Department of Decision Sciences
Statistics Seminar

Bocconi

Università Commerciale Luigi Bocconi

Building and fitting complex stochastic models for Historical data

Geoff Nicholls
University of Oxford
Thursday, 17th March 2016
12:30pm Room 3-E4-SR03 Via Rontgen 1 Milano

Abstract
We describe some new statistical methods developed in response to problems arising in the analysis of data from history and historical linguistics. Principal themes are model building with stochastic processes, Monte Carlo based Bayesian inference, and tools for Bayesian goodness of fit.

In the first problem we have data in which 12th C bishops are listed in order of social rank. Some pairs of bishops are strictly ordered in these lists, whilst other pairs swap order from one list to the next. This suggests the underlying hierarchy is not a total order. We represent the underlying hierarchy via a partial order on the bishops and model the bishop-list data as random linear extensions drawn uniformly at random from the set of orders allowed by the partially ordered hierarchy. Because the hierarchy evolves with time, we build and fit a Hidden Markov Model in which the hidden process is an evolving partial order (essentially, an evolving DAG). We use a particle filter and pMCMC to fit the model. The Monte Carlo based Bayesian inference is challenging. In the second example we have 14th C English texts sampled from across the UK. The texts show distinctive dialect spellings. Not all the text locations are known. We use the well-located texts to create a dialect map, and then use this map to locate the unknown texts. From the point of view of spatial statistics there is some novelty because the measurement locations are unknown. We simultaneously interpolate thousands of spatial word frequency fields and estimate model parameters and measurement locations. The dataset is both large and sparse, and Bayesian regularisation plays a key role. Goodness of fit tests suggest the model is a reasonable success. In the third example we have binary trait data for words in modern Polynesian languages and wish to infer the prehistoric relations between these languages, as that would inform the settlement history for the Pacific. In earlier work we have developed and fitted phylogenetic models for the evolution of these lexical traits in Indo-European. In our "Stochastic Dollo" models, vocabularies evolve in isolation. However the new data show a relatively high level of "borrowing" - the lateral transfer of traits across the tree from one language to another - so the tree supports a network of additional connections. There is very little work on likelihood based inference for network modelsfor evolving traits. We show that the network likelihood is given in terms of the solution of a very large linear system of ordinary differential equations. The sparsity structure of the ODE system is determined by the unknown phylogeny. We show the network model predicts withheld data whilst the purely tree-based model fails. We present preliminary results.

Department of Decision Sciences
Via Röntgen 1 - 20136Milano
Tel. 02 5836.5632
Fax 02 5836.5630