Geophysical Research Abstracts
Vol. 15, EGU2013-6140, 2013
EGU General Assembly 2013

A discrete perspective on nonlinear dimension reduction

Christina Bogner (1), Baltasar Trancón y Widemann (2), and Holger Lange (3)

(1) University of Bayreuth, Ecological Modelling, Bayreuth, Germany (christina.bogner@uni-bayreuth.de), (2) TU Ilmenau, Programming Languages and Compiler Technology, Ilmenau, Germany, (3) Norsk Institutt for Skog og Landskap, P.O. Box 115, 1431 As, Norway

Environmental data sets are often large and high-dimensional and thus difficult to visualize and analyze. In hydrology, for example, we often deal with time series from long-term physical and chemical monitoring of stream water, groundwater or soils. These data can be seen as a multivariate characteristic of chemico-physical properties of water or soil and are used to infer processes in ecosystems.

Despite their high dimensionality, ecological data are often assumed to have a simple underlying intrinsic structure. It means that despite their high-dimensional nature they can be summarized in less dimensions without a serious loss of information. Therefore, dimensionality reduction techniques are often the first step to data analysis. They are used to visualize data as well as to uncover the intrinsic (low-dimensional) structure.

As an example application, we use high-dimensional hydrochemical data at the headwater catchment level (ion concentrations from first-order streams). We investigate the Isometric Feature Mapping (Isomap), a popular method for non-linear dimension reduction. Here, the topology of the data set is approximated by constructing a local neighbourhood graph. However, the assumption of smoothness underlying this approximation is difficult to justify for many environmental data sets, and issues of measurement errors and sampling gaps render Isomap analyses questionable. Thus, we extend our methodology by an analogous, but more robust, discrete (non-smooth) transformation leading to a set of binary data. For the latter, a plethora of data-mining techniques, in particular unsupervised and semi-supervised machine learning algorithms, exists. These can be employed to automate or support classification and feature detection tasks, taking the non-linear structure of available data into account. First results of this newly developed analysis method will be presented.