Scalable maximum likelihood estimation for Gaussian processes

Scalable Statistical Analysis of Gaussian Models for Petascale Spatiotemporal Data

Michael Stein
University of Chicago
Department of Statistics
5734 University Ave., Chicago, IL 60637

Abstract

Gaussian processes are widely used throughout the statistical and machine learning communities for modeling natural processes with broad applications to fields such as nuclear engineering and climate science. Such models typically have some vector of unknown parameters that must be estimated to carry out further data analysis tasks such as regression, classification and interpolation. A standard statistical approach to estimating these parameters is to choose the parameter values that maximum the likelihood function (the joint density of the observations). Simple in principle, the technique is difficult to apply in practice in the case of very large data sets due to the need to work with the covariance matrices of the observations. In some cases, the covariance matrices (or their inverse) may have some exploitable property (sparseness, Toeplitz) to reduce computations and/or storage, but in many applications, the covariance matrices are dense and unstructured. The likelihood function for Gaussian processes involves a quadratic form in the inverse covariance matrix and the log determinant of the covariance matrix. Existing algorithms for maximizing the likelihood heavily rely on the Cholesky factorization, the computation of which is prohibitively costly for many problems of practical interest. We propose a sample average formulation of the optimization framework, which narrows down the several linear algebra challenges to solving a linear system with the covariance matrix for multiple right-hand sides. We further investigate two of the most important ingredients in the conjugate gradient solver: the conditioning of the covariance matrix and the multiplication of a vector to a covariance matrix. These two aspects require different designs of the preconditioner and the matrix-vector multiplication in different grid configurations. We demonstrate the successful scalable resolution of the maximum likelihood problem for data sizes as large as a million points on a grid on a single desktop machine, whereas the Cholesky factorization approach would have needed a moderate-size supercomputer. In addition, we have proved that in some circumstances optimal preconditioning is achievable by means of a filtering approach. Parallel programs are under development to solve the problem for much larger data sets and in higher dimensions. This is joint work with Mihai Anitescu and Jie Chen.