for this model is 0.9941, so the model explains nearly all of the variation in the data. The mean squared error of the fitted values (the average of the squared residuals) is 31.72. By definition, the β_{nm}'s and γ_{nm}'s have been estimated to minimize this mean squared error and it follows that any other fit to the observations that is in the column space of the design matrix will do worse, so in particular, treating the β_{nm}'s and γ_{nm}'s as random cannot improve this result. Although the R^2 of 0.9941 may seem impressive, the mean squared prediction error obtained by predicting each observation by using its nearest neighbor is 28.65 and by the average of its 10 nearest neighbors is 12.07.

Plotting separate empirical variograms for different latitude ranges makes it clear that the ozone process is not stationary in latitude. Nevertheless, let us consider modeling the process as having a

constant unknown mean and a covariance function that is an exponential with an unknown scale and range (with distance measured chordally) plus a nugget with unknown variance, for a total of four parameters. Because the data are not on a grid, calculating the likelihood function for them even once would be extremely challenging. Consider breaking up the data into 78 independent blocks, where the blocks are made up of the observations within latitude bands $[2b, 2b + 2)$ for $b = -45, \ldots, 32$. The largest block contains 2477 observations, so it is feasible to calculate likelihoods under this approximation. The increase in loglikelihood of this model relative to the regression model is greater than 34,000, which is a nontrivial fraction of the number of observations and is thus immense by any reasonable measure. By Proposition 5, treating the regression coefficients in (30) as random would only increase this difference in maximized loglikelihoods. If one combines the mean model (30) with this covariance function and maximizes the likelihood over all 217 parameters (214 mean parameters and 3 covariance parameters), the maximized loglikelihood increases by not quite 1500 over the model with a constant mean, still arguably large, but much smaller than 34,000. If one's goal is to increase the loglikelihood, there are other ways to do so by a larger amount without adding so many parameters, for example, by changing the blocks or by allowing some simple forms of nonstationarity in latitude in the covariance structure.

Consider spatial prediction with a constant unknown mean and covariance model of exponential plus nugget. Because the estimated model makes sense as a stationary model on the sphere, there is no need to predict separately within each block. However, predicting each observation given all 83,304 other observations using the fitted model would be computationally very challenging. Instead, for each observation, predict it on the basis of its 50 nearest neighbors using the best linear predictor assuming the fitted model is correct. The mean squared prediction error is then 10.44, which is much better than the regression model and moderately better than just averaging the 10 nearest neighbors.

The performance of the low rank method of course depends on the basis functions chosen. Because spherical harmonics form an orthogonal basis on the sphere and I have chosen the upper limits on n and m in (30) to provide a particularly good fit to the data for the number of parameters used, I am doubtful that a major improvement is possible using a comparable number of basis functions without "cheating" (i.e., using basis functions picked on the basis of the data). For example, I also tried basis functions of the form $e^{-|x-x_j|/\theta}$ with the x_j's given by the black $\times$ signs in Fig. 4. There are 244 such grid points, so, including the intercept, the rank of the mean space is 245. Despite having somewhat more regressors than the spherical harmonic model, the average squared residual is at least 29.16 for all values of θ, so only slightly better than the spherical harmonic model. It is interesting to note that the minimum of this mean squared error occurs as $\theta \to \infty$, which is the same as using $|x - x_j|$ as the basis functions. The slight advantage of these basis functions relative to the spherical harmonics appears to be due to the extra knots over Antarctica rather than the particular form for the functions. No doubt one can do better by choosing the knots and/or the form of the basis function more carefully (Cressie and Johannesson, 2008 suggest multiresolution basis functions), but it seems unlikely one could get anywhere near the value of 10.44 for the mean squared prediction error of the exponential plus nugget model without taking a much larger set of basis functions.

