

Information Enhanced Model Selection for High-Dimensional Gaussian Graphic Model with Application to Metabolomics Data

Jiang Gui

Dartmouth College

E-mail: jiang.gui@dartmouth.edu

Abstract: In light of the low signal-to-noise nature of many large biological data sets, we propose a novel method to identify the structure of association networks using a Gaussian graphical model combined with prior knowledge. Our algorithm includes the following two parts. In the first part we propose a model selection criterion called structural Bayesian information criterion (SBIC) in which the prior structure is modeled and incorporated into the Bayesian information criterion (BIC). It is shown that the popular extended BIC (EBIC) is a special case of SBIC. In second part we propose a two-step algorithm to construct the candidate model pool. The algorithm is data-driven and the prior structure is embedded into the candidate model automatically. Theoretical investigation shows that under some mild conditions SBIC is a consistent model selection criterion for high-dimensional Gaussian graphical model. Simulation studies validate the superiority of the SBIC over the standard BIC and show the robustness to the model misspecification. Application to untargeted profile data from infant feces collected from subjects enrolled in a large molecular epidemiologic cohort study validates that prior knowledge on metabolic pathway involvement is a significant factor for the conditional dependence among metabolites. More importantly new relationships among metabolites are identified through the proposed algorithm which can not be covered by standard pathway analysis. Some of them have been widely recognized in literature.