Chain event graphs for missing data: exploring informative missing data, with examples from longitudinal studies

Jane Hutton*
The University of Warwick, Coventry, UK - J.L.Hutton@warwick.ac.uk

Chain event graphs (CEGs) extend graphical models to address situations in which, after one variable takes a particular value, possible values of future variables differ from those following alternative values. This provides a framework for modelling discrete processes which exhibit strong asymmetric dependence structures, such as health status for smokers and non-smokers. These graphs are derived from probability trees by merging those vertices in the trees whose associated conditional probabilities are the same. It is possible to score each model efficiently and in closed form. Hence standard Bayesian selection methods can be used to search over a wide variety of models, each with its own explanatory narrative.

The expression 'missing at random' (MAR) was coined by Rubin in 1976, but the concepts of informative missingness compared to MAR are not simple to explain in studies with several or many variables which are only partially observed. In much research, knowledge of the design of data collection and the subject area will indicate that missing data is likely to be related to variables of interest. Problems caused by missingness can be especially acute in longitudinal data analyses when it is typical for substantial amounts of data about certain units in the sample to be missing at some of the observation times.

In order to use this background knowledge, we developed new classes of models in which the dependence of missingness, as well as independence conditional on observed values of variables can be displayed. The CEGs can capture data which are missing not at random but nevertheless exhibit context-specific symmetries. One of the advantages of this method is that the selected maximum a posteriori model and other closely scoring models can be easily read back to scientists in a graphically transparent way.

The efficacy of our methods are illustrated using a longitudinal study from birth to age 25 of children in New Zealand, and a simulation based on a study of weight loss.

Keywords: chain event graphs; graphical models; probability trees; missing data.