Geostatistical Modeling for Large Data Sets II

Whitney Huang

CLEMSON UNIVERSITY

wkhuang@clemson.edu

Clemson ENVR Group, September 30, 2020

Outline

Fixed Rank Kriging (Cressie & Johannesson 08)

J. R. Statist. Soc. B (2008)
70, *Part* 1, *pp.* 209–226

Fixed rank kriging for very large spatial data sets

Noel Cressie

The Ohio State University, Columbus, USA

and Gardar Johannesson

Lawrence Livermore National Laboratory, Livermore, USA

[Received May 2006. Final revision July 2007]

Summary. Spatial statistics for very large spatial data sets is challenging. The size of the data set, n, causes problems in computing optimal spatial predictors such as kriging, since its computational cost is of order n^3. In addition, a large data set is often defined on a large spatial domain, so the spatial process of interest typically exhibits non-stationary behaviour over that domain. A flexible family of non-stationary covariance functions is defined by using a set of basis functions that is fixed in number, which leads to a spatial prediction method that we call fixed rank kriging. Specifically, fixed rank kriging is kriging within this class of non-stationary covariance functions. It relies on computational simplifications when n is very large, for obtaining the spatial best linear unbiased predictor and its mean-squared prediction error for a hidden spatial process. A method based on minimizing a weighted Frobenius norm yields best estimators of the covariance function parameters, which are then substituted into the fixed rank kriging equations. The new methodology is applied to a very large data set of total column ozone data, observed over the entire globe, where n is of the order of hundreds of thousands.

Fixed Rank Kriging (FRK)

Fixed rank kriging outlines how kriging can be applied in the low-rank parameterization of a spatial process.

- $\boldsymbol{A}$: multiresolutional non-orthogonal basis functions

$$a(\boldsymbol{s})^{\mathrm{T}} = \left(\{1 - (\|\boldsymbol{s} - s^*_{1(l)}\|/\rho_{1(l)})^2\}^2_+, \cdots, \{1 - (\|\boldsymbol{s} - s^*_{r(l)}\|/\rho_{r(l)})^2\}^2_+ \right)$$

- $\boldsymbol{Z} = \left(Z\left(\boldsymbol{s}^*_1\right), \cdots, Z\left(\boldsymbol{s}^*_r\right)\right)^{\mathrm{T}}$, $\boldsymbol{\Sigma_Z}$ to be estimated "nonparametrically" from data

The fixed rank kriging is equivalent to the following low rank model

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{r} a(\boldsymbol{s} - \boldsymbol{s}^*_{j(l)})Z_j + \varepsilon(\boldsymbol{s}) + \xi(\boldsymbol{s})$$

