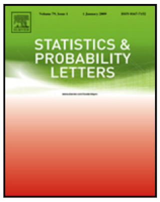
Contents lists available at ScienceDirect

Statistics and Probability Letters

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



Principles for statistical inference on big spatio-temporal data from climate models

Stefano Castruccio[a], Marc G. Genton[b,*]

[a] Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, 153 Hurley Hall, Notre Dame, USA
[b] Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Big Data
Climate model
Computational statistics
Spatio-temporal model

ABSTRACT

The vast increase in size of modern spatio-temporal data sets has prompted statisticians working in environmental applications to develop new and efficient methodologies that are still able to achieve inference for nontrivial models within an affordable time. Climate model outputs push the limits of inference for Gaussian processes, as their size can easily be larger than 10 billion data points. Drawing from our experience in a set of previous work, we provide three principles for the statistical analysis of such large data sets that leverage recent methodological and computational advances. These principles emphasize the need of embedding distributed and parallel computing in the inferential process.

1. Introduction

Data indexed in space and time for environmental applications have been greatly affected by the Big Data revolution. In particular, the increases in Volume, Variety, and Velocity (the three vs. of Big Data) have prompted statisticians working with spatio-temporal models to seek new methodologies and inferential approaches that combine flexibility with feasibility, and that leverage the latest advances in hardware and computer science.

In the case of Gaussian data in space and time, it is well known that a likelihood for a data set of size N requires $O(N^2)$ entries to store the covariance matrix and $O(N^3)$ flops to evaluate a log-determinant and a quadratic form. While most of the literature has traditionally focused on reducing the number of flops, the real constraint for fitting very large data sets is the storage of structured dependence in space and time. For example, storing a covariance matrix of 50,000 data points, which represents a standard-to-small data set in many environmental applications, in double precision requires $(50,000)^2 \times 8/(1024)^3 \approx 19$ Gb. Very few computers have sufficient RAM to store and perform operations with such matrices and hence to perform the linear algebra operations required to evaluate the likelihood.

Performing inference for very large data sets clearly requires more structure and assumptions on the statistical model in order to reduce the information to a more manageable scale. Low-rank methods seek to find a suitable subspace from the original space–time model where most of the information about the process is contained (Cressie and Johannesson, 2008). An important variant is the predictive process, which couples a low-rank method with a conditional approach (Banerjee et al., 2008). Other methods encourage sparsity either in the covariance matrix by tapering (Furrer et al., 2006) or in the precision matrix with Gaussian Markov random fields (Rue and Held, 2005). A powerful methodology has emerged in recent years that enforces sparsity in the precision matrix by expressing the spatio-temporal process as a solution of a stochastic partial different equation (Lindgren et al., 2011). All of the aforementioned methods allow the statistical community to still perform inference for space–time models despite the ever-increasing size of data, and hence to be able to serve practitioners despite increasing computational challenges; see the review by Sun et al. (2012) and references therein.

[*] Corresponding author.
E-mail addresses: scastruc@nd.edu (S. Castruccio), marc.genton@kaust.edu.sa (M.G. Genton).

https://doi.org/10.1016/j.spl.2018.02.026
0167-7152/© 2018 Elsevier B.V. All rights reserved.

In this work, we focus on the very end of the spectrum in terms of data size, i.e., on data sets generated from climate model ensembles, which are typically between 100 million to 10 billion points, and we provide three general principles that enable us to perform inference for nontrivial models within a reasonable time. These principles have emerged from a series of recent works on inference from extremely big spatio-temporal data (Castruccio and Stein, 2013; Castruccio and Genton, 2014; Castruccio, 2016; Castruccio and Genton, 2016; Castruccio and Guinness, 2017; Jeong et al., 2018) and ascribe to the general philosophy of the methods previously described, i.e., reduction of the information by exploiting the structure of a particular problem. These principles advocate designing a statistical model that fully leverages on parallel and high performance computing. While developed for climate model output, their reach extents well beyond this area, and they have already been applied to very large space–time data in neuroscience (Castruccio et al., 2018). We present and discuss this work in a frequentist setting, but most of the concepts can also be applied in a Bayesian setting.

