Identifying Latent Variables in Dynamic Bayesian Networks with Bootstrapping Applied to Type 2 Diabetes Complication Prediction

Leila Yousefi,Mashael Al-Luhaybi,
Lucia Sacchi, Luca Chiovato and Allan Tucker

THIS supplementary document is intended to explain and clarified the journal paper in more details.

I. Link Strength methodology

In this paper, we employed local and global sensitivity analysis [1] that consists of Mutual Information (MI) and Link Strength (LS). The Link Strength methodology finds a structure for locating latent variables within a Bayesian structure. We exploited the following measures to handle the uncertainty in the discovered model:

- Entropy introduced in [2] to measure the uncertainty in a single node and shown below:

$$U(X) = \sum_{x_i} P(x_i) \log_2 \frac{1}{P(x_i)}. \tag{1}$$

- Mutual Information is a way of inferring links in data and measuring the connection strength [2] [3]. The MI between node X and Y is uncertainties in Y that is decreased by knowing the state of X (or vice verse):

$$MI(X,Y) = U(Y) - U(Y|X), \tag{2}$$

  where $U(Y|X)$ is calculated by averaging $U(Y|x_i)$ over all possible states x_i of X, taking $P(x_i)$ into account in which:

$$MI(X,Y) = \sum_{x,y} P(x,y) \log_2 (\frac{P(x,y)}{P(x)P(y)}). \tag{3}$$

- The Link Strength [4] measure enables us to observe the impact of each discovered edge. Moreover, the percentage points of uncertainty reduction in Y are utilised by knowing the state of X if the states of all other parent variables are known. There are two types of LS in measuring uncertainties: True Average Link Strength (LSTA), and Blind Average Link Strength (LSBA).
- The LSTA calculates LS based on the average over the parent states using their actual joint probability. For a node with only one parent, MI Percentage and the LSTA Percentage yields the same value. LSTA of the edge $X \rightarrow Y$ is defined as the MI of (X,Y) conditioned on all other parents of Y, which shown as:

$$LSTA(X \rightarrow Y) \text{ ,X requires P(For all parents of Y)} \tag{4}$$

$$= MI(X,Y|Z) = U(Y|Z) - U(Y|X,Z) \tag{5}$$

  where $U(Y|X,Z)$ is the average over the states of all parents, and $U(Y|Z)$ is the average over all other parents.
- The LSBA is derived from the LSTA but ignores the actual frequency of occurrence of the parent states. Thus, in the LSBA measure, all parents are assumed to be independent of each other and uniformly distributed.

$$LSBA(X \rightarrow Y) = \text{requires no inference at all.} \tag{6}$$

The same probabilities as the corresponding absolute measure above are converted to each percentage measure. For removing all uncertainty, we require deterministic functions, in which the state of a child is completely known if the states of all of its parents are known. Representing all parents from Y in $MI(X,Y|Z)$ in Equation(5) essentially blocks all information flow through the other parents, Z. According to [5], we are confident that there are no other indirect open links between Y and X through descendants of Y, once all different parents are instantiated then there is a direct link from X to Y.

Theorem: Consider a BN (G,P) consisting of a DAG (G) and a joint probability (P). Let $X \rightarrow Y$ be an edge in G and denote the set of all other parents of Y as Z. Let $G\%$ be the modified DAG generated by deleting edge $X \rightarrow Y$ in G. Then X and Y are conditionally independent given Z in BN $(G\%, P\%)$ for any joint probability $P\%$. As indicated

TABLE I
PATIENT-BASED PREDICTION ACCURACY.

Complication	Retinopathy	Retinopathy	Liver disease	Liver disease
Method	B-K2-REVEAL	IC*LS	B-K2-REVEAL	IC*LS
Early	90	95	92	97
Late	2	2	0	0
Hit	0	2	2	1
Miss	8	1	8	2
FP	0	1	0	0
FN	8	2	8	2
TP-Early	1	0	2	2
TP-Late	24	25	18	18
TN-Early	68	73	75	81
TN-Late	2	2	0	0

by the LSTA, most links are quite strong, can be classified as significant, except for those with LSTA of less or equal to zero (removed from the final structure). The LS measure may be useful in the context of constraint-based structure learning algorithms to derive hypotheses of a system's primary causal pathways. In addition, it can be used to evaluate the quality of the structure learning algorithms. Currently, structure learning algorithms are evaluated by counting the number of incorrect arrows when identifying known systems. It may be more appropriate to weight those counts by the LS of the incorrect arrows. By setting the value of the LSTA greater than 20 percent, though, overfitting in the DAG can be reduced.