Multiresolutional basis functions

Motivation: To capture multiple scales of spatial variation

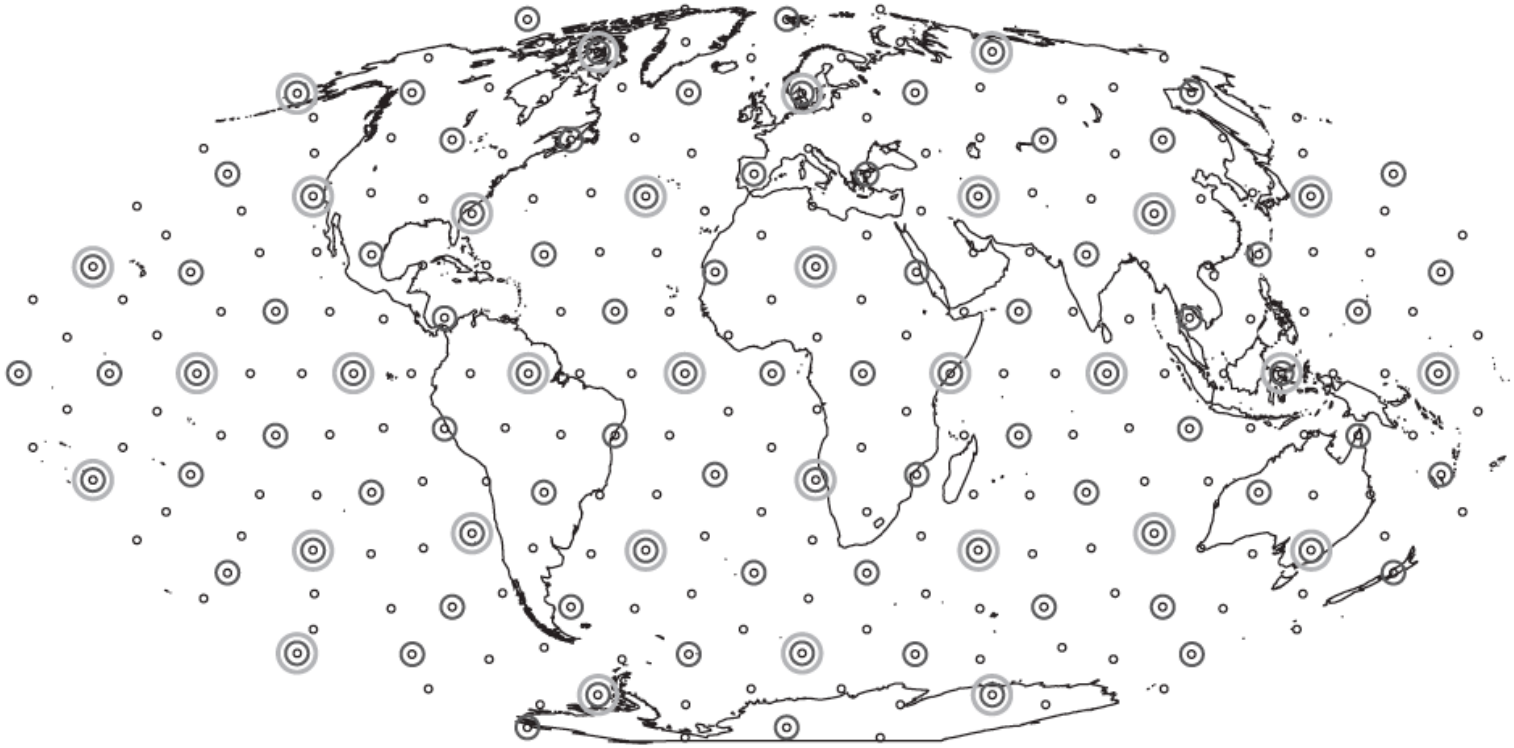


