Computational Statistics and Data Analysis 59 (2013) 82–94

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination

Oleksandr Gromenko[a,*], Piotr Kokoszka[b]

[a] Department of Mathematics and Statistics, Utah State University, Logan, UT 84322, United States

[b] Department of Statistics, Colorado State University, Fort Collins, CO 80523, United States

ARTICLE INFO

Article history:
Received 18 April 2012
Received in revised form 26 July 2012
Accepted 27 September 2012
Available online 4 October 2012

Keywords:
Functional data
Ionosphere
Long-term trend
Nonparametric inference
Spatial statistics

ABSTRACT

This paper is concerned with estimation and testing in data sets consisting of a small number (about 20–30) of curves observed at unevenly distributed spatial locations. Such data structures may be referred to as spatially indexed functional data. Motivated by an important space physics problem, we model such data as a mean function plus spatially dependent error functions. Given a small number of spatial locations, the parametric methods for the estimation of the spatial covariance structure of the error functions are not satisfactory. We propose a fully nonparametric estimator for the mean function. We also derive a test to determine the significance of the regression coefficients if the mean function is a linear combination of known covariate functions. In particular, we develop methodology for the estimation a trend in spatially indexed functional data, and for assessing its statistical significance. We apply the new tools to global ionosonde records to test the hypothesis of ionospheric cooling. Nonparametric modeling of the space–time covariances is surprisingly simple, much faster than those previously proposed, and less sensitive to computational errors. In simulated data, the new estimator and test uniformly dominate those based on parametric modeling.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Models for data which exhibit both space and time dependence have attracted increasing attention in geophysical and environmental research. This is a fast growing branch of statistics, for a general overview see Cressie and Wikle (2011) and Sherman (2011), for a fast, accessible introduction, we recommend (Gneiting et al., 2007). Space–time data could be roughly separated into several categories according to the amount of information contained, respectively, in their spatial and temporal components. One category is the data which have a very rich spatial component and relatively limited temporal component. Such data usually come from satellites, see e.g. Jun and Stein (2009), Cressie et al. (2010) and Katzfuss and Cressie (2011), among many others. Another category is data which have a rich temporal component and a relatively limited spatial component. Such data come typically as collections of long time series recorded at different spatial locations by ground based instruments. For example, the Irish wind data studied by Haslett and Raftery (1989) and consequently used in many other papers, the Canadian weather data extensively used in Ramsay and Silverman (2005) and Ramsay et al. (2009), pollution data studied by Bowman et al. (2009), and many others.

In this paper, we propose a flexible, fully nonparametric methodology for data of the latter type. It includes estimation of the mean function and is applied to testing the statistical significance of a linear trend. Our methodology builds on the theory of Hall et al. (1994) and Hall and Patil (1994) by (1) developing a practically applicable tool set for the estimation and

* Corresponding author. Tel.: +1 3152625416.
E-mail addresses: agromenko@gmail.com (O. Gromenko), Piotr.Kokoszka@colostate.edu (P. Kokoszka).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.09.016

testing in the spatial context with few data locations, (2) extending it to the framework of spatially indexed functional data, (3) developing suitable confidence bounds, and (4) applying it to an important space physics problem. The work presented in this paper is a direct result of our attempts to solve this important space physics problem in a fairly conclusive manner that would be satisfactory to the space physics community. Since the problem concerns the detection of a long term (many decades) trend, we hope that out methodology is general and useful enough to be applicable to other similar data sets and problems. Spatially indexed functional data have been the focus of several recent studies, see Delicado et al. (2010); Giraldo et al. (2011), Nerini et al. (2010), Gromenko et al. (2012) and Gromenko and Kokoszka (forthcoming). Existing approaches however often fail when the number of spatial locations is small because in such cases the numerical optimization required to fit a parametric spatial model may fail, or the fit may be poor. The research we report is, to a large extent, a result of computational difficulties we encountered with standard approaches. The resulting new methodology is computationally faster and the algorithms never fail to converge (in our data sets and simulations).

The paper is organized as follows. In Section 2, we develop a nonparametric covariance estimation procedure for scalar data. Next, in Section 3, a statistical model for spatially indexed functional data is introduced. Section 4 presents the estimation procedure for this model. In Section 5, we derive a test for assessing the significance of regression coefficients when the mean function is a linear combination of known covariate functions. The application of this test to the assessment of a long term cooling trend in the ionosphere is presented in Section 6. Section 7 presents the results of simulation studies that validate the methodology we propose and its application to the ionosonde data.

