# Linking Bayesian Networks and Bayesian Approach for Structural Equation Modeling

Sumaman Pankham[1] and Suchada Kornpetpanee[2]

[1]Faculty of Information Technology, Rangsit University, Thailand

[2]College of Research Methodology and Cognitive Science (RMCS), Burapha University, Thailand

**Abstract.** The Structural Equation Modeling (SEM) is not only constantly used in social science research, but also can move forward the quality of decision-making. Furthermore, it has a process of converting strategic objectives into effective actions. Several papers have emphasized the usefulness of Structural Equation Modeling. Maximum likelihood (ML) approach is one of the several popular for Structural Equation Modeling.

However, the problem often faced when we use maximum likelihood. Firstly, the assumption of maximum likelihood methods is made constant variance normal disturbances. So sample size must be a large number. Secondly, linking between attributes due to lack of knowledge of the background is unknown. To solve these problems, this paper proposes a method that links attributes using the Bayesian network and Bayesian approach modeling for Structural Equation Modeling. The assumption of Bayesian does not require constant variance normal disturbances. So sample a size can be a small number. An experimental study shown that maximum of standard error (STERR) is 0.019 and $R^2$ values 89%. This methodology is doing well in that the task of deciding the causal directions between attributes becomes easy through use of the proposed method and small sample size.

**Keywords:** Structure Equation Modeling, Bayesian network, Bayesian approach.

## 1. Introduction

The Structure Equation Modeling (SEM) has become a common technique of assessing the relationship between cause, and its use in research papers published in the most famous is involved. However, the use of the SEM is based on the possibility of knowledge. The SEM is a tool for analyzing multivariate data that has been long known in marketing to be especially appropriate for theory testing [1].

Major applications of structural equation modeling include [2]:

- Causal modeling or path analysis is causal relationship between variables and testing a causal model to the system of linear equations. Causal models can be related to the latent variable or variables that are manifested in both cases.
- Factor analysis confirmed is the extension of the analysis of factors specific assumptions about the structure of the factor loadings, and inter-correlations are tested.
- Second order factor analysis is a variation of factor analysis of the correlation matrix of the common factor which is the factor analysis to the second order factor.

The Bayesian network is directed acyclic graphical model. It is a probabilistic graphical model that shows a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). The DAG can show to help us decide the causal directions between constructs when using Bayesian approach. Bayesian has three outstanding points as following [3]:

- Bayesian methods can solve the problem of inference in maximum likelihood computed using numerical hessians, which are not always very good.

- Bayesian methods can be used to relax the assumption of constant variance normal disturbances made by maximum likelihood methods, resulting in extended models.
- Bayesian methods can be used to formally solve model comparison problems. We can compare models based on: different weight matrices, different explanatory variables (X), or different model specifications such as SAR, SDM, SEM.

This paper, we suggest a method that links the Bayesian network and Bayesian approach for Structural Equation Modeling. This study is shown to illustrate the application of the proposed method. This paper is organized follows. In Section 2, we proposed step of this paper. In Section 3, the Bayesian network, Bayesian approach and Structural equation modeling are discussed. In Section 4, an empirical study path analysis is shown by way of illustration. Finally, based on the findings of this research, conclusions and implications for management are shown.

## 2. The Proposed Method

Causal analysis can be presented by Cause Diagram. Fortunately, causal analysis can be used to produce causal maps [4]. The Bayesian network, Bayesian approach and the SEM are vital, trend-setting techniques. Many papers were mentioned about the SEM, it has been successfully applied to a variety of areas, such as: exploring the education research technology; relation with teacher burnout and job satisfaction.

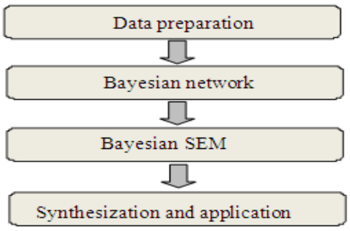

Fig.1: Structure equation model process.

However, no paper mentions Linking Bayesian networks and Bayesian Approach for Structural Equation Modeling. We proposed method shown in Fig.1. There are four steps in this paper. Firstly, we prepare data by using simulation data. Secondly, we use WEKA to obtain a DAG through Bayesian network classifiers with a TAN algorithm. Thirdly, Based on the DAG, the SEM path modeling phase can be implemented with AMOS program. Finally, application are conducted to support better decision-making and to apply in problem-solving.

## 3. Bayesian networks, Bayesian Approach and Structural Equation Modeling

The algorithm of data mining includes K-means clustering, decision trees, Bayesian networks, regression models, and so on. Data mining algorithms are supported by the WEKA program. Among data mining algorithm, the use of the Bayesian network can produce a DAG that models causal relations between attributes. Additionally, the AMOS program can produce graphic path modeling with latent variables. It is a software application to be used for structural equation modeling with a user-friendly graphical interface.