Figure: courtesy of Cressie & Johannesson 08

FRK Can Accommodate Nonstationary covariance

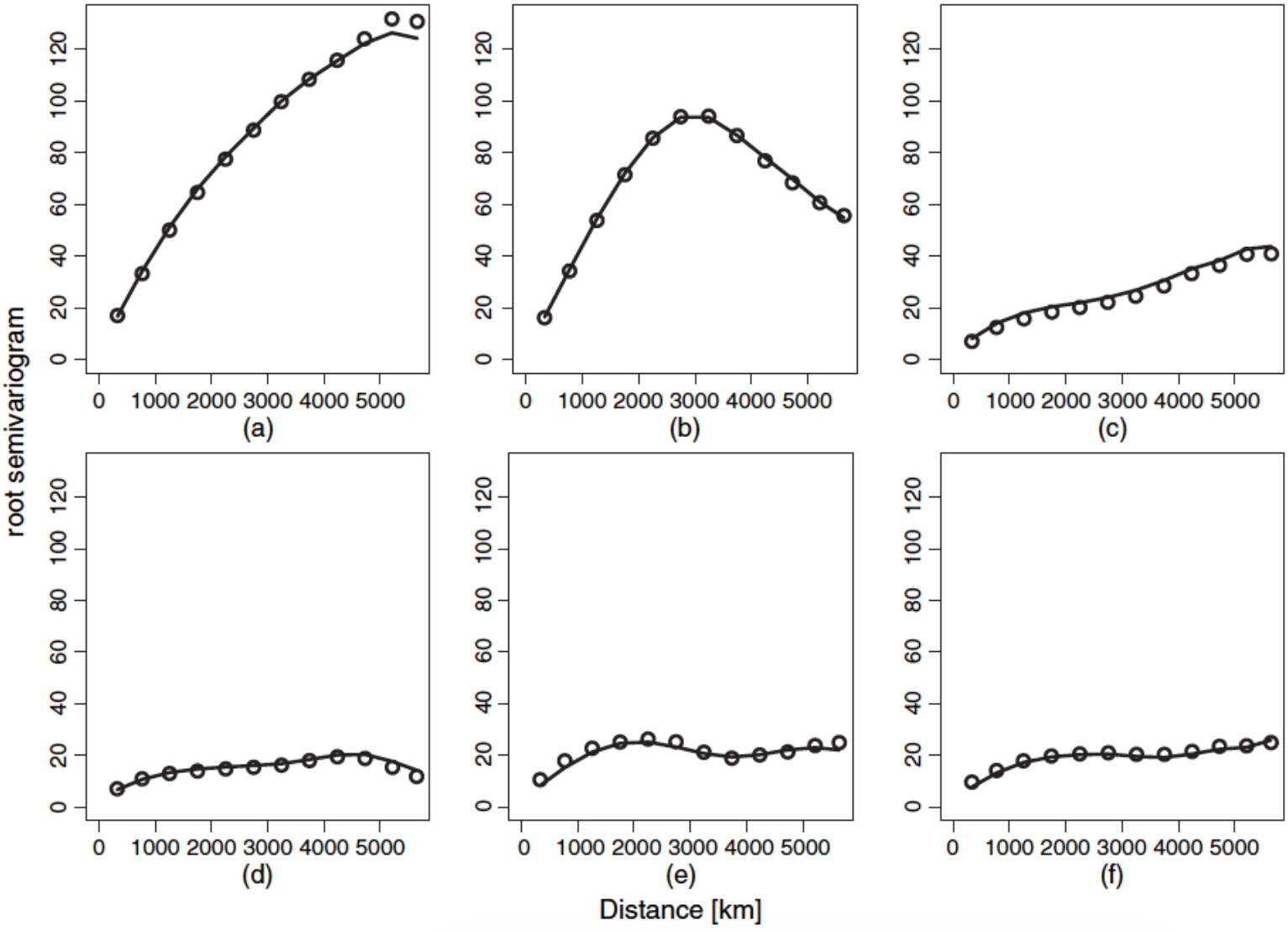
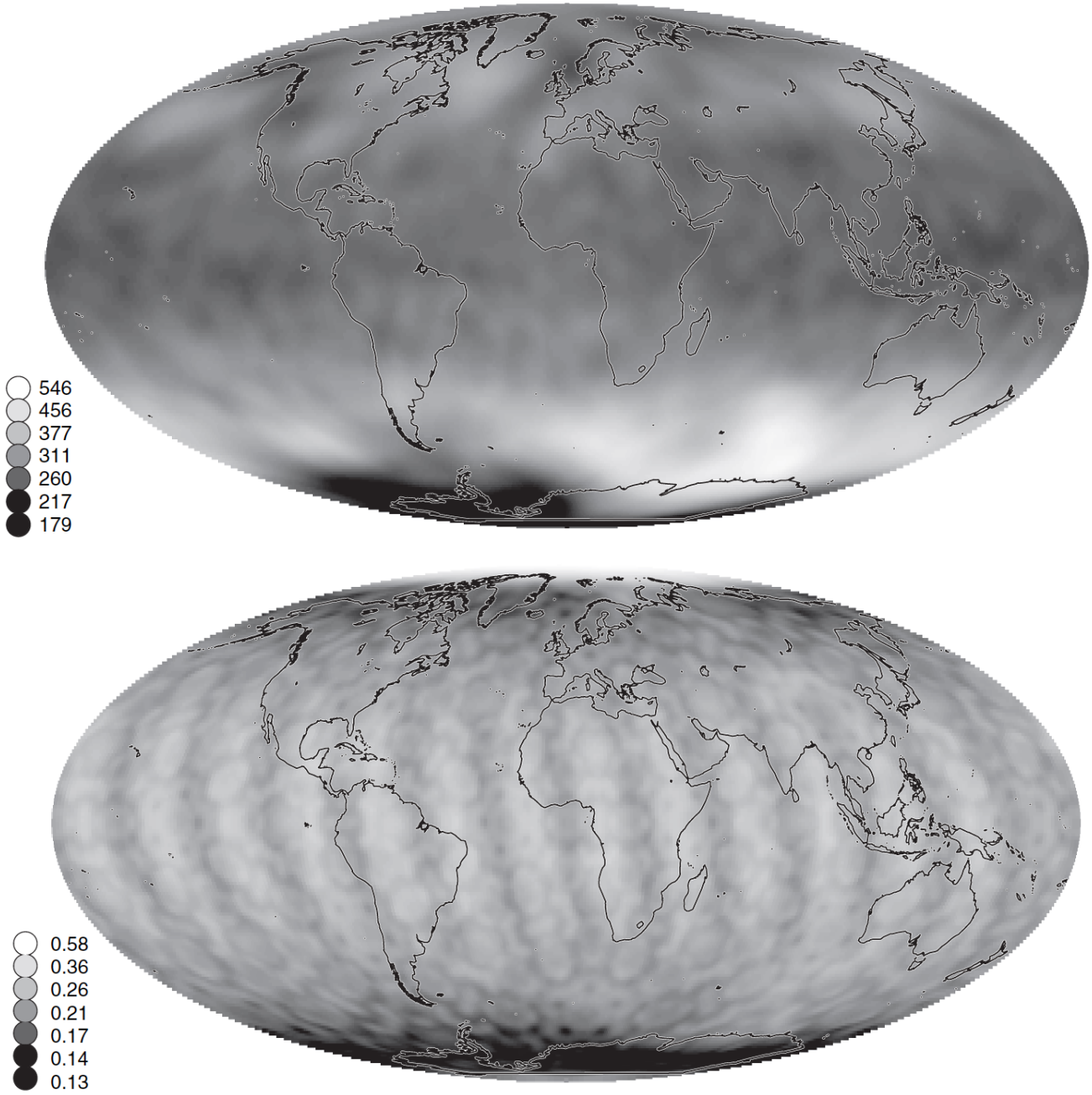