2. Description of the method for scalar data

In this section, we assume that ζ is a mean zero stationary and isotropic *scalar* random field observed at locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N$, and $\boldsymbol{\Gamma}$ is the $N \times N$ matrix of covariances

$$\gamma(d_{k\ell}) = \mathrm{Cov}(\zeta(\mathbf{s}_k), \zeta(\mathbf{s}_\ell)) = E[\zeta(\mathbf{s}_k)\zeta(\mathbf{s}_\ell)],$$

where $d_{k\ell}$ is the distance between $\mathbf{s}_k$ and $\mathbf{s}_\ell$. Estimation of $\boldsymbol{\Gamma}$ is not trivial for small samples. A standard variogram based estimator for small spatial data sets is generally unstable, and the the optimization often fails to converge. It is recommended that every lag interval should contain at least 30 distinct distances, but for small sample sizes, it is difficult to meet this condition without reducing the number of intervals to a level which makes fitting a parametric model difficult. We therefore develop nonparametric methodology, based on the work of Hall et al. (1994) and Hall and Patil (1994), which is suitable for small data sets. It forms the basis of the estimation and testing procedures for functional spatially indexed data, but can also be used for different spatio-temporal models, as illustrated in Example 7.1.

Recall that $d_{k\ell}$ is the distance between $\mathbf{s}_k$ and $\mathbf{s}_\ell$, and consider the preliminary estimator

$$\tilde{\gamma}(d_{k\ell}) = \zeta(\mathbf{s}_k)\zeta(\mathbf{s}_\ell). \tag{1}$$

It is possible that for some distances there exist several distinct estimators $\tilde{\gamma}(d_{k\ell})$, in fact for $d_{k\ell} = 0$ there are always N different preliminary estimators. The estimated covariances are ordered so that the corresponding distances do not decrease: denoting the $d_{k\ell}$ by d_i, we thus have $d_i \le d_{i+1}$, $1 \le i \le N(N+1)/2$. The resulting sequence $\{\tilde{\gamma}(d_i) : 1 \le i \le N(N+1)/2\}$ is very noisy and must be smoothed. We use local linear regression, see Fan and Gijbels (1996), rather than the kernel smoother suggested by Hall et al. (1994). The reason for using the local linear regression is that it introduces a slightly smaller bias for small and large distances d_i. Let $\kappa(x)$ be a compactly supported symmetric probability density function. The smoothed value of $\gamma(d)$ is thus estimated by $\hat{m}(d)$ computed by minimizing

$$(\hat{m}(d), \hat{m}_1) = \arg\min_{m, m_1} \sum_{i=1}^{N(N+1)/2} \kappa\left(\frac{d - d_i}{h}\right) \left\{\tilde{\gamma}(d_i) - m(d) - m_1(d - d_i)\right\}^2 . \tag{2}$$

We performed simulations using several popular kernels (triangular, quadratic, Epanechnikov, triweight, tricube), and found that they produce practically the same estimates. The results reported in this paper are based on the Epanechnikov kernel. As with all problems of this type, the most difficult issue is the selection of the bandwidth h; Hall et al. (1994) do not recommend any specific procedure. They developed an interactive software which allows the user to choose the bandwidth and visually compare the resulting estimates. We describe our method of bandwidth selection in the Appendix.

To construct a positive definite covariance function, we use Bochner's theorem: We compute the Fourier transform of $\hat{m}$ and delete all negative lobes. The inverse Fourier transform is then our final estimator $\hat{\gamma}(d)$. We enhance the idea of Hall et al. (1994) by providing a procedure to construct functional confidence intervals for $\hat{\gamma}(\cdot)$, see the Appendix. The application of the procedure to simulated data is illustrated in Fig. 1.

Hall et al. (1994) showed that to achieve consistency in the estimation of $\gamma(d)$, the distance between $\min(d_i)$ and $\max(d_i)$ (the range) must grow much slower than the number of the d_i. This condition is naturally satisfied in the spatial setting because adding one more $\mathbf{s}_k$ roughly increases the range at most by a unit, but increases the number of the d_i by N.

Download English Version:

https://daneshyari.com/en/article/416828

Download Persian Version:

https://daneshyari.com/article/416828

Daneshyari.com