Dealing with unknown discontinuities in data and models

Kerry Gallagher [(1)] , John Stephenson[(1)], Chris Holmes[(2)]

(1) Dept. of Earth Sciences and Engineering, Imperial College London, London, England
(2) Dept. of Statistics, University of Oxford, Oxford, England.

Abstract

The Earth is characterised by variability on many scales, and the significance of these scales depends on the particular problem under consideration. Natural spatial discontinuities, such as faults and lithological boundaries, separate regions within which the physical properties, processes or geological evolution may be similar. Similarly, time series may show very rapid changes (effectively discontinuous) either in the actual signal, or the underlying process over time, such as proxy records for palaeoclimate and climate itself. In the most general problem, we may want to deal with both spatial and temporal discontinuities, or where the relationship between spatial discontinuities may change with time. Another situation is where we may want to improve inference of a model or process by combining analyses samples collected at different locations. However, it is generally not be obvious how best to cluster these samples, given they may contain an unknown (and so potentially different) record about the processes of interest.

We can propose a solution to such problems in terms of inferring the locations (in space or time) of an unknown number of discontinuities in either data or a model of the underlying process. This is known as partition modelling (or changepoint modelling in 1 dimension). It is conveniently posed in a Bayesian formulation, and solved used reversible jump Markov chain Monte Carlo (RJMCMC), the transdimensional form of conventional MCMC. This approach allows efficient sampling of variable dimensional model spaces, in which we only need to specify the maximum number of dimensions. The Bayesian approach has the advantage of parsimony, so that we avoid producing overly complex models, while still achieving a satisfactory fit to the data. Furthermore, we can average models over variable of fixed dimensions, which provides a natural smoothing to the ensemble of discontinuous models, avoiding the need to specify smoothing functions.