Figure: courtesy of Cressie & Johannesson 08

FRK: Predictions and Their Associated Uncertainties



Outline

Gaussian Predictive Process (Banerjee et al. 08)

J. R. Statist. Soc. B (2008)
70, *Part* 4, *pp.* 825–848

Gaussian predictive process models for large spatial data sets

Sudipto Banerjee,

University of Minnesota, Minneapolis, USA

Alan E. Gelfand,

Duke University, Durham, USA

Andrew O. Finley

Michigan State University, East Lansing, USA

and Huiyan Sang

Duke University, Durham, USA

[Received April 2007. Final revision February 2008]

Summary. With scientific data available at geocoded locations, investigators are increasingly turning to spatial process models for carrying out statistical inference. Over the last decade, hierarchical models implemented through Markov chain Monte Carlo methods have become especially popular for spatial modelling, given their flexibility and power to fit models that would be infeasible with classical methods as well as their avoidance of possibly inappropriate asymptotics. However, fitting hierarchical spatial models often involves expensive matrix decompositions whose computational complexity increases in cubic order with the number of spatial locations, rendering such models infeasible for large spatial data sets. This computational burden is exacerbated in multivariate settings with several spatially dependent response variables. It is also aggravated when data are collected at frequent time points and spatiotemporal process models are used. With regard to this challenge, our contribution is to work with what we call predictive process models for spatial and spatiotemporal data. Every spatial (or spatiotemporal) process induces a predictive process model (in fact, arbitrarily many of them). The latter models project process realizations of the former to a lower dimensional subspace, thereby reducing the computational burden. Hence, we achieve the flexibility to accommodate non-stationary, non-Gaussian, possibly multivariate, possibly spatiotemporal processes in the context of large data sets. We discuss attractive theoretical properties of these predictive processes. We also provide a computational template encompassing these diverse settings. Finally, we illustrate the approach with simulated and real data sets.

Gaussian Predictive Process (Banerjee et al. 08)

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \eta(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}), \quad \eta \sim \mathrm{GP}(0, C(\cdot,\cdot;\boldsymbol{\theta})$$

Reasoning: Given $\boldsymbol{Z}$, we seek a linear combination $a(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{Z}$ such that

$$\int_{\mathcal{S}} \mathbb{E}\left[\eta\left(\boldsymbol{s}\right) - a\left(\boldsymbol{s}\right)^{\mathrm{T}}\boldsymbol{Z}\right]^2 d\boldsymbol{s}$$

is minimized.

In predictive process one let $\boldsymbol{Z} = \boldsymbol{\eta}^* = \{\eta(\boldsymbol{s}_i^*)\}_{i=1}^r \sim N_r(\boldsymbol{0}, C^*(\boldsymbol{\theta}))$ where $C^*(\boldsymbol{\theta}) = \left[C(s_i^*, s_j^*; \boldsymbol{\theta})\right]_{i,j=1}^r$ and

$$\boldsymbol{A} = \left[C(\boldsymbol{s}_i, \boldsymbol{s}_j^*; \boldsymbol{\theta})\right]_{i=1,\cdots,n}^{j=1,\cdots,r} [C^*]^{-1}$$

$$\Rightarrow C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C^{\mathrm{T}}(\boldsymbol{s}_1, \boldsymbol{s}^*)\left[C^*(\boldsymbol{\theta})\right]^{-1} C(\boldsymbol{s}_2, \boldsymbol{s}^*), \qquad \boldsymbol{s}^* = (\boldsymbol{s}_1^*, \cdots, \boldsymbol{s}_r^*)^{\mathrm{T}}$$

