2014 Seventh International Joint Conference on Computational Sciences and Optimization

Construction of Probabilistic Boolean Network for Credit Default Data

Ruochen Liang Yushan Qiu* Wai-Ki Ching
The Advanced Modeling and Applied Computing Laboratory
Department of Mathematics
The University of Hong Kong, Hong Kong, China
Email: h1181227@hku.hk
Email: yushanqiu2526374@163.com
Email: wching@hku.hk

***Abstract*—In this article, we consider the problem of construction of Probabilistic Boolean Networks (PBNs). Previous works have shown that Boolean Networks (BNs) and PBNs have many potential applications in modeling genetic regulatory networks and credit default data. A PBN can be considered as a Markov chain process and the construction of a PBN is an inverse problem. Given the transition probability matrix of the PBN, we try to find a set of BNs with probabilities constituting the given PBN. We propose a revised estimation method based on entropy approach to estimate the model parameters. Practical real credit default data are employed to demonstrate our proposed method.**

***Keywords*-Boolean Networks; Probabilistic Boolean Networks; Inverse Problem; Transition Probability Matrix;**

I. Introduction

In this article, we focus on the problem of construction of Probabilistic Boolean Networks (PBNs). It is well known that Boolean Networks (BNs) [12], [13] and PBNs have a lot of applications in the field of modeling genetic regulatory networks [14], [15], [16] and credit default data [11]. BN, as a deterministic model, was first proposed by Kauffman [12], [13]. Later, Shmulevich extended the BN model to a stochastic setting, PBNs [14], [15]. A PBN can be regarded as a Markov chain process. Both construction and control of PBNs are important issues and they have been well studied in [3], [4]. However, the construction of a PBN is an ill-posed inverse problem. This means that it may have many solutions or no solution. Our goal is to identify a set of BNs with probabilities constituting the given PBN based on the given transition probability matrix of the PBN. We propose a revised estimation method integrating with entropy approach to estimate parameters of the model. Furthermore, numerical examples are given to demonstrate the effectiveness and efficiency of our proposed method by utilizing practical real credit default data.

The investigation of the relationship between correlated defaults of different industrial sectors and business cycles has become an important challenge in financial risk, especially after the financial credit crisis in 2007-08. Thus, a number of infectious models [5], [6], [7], [8] and multivariate Markov chain models [17] have been proposed to tackle the problem. PBN approach was first proposed by Gu et al. [11] to study the correlated defaults in a credit default data set. Using the credit default data, they formulate a PBN model for explaining the default structure and making reasonably good predictions of joint defaults in different sectors. The key idea is to decompose the transition probability matrix of the PBN inferred from the real data into a weighted average of several deterministic BNs, which contain useful information about business cycles. It is well known that given an initial state, a BN will eventually enter into a cycle of states, called attractor cycle or limit cycle. Thus, the business cycle can be described by using such limit cycles. Furthermore, heuristic algorithm have been proposed to solve this inverse problem effectively [11]. In this article, we shall modify the algorithm and show that our proposed algorithm is more efficient and better results can be achieved.

II. Introduction to Boolean Networks (BNs) and Probabilistic Boolean Networks (PBNs)

In this section, we give a brief introduction to Boolean Networks (BNs) and Probabilistic Boolean Networks (PBNs).

A. Boolean Networks

BNs are deterministic models proposed by Kauffman [12], [13]. BN models are popular mathematical model for formulating genetic regulatory networks. In a BN, the genes are regarded as vertices and each vertex has two possible states: 0 (not expressed) and 1 (expressed). The output (target vertex) of each gene is determined by several genes called its input genes by using a Boolean function. A BN is said to be well defined if its input vertices and their corresponding Boolean functions are also given.

Generally speaking, a BN $G(V, F)$ is represented by a set of vertices

$$V = \{v_1, v_2, \ldots, v_n\}$$

and a set of Boolean functions

$$F = \{f_1, f_2, \ldots, f_n\}.$$

Here we have $f_i : \{0,1\}^n \to \{0,1\}$. And we define $v_i(t)$ to be the state of vertex i at time t, taking 0 or 1. The rules of

978-1-4799-5372-1/14 $31.00 © 2014 IEEE
DOI 10.1109/CSO.2014.11

States	$v_1(t)$	$v_2(t)$	$f^{(1)}$	$f^{(2)}$
1	0	0	1	1
2	0	1	1	0
3	1	0	1	1
4	1	1	0	0

Table I
THE TRUTH TABLE OF A BN.

the network of vertices can be represented by the Boolean functions:

$$v_i(t+1) = f_i(\mathbf{v}(t)),\ i = 1, 2, \ldots, n.$$

The Boolean vector

$$\mathbf{v}(t) = (v_1(t), v_2(t), \ldots, v_n(t))$$

can take any possible states in the Gene Activity Profile (GAP) set

$$S = \{(v_1, v_2, \ldots, v_n)^T : v_i \in \{0, 1\}\}.$$

We give an example of a two-vertex BN and it is described in Table 1.

