Statistics and Probability Letters 136 (2018) 92–96

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 24 February 2018

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 differential 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/

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 extends 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

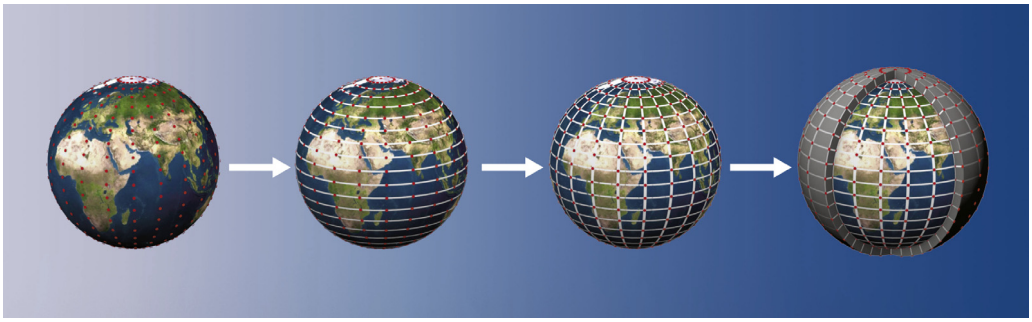

Fig. 1. Example of an inference scheme for three-dimensional global data; each step is conditional on the previous one (Castruccio and Genton, 2016). In the first step, the temporal dependence is analyzed according to (3); in the second step the longitudinal dependence; in the third step the latitudinal dependence; and in the final step, the altitudinal dependence.

higher spatial resolution requires more complex nonstationarities to be factored into the model), larger data sets necessitate not just the methods to efficiently evaluate the likelihood, but also the means to optimize it over an increased parameter space.

By sequentially (and conditionally) estimating different data subsets, we obtain an approximate global solution without processing the entire data set. The sequence of estimations is not unique and is suggested by the particular problem at hand.

For the LENS, the natural sequence is temporal, longitudinal and latitudinal dependence (then possibly altitudinal dependence). Here, we show only the temporal dependence; we defer the longitudinal and latitudinal step to the next section. We apply Principle 2 to model (3) below; the diagnostics can be found for annual wind speed in Jeong et al. (2018) and for temperature in Castruccio and Genton (2016). If the vector of all values of $\boldsymbol{\epsilon}_r$ at time t_k is denoted by $\boldsymbol{\epsilon}_r(t_k)$, then

$$\boldsymbol{\epsilon}_r(t_k) = \sum_{p=1}^{P} \boldsymbol{\Phi}_p \boldsymbol{\epsilon}_r(t_{k-p}) + \mathbf{S}\mathbf{H}_r(t_k), \tag{3}$$

where the error has a vector autoregressive, VAR(p), structure (simple diagnostics show that $p = 1, 2$ is sufficient in many cases), $\boldsymbol{\Phi}_p$, $p = 1, \ldots, P$, are matrices representing the temporal evolution, $\mathbf{S}$ is the (diagonal) matrix of standard deviations, and $\mathbf{H}_r(t_k)$ is the vector of the standardized innovations, which are assumed to be independent and identically distributed (i.i.d.) in time.

We further assume that $\boldsymbol{\Phi}_p = \text{diag}(\phi_{i;p})$, $i = 1, \ldots, NM$, $p = 1, \ldots, P$; that is, there is no cross-temporal dependence. Such an assumption implies separability in space and time, and allows the temporal dependence in model (3) to be estimated in parallel for each location. In model (3), we also assume that the temporal structure does not change across time. Such properties are reasonable for many variables in LENS at a relatively coarse resolution (monthly to yearly), but they are not suitable for analyses at fine temporal resolutions (e.g. daily), because the separability assumption would not hold.

An example of the stepwise approach for three-dimensional global data is depicted in Fig. 1. Details about this model can be found in Castruccio and Genton (2016).

3.3. Spectral methods

Principle 3: *Use spectral methods to save on storage and computation, and to build nonstationary models.* Spectral methods can provide an alternative characterization of the likelihood, i.e., a Whittle approximation (Whittle, 1951), when the process is stationary and on a regular lattice.

In our case, we assume that the process is stationary for each latitudinal band, and we model

$$H_r(L_m, \ell_n, t_k) = \sum_{c=0}^{N-1} \exp(i\ell_n c)\sqrt{f(c)}\widetilde{H}_r(c, L_m, t_k), \tag{4}$$

where $\widetilde{H}_r(c, L_m, t_k)$ is the spectral process assumed to be independent across wavenumbers, c. Hence, for the storage and computation of $f(c)$, model (4) requires $\mathrm{O}(N)$ entries and $\mathrm{O}(N\log N)$ operations, respectively, as opposed to the covariance function, which requires $\mathrm{O}(N^2)$ entries and $\mathrm{O}(N^3)$ operations, respectively. We can also generalize (4) to allow for changing behavior across domains such as land and ocean (Castruccio and Guinness, 2017) or altitude (Jeong et al., 2018). Besides diminishing the storage and computational burdens, this approach allows the inference to be performed in parallel across m, i.e., across latitude, following Principle 2. In the final step, a model joining the different bands is then provided for the latitudinal dependence (Castruccio and Stein, 2013).

While (4) predicates longitudinally-varying Fourier coefficients, this approach can be generalized to other basis decomposition. In particular, a similar approach with wavelets would be beneficial for data sets with localized behavior near sharp geographical descriptors such as land and ocean.