Although it is not a focus of this work, it is worthwhile to consider what happens to the predictions when using the exponential plus nugget model with parameters estimated by the independent block approximation with blocks of different sizes. If, instead of using estimated parameters based on blocks made up of 2° latitude bands, which yielded a mean squared prediction error of 10.44, using blocks of width 4° yields a mean squared prediction error of 10.43 and using blocks of width 1° yields a mean squared prediction error of 10.46. In all cases, the linear predictors were based on the 50 nearest neighbors. These nearly identical mean squared errors are not surprising because it is well-known that small changes in the covariance function can often lead to negligible changes in the resulting spatial predictions (Stein, 1999). Since every doubling of the size of the blocks approximately quadruples the computational effort needed to obtain the independent block loglikelihood, one could argue that there is no point in using large blocks, at least in this setting. In particular, if the only interest is in the mean squared error of linear predictions within the region covered by the observations, there is likely to be no meaningful advantage in trying to estimate the covariance parameters of a stationary model via exact maximum likelihood. If, on the other hand, one were interested in an accurate estimate of

the range parameter, then exact maximum likelihood or an approximation that made good use of the information in the data at larger scales might help quite a bit.

6. Discussion

This work has focused on KL divergence from some specified model for a covariance function as a measure of model quality, although the ozone example also examines spatial prediction. The problem of how to assess the quality of an approximation is a subtle one and depends in part on the use for the approximate model. For example, if the goal is just to estimate the parameters of the original full rank model, then what counts is the quality of the inference about these parameters. The independent block approximation and others that are based on composite likelihoods (Vecchia, 1988; Stein et al., 2004; Caragea and Smith, 2007) yield unbiased estimating equations (Varin et al., 2011), and thus are clearly estimating the parameters of the original model. If one were only comparing estimation methods based on unbiased estimating equations, it would be appropriate to use the Godambe information measure (Varin et al., 2011; Stein, 2013) to compare their statistical efficiencies, but since the low rank methods do not generally yield unbiased estimating equations, that approach was not possible here. The KL divergence of the best possible low rank approximation puts a bound on how well the low rank approximations can do. The KL divergences for the independent block approximation serve as a reference for assessing the quality of the low rank approach and are not the best tools for evaluating the independent block approximation.

An important distinction between the low rank and independent block approximations is in how they are subsequently used in prediction. Thus, as we saw with the ozone data, there was no need to respect the blocks when predicting with the estimated parameters from the independent block approximation. Instead, we could predict at any location using nearest neighbors, thus achieving fast computation without having to use completely different observations for prediction on opposite sides of a block boundary. In contrast, at least as it is generally formulated, the low rank approach to analysis of spatial data uses the resulting low rank model for spatial prediction as well.

Other measures of model quality have been used in the literature, especially the Frobenius norm (Cressie and Johannesson, 2008) (the square root of the sum of squared elements) and the relative Frobenius (Sang and Huang, 2012) norm, on the differences of covariance matrices. This norm will often make low rank approximations look much better than independent block approximations. However, the choice of focusing on KL divergence is not arbitrary. The KL divergence has a direct statistical interpretation in terms of expected loglikelihoods. Likelihoods also have a close connection to prediction. Suppose P is the actual law for the spatial process Z, $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ is a set of observations from Z, p_1 is the density for Z_1 and, for $j > 1$, p_j is the conditional density of Z_j given $Z_1, \ldots, Z_{j-1}$. By the law of successive conditioning, the joint density of $\boldsymbol{Z}$ is

$$p_1(z_1) \prod_{j=2}^{n} p_j(z_j \mid z_1, \ldots, z_{j-1}).$$

If Q is another possible law for Z, then if $\mathrm{KL}(P_n, Q_n)$ is small, predictive statements under the two laws are similar, at least where there are observations. Thus, as long as the places at which we want to predict Z are not too different from the places at which we observe Z, we might then expect P and Q to yield similar predictive distributions. Indeed, the results on asymptotically optimal interpolants for equivalent Gaussian measures (Stein, 1999, Chapter 4) can be viewed as a mathematical embodiment of this notion. In contrast, the Frobenius norm has no comparable interpretations when applied to covariance matrices. In particular, it does not specifically penalize for lack of positive definiteness in a covariance matrix. Furthermore, even the relative Frobenius norm, unlike the KL divergence, does not possess the natural property of invariance under nonsingular linear transformations of the observations. Cressie and Johannesson (2008) develops a clever algorithm for fitting low rank covariance matrices to data based on the method of moments and the Frobenius norm that ensures that the resulting estimate is positive definite. Cressie and Johannesson (2008) directly assumes a low rank model and does not motivate it as an approximation to some parametric class of covariance functions. In particular, no assumption of stationarity is made in this approach. The fitted low rank

model is able to track variations in the empirical variogram at different locations across the globe, but, considering the results in Section 3, I am not persuaded that this approach is any better than the simpler and faster approach of using least squares with these same basis functions as linear regressors. Specifically, because all of the observations are from a single day, it seems to me that the estimates of what is called here the matrix V should be highly unstable. It would be interesting to run a large simulation study under known stationary and nonstationary models to see how the various low rank methods compare to each other and to full rank methods.

