Detecting Influential Observations for a Graphical Model

Avner Bar-Hen*
Laboratoire MAP5, Université Paris Descartes,
Paris, France – Avner.Bar-Hen@mi.parisdescartes.fr

Jean-Michel Poggi
Laboratoire de Mathématiques, Université Paris Sud, Orsay, France and
Université Paris Descartes, Paris, France – Jean-Michel.Poggi@math.u-psud.fr

Graphical models represent a set of random variables and encode their probabilistic conditional dependencies as a graph in which nodes represent random variables and edges represent conditional dependencies among them.

The interest in such models, since the graph can represent the scientific content of a given model, is twofold. On the one hand, from the applied side, they capture knowledge from multiple experts and experience from knowledge and data. On the other hand, from a statistical viewpoint we are interested to examine issues of stability, sensitivity, and scalability to cope with massive data. Therefore graphical models infer probabilistic relationships among variables and conditional dependence probabilities are estimated from data. Various algorithms have been proposed to estimate the topology and we focus on Maximum Likelihood Estimation.

Sensitivity issues are naturally of interest since the topology of the network and new relationships are estimated from data through various algorithms. We focus on a different viewpoint centered on individuals rather than parameters or model stability. The question of measuring influence of observations on the results obtained with a graphical model is of interest. A key tool is such a direction can be the use of an influence measure, which is a classical diagnostic method to measure the perturbation induced by a single element, in other terms we examine stability issue through jackknife highlighting influential observations.

To define the influence of individuals on the analysis, we propose various criterions to measure the sensitivity of the graphical model using jackknife network. More precisely, to compute the influence of one observation we compare the network based on all observations except the concerned observation with the reference network based on all observations. This can be done by iterative computation or directly by handling the inverse covariance estimation and derive from the variation of the inverse, some criterions.

All along the paper an application to gene expression data set is carried out. This dataset provided by Hess et al. concerns 133 patients with stage I-III breast cancer. The patients were treated with chemotherapy prior to surgery. Patient response to the treatment is classified as either a pathologic complete response or a residual disease.

This dataset allows us to explore some clustering issues using influence-based tools introduced in the first part of the paper.

Keywords: Graphical model; Influence; Jackknife.