PC algorithm for complex survey data via resampling

Abstract

The association structure of a Bayesian network can be known in advance by subject matter knowledge or have to be learned from a database. In case of data driven learning, one of the most known procedures is the PC algorithm where the structure is inferred carrying out several independence tests on the database and building a Bayesian network in agreement with the tests results.

The PC algorithm uses conditional independence tests and it is based on the assumption of independent and identically distributed observations, which is equivalent to the assumption of simple random sampling. It is well known, that for i.i.d. data the standard Pearson chi-squared test statistic is distributed asymptotically as a chi-squared random variable when the null hypothesis holds. For complex designs, the test procedure is not valid even asymptotically. In general, sample selection in surveys involves more complex sampling designs based on stratification, different level of clustering and inclusion probabilities proportional to an appropriate measure of size. The impact of complex designs on i.i.d. based methods can be severe.

Here, a novel approach for inferring casual structure from complex survey data is investigated. The complexity of sampling design is accounted for via a design-based approach by including the sampling weights in the BN parameters estimates. A modified version of the PC algorithm is proposed; it is named PC complex. First of all, since the PC algorithm uses conditional independence tests for model selection, a procedure for testing the association in a two-way table in complex sample surveys is proposed. The procedure is based on resampling techniques for finite population. Secondly, the procedure is applied to BN structural learning and its robustness with respect to the departure from faithfulness assumption and causal Markov condition is studied.