Another general approach for reducing the computational burden of estimation and prediction for large spatial datasets is covariance tapering (Furrer et al., 2006; Kaufman et al., 2008), in which a covariance function of interest is replaced by one that equals 0 beyond some relatively small interpoint distance. If the resulting sparse covariance matrix has approximately r nonzero elements per row, the situation might be viewed as computationally comparable to a low rank model with rank r or r independent blocks. However, this is not the case, at least if, as is recommended in Furrer et al. (2006), one uses exact likelihood calculations based on a sparse Cholesky decomposition (Davis, 2006). The sparse Cholesky decomposition generally requires a reordering of the columns of the covariance matrix to be effective and this reordering can require a substantial computational effort. Even after this reordering, the storage and number of flops required may be quite a bit greater than for an independent block approximation, depending on the exact locations of the observations. The tapered covariance function can yield severely biased parameter estimates, but Kaufman et al. (2008) and Stein (2013) describe various ways of obtaining unbiased estimating equations based on tapered covariance matrices. Although these approaches can yield good parameter estimates, for a given value of r, they are often outperformed by the independent block approach (Stein, 2013).

The focus here has been on irregularly sited observations. For stationary processes observed on a regular lattice, spectral approximations to the likelihood, first suggested by Whittle (1954), can be both fast and accurate (Dahlhaus and Künsch, 1987; Guyon, 1982). If observation locations form a subset of a regular grid, then spectral methods can still be effective if the locations of grid points without observations are not too irregular (Fuentes, 2007). When observations do not form a substantial subset of a regular grid, Fuentes (2007) describes a spectral approach based on averaging observations within grid cells. This approach is fast computationally, but Matsuda and Yajima (2009) shows that the resulting estimates can be substantially biased and instead suggests a bias correction to the usual Whittle approximation that takes account of a varying density of observations in the observation domain. The theoretical and simulation results in Matsuda and Yajima (2009) suggest that this approach sometimes provides estimates with reasonable statistical efficiency (but is not generally as good as the approach in Stein et al., 2004) and quite fast computation for moderate sample sizes. However, the largest sample size considered in this paper is 2000 and since one cannot use the fast Fourier transform with irregular observations, approaches based on computing the discrete Fourier transform at a number of frequencies comparable to the number of observations will likely run into computational problems for massive datasets.

The theoretical and numerical results in this work only consider multiples of the identity for the diagonal matrix. More general diagonal matrices (see, e.g., Eidsvik et al., 2012) could work substantially better, but I do not see how they can solve the fundamental problem that neither a low rank term nor a diagonal term can effectively capture the local fluctuations of a continuous but not too smooth process. Still, further investigation is warranted. Perhaps more promising, at least from a statistical perspective, is to take $D(\boldsymbol{\theta})$ in (3) to be sparse but not diagonal (Quiñonero-Candela and Rasmussen, 2005; Stein, 2007a). Sang and Huang (2012) combines covariance tapering with the predictive process approach of Banerjee et al. (2008) to good effect in a US precipitation dataset of size around 7000, but it is not so clear that this approach will scale well computationally or statistically to much larger datasets. Furthermore, these data show a range of around 50 km (Kaufman et al., 2008), which is quite small compared to the size of the region (the 48 contiguous United States), and the approximation may work less well for longer ranges (Sang and Huang, 2012, Fig. 2). Along somewhat similar lines, Sang et al. (2011) combines low rank and independent block approximations and shows that this combination can work better than a low rank term plus a tapered covariance function.

In some applications, one may not be much interested in modeling the fine scale structure of the process or in predicting the process. For example, if one were interested in the long term trends in a

spatial–temporal process, then fine scale spatial features might not matter. However, even in these situations, I would suggest great caution in using any likelihood-based methodology together with a low rank approximation because the likelihood's behavior will be dominated by the fine scale behavior whether or not that is of interest.