Covaraince Function Approximation

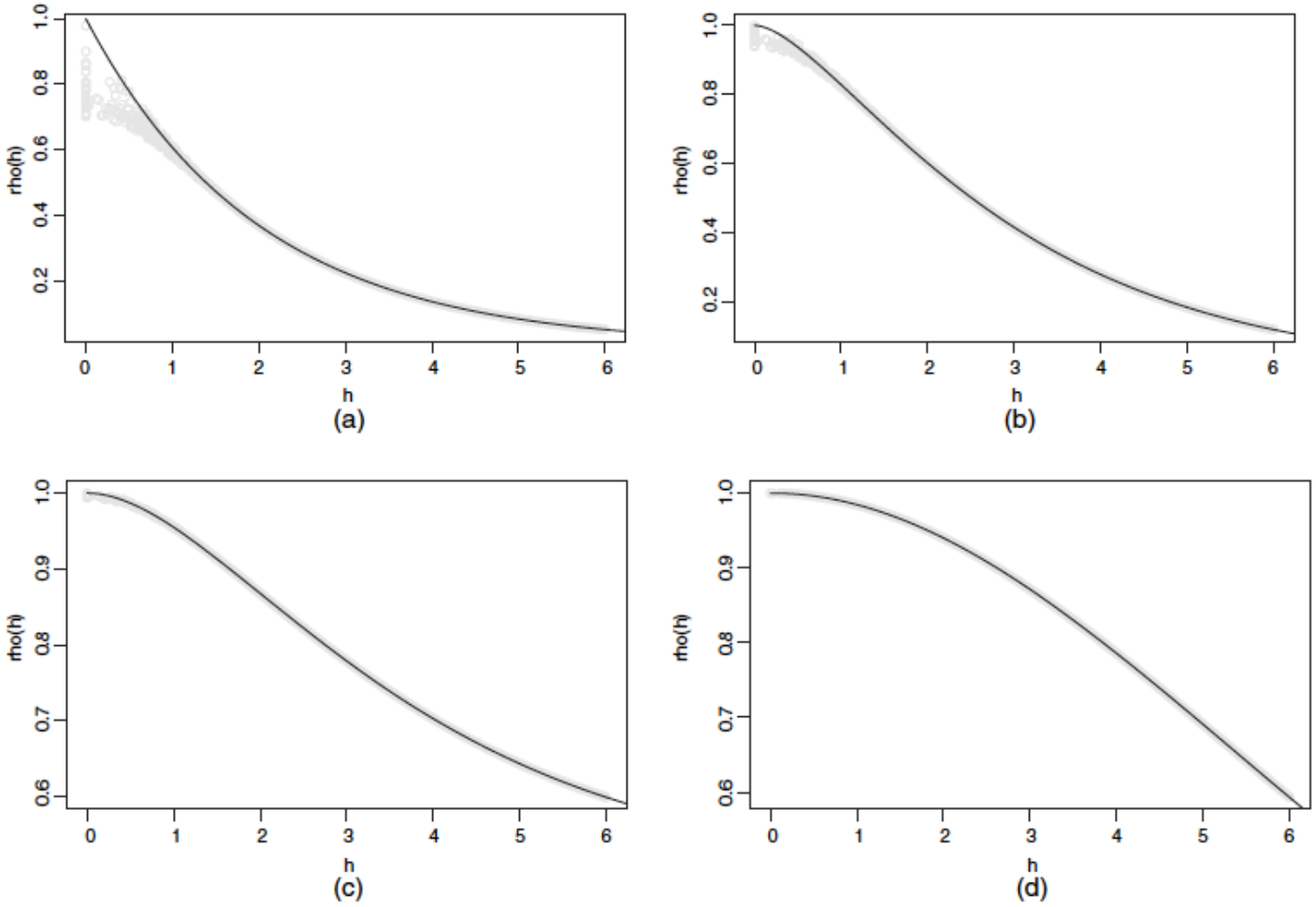


Fig. 1. Covariances of $w(\mathbf{s})$ against distance (———) and covariances of $\tilde{w}(\mathbf{s})$ against distance (●): (a) smoothness parameter 0.5; (b) smoothness parameter 1; (c) smoothness parameter 1.5; (d) smoothness parameter 5