There are four states in the BN and they are $(0,0), (0,1), (1,0)$ and $(1,1)$. One may label them by $1, 2, 3$ and 4. We note that if the current state of the network is 1, then the network will go to State 4 in the next step (with probability one) and vice versa. If the current state is 2, the network will go to State 3 in the next step (with probability one). Eventually, the BN will go into the cycles of two states: $(0,0)$ (State 1) and $(1,1)$ (State 4).

The network dynamics of the BN (truth table) can be represented by using a Boolean transition probability matrix, we call it a BN matrix, as follows:

$$B = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}. \tag{1}$$

The matrix B is a column stochastic matrix, each column sum is one. We remark that there is a one-to-one relation between a BN (truth table) and its corresponding BN matrix. It is straightforward to see that for a given initial state, a BN will eventually evolve into a cycle of states called its attractor cycle.

B. Probabilistic Boolean Networks

Unavoidably, data used for inferring a BN may have significant level of noise, therefore it is more appropriate and desirable to employ a stochastic model. In [14], [15], [16], the concept of a BN, a deterministic model, is extended to a PBN, a probabilistic model. The main idea can be explained as follows. For each of the network vertices, we allow more than one Boolean function to be associated with it. We also assume that there is selection probability distribution associate with the Boolean functions. We remark that the dynamics (transitions) of a PBN can be investigated by using Markov chain theory [2].

To extend a BN to a stochastic model, for each vertex v_j in a PBN, instead of allowing only one Boolean function, we assume that there are a number of Boolean functions $f_i^{(j)}(i = 1, 2, \ldots, l(j))$ to be selected for determining the next state of the gene v_j. The probability of selecting $f_i^{(j)}$ to be the Boolean function is assumed to be $c_i^{(j)}$ and clearly we have

$$0 \le c_i^{(j)} \le 1 \quad \text{and} \quad \sum_{i=1}^{l(j)} c_i^{(j)} = 1 \quad \text{for} \quad j = 1, 2, \ldots, n.$$

Let f_i be the ith possible realization and

$$f_i = (f_{i_1}^{(1)}, f_{i_2}^{(2)}, \ldots, f_{i_n}^{(n)}), \quad 1 \le i_j \le l(j), \quad j = 1, 2, \ldots, n. \tag{2}$$

If the selection process of the Boolean functions f_{i_j} for each gene j is an independent process, then the probability of getting the BN having Boolean functions

$$(f_{i_1}, f_{i_2}, \ldots, f_{i_n})$$

is given by

$$q_{i_1 i_2 \cdots i_n} = \prod_{j=1}^{n} c_{i_j}^{(j)}.$$

For the underlying PBN, there are $N = \prod_{j=1}^{n} l(j)$ different possible BN realizations. The transition process of the network states forms a Markov chain process. If we let $\mathbf{a}$ and $\mathbf{b}$ be any two states in S, then the transition probability is given by

$$\mathrm{P}\{\mathbf{v}(t+1) = \mathbf{a} \mid \mathbf{v}(t) = \mathbf{b}\} = \sum_{i=1}^{N} \mathrm{P}\{\mathbf{v}(t+1) = \mathbf{a} \mid \mathbf{v}(t) = \mathbf{b}, \text{the } i\text{th BN is selected}\} \cdot q_i.$$

If we let

$$q_i = q_{i_1 i_2 \cdots i_n} \quad \text{and} \quad i = i_1 + \sum_{j=2}^{n} \left((i_j - 1)(\prod_{k=1}^{j-1} l(k)) \right)$$

then it can be shown that the transition probability matrix of the Markov chain can be written as

$$A = \sum_{i=1}^{N} q_i A_i \tag{3}$$

where A_i is the BN matrix of the ith BN and q_i is the selection probability.

III. CONSTRUCTION OF PBN

In this section, we first introduce the problem of construction of PBNs, then present a revised heuristic construction algorithm. PBN has achieved many attention as it can be used to model genetic regulatory networks and credit default data. Ching et al. [3] proposed algorithms of generating PBNs from a given transition probability matrix A which can be written as the sum of the BN matrices A_i as in Eq.

