

Computing mutual information based nonlinear dependence among noisy and finite geophysical time series

Shiraj Khan

Civil & Environmental
Engineering; University of
South Florida; Tampa FL

Sharba Bandyopadhyay

Department of Biomedical
Engineering; The Johns Hopkins
University; Baltimore, MD

Auroop R Ganguly ‡

Computational Sciences &
Engineering; Oak Ridge National
Laboratory; Oak Ridge, TN

Abstract

Linear correlation measures are widely used but may not be adequate for many geophysical problems, especially those that are dominated by nonlinear dynamics and nonlinear interactions. Mutual information (MI), which originated in communications and information theory, can be utilized to obtain measures of complete dependence, or “nonlinear correlation” (NLC). However, while the computation of MI is conceptually straightforward when the full probability density function (PDF) is available, there is no one best approach to compute MI or NLC from finite data sets. The state-of-the-art and emerging approaches used to compute the MI or NLC range from methods that are based on ranking of variables (RANKS), kernel density estimation (KDE), k -nearest neighbors (KNN) and what is called the “Edgeworth approximation” (Edgeworth). However, the emerging literature does not point to a clear winner that outperforms the other methods for real data sets and all the methods can be extremely sensitive to the presence of significant amount of noise. Thus, developing a better estimate often reduces to a better judgmental choice of the model parameters like the number of kernels or neighbors, even though preliminary guidelines may be available. The estimation problem, especially for uncertainty bounds, becomes even more difficult for time series data, where approaches like bootstrapping need to be applied with care. This study implements the four approaches (RANKS, KDE, KNN, Edgeworth) and investigates their relative performance, specifically for limited amount of noisy time series data, as a function of the signal-to-noise ratios and the size of the data. The datasets range from simulations (e.g., time series generated from the Lorenz system of equations contaminated with various noise levels) to real geophysical problems. The relative performance of the methodologies, as well as the insights gained over and above linear correlation approaches, is presented. The impacts of these insights on predictive modeling and scientific understanding are discussed.

***Acknowledgment:** Shiraj Khan would like to thank Professor Sunil Saigal at the University of South Florida. Auroop R Ganguly gratefully acknowledges the Laboratory Directed Research and Development Program (SEED money funds) of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. DOE under Contract No. DE-AC05-00OR22725.*

‡ Corresponding Author:

Oak Ridge National Laboratory; 1 Bethel Valley Road, P.O. Box 2008; Mail Stop 6085
Room B-106, Building 5700, Oak Ridge, TN 37831
Phone: (865) 241-1305; Fax: (865) 241-6261; Email: gangulyar@ornl.gov