Although independent blocks can work quite well from both computational and statistical standpoints, the assumption of independence of blocks is so obviously wrong as to suggest that there must be something better one can do. The Vecchia approach to likelihood approximation (Vecchia, 1988; Stein et al., 2004) is an attractive alternative; Anitescu et al. (2012) describes another possible approach to parameter estimation based on stochastic approximations to the score equations. I prefer these approximate computations to exact computations under models whose KL divergences from natural full rank models are large. Thus, I disagree with the premise of Simpson et al. (2012), which advocates abandoning covariance functions as the basis of spatial modeling because of the computational difficulties that arise. Simpson et al. (2012) recommends using stochastic partial differential equations to motivate models, which is conceptually attractive in that it allows one to define nonstationary models and models on curved manifolds naturally. The equations are then discretized, leading to a Markov random field approximation. It would be interesting to evaluate the KL divergences of these Markov random field models from Matérn models in a broad set of situations, including using a range of integer and noninteger values for their parameter α (which corresponds to $\nu = \alpha - \frac{1}{2}d$ in the Matérn model), a range of nugget effects, and gridded and irregularly sited observations.

Given that exact likelihoods will often not be computable for spatial data, an important open challenge is then how to perform Bayesian inference when the likelihood cannot be accurately approximated. Banerjee et al. (2008), Finley et al. (2009) and Sang and Huang (2012) just use their approximate likelihoods as if they were the exact likelihoods in their Bayesian implementations, which has the risk of leading to overoptimistic inferences. Shaby and Ruppert (2012) show how one can use the results from an MCMC method to get asymptotically valid confidence intervals for parameters of a covariance function when using covariance tapering together with a corrected likelihood suggested in Kaufman et al. (2008). The method is based on an assumption of asymptotic normality and, as Shaby and Ruppert (2012) notes, it is not clear whether this approach has any inferential advantage over the asymptotic frequentist approach that they describe. Approximate Bayesian computation based on simulating from the prior (Marjoram et al., 2003; Blum et al., 2013) does not appear to be workable for large, irregularly sited spatial datasets for which it is not clear how one could choose a low dimensional summary statistic that contains most of the information about the unknown covariance structure.

Acknowledgments

This research was supported by US Department of Energy grant No. DE-SC0002557. The data used in this paper were acquired as part of the activities of NASA's Science Mission Directorate, and are archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC). The author thanks an associate editor and two referees for several helpful suggestions, especially that of looking at the effect of varying the range in the numerical study.

Appendix. Proofs

Proof of Proposition 1. For Cases 2 and 3,

$$\sum_{j=r_n'+1}^{n'} S_\alpha\left(\frac{j}{n}\right) = \sum_{j=r_n'+1}^{n'} \sum_{q=-\infty}^{\infty} \left(\left| q + \frac{j}{n} \right|^{-\alpha-1} + \left| q + 1 - \frac{j}{n} \right|^{-\alpha-1} \right)$$

$$= \sum_{q=-\infty}^{\infty} \sum_{j=r_n'+1}^{n-r_n'-1} \left| q + \frac{j}{n} \right|^{-\alpha-1}$$

$$
\begin{aligned}
&= 2\sum_{j=r_n'+1}^{\infty}\left(\frac{j}{n}\right)^{-\alpha-1} - 2\sum_{q\neq 0}\sum_{j=0}^{r_n'}\left|q+\frac{j}{n}\right|^{-\alpha-1}\\
&= 2n^{\alpha+1}\sum_{j=r_n'+1}^{\infty} j^{-\alpha-1} + O\left(r_n\sum_{q=1}^{\infty} q^{-\alpha-1}\right)\\
&= 2n^{\alpha+1}\sum_{j=r_n'+1}^{\infty} j^{-\alpha-1} + O\left(n^{\delta}\right)
\end{aligned}
$$

and (10) follows since $n' - r_n' = \frac{1}{2}n + O(1)$ for Case 3. For nonnegative sequences a_n and b_n, write $a_n \ll b_n$ if there exists a finite constant C such that $a_n \le Cb_n$ for all n. To obtain (9), use