Figure: courtesy of Banerjee et al. 08

Covaraince Function Approximation Cont'd

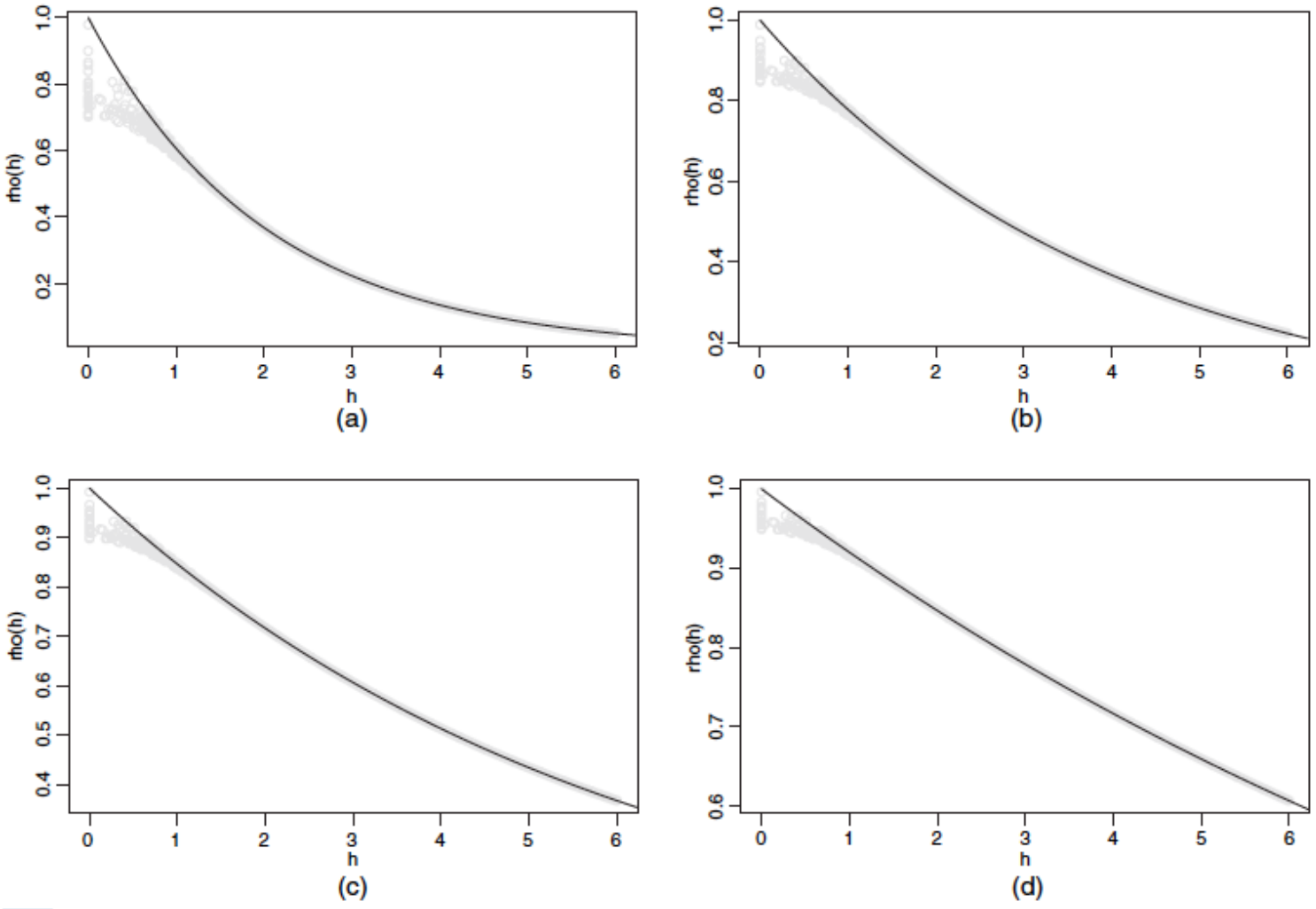


Fig. 2. Covariances of $w(\mathbf{s})$ against distance (———) and covariances of $\tilde{w}(\mathbf{s})$ against distance: (a) range parameter 2; (b) range parameter 4; (c) range parameter 6; (d) range parameter 12

