Computational Statistics and Data Analysis 104 (2016) 233–246

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Modeling nonstationary covariance function with convolution on sphere

Yang Li [a,*], Zhengyuan Zhu [b]

[a] Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth MN 55812, United States
[b] Department of Statistics, Iowa State University, Ames IA 50011, United States

ARTICLE INFO

Article history:
Received 10 April 2015
Received in revised form 21 June 2016
Accepted 4 July 2016
Available online 19 July 2016

Keywords:
Spatial statistics
Kernel convolution
Axial symmetry
Nonstationary

ABSTRACT

The wide use of satellite-based instruments provides measurements in climatology on a global scale, which often have nonstationary covariance structure. The issue of modeling a spatial random fields on sphere which is stationary across longitudes is addressed with a kernel convolution approach. The observed random field is generated by convolving a latent uncorrelated random field with a class of Matérn type kernel functions. By allowing the parameters in the kernel functions to vary with locations, it is possible to generate a flexible class of covariance functions and capture the nonstationary properties. Since the corresponding covariance functions generally do not have a closed form, numerical evaluations are necessary and a pre-computation table is used to speed up the computation. For regular grid data on sphere, the circulant block property of the covariance matrix enables us to use Fast Fourier Transform (FFT) to get its determinant and inverse matrix efficiently. The proposed approach is applied to the Total Ozone Mapping Spectrometer (TOMS) data for illustration.

1. Introduction

The need to model a large-scale spatial data has been increasing in the past decades. Due to the wide use of high-tech instruments and accumulation of observed data over time, it is not uncommon to have large data sets which have nonstationary dependence structure, especially for global data. As an example, the Level 3 data set from the Total Ozone Mapping Spectrometer (TOMS), a satellite measurement on the global ozone level, has 28,800 daily observations around the globe and the spatial structure is far from being stationary (Cressie and Johannesson, 2008; Jun and Stein, 2008; Stein, 2008).

Statisticians have recognized the necessity to model nonstationary spatial random processes and have proposed different methodologies. Haas used lognormal and moving window methods to model acid deposition, where only the data in a local window were used in both estimation and prediction (Haas, 1990). Sampson and Guttorp used a smooth deformation of the spatial space, which is equivalent to a nonlinear transformation, to generate nonstationarity (Sampson and Guttorp, 1992). Paciorek and Schervish introduced a class of nonstationary covariance functions with closed forms (Paciorek and Schervish, 2006). In spectral domain, Fuentes proposed a method where the random field is represented locally as stationary and isotropic, but allowing the parameters to vary across space (Fuentes, 2001, 2002, 2005).

* Corresponding author.
E-mail addresses: yangli@d.umn.edu (Y. Li), zhuz@iastate.edu (Z. Zhu).

http://dx.doi.org/10.1016/j.csda.2016.07.001
0167-9473/

Another approach to model nonstationarity is the process convolution approach introduced by Higdon et al. (1999). In this approach, the random process $Z(\mathbf{s})$ is defined as a kernel convolution of the underlying excitation field as

$$Z(\mathbf{s}) = \int_{\mathbb{R}^d} k(\mathbf{s}-\mathbf{u};\, \eta_{\mathbf{s}})X(\mathbf{u})d\mathbf{u}, \tag{1}$$

where $X(\mathbf{u})$ is an independent random process, and $k(\mathbf{s};\, \eta_{\mathbf{s}})$ is a nonrandom, square-integrable kernel function with $\eta_{\mathbf{s}}$ being the parameters at location $\mathbf{s}$. It is easy to see that $Z(\mathbf{s})$ has a constant mean zero and its covariance function $C(\mathbf{u},\mathbf{v}) \equiv \mathrm{Cov}(Z(\mathbf{u}), Z(\mathbf{v}))$ is

$$C(\mathbf{u},\mathbf{v}) = \int_{\mathbb{R}^d} k(\mathbf{u}-\mathbf{w};\, \eta_{\mathbf{u}})k(\mathbf{v}-\mathbf{w};\, \eta_{\mathbf{v}})d\mathbf{w}. \tag{2}$$

If $\eta_{\mathbf{s}} = \eta$ is a constant across all locations, the random process $Z(\mathbf{s})$ generated by (1) is stationary and the covariance function in (2) only depends on the difference $\mathbf{u}-\mathbf{v}$. By allowing $\eta_{\mathbf{s}}$ to vary at different locations, it is possible to generate a nonstationary field on $\mathbb{R}^d$. Convolution-based methods have the appealing features in nonparametric modeling since one only needs to model the smoothing kernel $k(\cdot)$ instead of the covariance function which is restricted to be non-negative definite. It is also not difficult to augment the space with time so that the kernel function and the excitation field are both spatio-temporally related.