$$
\begin{aligned}
\left|\sum_{j=r_n'+1}^{\infty} j^{-\alpha-1} - \int_{\frac{1}{2}Dn^{\delta}}^{\infty} x^{-\alpha-1}\,dx\right| &\ll \sum_{j=r_n'+1}^{\infty}\left|j^{-\alpha-1} - \int_{j-1}^{j} x^{-\alpha-1}\,dx\right| + n^{-\delta(\alpha+1)}\\
&\ll \sum_{j=r_n'+1}^{\infty} j^{-\alpha-2} + n^{-\delta(\alpha+1)}\\
&\ll n^{-\delta(\alpha+1)},
\end{aligned}
$$

which gives

$$
\sum_{j=r_n'+1}^{n'} S_\alpha\left(\frac{j}{n}\right) = \frac{2^{\alpha+1}}{\alpha D^{\alpha}} n^{\alpha(1-\delta)+1} + O\left(n^{(1-\delta)(\alpha+1)} + n^{\delta}\right)
$$

and (9) follows since $n' - r_n' = \frac{1}{2}n + O(n^{\delta})$. To obtain (8), use

$$
\begin{aligned}
\left|\sum_{j=r_n'+1}^{n'} S_\alpha\left(\frac{j}{n}\right) - n\int_{D/2}^{1/2} S_\alpha(x)\,dx\right| &\ll 1 + \sum_{j=r_n'+1}^{n'}\left|S_\alpha\left(\frac{j}{n}\right) - \int_{j-1}^{j} S_\alpha(x)\,dx\right|\\
&\ll 1 + \sum_{j=r_n'+1}^{n'}\left|S_\alpha\left(\frac{j}{n}\right) - S_\alpha\left(\frac{j-1}{n}\right)\right|,
\end{aligned}
$$

which is bounded in n since S has a bounded first derivative in $\left[t, \frac{1}{2}\right]$ for any $0 < t < \frac{1}{2}$.

Proof of Proposition 2. The proofs of (12)–(17) are all similar and are based on the fact that $\log S_\alpha$ is integrable on $\left(0, \frac{1}{2}\right)$. For example, to prove (16), note that

$$
\sum_{j=r_n'+1}^{n'} \log g_{j,n} = -\alpha(n' - r_n') + \sum_{j=r_n'+1}^{n'} \log\left\{B + CS_\alpha\left(\frac{j}{n}\right)\right\}
$$

and

$$
\begin{aligned}
\sum_{j=r_n'+1}^{n'} \log\left\{B + CS_\alpha\left(\frac{j}{n}\right)\right\} &\approx_n \sum_{j=1}^{n'} \log\left\{B + CS_\alpha\left(\frac{j}{n}\right)\right\}\\
&\approx_n n\int_0^{1/2} \log\{B + CS_\alpha(x)\}\,dx,
\end{aligned}
$$

and (16) follows.

Proof of Proposition 4. The diagonal elements of Σ_n^r are given by $K(0) + Bn^{-\beta}$, where $K(0) = \sum_{j=-\infty}^{\infty} f_j$. By using an Abelian theorem for Fourier transforms (Pitman, 1968; see also Stein, 1999, p. 34), it is possible to show that as $x \to 0$,

$$K(x) - K(0) \sim -C_1|x|^\alpha,$$

where $C_1 = \pi C(2\pi)^\alpha / \left\{\Gamma(\alpha+1)\sin\left(\frac{1}{2}\pi\alpha\right)\right\}$. Thus, for $i \neq j$, $\sigma_n(i,j)$, the ijth element of Σ_n^r is of the form

$$\sigma_n(i,j) = K(0) - C_1\left|\frac{i-j}{n}\right|^\alpha + o\left(n^{-\alpha}\right).$$

For $0 < \alpha < 2$, the function $-|x|^\alpha$ has a locally stationary equivalent version on any bounded interval (Chilès and Delfiner, 2012, (4.33)). This means that for any given finite T, there exists a constant C_0 (depending on T) such that $C_0 - |x|^\alpha$ is a positive definite function on $[-T, T]$. Thus, for any given r, we can find C_0 such that the $r \times r$ matrix with ijth element $C_0 - C_1|i-j|^\alpha$ is positive definite. Let us call this matrix M_r and write $\mathbf{1}_r$ for the vector of entries 1 of length r.

