Final report for "Scalable Statistical Analysis of Gaussian Models for Petascale Spatiotemporal Data"

Award Number: DE-SC002557

Title: Scalable Statistical Analysis of Gaussian Models for Petascale Spatiotemporal Data

Principal Investigator and co-PIs: Michael Stein (stein@galton.uchicago.edu) is the principal investigator for this grant.

Dates of Performance: 9/15/2009–9/14/2013

This grant was part of a joint project with Mihai Anitescu and Emil Constantinescu at Argonne National Laboratory, who received separate support for this work from DOE. Only those parts of the joint project in which Michael Stein was involved are reported on here.

Objectives

Gaussian process models for spatial and spatiotemporal data are ubiquitous in scientific and engineering applications. Likelihood-based methods for fitting such models, which includes all Bayesian methods, have desirable statistical properties and are thus the methods of choice when feasible. However, calculating the likelihood function even once can be a challenge for N irregularly sited observations from such a process, requiring $O(N^2)$ memory and $O(N^3)$ complexity. Thus, there is a need for fast and accurate likelihood approximation even for N on the order of 10^5, not to mention the 10^{15} implied by petascale data. The main objective of this research was to develop computational tools for the statistical analysis of massive spatiotemporal datasets based on Gaussian process models. Another important objective was the development of appropriate statistical models for specific spatiotemporal processes arising in the physical sciences.

Accomplishments

We explored the suitability of a large number of approaches to reducing computations related to the statistical analysis of Gaussian processes. Our work has led to a better understanding of the strengths and weaknesses of various existing approaches and the development of a number of new approaches. For example, we showed in [8] that low rank methods, while efficient in terms of both computation and storage, can be disastrous statistically in situations that frequently occur in practice. Furthermore, in these situations, very simple composite likelihood methods can be equally computationally efficient while producing much more accurate estimates of the parameters describing the covariance structure of the Gaussian process. In [3] and [6], we showed via both theory and numerical experiments, that stochastic approximations to the score function (the gradient of the loglikelihood function) yields procedures that are both computationally and statistically efficient.

Our work has included theoretical investigations of the properties of spatiotemporal covariance functions [1], [4]. However, even these works have a bearing on computational issues, as the types of models these papers advocate, in which spectral densities are "well-behaved" at high frequencies, are also models for which certain approximations such as composite likelihoods and fast multipole methods will perform well.

[6], [7], [9], [10] and [11] all consider applications to environmental data, with [7], [9] and [10] having the strongest applied components. [7] deals with the development of appropriate nonstationary process models and computational methods for high-frequency meteorological data taken from the DOE-run ARM SGP site. [10] also looks at ARM data, developing spectral methods for the simultaneous modeling of temperature and dew point. [9] develops an extension of composite likelihood methods that is suitable for data taken from a polar-orbiting satellite, which is a major and expanding modality for environmental monitoring.

Research contributions

Research supported by this grant has resulted in twelve papers, seven of which have been published, two more are in press and the last three have been submitted. These papers represent a range of efforts to address computational and modeling issues related to large spatiotemporal datasets.

[3], [6] and [12] describe a major focus of our research: matrix-free methods for approximating score functions with a variety of unbiased estimating equations. The first two papers make heavy use of iterative methods to solve systems of linear equations. As in many problems with iterative solvers, preconditioning is crucial to the success of these methods. In [3], results on equivalence of Gaussian measures are used to prove that simple filters can sometimes yield a preconditioner that transforms a sequence of matrices with rapidly growing condition numbers to matrices with bounded condition number. In [6], we prove that the statistical efficiency of our method is related to the condition number of the covariance matrix of the filtered observations. Thus, effective preconditioning plays a central role in both the computational and statistical properties of the stochastic score approximations. [12] describes unbiased estimating equations that do not require solving large systems of linear equations as well as leading to easier optimization problems in some circumstances. Thus, this procedure can lead to much faster parameter estimation schemes than the procedures in [3] and [6], although the losses in statistical efficiency may be greater.

[11] represents somewhat related work on likelihood methods for stationary processes observed on a partial grid. The idea here is that if this stationary process can be embedded in a periodic process, then Monte Carlo methods become natural because all likelihood computations for a periodic process on a grid can be done exactly using the fast Fourier transform. This approach is also matrix-free and was used successfully both to find maximum likelihood estimates and MCMC simulations of posterior distributions for problems for which direct methods would not be feasible. In contrast, it is not clear how our methods that only

compute the score function and not the likelihood function itself could be used in a feasible MCMC scheme.