II. LATENT STRUCTURES

Some variables are unmeasured but have potential to be developed, called hidden or latent variables. Based on Pearl's causality, a latent structure is a pair $L =< D, O >$ in which D is a causal structure over variables that O is a set of observed variables. The IC* algorithm returns a marked pattern, a partially DAG that contains four types of edges:

1) A marked arrow $O_1 \overset{*}{\rightarrow} O_2$, signifies a directed path between observed nodes (O_1 and O_2) in the underlying latent structure (and there is no latent common cause for these two nodes).
2) A bi-directed edge $O_1 \leftarrow\rightarrow O_2$ signifies a latent common cause ($O_1 \leftarrow L \rightarrow O_2$) in the underlying latent structure, or there is an inducing path between two variables, thus there is no directed path between them.
3) An unmarked arrow $O_1 \rightarrow O_2$, signifies either a directed path from O_1 to O_2 or a bi-directed edge.
4) An indirect edge $O_1 -\!- O_2$, stands for either $O_1 \rightarrow O_2$ or $O_1 \leftarrow O_2$ or $O_1 \leftarrow\rightarrow O_2$.

III. PATIENT-BASED VALIDATION

Table. I demonstrates the effect of early/late time prediction of comorbidities. If a patient is diagnosed with a complication earlier than their actual observed occurrence time, it is called an early prediction otherwise known as a late prediction. Patient-based prediction performance for two complications (retinopathy and liver disease) is assessed in Table. I, representing percentage of patients in which they are diagnosed early and late based on different structure learning methods (i.e., the K2 and IC*LS). We identify the diagnosis point as a comorbidity is diagnosed. The results of class confusion matrices and the prediction accuracy are shown in terms of percentage of patients who are diagnosed correctly (Hit) and incorrectly (Miss) in the same time as the diagnosis time. Moreover, results from confusion matrices are retrieved as number of False Positives (FP), False Negatives (FN), True positives in the early prediction (TP-Early), True Positives in the late prediction (TP-Late), True Negatives in the early prediction (TN-Early), True Negatives in the late prediction (TN-Late). These results have been used to reveal that a better prediction accuracy is generally achieved by using the IC*LS method and interpreted in the body of the paper.

IV. LATENT VARIABLE AS EVIDENCE

In this section, we look at the influence of the latent variable as evidence to predict the complications. As mentioned earlier, the aim of this research is to explore the impact of the targeted latent variable on prediction of the T2DM complications. Here, we validate the results to uncover influential factors in the diagnosis with regard to a set of diagnosis targets given the evidence. This target was possible to achieve due to the nature of the Bayesian inference, where any T2DM comorbidity or risk factor could be queried using a joint probability distribution. The bar chart in Fig. 1-a shows how the probability of retinopathy being

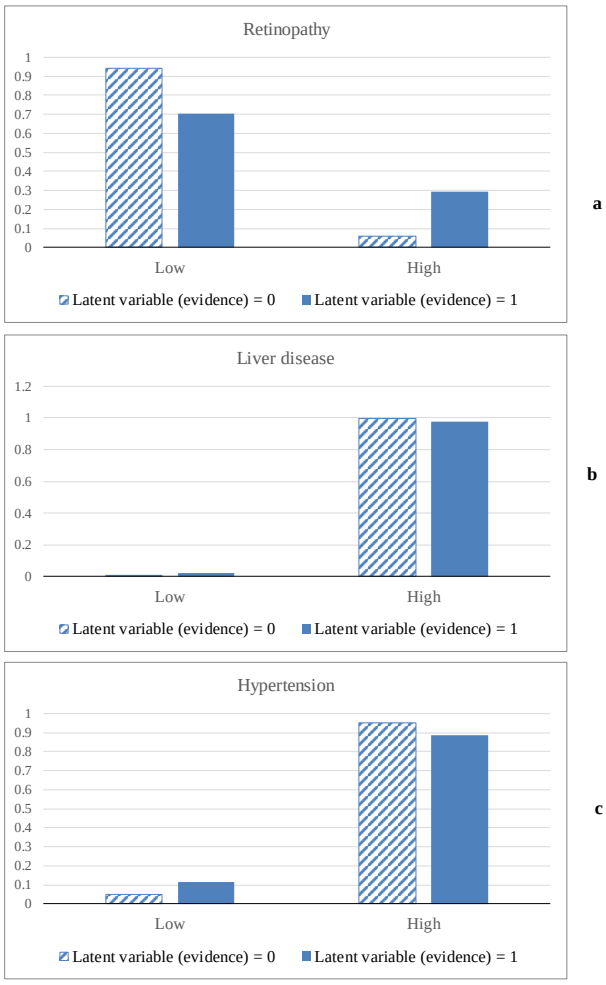

Fig. 1. Prediction probabilities for retinopathy, liver disease and hypertension using the Latent variable as the evidence.