### 3.1. Bayesian networks

Bayesian network or Bayesian belief network or a Belief Network is a short form of graphical models. It is caused by a combination of probability theory and graph theory is connected together using the theory of probability [5,6,7].

A Bayesian network (BN) was configured as a probability distribution condition (Condition Probability Distribution) of each node. A Bayesian network is a directed acyclic graph (DAG) that consists of a set of nodes connected by arcs. The nodes represent the attributes, and the arcs stand for relationships among the connected attributes [8].

If node A has a branch connected to node B, so the variable's B will report directly to the variable's A, and A is the parent of B and for each of the variable's $X_i$, i = 1 to n, so the joint distribution of the variables is the product of the local distribution is shown in the following equation.

$$\Pr(X_1,\ldots,X_n) = \prod_{i=1}^{n} \Pr(X_i \mid parents(X_i)) \qquad (1)$$

$\Pr(X_1,\ldots,X_n)$ = The probability of a joint distribution.

$\Pr(X_i \mid parents\ (X_i))$ = Conditional probability: for variable $X_i$ and the set of parent nodes of the variable $X_i$

$parents(X_i)$ = The set of parent nodes of the variable $X_i$.

In this paper, we construct Bayesian network by using the WEKA program. It offers various algorithms such as: K2, HillCiimber, RepeatedHillClibe, SimulatedAnnealing, TabuSearch, GeneticSearch and TAN. The TAN algorithm can produce a graph in which the class attribute treated as the only and greatest parent node of all other nodes is located at the top in the DAG and determines the maximum weight spanning tree. [9].

In general, there are five classes of BN classifiers: Naïve-Bayes, Tree augmented Naïve-Bayes(TANs), Bayesian network augmented Naïve-Bayes (BANs), Bayesian multi-nets and general Bayesian networks (GBNs).

### 3.1.1 Tree Augmented Naïve-Bayes (TAN)

TAN classifiers extend Naïve-Bayes by allowing the attributes to form a tree. An attribute node can have only one parent from another attribute node. In Fig. 3, *A* is the parent node in which it directly points to all attribute nodes such as $B_1$, $B_2$, $B_3$, $B_4$. The child node can point to other children [10,11].

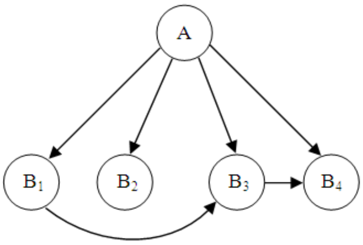

Fig. 3: A simple TAN structure

## 3.2. Bayesian Approach

The Bayesian refers to the 18th century mathematician and theologian Thomas Bayes that provided the first mathematical treatment of a non-trivial problem of Bayesian inference [14]. Bayesian Approach is based on the probability theory. It is applied to calculate the posterior from the prior and the likelihood, because the latter two is generally easier to be calculated from a probability model. The joint probability of two events, A&B, can be expressed as

$$\begin{aligned} P(A \cap B) &= P(A \mid B)P(B) \\ &= P(B \mid A)P(A) \end{aligned} \qquad (2)$$

(3)

In Bayesian probability theory, one of these "events" is the hypothesis, H, and the other is data, D, and we wish to judge the relative truth of the hypothesis given the data. According to Bayes' rule, we do this via the relation

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)} \qquad (4)$$

The term P(D|H) is called the likelihood function, and it assesses the probability of the observed data arising from the hypothesis. The term P(H) is called the prior probability, as it reflects one's, prior knowledge before the data are considered. The term P(D) is obtained by integrating (or summing) P(D|H)P(H) over all H, and usually plays the role of an ignorable normalizing constant. Finally, the term P(H|D) is known as the posterior probability, and as its name suggests, reflects the probability of the hypothesis after consideration of the data.

## 3.3. Structural Equation Modeling (SEM)[12]

SEM models take account of statistical methodologies that allow us to estimate the causal relationships linking two or more latent composite indicators. Several observed indicators usually defined as Manifest Variables (MV).Causal relationships among the latent concepts called Latent Variables (LV).SEM represents a joint-point between the path analysis and the Confirmatory Factor Analysis.

## 4. Empirical Study Path Analysis

In this section, the empirical studies that demonstrate the application of the method proposed for the analysis of the causes. Phase 1 is data preparation. This paper adopts the Computer Hardware Data Set selected from the UCI Machine Learning Repository [13]. As shown in Table 1, this data set has 50 instances regarding relative CPU performance between sellers. This paper sets out to from model the relationships between six independent attributes (MYCT, MMIN, MMAX, CACH, CHMIN, and CHMAX) and the dependent attribute (PRP).