(3). The construction problem itself is known to be an ill-posed inverse problem owing to the fact that there are many possible solutions or no solution.

A. The Inverse Problem

Given a transition probability matrix A of size $2^n \times 2^n$ (n is the number of genes), we assume the following representation:

$$A = \sum_{i=1}^{M} q_i A_i + \epsilon.$$

Here $\{A_i\}_{i=1}^{M}$ is a set of BNs and q_i is the corresponding selection probability of A_i and ϵ is the residual part of A. One may regard A_i as the major component of the transition probability matrix A associated with a weighting of q_i. The residual ϵ is the noise for the transition probability matrix A and $||\epsilon||_F$ shall be sufficiently small.

In [3], the construction was formulated as the following minimization problem:

$$\min_{q_i} \left\{ -\sum_{i=1}^{M} q_i \log q_i \right\}$$

subject to the constraints in Eq. (3) and also

$$\sum_{1}^{M} q_i = 1 \quad \text{and} \quad q_i \geq 0.$$

The inverse problem has been shown to have an unique solution [1], [18], however, the computational cost is huge. Therefore heuristic algorithms based on optimizing the entropy have been proposed to better address the problem [3], [9], [11].

B. The Heuristic Algorithms

The following algorithm in Table 2, we call it "Uniform" which has been proposed in [3] and also adopted in [11] to solve the problem with LIMIT and THRESHOLD being 0.

The algorithm is allowed to iterate for a fixed number of times (say for example 1000), and we compute and record the best entropy of solution $\mathbf{q}$ obtained. Furthermore, we suggest to set a "LIMIT" for the residual of the PBN and also a "THRESHOLD" level for the selection probability. These can both avoid the algorithm from constituting of BNs with too small selection (not significant) probabilities. Most importantly, we suggest to modify (*) in the above algorithm by

$$p_{kij} = \frac{[R_k]_{ji}^{\alpha}}{[R_k]_{1i}^{\alpha} + [R_k]_{2i}^{\alpha} + \ldots + [R_k]_{mi}^{\alpha}}$$

with $\alpha > 1$. It should be noted that $\max\{p_{kij}\}$ increases as α increases. In our numerical experiments, we adopt

$$\text{LIMIT} = 0.001, \text{THRESHOLD} = 0.0001, \alpha = 2.$$

We shall call our newly revised algorithm as "Quadratic".

Step 0: Set $R_1 = A, k = 0$, (initial condition)
Input LIMIT and THRESHOLD,
Step 1: $k := k + 1$
Step 2: We assume in the ith column of matrix R_k,
there are totally m non-zero entries
$[R_k]_{1i}, [R_k]_{2i}, \ldots, [R_k]_{mi}$.
Then we define the probability of choosing $[R_k]_{ji}$ to be p_{kij}
and $p_{kij} = \frac{[R_k]_{ji}}{[R_k]_{1i}+[R_k]_{2i}+\ldots+[R_k]_{mi}} - -(*)$.
After choosing entries based on the probablity defined above,
assume the concerned entries are given by
$[R_k]_{k_1,1}, [R_k]_{k_2,2}, \ldots, [R_k]_{k_{2^n},2^n}$,
we choose the smallest entry q_k from $[R_k]_{k_i,i} (i = 1, \ldots, 2^n)$.
Check: If $q_k <$ THRESHOLD, go to **Step 5**.
Then we define the following BN matrix:
$A_k = [e_{k_1,1}, \ldots, e_{k_{2^n},2^n}]$.
Here $e_{j,i}$ is the unit column vector whose jth entry is 1
for $i = 1, \ldots, 2^n$.
Step 3: $R_{k+1} = R_k - q_k A_k$
Step 4: If $||R_{k+1}||_F <$ LIMIT go to **Step 5**,
otherwise go to **Step 1**.
Step 5: $M = k$ and $A = \sum_{k=1}^{M} q_k A_k$.

Table II
THE HEURISTIC ALGORITHM ([3], [11]).

