Causation: Foundation to Application Workshop
UAI, 29 June, 2016, NY

Pairwise Cluster Comparison for Learning Latent Variable Models

Nuaman Asbeh and Boaz Lerner

Industrial Engineering and Management , Ben-Gurion University of the Negev, Israel
{nuamana@yahoo.com; boaz@bgu.ac.il}

Introduction

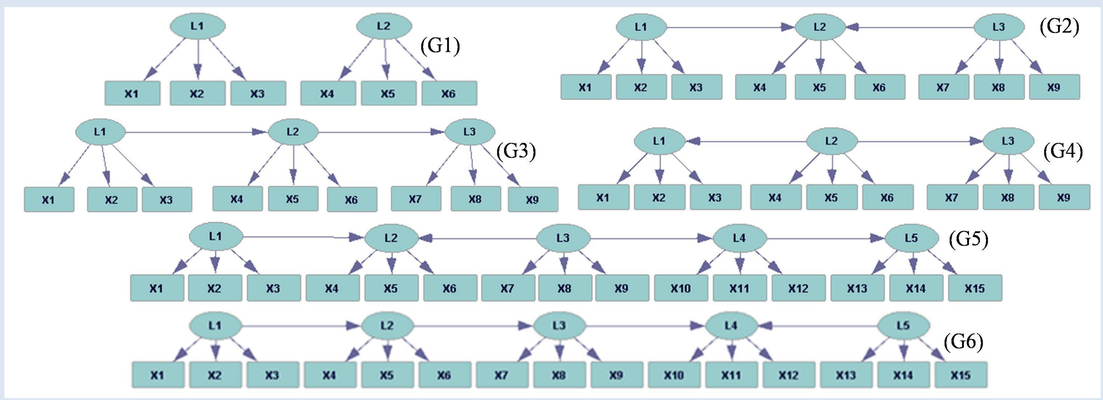
- Statistical methods (e.g., FA) reveal the existence/influence of latent variables, accomplish dimensionality reduction, and may fit the data reasonable well, but the resulting models might have no correspondence to real causal mechanisms
- Bayesian networks (BNs) focus on observed variables and cannot represent latents (Exceptions are IC* [1] and FCI [2] that indicate potential hidden common causes, Elidan's heuristic search [3] for "structural signatures" that suggest latent existence, and binary and Gaussian latent trees [1] and discrete latent trees [4])
- However, BN methods are not appropriate to learn multiple indicator models (MIMs).
  - Silva et al. [5] fill the gap for continuous variables under the assumption of linearity

Objectives

Develop a concept and an algorithm for learning latent variable models (LMV) that:

- provide causal explanation, which is overlooked by statistical methods
- identify latent variables and their interrelations, which BN cannot do
- are not limited to learning latent trees
- can learn MIMs but without too restricting assumptions

Preliminaries



O, EX, EN, L – observed, exogenous, endogenous, latent variables in G. $\mathbf{V}=\mathbf{L}\cup\mathbf{O}$, $\mathbf{EX}\subset\mathbf{L}$, $\mathbf{EL}\subset(\mathbf{L}\cap\mathbf{EN})$, $\mathbf{O}\subset\mathbf{EN}$. **ex, el**, **en**, **o** value configurations

Figure 1. Four basic LVMs (G1-G4) and two larger graphs (G5 and G6) that challenge our concept and algorithm.

Data is generated by BN

Proposition 1 The joint probability over **V** due to an assignment **ex** to **EX** is determined only by this assignment and BN,

$$\mathrm{P}(\mathbf{V}|\mathbf{EX}=\mathbf{ex})=\mathrm{P}(\mathbf{EX},\mathbf{EL},\mathbf{O}|\mathbf{EX}=\mathbf{ex})=\prod_{EX_i\in\mathbf{EX}}P(EX_i=ex_i)\prod_{EL_j\in\mathbf{EL}}P(EL_j=el_j|\mathbf{L}_j=\mathbf{l}_j^{\mathbf{ex}})\prod_{O_k\in\mathbf{O}}P(O_k=o_k|L_k=l_k^{\mathbf{ex}})$$

Local major/minor effect

- A *local effect* on *EN* is the influence of *EN's direct latent parents* on any of EN values
- The *major local effect* is the largest local effect on *EN*, $MAE_i(\boldsymbol{l}_i)=max_{en_i'}\{P(EN_i=en_i'|\boldsymbol{L_i}=\boldsymbol{l}_i)\}$
- The *major local value* is the en_i corresponding to $MAE_i(\boldsymbol{l}_i)$
- *Minor local effect & minor value* are any non-major local effect on *EN* and corresponding en_i