There are three cases to consider. When $\beta > \alpha$,

$$n^\alpha \Sigma_n^r = M_r + \{K(0) - C_0 n^{-\alpha}\}\mathbf{1}_r\mathbf{1}_r' + R_{nr},$$

where here and below the remainder term R_{nr} has all terms $o(1)$ as $n \to \infty$. By the matrix determinant lemma,

$$|M_r + \{K(0) - C_0 n^{-\alpha}\}\mathbf{1}_r\mathbf{1}_r' + R_{nr}| \quad = |M_r + R_{nr}|\left[1 + \left\{n^\alpha K(0) - C_0\right\}\mathbf{1}_r'(M_r + R_{nr})^{-1}\mathbf{1}_r\right].$$

Since $|M_r + R_{nr}| \to |M_r|$ and $\mathbf{1}_r'(M_r + R_{nr})^{-1}\mathbf{1}_r \to \mathbf{1}_r' M_r^{-1}\mathbf{1}_r$ as $n \to \infty$, it follows that

$$|\Sigma_n^r| \sim n^{-\alpha(r-1)}|M_r|\, K(0)\, \mathbf{1}_r' M_r^{-1}\mathbf{1}_r,$$

which is more than sufficient to prove Proposition 4 when $\beta > \alpha$. When $\beta = \alpha$, this argument can be slightly modified to prove

$$|\Sigma_n^r| \sim n^{-\alpha(r-1)}|M_r + BI_r|\, K(0)\, \mathbf{1}_r'(M_r + BI_r)^{-1}\mathbf{1}_r.$$

Finally, when $\beta < \alpha$,

$$n^\beta \Sigma_n^r = BI_r + n^\beta K(0)\, \mathbf{1}_r\mathbf{1}_r' + R_{nr}$$

and it follows from the matrix determinant lemma that

$$|\Sigma_n^r| \sim n^{-\beta(r-1)} r B^{r-1} K(0),$$

yielding Proposition 4 when $\beta < \alpha$.

References

Anitescu, M., Chen, J., Wang, L., 2012. A matrix-free approach for solving the parametric Gaussian process maximum likelihood problem. SIAM Journal on Scientific Computing 34, A240–A262.

Banerjee, S., Finley, A.O., Waldmann, P., Ericsson, T., 2010. Hierarchical spatial process models for multiple traits in large genetic trials. Journal of the American Statistical Association 105, 506–521.

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society. Series B 70, 825–848.

Blum, M.G.B., Nunes, M.A., Prangle, D., Sisson, S.A., 2013. A comparative review of dimension reduction methods in approximate Bayesian computation. Statistical Science 28, 189–208.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, second ed. Springer-Verlag, New York.

Caragea, P.C., Smith, R.L., 2007. Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. Journal of Multivariate Analysis 98, 1417–1440.

Chalupka, K., Williams, C.K.I., Murray, I., 2013. A framework for evaluating approximation methods for Gaussian process regression. The Journal of Machine Learning Research 14, 333–350.

Chilès, J., Delfiner, P., 2012. Geostatistics: Modeling Spatial Uncertainty, second ed. John Wiley, New York.

Cressie, N., Johannesson, G., 2006. Spatial prediction of massive datasets. In: Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference. Australian Academy of Science, Canberra, pp. 1–11.

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society. Series B 70, 209–226.

