Spatial Statistics 21 (2017) 42–65

Contents lists available at ScienceDirect

Spatial Statistics

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

Multivariate bias adjusted tapered predictive process models

Rajarshi Guhaniyogi

Department of Applied Mathematics & Statistics, University of California Santa Cruz, SOE 2, 1156 High Street, Santa Cruz, CA 95064, USA

ARTICLE INFO

Article history:
Received 13 May 2016
Accepted 10 April 2017
Available online 31 May 2017

Keywords:
Bayesian inference
Covariance-tapering
Low rank models
Mean square differentiability
Predictive process
Spatial smoothness

ABSTRACT

We extend prior work on multivariate "low-rank" methods for the analysis of large multivariate spatial datasets. "Low-rank" methods usually operate on lower-dimensional subspaces and induce biases in the residual variance component as a result of over-smoothing or model mis-specification. Our current work attempts to characterize these biases, demonstrates their presence as a systemic phenomena, and explores remedial models without incurring computational costs. Our methodological contribution lies in the development of the multivariate tapered predictive process model that accounts for spatial correlations among multivariate components by the recently proposed *multivariate matern* correlation kernel. Both the proposed framework and the multivariate tapered predictive process model using linear model co-regionalization (LMC) (Sang et al., 2011) have been found to rectify bias in parameter estimation. We also prove novel theoretical results comparing *smoothness* properties of multivariate tapered predictive process models and classes of low rank models, including predictive processes. Finally, we illustrate our work using synthetic experiments as well as an application to forestry.

1. Introduction

With the advent and expansion of Geographical Information Systems (GIS) along with related software, statisticians today routinely encounter large spatial or spatiotemporal datasets containing multiple variables observed across thousands of locations. This has, in turn, generated considerable interest in statistical modeling for location-referenced spatial data; see, for example, the books

E-mail address: rguhaniy@ucsc.edu.

http://dx.doi.org/10.1016/j.spasta.2017.04.008
2211-6753/

by Cressie (1993), Banerjee et al. (2014) and Schabenberger and Gotway (2004) for a variety of methods and applications. Fitting Bayesian hierarchical spatial models to these datasets involves matrix decomposition of complexity n^3 with number of locations n for a univariate outcome. This makes computation of these models infeasible for large datasets. In popular spatial literature, this problem is generally referred to as the "big-n" problem. Evidently, multivariate and spatiotemporal processes exacerbate this problem.

Low rank or reduced rank models have become popular for analyzing large spatial and spatiotemporal datasets (see Higdon et al., 2002; Stein, 2008; Cressie and Johannesson, 2008; Banerjee et al., 2008). The basic idea behind low rank models is to set a few "basis functions" in space, usually taken to be much smaller in number compared to the number of data locations, and to express the spatial process realizations over the entire set of observed locations in terms of only the few basis functions and the associated coefficients. Depending on various choices of basis functions, a wide class of low rank models have emerged in the recent past. In what follows, we will focus mostly on one specific class of low rank models, known as the predictive process model. In predictive process, one considers a set of locations in the spatial domain, or "knots", and constructs basis functions based on these knots. In this process, one must ensure that the spatial information available from the entire set of locations can be summarized with the set of knots allowing some acceptable loss of information. This is achieved by pursuing a rich and flexible framework that integrates knot selection into the modeling (Guhaniyogi et al., 2011).

Low rank spatial models have been widely deployed in the environmental and natural sciences to develop highly competent inferential frameworks for large spatial databases (Banerjee et al., 2008; Cressie and Johannesson, 2008). However, Stein (2014) reports potential problems in prediction and inference arising from low-rank models. More specifically, Finley et al. (2009) demonstrate how the *predictive process* yields biased estimates of certain variance components in spatial progeny trials and develop one remedy. Though multiple articles have been written on univariate spatial low-rank models, multivariate low rank models have found relatively less emphasis.

This manuscript embarks upon characterizing and understanding biases in multivariate low rank models. Assuming the geostatistical model with a Gaussian process prior on the spatial components as the "gold standard", we intend to investigate how low rank models approximate the "gold standard" in terms of spatial surface recovery and parameter estimation. In particular, it is observed that low rank models underestimate spatial variability and overestimate noise variability, thereby yielding "smoother" spatial surfaces. Such inaccuracies in terms of parameter estimation that result in smoothed surface estimation are referred to as *bias* in this article. We show their presence in low rank models as a systematic phenomenon, discuss their potential impact on spatial inference and explore remedies applicable to a wide range of hierarchical multivariate low-rank spatial process models. Our methodological contribution lies in proposing a multivariate extension of the univariate tapered predictive process (Sang and Huang, 2012) that induces correlations among components using the multivariate Matern kernel. We compare our proposed approach with the multivariate tapered predictive process model based on the linear model co-regionalization (LMC) (Sang et al., 2011) and find that both of these are satisfactory tools for bias adjustment, with mixed relative performance. This article also highlights some of the desirable theoretical features of the multivariate tapered predictive process. In particular, we show that the class of low rank processes (including the predictive process) is infinitely mean square differentiable while the tapered predictive process is mean square differentiable up to a certain order depending upon the tapering correlation kernel and the parent Gaussian process. This is a novel theoretical result that concurs with practical findings on the local behavior ("smoothness") of estimated spatial surfaces from these models. Such observations deem the choice of multivariate tapered predictive process attractive as a bias adjustment tool for multivariate spatial applications.

The remainder of the article evolves as follows. Section 2 discusses low-rank spatial modeling in general, while Section 3 discusses how hierarchical Gaussian predictive process models help in quantifying bias in residual variability. Section 4 discusses multivariate tapered predictive process models to remedy such biases and also offers some theoretical comparisons. Section 5 talks about estimation and inference. Section 6 illustrates the different bias-adjusted models using two simulation studies, followed by a forestry application. Finally, Section 7 concludes the paper with some discussion and general conclusions.