[5] and [8] look at two popular methods for reducing computational requirements for large spatial datasets: covariance tapering and low rank methods. The common message in these papers is that they are often dominated by a very simple likelihood approximation based on dividing the data into contiguous blocks, computing the loglikelihood for the observations in each block and then adding these blockwise loglikelihoods. This simple approach would give the exact loglikelihood if observations in different blocks were independent. Despite the simplicity of this approach, for purely spatial data, it often works quite well both statistically and computationally. Its extension to spatiotemporal data is not so clear, however, due in part to the lack of commensurability of space and time.

[2], [7] and [10] consider nonstationary Gaussian processes. [2] develops an analog to nonparametric regression estimation via kernel methods for Gaussian processes that are nearly stationary over small regions. This method was motivated by a problem in cosmology in which the presence of a massive object will create small distortions in the observed cosmic microwave background. [7] considers modeling temperature data measured every minute by nonstationary Gaussian processes in time, using a spectral representation with a time-varying mixture of spectral densities. Computational innovations include accurate and fast approximations to the likelihood function for such processes and the use of genetic algorithms to handle the rather nasty optimization problem that results from this model. [10] models temperature and dew point in a way that respects the constraint that dew point cannot be higher than temperature. It then describes changes in the time-evolving bivariate spectrum of dew point and temperature in terms of time of day, variability in wind direction and a smoothed version of relative humidity. This work illustrates some of the difficulties that will arise in modeling of multivariate environmental processes, especially at high frequencies.

[9] develops an approach to likelihood approximation that may have broad scope for ungridded massive datasets. The basic idea is to estimate those parameters of the process relating to local behavior using moderately sized subsets of the data, then interpolate these observations to a sparse grid and use these gridded observations to estimate parameters related to large-scale behavior. Because these interpolated "pseudo-observations" are gridded, methods that can take advantage of this regularity can then be used to reduce computations and storage. The application (to Level 2 total column ozone levels as measured by the OMI instrument) only considers observations in a narrow latitute band, but our approach offers what I believe to be the only realistic option to fitting suitably rich spatiotemporal statistical models via likelihood-based methods to polar-orbiting satellite data at the global level.

[1] and [4] focus on theoretical aspects of spatiotemporal models. [1] studies the screening effect, which says that the conditional distribution of an observation given its nearest neighbors should be nearly independent of more distant observations. Such a property seems quite natural, but it does not always hold.

[1] proves that if the spectral density of a stationary Gaussian process is well-behaved (in the sense that it changes relatively slowly at high frequencies), then a screening effect will hold. When modeling spatiotemporal processes, it may often be important to allow for the degree of smoothness in space to differ from the degree of smoothness in time. It is not difficult to write down spectral densities for processes that allow these different degrees of smoothness and are well-behaved at high frequencies. However, it is generally then not possible to give explicit expressions for the corresponding covariance functions, limiting the applicability of these models. [4] shows that there is one fairly general class of such spectral densities for which the (generalized) covariance functions can be written in terms of H functions, which are a generalization of generalized hypergeometric functions.

Impact on DOE

This work substantially expands the scope of problems to which likelihood-based methods might reasonably be applied to the fitting of Gaussian process models. In addition to the environmental and meteorological applications considered by us, Gaussian process models can be used in engineering applications such as fluid flow in tanks and are commonly used to model the output of computer experiments (e.g., by Katrin Heitmann at ANL and David Higdon at LANL). Our work with ARM data illustrates the power of Gaussian process models and models based on Gaussian processes to capture the high-frequency behavior of these data which, we believe, will eventually lead to a better understanding of meteorological processes at fine scales. In the long run, we hope that our connection of computational problems and the theoretical properties of spatiotemporal models will lead to fundamental changes in how spatiotemporal processes are modeled and in how one addesses the computational challenges that arise in the statistical analysis of massive spatiotemporal datasets.

Presentations

Kansas State University, Department of Statistics, 2010. Invited as part of the ADVANCE Distinguished Lecture Series, which was supported by an ADVANCE grant from NSF to KSU to encourage the full participation of women in academic science and engineering.