Table 1 Attribute information.

| Attribute Name | Data Type | Description |
|---|---|---|
| MYCT | Integer | Machine cycle time in nanoseconds |
| MMIN | Integer | Minimum main memory in kilobytes |
| MMAX | Integer | Maximum main memory in kilobytes |
| CACH | Integer | Cache memory in kilobytes |
| CHMIN | Integer | Minimum channels in units |
| CHMAX | Integer | Maximum channels in units |
| PRP | Integer | Published relative performance |

In phase 2, the Bayesian network classifier with the TAN search algorithm is implemented with WEKA. As a result, it is shown in Fig. 4. It is apparent that CACH (Cache memory in kilobytes) is the most important factor, because it plays a core role, and one that further affects CHMAX (Maximum channels in units) and CHMAX directly affect CHMIN (minimum channels in units). However, the CACH does not directly affect MMIN (Minimum main memory in kilobytes), MMAX (Maximum main memory in kilobytes), and MYCT (Machine cycle time in nanoseconds).

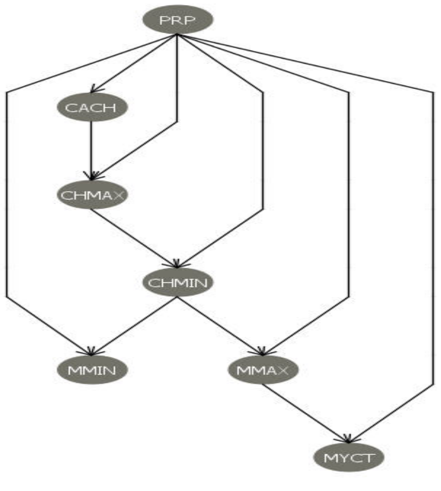

Fig. 4: The causal attribute relationship diagram.

Phase 3, we use the causal attribute relationship diagram (Fig.4). It helped us to decide the causal directions between attributes. Consequently, after setting up the causal directions, implementation of Bayesian SEM can be promptly performed. As a result, Fig. 5 is produced to display the relations between attributes. Each of the independent attributes and the dependent attribute is measured by direct effect value. Table 2 shows path coefficients between attributes. In particular, the highest path coefficient is the MMAX → MMIN of 0.89 and MMIN → PRP of 0.72. The PRP shows $R^2$ value (91%) in this model. On the whole, the combination of MMAX, MMIN, CHMIN, CACH, CHMAX, and MYCT has predictive ability for 91% of the PRP. If non-significant path coefficients are removed such as MMAX → PRP and

CHMAX → PRP, then Fig. 6 exhibits relations between attributes and PRP shows $R^2$ value (89%) in this model. We can see the $R^2$ values decreased, but we can accept it because of its high value.

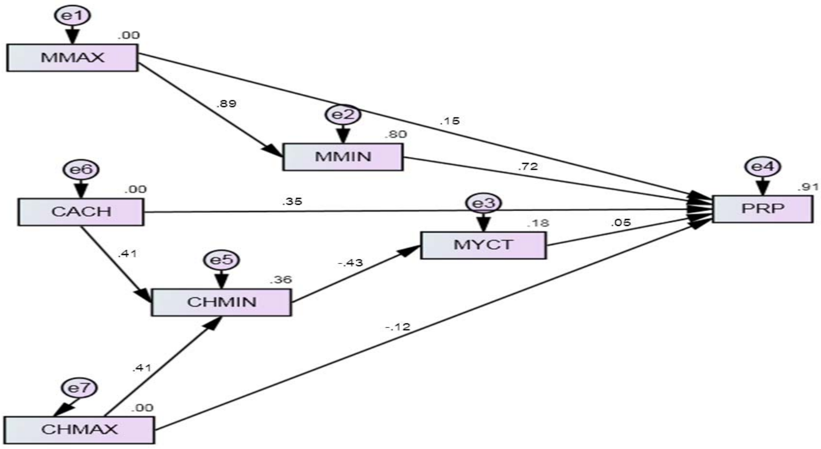

Fig. 5: Relations between attributes.

Table 2: Path coefficients

| | Original sample (O) | Sample mean (M) | Standard deviation (STDEV) | Standard error (STERR) |
|---|---|---|---|---|
| MMAX→PRP | 0.15 | 0.002 | 0.002 | 0.000 |
| MMAX→MMIN | 0.89 | 0.363 | 0.028 | 0.000 |
| CACH→PRP | 0.35 | 1.374 | 0.272 | 0.001 |
| CACH→CHMIN | 0.41 | 0.047 | 0.019 | 0.000 |
| CHMAX→PRP | -0.12 | -0.977 | 0.579 | 0.003 |
| CHMAX→CHMIN | 0.41 | 0.101 | 0.041 | 0.000 |
| MMIN→PRP | 0.72 | 0.024 | 0.004 | 0.000 |
| CHMIN→MYCT | -0.43 | -13.472 | 3.886 | 0.019 |
| MYCT→PRP | 0.05 | 0.055 | 0.060 | 0.000 |