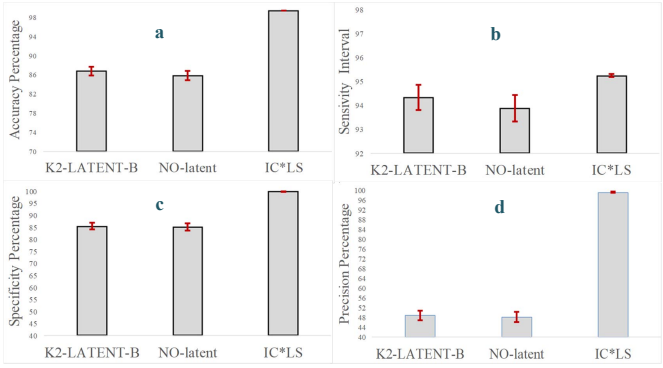

Fig. 2. Bootstrap Confidence Interval represents accuracy, sensitivity, specificity, and precision to predict liver disease.

in its high value (the diagnosis point) changes by setting the evidence on the comorbidity-related risk factor. For instance, the marginal probability of one or more patients being diagnosed with retinopathy is 0.05 (no evidence set). According to the DBN model, setting evidence to the latent variable increases the marginal probability of retinopathy from 0.05 to 0.29. Thus, we guarantee that the latent variable has a significant impact on the target comorbidity posterior probability. However, setting evidence on the latent variable changes the posterior probability of liver disease and hypertension from 0.99 to 0.97 and from 0.95 to 0.88, respectively. This shows the latent variable has a negative impact on developing liver disease and hypertension

whilst a much stronger positive effect on retinopathy.

V. Confidence Interval Results

The experimental evidence on bootstrapping are explored in this section. In the bootstrap related literature, there are some articles that deal with bias issues in the data in order to enhance convergence accuracy [6] [7] [8]. For example, in a situation in which asymptotic confidence intervals are known and correct, bias-corrected and accelerated (BCa) confidence intervals have been demonstrated to show faster convergence and increased accuracy over ordinary percentile-based methods whilst retained the robustness [9]. The findings provided in here reveal the influence of using a latent variable on the bootstrapped data in predicting liver disease. We compare the accuracy percentage average, among 250 times bootstrap, of the three methods (K2-LATENT-B, IC*LS-NO-LATENT-B, and IC*LS-LATENT-B), which are 88, 86 and 99 per cent, respectively. We also illustrate error bars on the top of the bar charts. These results reveal that IC*LS-LATENT-B has higher accuracy than IC*LS-NO-LATENT-B and K2-LATENT-B, while IC*LS-NO-LATENT-B error bar is larger than others. The error bar in K2-LATENT-B is quite big due to the corresponding confidence interval of IC*LS-NO-LATENT-B. However, a smaller confidence interval in IC*LS-LATENT-B makes the corresponding error bar consistent. Whenever the error bars overlap, as in the sensitivity analysis in Fig. 2 b, which lower bottom of the error bar in K2-LATENT-B is lower than the top of error bar in IC*LS-LATENT-B, then statistically speaking these two averages are not different, even though the means themselves are totally different. In contrast, the error bars in Fig. 2 IC*LS-LATENT-B and IC*LS-NO-LATENT-B do not overlap, whereas the results of means show that they are pretty much the same average, then we can say they are statistically different. We are confident that without using a latent variable the performance measures is reduced considerably, as seen in Fig. 2 IC*LS-NO-LATENT-B compared to K2-LATENT-B. Overall, we are at least 95 per cent sure that exploiting IC*LS causes a huge and significant improvement in the classification precision, accuracy, sensitivity, and precision compared to the K2 and no latent variable results (see Fig. 2).

References

[1] J. Khoo, T.-L. Tay, J.-P. Foo, E. Tan, S.-B. Soh, R. Chen, V. Au, B. J.-M. Ng, and L.-W. Cho, "Sensitivity of a1c to diagnose diabetes is decreased in high-risk older southeast asians," *Journal of Diabetes and its Complications*, vol. 26, no. 2, pp. 99–101, 2012.

[2] C. E. Shannon, W. Weaver, and A. W. Burks, "The mathematical theory of communication," 1951.

[3] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[4] N. Jitnah, *Using Mutual Information for Approximate Evalutation of Bayesian Networks*. Monash University, 1999.

[5] I. Ebert-Uphoff, "Measuring connection strengths and link strengths in discrete bayesian networks," Georgia Institute of Technology, Tech. Rep., 2007.

[6] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statistical science*, pp. 189–212, 1996.

[7] F. A. Wichmann and N. J. Hill, "The psychometric function: Ii. bootstrap-based confidence intervals and sampling," *Perception & psychophysics*, vol. 63, no. 8, pp. 1314–1329, 2001.

[8] H. DAVID, "Bootstrap estimates of the statistical accuracy of thresholds obtained from psychometric functions*ˆ," *Spatial Vision*, vol. 11, no. 1, pp. 135–139, 1997.

[9] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.