Figure: courtesy of Banerjee et al. 08

Selection of Spatial Knots

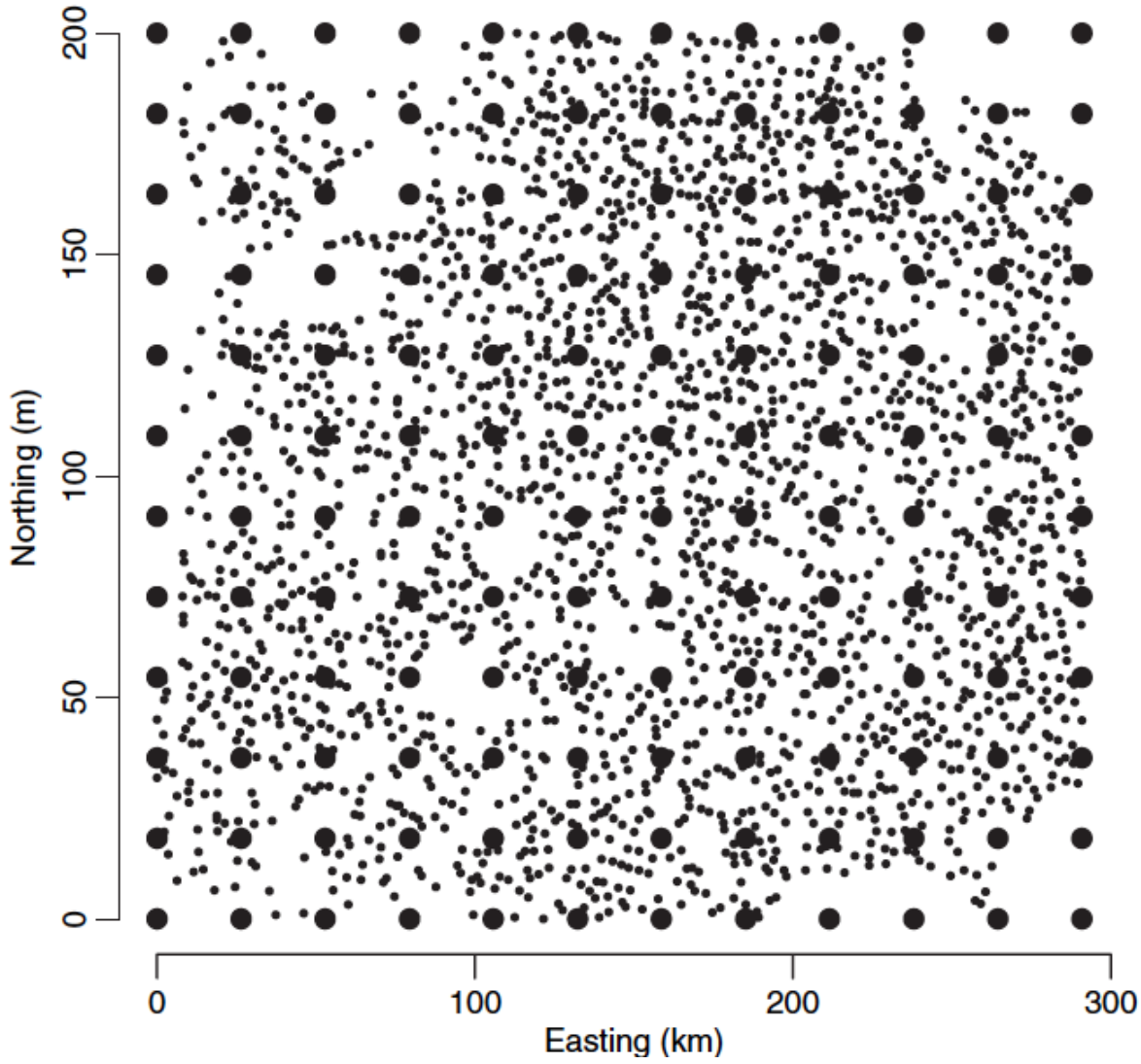


Figure: courtesy of Banerjee et al. 08

Recap: Low-Rank Approximation

- Use

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{r} a_j(\boldsymbol{s}) Z_j + \varepsilon(\boldsymbol{s})$$

  to approximate

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})^{\mathrm{T}}\boldsymbol{\beta} + \eta(\boldsymbol{s}), \qquad \forall \boldsymbol{s} \in \mathcal{S}$$

- Why: to gain computational efficiency
- Question: how well the low rank approach work statistically

Limitation on Low-Rank Approximation (Stein 2014)

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.

How to Improve Low-Rank Approach?