Major/minor effect

- An *effect* on **EN** is the influence of a configuration **ex** of **EX** on **EN**
- *Major/minor effect* (MAE/MIE) - largest/any non-MAE effect of **ex** on **EN**
- *Major/minor value configuration* (MAV/MIV) - **en** corresponding to MAE/any MIE
- Based on Proposition 1 and the BN Markov property, we quantify the major effect

$$MAE(\mathbf{ex})=\prod_{EX_i\in\mathbf{EX}}P(EX_i=ex_i)\cdot\prod_{EL_j\in\mathbf{EL}}MAE_j(\mathbf{l}_j^{\mathbf{ex}})\cdot\prod_{O_k\in\mathbf{O}}MAE_k(l_k^{\mathbf{ex}})$$

Proposition 2 A single observed value configuration (parts in **en** of **o**) due to **ex** is major

Linking effects with clustering

Due to the BN probabilistic nature, each observed value configuration due to **ex** is represented by several data patterns. Data clustering produces several clusters per each **ex**, where one of them – the *major cluster* – corresponds to the observed major value configuration (that represents the major effect), and the other are *minor clusters* representing minor effects.

Pairwise cluster comparison (PCC)

- Comparison of cluster (centroid) pairs ⇒ Binary vector: 1 (0) if there is (not) a difference

Centroid	X1	X2	X3	X4	X5	X6
C1	0	0	0	1	1	1
C2	1	1	1	1	1	1
C3	0	0	0	0	0	0
C4	1	1	1	0	0	0

(a)

PCC	δX1	δX2	δX3	δX4	δX5	δX6
PCC1,2	1	1	1	0	0	0
PCC1,3	0	0	0	1	1	1
PCC1,4	1	1	1	1	1	1
PCC2,3	1	1	1	1	1	1
PCC2,4	0	0	0	1	1	1
PCC3,4	1	1	1	0	0	0

(b)

Table 1. (a) Centroids of major clusters for G1. (b) PCCs between these major clusters..

Learning PCC (LPCC) – An Algorithm Overview

I. Identification of latent variables and their descendants

X1, X2, and X3 (G1) change together in all PCCs (maximal set of observed variables, ***MSO***), and thus are descendants of the same latent variable (L1)

II. Identification of collider latent variables and their parents

{X4, X5,X6} (G2) change with {X1,X2,X3} in part of the PCCs and with {X7,X8,X9} in other PCCs. This is a clue that X4, X5, and X6 are descendants of a collider (i.e., L2 for L1 and L3)

III. Strategy for choosing major clusters

1st iteration: select clusters larger than the average cluster size. (t+1) iteration: assume the learned graph Gt is true, and learn latents' cardinalities. Find the set of all possible **ex.** For each **ex**, find the most probable cluster with a centroid $c^*=argmax_{C_i\in\boldsymbol{C}}P(C_i|\boldsymbol{ex})$(EM style [6])

IV. Identification of non-collider latent variables

Split them from their exogenous latents using major-(first-order) minor PCCs to identify latent non-colliders and observed variables

Table 2. Learning G3. (left) Largest clusters. (right) PCCs for C3 – a first order minor cluster – with C1 and C2, which are major clusters.

Centroid	X1	X2	X3	X4	X5	X6	X7	X8	X9	size
C1	1	1	1	1	1	1	1	1	1	49
C2	0	0	0	0	0	0	0	0	0	47
C3	1	1	1	1	1	1	1	1	0	28
C4	0	0	0	0	0	0	0	1	0	24
C5	0	1	0	0	0	0	0	0	0	22

Centroid	X1	X2	X3	X4	X5	X6	X7	X8	X9
PCC1,3	0	0	0	0	0	0	0	0	1
PCC2,3	1	1	1	1	1	1	1	1	0

