Physics-based Covariance Models for Processes with Multiple Outputs

Scalable Statistical Analysis of Gaussian Models for Petascale Spatiotemporal Data

Emil M. Constantinescu
Argonne National Laboratory
Mathematics and Computer Science Division
9700 S. Cass Avenue
Argonne, IL 60439

Abstract

Petascale spatiotemporal data pose a tremendous challenge for state-of-the art statistical techniques. In particular, an important issue is determining models that appropriately account for the major features in the data, such as physical or nonstationarity characteristics. Current techniques rely on local fitting of ad hoc statistical models, which may not reveal a robust characterization of the data statistics, or empirical sampling techniques that for large-scale applications result in low-rank covariance matrices and may lead to difficulties in sampling and interpretation. Our focus is on covariance modeling for Gaussian process regression with multiple outputs. Hitherto, such Gaussian process analysis with multiple outputs was limited by the fact that far fewer good classes of covariance functions exist compared with the scalar (single-output) case.

To address this difficulty, we turn to covariance function models that take a form consistent in some sense with physical laws that govern the underlying simulated process. Models that incorporate such information are suitable when performing uncertainty quantification or inferences on multidimensional processes with partially known relationships among different variables, also known as co-kriging. One example is in atmospheric dynamics where pressure and wind speed are driven by geostrophic assumptions (wind $\propto \partial/\partial x$ pressure). We develop both analytical and numerical auto-covariance and cross-covariance models that are consistent with physical constraints or can incorporate automatically sensible assumptions about the process that generated the data.

We use these models to study Gaussian process regression for processes with multiple outputs and latent processes (i.e., processes that are not directly observed and predicted but interrelate the output quantities). In addition to deriving a systematic approach for describing the construction of covariance models governed by linear processes, we ask what happens if the process is not linear. In this case. We find that high-order closures are necessary to correctly specify the resulting covariance models. We have demostrated that such a strategy, can be very important for nonlinear models by comparing the fine approximation of the covariance structure resulting from a nonlinear process with low- and high-order closure assumptions. The latter proves to be significantly more accurate. Moreover, the strategy that we introduce in this study provides a physically consistent approach to introduce nonstationarity in the structure of the covariance matrix.

Our results demonstrate the effectiveness of the approach on both synthetic and realistic data sets. We consider Gaussian process regression experiments with a covariance model that has the correct physically consistent structure, which demonstrates significant improvements in the forecast efficiency. This strategy is validated on various synthetic and realistic data sets. The analytic covariance functions are validated by comparing results obtained with the models introduced in this study and covariance structures obtained through sampling strategies. We introduce new nonstationary covariance models that are generated directly through the physical process. For instance, we use a differential model on a nonuniform grid to generate nonstationary covariance kernels. These models have properties that are appropriate for processes that take place on adaptive grids or have various degrees of anisotropy.