4. Discussion

In this work, we provided general principles to perform statistical inference on extremely big spatio-temporal climate data sets.

Principle 1 is predicated on the use of REML on the Gaussian likelihood whenever multiple, conditionally independent data sets occur. This particular setting leads to a REML expression that does not depend on the choice of contrasts and allows us to decouple the inference on the mean and the error structure according to Eq. (1). The extension of Principle 1 to non-Gaussian processes remains unclear, as one of the key assumptions is closure under linear combinations, a property enjoyed only by a small subset of distributions.

According to Principle 2, a large and complex data set such as the output from a climate model, should be divided into sequential steps, with each one comprised of inferences from smaller subsets of data that can be analyzed in parallel. This both decreases computational time and allows for an approximate optimum to be found when the parameter space is too large to be optimized all at once. However, this step-wise approach introduces the need to monitor for bias and uncertainty propagation. While examples of control over bias and uncertainty exist in some special cases (Castruccio and Guinness, 2017), a general framework has not yet been developed. Principle 2 is expected to be readily generalizable to non-Gaussian processes, as long as the parameter space can be partitioned into non-communicating subsets.

The spectral methods proposed in Principle 3 allow considerable savings in storage and computation under the assumption of stationarity, and can be used as a reference to build nonstationary models with changing spectral behavior across geographical descriptors. Spectral methods are best used for data on a regular lattice, which is typical of climate model data, but they have been used in the case of irregular designs (Fuentes, 2007), and a similar method could be used to fit data from hexagonal or pentagon-based grids available for other variables in climate models. Another method to fit data in other grid geometries would be to use kriging to regrid the points, as proposed in Horrell and Stein (2015). Hence, even if this class of models is fine-tuned for climate models with extremely large data sets, there are possible extensions to be considered for more general designs.

This class of models is considerably flexible, but a comprehensive set of comparison with other approximation methods for large data sets still needs to be performed. While a comparison with other models for global data can be found in Castruccio and Stein (2013), a full comparison with other non-global approximation methods would highlight the relative value of our approach against other currently available methods.

Acknowledgments

This publication is based upon work supported by King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No: OSR-2015-CRG4-2640. Fig. 1 was produced by Heno Hwang, scientific illustrator at King Abdullah University of Science and Technology (KAUST).

References

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (4), 825–848.

Branstator, G., Teng, H., 2010. Two limits of initial-value decadal predictability in a CGCM. J. Clim. 23 (23), 6292–6311.

Castruccio, S., 2016. Assessing the spatio-temporal structure of annual and seasonal surface temperature for CMIP5 and reanalysis. Spat. Stat. 18, 179–193.

Castruccio, S., Genton, M.G., 2014. Beyond axial symmetry: An improved class of models for global data. Stat 3 (1), 48–55.

Castruccio, S., Genton, M.G., 2016. Compressing an ensemble with statistical models: An algorithm for global 3D spatio-temporal temperature. Technometrics 58 (3), 319–328.

Castruccio, S., Guinness, J., 2017. An evolutionary spectrum approach to incorporate large-scale geographical descriptors on global processes. J. R. Stat. Soc. Ser. C. Appl. Stat. 66 (2), 329–344.

Castruccio, S., Ombao, H., Genton, M.G., 2018. A multi-resolution spatio-temporal model for brain activation and connectivity in fMRI data. Biometrics. http://dx.doi.org/10.1111/biom.12844. in press.

Castruccio, S., Stein, M.L., 2013. Global space-time models for climate ensembles. Ann. Appl. Stat. 7 (3), 1593–1611.

Collins, M., 2002. Climate predictability on interannual to decadal time scales: the initial value problem. Clim. Dynam. 19 (8), 671–692.

Collins, M., Allen, M.R., 2002. Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. J. Clim. 15 (21), 3104–3109.

Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1), 209–226.

Fuentes, M., 2007. Approximate likelihood for large irregularly spaced spatial data. J. Amer. Statist. Assoc. 477 (102), 321–331.

Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. J. Comput. Graph. Statist. 15 (3), 502–523.

Horrell, M.T., Stein, M.L., 2015. A covariance parameter estimation method for polar-orbiting satellite data. Statist. Sinica 25, 41–59.

Jeong, J., Castruccio, S., Crippa, P., Genton, M.G., 2018. Reducing storage of global wind ensembles with stochastic generators. Ann. Appl. Stat. 12, 490–509.

Kay, J.E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J.M., Bates, S.C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., Vertenstein, M., 2015. The community earth system model (CESM) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. Bull. Am. Meteorol. Soc. 96 (8), 1333–1349.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (4), 423–498.

Lorenz, E.N., 1963. Deterministic nonperiodic flow. J. Atmos. Sci. 20 (2), 130–141.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. CRC Press, Boca Raton, FL.

Sun, Y., Li, B., Genton, M.G., 2012. Geostatistics for large datasets. In: Porcu, E., Montero, J.M., Schlather, M. (Eds.), Space-Time Processes and Challenges Related to Environmental Problems, Vol. 207. Springer, pp. 55–77 (Chapter 3).

van Vuuren, D.P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G.C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S.J., Rose, S.K., 2011. The representative concentration pathways: An overview. Clim. Change 109, 5–31.

Whittle, P., 1951. Hypothesis Testing in Times Series Analysis (Ph.D. thesis), Uppsala University, Almqvist and Wiksells Boktryckeri AB.