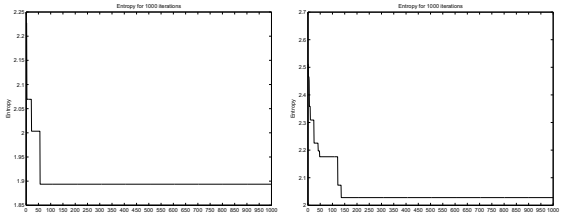
IV. NUMERICAL RESULTS

In this section, we present some numerical results of applying the proposed revised algorithm presented in Section 3 to the PBN construction problem. We employ real default data extracted from the figures in [10]. The default data come from four different sectors including consumer sector, energy sector, media sector and transportation sector. We can see the default data taken from [10] in Table 3. All the data sets are quarterly (88 quarters) time series on the number of defaults in the captured sectors. To construct a PBN, we need to consider binary data: 0 representing having no default observed and 1 stands for at least one default. In [11], they only consider 4 possible combinations among the 4 sectors. Here we also present the results for the case of 4 sector, respectively.

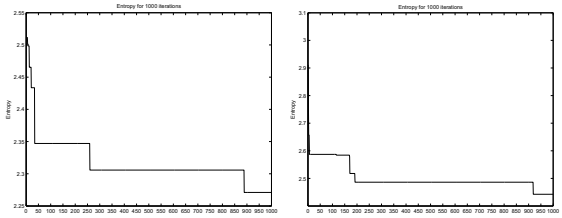
Sectors	Total	Defaults
Consumer	1041	251
Energy	420	71
Media	650	133
Transport	281	59

Table III
THE DEFAULT DATA (TAKEN FROM [10]).

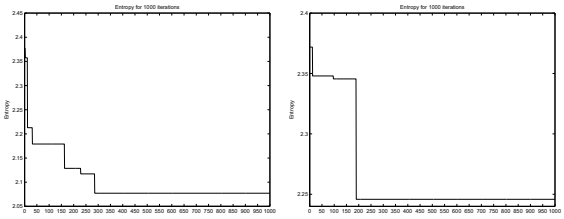
Figure 1 presents the optimal entropy values obtained by QUADRATIC and UNIFORM against the number of iterations in all the five cases. From the numerical results, we



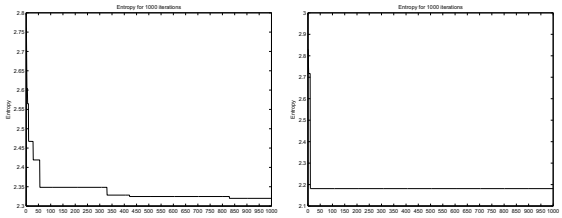
Entropy results (3-sector Consumer-Energy-Media): Quadratic (left) Uniform (right)



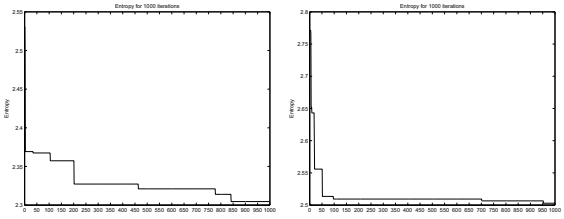
Entropy results (3-sector Consumer-Energy-Transport): Quadratic (left) Uniform (right)



Entropy results (3-sector Consumer-Media-Transport): Quadratic (left) Uniform (right)

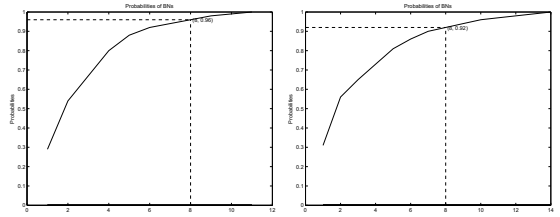


Entropy results (3-sector Energy-Media-Transport): Quadratic (left) Uniform (right)

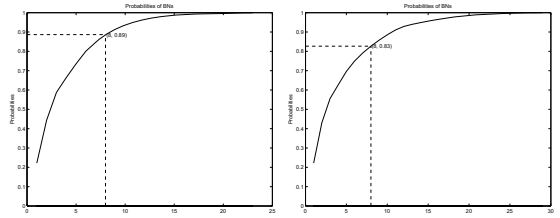


Entropy results (4-sector) Quadratic (left) Uniform (right)