- Low rank does not perform well for short range dependence
- Recall the low rank representation:

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\xi}$$

  where $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ is assumed to be diagonal

- Sang and Huang (2012): replace $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ by a sensible spatial model while keep the computation tractable $\Rightarrow$ covariance tapering.

Outline

Fixed Rank Kriging

Gaussian Predictive Process

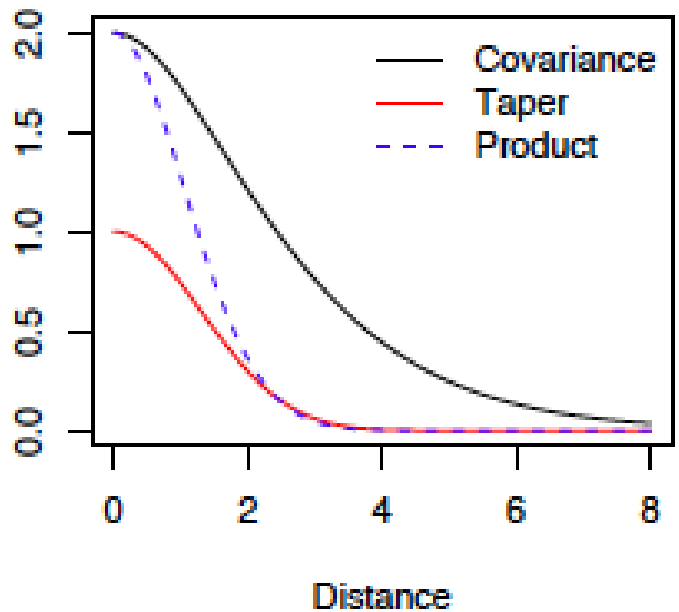
Covariance Tapering

Covariance Tapering (Furrer *et al.* 06)

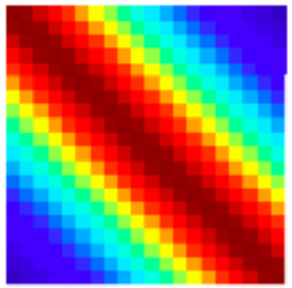
Replace the covariance function $C(h)$ by

$$C_{\text{tap}}(h;\gamma) = C(h)\rho_{\text{tap}}(h;\gamma)$$

where $\rho_{\text{tap}}(h;\gamma)$ is an isotropic correlation function with compact support ($\rho_{\text{tap}}(h) = 0$ if $h \geq \gamma$)



Covariance Tapering Cont'd



$$\begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix} \circ \begin{bmatrix} T_{11} & \cdots & T_{1n} \\ \vdots & \ddots & \vdots \\ T_{n1} & \cdots & T_{nn} \end{bmatrix} = \begin{bmatrix} \Sigma_{11}T_{11} & \cdots & \Sigma_{1n}T_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1}T_{n1} & \cdots & \Sigma_{nn}T_{nn} \end{bmatrix}$$

- $C_{\mathsf{tap}}(h)$ is a valid covariance function
- Sparse matrix algorithm can be used

A Case Study Competition Among Methods for Analyzing Large Spatial Data

Matthew J. HEATON, Abhirup DATTA, Andrew O. FINLEY, Reinhard FURRER, Joseph GUINNESS, Rajarshi GUHANIYOGI, Florian GERBER, Robert B. GRAMACY, Dorit HAMMERLING, Matthias KATZFUSS, Finn LINDGREN, Douglas W. NYCHKA, Furong SUN, and Andrew ZAMMIT-MANGION

The Gaussian process is an indispensable tool for spatial data analysts. The onset of the "big data" era, however, has lead to the traditional Gaussian process being computationally infeasible for modern spatial data. As such, various alternatives to the full Gaussian process that are more amenable to handling big spatial data have been proposed. These modern methods often exploit low-rank structures and/or multi-core and multi-threaded computing environments to facilitate computation. This study provides, first, an introductory overview of several methods for analyzing large spatial data. Second, this study describes the results of a predictive competition among the described methods as implemented by different groups with strong expertise in the methodology. Specifically, each research group was provided with two training datasets (one simulated and one observed) along with a set of prediction locations. Each group then wrote their own implementation of their method to produce predictions at the given location and each was subsequently run on a common computing environment. The methods were then compared in terms of various predictive diagnostics.

Supplementary materials regarding implementation details of the methods and code are available for this article online.

Key Words: Big data; Gaussian process; Parallel computing; Low-rank approximation.