The paper proceeds as follows: Section 2 presents an example of a typical data set, Section 3 introduces the three principles, and Section 4 ends with a discussion about their general applicability and limitations.

2. Big climate model output

Climate models can generate data sets of extremely large size. As an example, we take the Large ENSemble (LENS), a collection of 35 runs from the Community Atmosphere Model from the National Center for Atmospheric Research (NCAR) (Kay et al., 2015). We consider relatively low resolution in time, e.g., monthly values of some physical quantity $\mathbf{Y}_r$ for a run r, and assume that the data are on a regular $N \times M \times H$ grid, where N is the number of longitudinal bands, M is the number of latitudinal bands, H is the number of pressure levels, and the resolution is approximately 1 degree in latitude and longitude. Here, $N = 288$, $M = 192$, and $H = 17$. We consider all 35 realizations from the LENS, run under the Representative Concentration Pathways 8.5 (RCP 8.5, van Vuuren et al. (2011)) scenario with high greenhouse gas emissions, from 2006 to 2100, for a total of $K = 95 \times 12 = 1140$ months. The resulting data set is comprised of $192 \times 288 \times 1140 \times 17 \times 35 \approx 10$ billion data points. To simplify the notation in the rest of the paper, we assume that the data are only observed over latitude and longitude, though similar principles also hold for three-dimensional data (Castruccio and Genton, 2016).

3. Three principles

In this section, we detail our principles for statistical inference from large data sets. Section 3.1 introduces conditional independence across runs and its implications, Section 3.2 presents the stepwise approach for optimizing high-dimensional functions, and Section 3.3 discusses the spectral approach and its benefits in terms of computation and storage.

3.1. Conditional independence and restricted maximum likelihood

Principle 1: *When possible, use conditional independence across data sets to decouple inference for the mean and the error.* Despite the very large quantity of data, the structure of the LENS facilitates inference. Indeed, we assume that

$$\mathbf{Y}_r = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_r, \qquad \boldsymbol{\varepsilon}_r \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{1}$$

which implies that each run is independent of the others, conditional on the climate $\boldsymbol{\mu}$. The assumption of random fluctuation around a climatological mean is rooted in the deterministically chaotic nature of climate models (Lorenz, 1963). Collins (2002), Collins and Allen (2002), and Branstator and Teng (2010) discussed this assumption in different contexts. Since atmospheric processes mix efficiently, they comply well with assumption in (1); slow-mixing processes such as deep ocean temperature, on the other hand, are not guaranteed this property.

Conditional independence allows us to decouple the estimations of $\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}_r$ without incurring additional computational costs. If we assume that the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and contrasts are denoted by $\mathbf{D}_r = \mathbf{Y}_r - \bar{\mathbf{Y}}$, where $\bar{\mathbf{Y}} = 1/R\sum_{r=1}^{R}\mathbf{Y}_r$, then the negative restricted log-likelihood function can be written as follows (Castruccio and Stein, 2013):

$$2l(\boldsymbol{\theta};\mathbf{D}) = KNM(R-1)\log(2\pi) + KNM\log R + (R-1)\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \sum_{r=1}^{R} \mathbf{D}_r^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{D}_r, \tag{2}$$

which leads to a restricted maximum likelihood (REML) estimator of $\boldsymbol{\theta}$.

This expression (2) allows us to focus on the inference of $\boldsymbol{\varepsilon}_r$ without providing any (parametric or nonparametric) expression for $\boldsymbol{\mu}$. Moreover, the evaluation of the restricted log-likelihood function is no more computationally onerous than the likelihood; it requires an evaluation of the quadratic forms in $\mathbf{D}_r$ instead of $\mathbf{T}_r$, and each quadratic form can be evaluated in parallel by different cores of a workstation or cluster.

3.2. Stepwise inference

Principle 2: *Stepwise inference.* When the data set is very large and possibly indexed on a complex geometry, the simultaneous estimation of the model parameters is prohibitive. Since the model complexity generally increases with the data size (e.g., a

Please cite this article in press as: Castruccio S., Genton M.G., Principles for statistical inference on big spatio-temporal data from climate models. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.026.