Exploring the data-space using trans-dimensional Monte Carlo sampling: the next frontier in (geo)data-mining.

Piana Agostinetti N.
Department of Geodynamics and Sedimentology, University of Vienna, Austria

In recent years, a number of algorithms have been developed to tackle the so called "Curse of Dimensionality" in the model space (i.e. the exponential increasing number of parameters needed to represent a physical model when a large number of observations are taken into account), exploiting the potentiality of trans-dimensional (trans-D) Monte Carlo sampling. In classical trans-D algorithms, the number of unknown parameters in the model is an unknown itself, and is completely defined by the amount of information contained in the observed data, avoiding dangerous over-parameterization of the investigated model. Nowadays, the scientific community is facing a new and emerging issue: the "Curse of Dimensionality" in the data space (i.e. how to handle the huge amount of data available without any subjective selection of sub data-sets, possibly integrating data from different sources and observations). To find appropriate theoretical and practical solutions to this key issue, I applied trans-D algorithm to the data selection problem, in the field of geophysical inverse problems. Preliminary results indicate that the theoretical framework developed can positively help in diving into huge data-sets (order of millions of data) and can be used to avoid subjective partition of data-set based on expert opinion. The approach presented is inherently cross-disciplinary and, while I applied it to seismological data, different geological and geophysical observations, from surface geology to geochemistry to geodesy can be easily tested.