Dahlhaus, R., Künsch, H., 1987. Edge effects and efficient parameter estimation for stationary random fields. Biometrika 74, 877–882.
Davis, T.A., 2006. Direct Methods for Sparse Linear Systems. SIAM, Philadelphia.
Eidsvik, J., Finley, A.O., Banerjee, S., Rue, H., 2012. Approximate Bayesian inference for large spatial datasets using predictive process models. Computational Statistics & Data Analysis 56, 1362–1380.
Finley, A.O., Sang, H., Banerjee, S., Gelfand, A.E., 2009. Improving the performance of predictive process modeling for large datasets. Computational Statistics & Data Analysis 53, 2873–2884.
Fuentes, M., 2007. Approximate likelihood for large irregularly spaced spatial data. Journal of the American Statistical Association 102, 321–331.
Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. Journal of Computational and Graphical Statistics 15, 502–523.
Golub, G.H., van Loan, C.F., 1996. Matrix Computations, third ed. The Johns Hopkins University Press, Baltimore.
Guyon, X., 1982. Parameter estimation for a stationary process on a d-dimensional lattice. Biometrika 69, 95–105.
Katzfuss, M., Cressie, N., 2012. Bayesian hierarchical spatio-temporal smoothing for very large datasets. Environmetrics 23, 94–107.
Kaufman, C.G., Schervish, M.J., Nychka, D.W., 2008. Covariance tapering for likelhood-based estimation in large spatial data sets. Journal of the American Statistical Association 103, 1545–1555.
Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society. Series B 63, 425–464.
Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S., 2003. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America 100, 15324–15328.
Matsuda, Y., Yajima, Y., 2009. Fourier analysis of irregularly spaced data on R^d. Journal of the Royal Statistical Society. Series B 71, 191–217.
Pitman, E.J.G., 1968. On the behaviour of the characteristic function of a probability distribution in the neighbourhood of the origin. Journal of the Australian Mathematics Society. Series A 8, 422–443.
Quiñonero-Candela, J., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. Journal of Machine Learning Research 6, 1939–1959.
Quiñonero-Candela, J., Rasmussen, C.E., Williams, C.K.I., 2007. Approximation methods for Gaussian process regression. In: Bottou, L., Chapelle, O., DeCoste, D., Weston, J. (Eds.), Large-Scale Kernel Machines. In: Neural Information Processing, MIT Press, Cambridge, MA, pp. 203–223.
Sang, H., Huang, J.Z., 2012. A full scale approximation of covariance functions for large spatial data sets. Journal of the Royal Statistical Society. Series B 74, 111–132.
Sang, H., Jun, M., Huang, J.Z., 2011. Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. Annals of Applied Statistics 5, 2519–2548.
Shaby, B., Ruppert, D., 2012. Tapered covariance: Bayesian estimation and asymptotics. Journal of Computational and Graphical Statistics 21, 433–452.
Simpson, D., Lindgren, F., Rue, H., 2012. In order to make spatial statistics computationally feasible, we need to forget about the covariance function. Environmetrics 23, 65–74.
Smola, A.J., Bartlett, P.L., 2001. Sparse greedy Gaussian process regression. In: Leen, T.K., Diettrich, T.G., Tresp, V. (Eds.), Advances in Neural Information Processing Systems, vol. 13. MIT Press, Cambridge, MA, pp. 619–625.
Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.
Stein, M.L., 2007a. A modeling approach for large spatial datasets. Journal of the Korean Statistical Society 37, 3–10.
Stein, M.L., 2007b. Spatial variation of total column ozone on a global scale. Annals of Applied Statistics 1, 191–210.
Stein, M.L., 2013. Statistical properties of covariance tapers. Journal of Computational and Graphical Statistics http://dx.doi.org/10.1080/10618600.2012.719844. (in press).
Stein, M.L., Chi, Z., Welty, L.J., 2004. Approximating likelihoods for large spatial datasets. Journal of the Royal Statistical Society. Series B 66, 275–296.
Stoica, P., Jansson, M., 2009. On maximum likelihood estimation in factor analysis—an algebraic derivation. Signal Processing 89, 1260–1262.
Sun, Y., Li, B., Genton, M.G., 2012. Geostatistics for large datasets. In: Porcu, E., Montero, J.M., Schlather, M. (Eds.), Advances and Challenges in Space–Time Modelling of Natural Events, vol. 207. Springer, Berlin, pp. 55–77.
Varin, C., Reid, N., Firth, D., 2011. An overview of composite likelihood methods. Statistica Sinica 21, 5–42.
Vecchia, A.V., 1988. Estimation and identification for continuous spatial processes. Journal of the Royal Statistical Society. Series B 50, 297–312.
Whittle, P., 1954. On stationary processes in the plane. Biometrika 41, 434–449.
Williams, C., Seeger, M., 2001. Using the Nyström method to speed up kernel machines. In: Leen, T.K., Diettrich, T.G., Tresp, V. (Eds.), Advances in Neural Information Processing Systems, vol. 13. MIT Press, Cambridge, MA, pp. 682–688.