Graphical models for categorical data with R

Monia Lupparelli

Department of Statistics, Computer Science, Applications
University of Florence

Graphical models represent a class of statistical models which result from the combination of probability theory, graph theory and statistical methodology. Specifically, these models combine a graph structure with the joint probability distribution of a set of variables. The nodes of the graph are associated to random variables and the edges linking pairs of nodes provide a picture of the relationships in a multivariate system, in the sense that missing edges correspond to conditional independencies for the joint distribution. The independence model can be directly read from the graph using specific rules known as Markov properties. In essence, the merit of graphical models is to provide a graph representation of a complex multivariate setting exploiting probability rules supported by the data.

Multidimensional tables often represent the starting point for data analysis in biomedicine since they are the joint distribution of a set of categorical (binary, qualitative) data. In this case the association structure might be arduous to interpret as the discrete nature of the variables rises some challenges. Log-linear and logistic regression models are typically used for the analysis of cross-classified data collected in multy-way contingency tables and for the prediction of the probability of a discrete outcome. Under suitable constraints on the parameter space, log-linear and non-linear regression models specify independence graphical models, specifically, undirected graphical models and directed acyclic graph/Bayesian network models.

The course aims to provide foundations of graphical models with special focus for the analysis of categorical data and to enable the participants to conduct their graphical model analyses with the use of R. The methodology will be gradually introduced and motivated mainly by biomedical applications. All participants are expected to have basic knowledge of log-linear, logistic regression models and familiarity with R.

Outline and scheduling

September 22 - h 10:00-12:00
-   Introduction to graphical models
-   Undirected graphs and their Markov properties
-   Log-linear conditional independence models; model specification and parameter interpretation

September 22 - h 14:00-16:00
-   Directed acyclic graphs/Bayesian networks and their Markov properties
-   Sequences of logistic regression models

September 22 - h 16:30-17:30
-   Virtual coffee and question time

Reference

The course is mainly inspired on this book:

Højsgaard, S., Edwards, D. and Lauritzen, S. (2012), Graphical Models with R. Springer