Evaluation

I. Simulated data sets

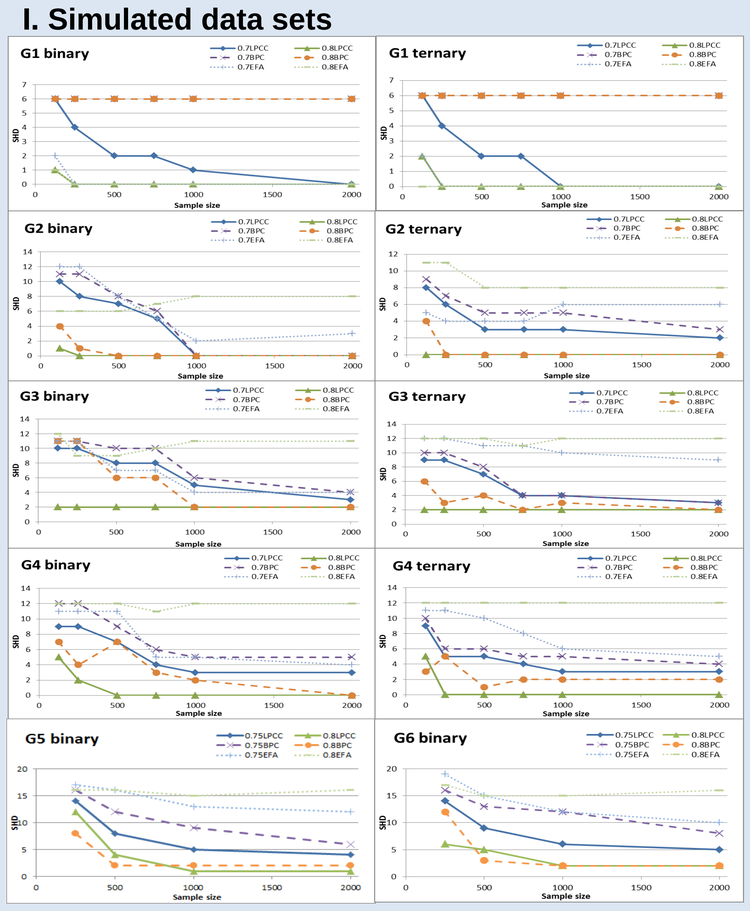


Figure 2. Structural Hamming Distance (SHD) [7] of LPCC, EFA, and BPC [5] for all graphs of Figure 1 for increasing sample sizes.

LPCC versus BPC and EFA.

Two parameterization schemes , p_j=0.7 and 0.8.

For all graphs in the binary case, except L2 in G2, $P(EN_i=v|L_k=v)=p_j$, $v=0$ or 1.

For all graphs in the ternary case, except L2 in G2, $P(EN_i=v|L_k=v)=p_j$, $P(EN_i\neq v|L_k=v)=(1-p_j)/2$, $v=0,1,\text{or }2$.

For L2 in G2, $P(L_2=0|L_1L_3=00,01,10)=P(L_2=1|L_1L_3=11)=p_j$ in the binary case, and $P(L_2=v|\max\{L_1,L_3\}=v)=p_j$ $P(L_2\neq v|\max\{L_1,L_3\}=v)=(1-p_j)/2$ in the ternary case.

II. Real-world data sets

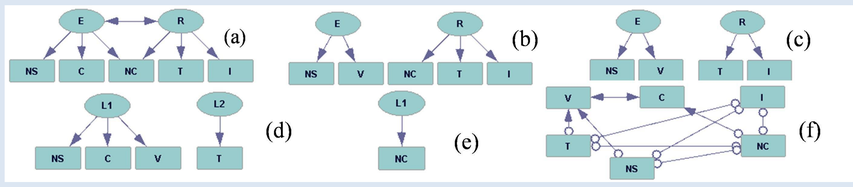


Figure 3. (a) A theoretical model [8]; (b) LPCC; (c) BSPC [5]; (d) BPC [5] for alpha=0.01or 0.05; (e) BPC for alpha=0.1; and (f) FCI [2] for alpha=0.05.

Simplified political action survey data:
- Six observed variables: NOSAY (NS), VOTING (V), COMPLEX (C), NOCARE (NC), TOUCH (T), and INTEREST (I)
- Ordinal variables of 4 values
- Two latents correspond to a theoretical trait: Efficacy (E) and Responsiveness (R) [8]
- Sample size is 1,079

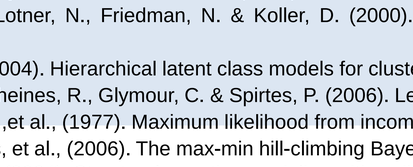


Figure 4. Model learned for HIV using LPCC is identical to true model (BPC returns empty graph for any alpha).

HIV test data [4] consists of results on 428 subjects of four diagnostic tests for the human HIV virus: radioimmunoassay of antigen ag121 (X1); radioimmunoassay of HIV p24 (X2); radioimmunoassay of HIV gp120 (X3); and enzyme-linked immunosorbent assay (X4).

Discussion and Future Research

- LPCC – a concept and an algorithm to learn LVM – improves performance with the sample size, can learn large LVMs, and has consistently good results compared to expert-based models and state-of-the-art algorithms
- LPCC dispenses with the linearity assumption but assumes a pure model
- LPCC is suitable to MIM and not only latent trees
- LPCC does not require pre-setting of a significance level as BPC/FCI
- Further research will: 1) identify interrelations among observed variables; 2) analyze model identification conditions; 3) analyze LPCC complexity; and 4) explore the impact of clustering.

References

1. Pearl, J. (2000). *Causality: Models, Reasoning, & Inference*. NY: Cambridge Univ. Press.
2. Spirtes, P., Glymour, C. & Scheines, R. (2000). *Causation, Prediction, and Search, second edition*. NY: MIT Press.
3. Elidan, G., Lotner, N., Friedman, N. & Koller, D. (2000). Discovering hidden variables: A structure-based approach. In *NIPS* 13:479-485.
4. Zhang, N. (2004). Hierarchical latent class models for cluster analysis. *JMLR* 5:697–723.
5. Silva, R., Scheines, R., Glymour, C. & Spirtes, P. (2006). Learning the structure of linear latent variable models. *JMLR* 7:191–246.
6. Dempster, A.,et al., (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. of Royal Stat. Soc.*, B 39:1–39.
7. Tsamardinos, et al., (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn* 65:31–78.
8. Joreskog, K. (2004). Structural equation modeling with ordinal variables using LISREL. Scientific Software International Inc.