Project No.	Project Title	
2021-01-131	Algorithms for Learning Latent Variable Models	
Academic Advisor		**Co-Advisor**
Prof. Boaz Lerner		
Team Members		
Shoham Shabat		
shohas@post.bgu.ac.il		

Abstract

In many if not in all domains, there are two types of variables: observed and unobserved (latent). Latent variables refer to abstract concepts that cannot be measured directly. In contrast to modeling with observed variables, which may be straightforward, learning a model that combines the two types of variables is more difficult. In the presence of latent variables, modeling assuming only observed variables will likely lead to a wrong model that does not describe the true realm and may also be misleading. Hence, the identification of latent variables and the relationships among them and with the observed variables is crucial.

Latent variable models (LVMs) aim to model latent variables as root causes of the problem and affect some indicators, which are the observed variables. The learning pairwise cluster comparison (LPCC) algorithm learns an LVM based on cluster analysis with two steps. At the first step, it finds exogenous and collider latent variables and their observed descendants by pairwise comparison of appropriate major clusters (MCs), which are meaningful data-point clusters in the domain. In the second step, the LPCC splits exogenous latent variables, if necessary, to learn serial or diverging connections if exist in the domain. Thus, correct selection of the MCs is essential for the success of step one and consequently for that of the whole algorithm.

In this work, we examined methods to accurately identify the MCs. Using statistical methods, we identified "matched groups" in which any pair of observed variables are interdependent. To find the matched groups, we compared two types of statistical tests: the Hilbert-Schmidt independence criterion (HSIC) test that is based on an approximation to the gamma distribution using the radial basis function (RBF) kernel, and a t-test on the Spearman's correlation between observed variables. In addition, we proposed and experimentally validated three formulas for the number of MCs that depend on the number of matched groups and cardinality of the observed variables.

Our empirical findings show that for six synthetic networks posing different challenges to the learning algorithm, procedures combing any of the two statistical tests with any of the three formulae greatly improve the LPCC performance, although with no statistical difference between them. The analysis shows that the statistical tests are comparable, but the formulas differ in learning the structure of the true graph. To further distinguish among the procedures, more experiments using more complex synthetic graphs with higher levels of cardinality as well as real networks are needed.

Keywords: Latent variables, LVMs, LPCC, HSIC, Spearman's correlation