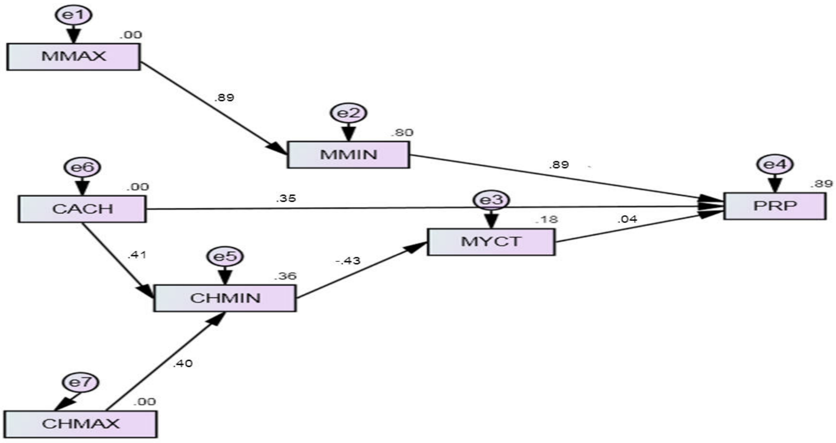

Fig. 6: Significant relations between attributes.

## 5. Implications and conclusions

The causes of this analysis can improve the quality of decision-making and facilitate the change and the strategic objectives into effective action. Many papers have suggested cause analysis techniques to achieve knowledge about the cause. Among the analytical techniques, the SEM is one of the most effective and popular techniques to diagnose. When using the SEM in a lack of support for this theory, under the serious problems that may occur. Especially, we do not know linking between attributes due to lack of knowledge of the background. The assumption of maximum likelihood methods is made constant variance normal disturbances. So sample size must be a large number, but we need small sample size. In an effort to solve this problem, this article offers an effective way of pursuing the cause analysis, as the authors point out the Bayesian network and Bayesian approach prior to the SEM. This paper was to demonstrate the successful application of the proposed method.

We proposed method combination of the Bayesian network with the TAN search algorithm and Bayesian approach in SEM for causal analysis. This methodology is doing well in that the task of deciding the causal directions between attributes becomes easy through use of the proposed method and small sample size. This study has some limitations when we use the TAN search algorithm. Data type of TAN algorithm must be an integer.

## 6. References

[1] R.P. Bagozzi. *Causal models in marketing.* New York: Wiley, 1980.

[2] http://www.statsoft.com/textbook/structural-equation-modeling/

[3] J.P. LeSage. Bayesian estimation of spatial regression models.np, 2004.

[4] C. J. Lin, and W.W. Wu. A causal analytical method for group decision-making under fuzzy environment. 2008, 34(1): 205–213.

[5] G. Quer, H. Meenakshisundaram, B. Tamma, B.S. Manoj, R. Rao, and M. Zorzi. Cognitive Network Inference through Bayesian Network Analysis*. GLOBECOM 2010, 2010 IEEE Global Telecommunications Conference*. 2010: 1-6.

[6] Bo. Chen, Liao,Qin and Tang, Zhonghua. A Clustering Based Bayesian Network Classifier. *Fourth International Conference on* **.** 2007, 4(1): 444-448.

[7] Wong, Man Leung and Leung, Kwong Sak . An efficient data mining method for learning Bayesian networks using an evolutionary algorithm-based hybrid approach. *IEEE Transactions on Evolutionary Computation.* 2004, 8(1): 378-404.

[8] [8] E. R. Hruschka, and N. F. F. Ebecken. Towards efficient variables ordering for Bayesian networks classifier. *Data and Knowledge Engineering*. 2007, 63(2): 258–269.

[9] [9] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory.*1968, IT-14: 426-467.

[10] S. Hong-bo, W. Zhi-Hai, H. Hou-Kuan, and J. Li-Ping. Text classification based on the TAN model*. TENCON '02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering.* 2002*,* 1(2): 43-46.

[11] *http://webdocs.cs.ualberta.ca/~jcheng/Doc/cscsi.pdf*

[12] Q. Yue; L. Hai-lin and J. Luan. Data Mining in the Relation between Ownership Structure and Firm Performance: A Structural Equation Model Analysis of Chinese Public Companies in Manufacturing Industry. *International Conference on Computer Science and Software Engineering*. 2008, 4(1): 297-300.

[13] http://archive.ics.uci.edu/ml/

[14] Stigler, and M. Stephen. *The history of statistics.* Harvard University press, 1986.