International Chinese Statistical Association 2010 Applied Statistics Symposium, Indianapolis. Invited talk in session Spatial Statistics and Computation.

Rietz Lecture, Institute of Mathematical Statistics (IMS) Annual Meeting, Gothenburg, Sweden, 2010. This named lecture is given once every three years and is intended "to clarify the relationship of statistical methodology and analysis to other fields" (quote taken from IMS website). Only two named lectures are presented at IMS meetings in any given year. This talk turned into publication [1].

Keynote speaker, International Symposium on Statistical Analysis of Spatio-Temporal Data, Kamakura City, Japan, 2010.

Invited speaker, New Trends in Kriging, meeting sponsored by ANR Costa Brava, Toulouse, 2011.

Invited speaker, Modélisation pour l'environnement et expérimentation numérique, meeting sponsored by GDR MASCOT-NUM (Research Group on Stochastic Analysis Methods for COdes and NUMerical Treatments, Paris), 2011.

Invited speaker, 2011 Joint Statistical Meetings (Miami), session on Computational and Inferential Issues in Spatio-Temporal Modeling.

University of Miami Spatial Statistics Conference, 2012. Opening lecture. This talk turned into publication [8].

Invited speaker, 8th World Congress in Probability and Statistics, session on Composite Likelihood Inference. Istanbul, Turkey, 2012.

Short course at 2012 ENVR Workshop on Environmetrics.

Invited speaker, 2013 Joint Statistical Meetings (Montréal), session on Computational Statistics in the Atmospheric and Oceanic Sciences.

Invited speaker, 2013 American Geophysical Union (San Francisco), session on Closing the Loop: Integrating Socio-Economic and Climate Scenarios in the Assessment of Global Change Impacts.

Invited speaker, 2014 AAAS meeting (Chicago), session on Statistical Methods for Large Environmental Datasets.

Departmental/institutional seminars at Texas A&M, University of Minnesota (joint seminar for Departments of Statistics and Biostatistics), University of Michigan (student-invited seminar), Northwestern University, Georgia Tech, Illinois Institute of Technology, Depaul, North Carolina State University, Ohio State University, University of Illinois, Columbia University, NCAR (two talks).

Awards

Rietz Lecturer for IMS, 2010.

University of Chicago Faculty Award for Excellence in Graduate Teaching, 2011.

Elected fellow of AAAS, 2013.

References

[1] M. L. Stein (2012), "When does the screening effect hold?" *Annals of Statistics*, **39**, 2795–2819.
[2] E. Anderes and M. L. Stein (2011), "Local likelihood estimation for nonstationary random fields," *Journal of Multivariate Analysis*, **102**, 506–520.
[3] M. L. Stein, J. Chen and M. Anitescu (2012), "Difference filter preconditioning for large covariance matrices," *SIAM Journal on Matrix Analysis and Applications*, **33**, 52–72.

[4] M. L. Stein (2013), "On a class of space-time intrinsic random functions," *Bernoulli*, **19**, 387–408.
[5] M. L. Stein (2013), "Statistical properties of covariance tapers," *Journal of Computational and Graphical Statistics*, **22**, 866-885.
[6] M. L. Stein, J. Chen and M. Anitescu (2013), "Stochastic Approximation of Score Functions for Gaussian Processes," *Annals of Applied Statistics*, **7**, 1162–1191.
[7] J. Guinness and M. L. Stein (2013), "Transformation to approximate independence for locally stationary Gaussian processes," *Journal of Time Series Analysis*, **34**, 574–590.
[8] M. L. Stein, "Limitations on low rank approximations for covariance matrices of spatial data," *Spatial Statistics*, in press.
[9] M. Horrell and M. L. Stein, "A covariance parameter estimation method for polar-orbiting satellite data," *Statistica Sinica*, in press.
[10] A. Poppick and M. L. Stein, "Using covariates to model dependence in non-stationary, high frequency meteorological process," under revision for *Environmetrics*.
[11] J. R. Stroud, M. L. Stein and S. Lysen, "Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice," submitted for publication. arXiv:1402.4281v1
[12] M. Anitescu, J. Chen and M. L. Stein, "An inversion-free estimating equation approach for Gaussian process models," submitted for publication.