The choice of the kernel function is important in the modeling since it controls the properties of the resulting covariance structure, including the range, variance, and smoothness. An intuitive choice would be the Gaussian kernel (Higdon et al., 1999). It has the advantage of being evaluated analytically since the covariance function also has an exponential quadratic form. However, as described in Stein (1999), the exponential quadratic covariance function is infinitely differentiable which may not be realistic for physical processes. It is known that Matérn covariance function has a smoothness parameter which can be estimated from the data. It has been shown that if the kernel function is chosen to be a modified Bessel function as Zhu and Wu (2010) and Xia and Gelfand (2006)

$$k(x;\sigma,\kappa,\nu) = \frac{2\Gamma(\nu+d/2)^{1/2}\nu^{\nu/4+d/8}\sigma^{1/2}|x|^{\nu/2-d/4}}{\pi^{d/4}\Gamma(\nu/2+d/4)\Gamma(\nu)^{1/2}\kappa^{\nu/2+d/4}} K_{\nu/2-d/4}\left(\frac{2\nu^{1/2}|x|}{\kappa}\right), \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is the modified Bessel function of the second kind with order ν (Abramowitz and Stegun, 1965), the corresponding covariance function takes the familiar Matérn form

$$C(u;\sigma,\kappa,\nu) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\nu^{1/2}u}{\kappa}\right)^{\nu} K_\nu\left(\frac{2\nu^{1/2}u}{\kappa}\right). \tag{4}$$

All the methods mentioned up to now are for Euclidean space $\mathbb{R}^d$ where the covariance structure is based on the Euclidean distance. Recently, however, more and more large-scale data in climatology and environmental science are collected where the curvature of the Earth cannot be simply neglected. The aforementioned Level 3 TOMS data are observed globally along satellite tracks. It is apparently not appropriate to use Euclidean distance if the two locations are far apart from each other.

Some analysis methods designed specifically to handle global data, such as the TOMS data, are already available. Cressie and Johannesson expressed the covariance matrix in terms of a diagonal matrix plus a fixed low rank matrix, which makes it possible to compute the likelihood function exactly with massive spatial data (Cressie and Johannesson, 2008). Stein further replaced the diagonal matrix with a sparse matrix hoping to capture both the small-scale and large-scale spatial dependence structures (Stein, 2008). Stein also discussed the limitations of using reduced rank methodology to reproduce the high-frequency spatial structure in the data (Stein, 2014). Moreover, Jun and Stein proposed an approach to producing space–time covariance functions on sphere by applying differential operators to fully symmetric processes (Jun and Stein, 2007). In this way, nonstationary spatial random fields can be produced with a closed form on sphere and time. They applied this method to the analysis of TOMS data (Jun and Stein, 2008). With the aid of Discrete Fourier Transform (DFT), they were able to calculate the exact likelihood for large data sets on regular grids.

Spatial random fields observed within a local region can often be approximated to be stationary or isotropic. On the other hand, large scale or global processes usually show the pattern of nonstationarity since the factors driving the characteristics of the random field typically vary at different locations. A special kind of nonstationarity is the axial symmetry as described in Jones (1962). For an axially symmetric process, the first two moments are invariant to rotations with respect to the Earth's axis. Their covariance function depends on longitude only through their difference. Jun and Stein applied this approach to model the ozone data described above on a global scale, where they consider the axially symmetric process by applying differential operators to an isotropic process (Stein, 2007; Jun and Stein, 2007, 2008). The parameters of the random field at each latitude are homogeneous.

Castruccio and Stein defined a spectral model at each latitude and utilized a coherence model across latitudes which leads to a computationally efficient estimation procedure (Castruccio and Stein, 2013). Castruccio and Genton introduced a flexible class of models by relaxing the assumption of longitudinal stationarity in the context of regularly gridded climate model output (Castruccio and Genton, 2015). They also proposed a method of compressing the ensemble and was able to fit a non-trivial model to a data set of one billion data points (Castruccio and Genton, 2016).