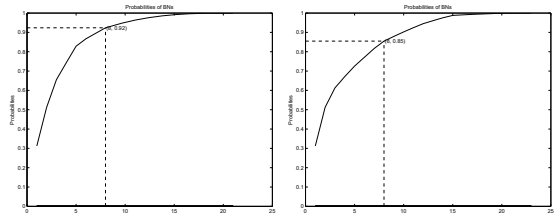
Figure 1



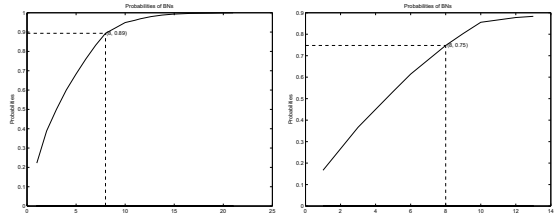
Probability results (3-sector Consumer-Energy-Media): Quadratic (left) Uniform (right)



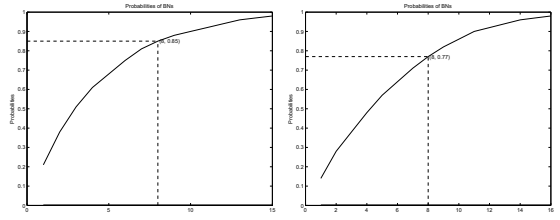
Probability results (3-sector Consumer-Energy-Transport): Quadratic (left) Uniform (right)



Probability results (3-sector Consumer-Media-Transport): Quadratic (left) Uniform (right)



Probability results (3-sector Energy-Media-Transport): Quadratic (left) Uniform (right)



Probability results (4-sector) Quadratic (left) Uniform (right)

Figure 2

can see that the "optimal" entropy values obtained by our Quadratic method are small (better) in most cases. Figure 2 shows the solutions (the accumulated probabilities from the largest to the smallest) obtained by both methods. It has been shown that our method can produce solution with larger weighting as the curve is more "concave". The dashed lines and text labels on the plots in Figure 2 show the proportion captured by the first eight BNs. From these labels we can see that Quadratic method outperforms by capturing at least 85% using eight BNs. From the perspective of computational times, Table 4 has shown that our proposed algorithm is

Table IV
COMPUTATIONAL TIME IN SECONDS

	Quadratic	Uniform
3-sector Consumer-Energy-Media	2.0	4.0
3-sector Consumer-Energy-Transport	5.0	6.0
3-sector Consumer-Media-Transport	3.0	4.0
3-sector Energy-Media-Transport	5.0	5.0
4-sector	6.0	8.0

more efficient.

V. CONCLUSION

The effect of control variables LIMIT and THRESHOLD can reduce both running time and storage used for iteration computation. Meanwhile, the proportion of all BNs captured by our algorithm is higher than 98% (The sum of all probabilities in a test). Since we are dealing with mostly very sparse matrices, the "Quadratic" algorithm actually enhances the probability of non-zero entries, i.e., larger numbers having larger probabilities are chosen in the comparison to the "Uniform" algorithm. This helps us efficiently in picking up the "dominant" BNs in each iteration and hence further reduces running time and the number of BNs used to approximate. Our proposed revised method outperform the other methods in terms of the numerical results [11].

However, this "Enhance" effect is very closely related to the distribution of non-zero entries in the columns of the transition matrix. The more sparse the matrix is, the better the effect will be.

As a further research issue, we shall study the effect of the parameters, "LIMIT, "THRESHOLD" and α, on the overall performance of the revised algorithm. We shall also develop theory for the selection of these parameters.

ACKNOWLEDGMENT

Research supported in part by HKRGC Grant, HKU-CRGC Grants, HKU Strategy Research Theme fund on Computational Sciences, Hung Hing Ying Physical Research Sciences Research Grant.

REFERENCES

[1] X. Chen, W. Ching, X.S. Chen, Y. Cong and N. Tsing, "Construction of Probabilistic Boolean Networks from a Prescribed Transition Probability Matrix: A Maximum Entropy Rate Approach," East Asian Journal of Applied Mathematics, vol. 1, 2011, pp. 132-154.

[2] W. Ching and M. Ng, "Markov Chains : Models, Algorithms and Applications", International Series on Operations Research and Management Science, Springer, New York, 2006.

[3] W. Ching, X. Chen and N. Tsing, (2009), "Generating Probabilistic Boolean Networks from a Prescribed Transition Probability Matrix", IET Systems biology, vol.3, 2009, pp. 453-464.

[4] W. Ching, S. Zhang, Y. Jiao, T. Akutsu, N. Tsing and A. Wong, "Optimal Control Policy for Probabilistic Boolean Networks with Hard Constraints", IET on Systems Biology, vol. 4, 2009, pp. 90-99.

[5] W. Ching, H. Leung, H. Jiang, L. Sun and T. Siu, "A Markovian Network Model for Default Risk Management", International Journal of Intelligent Engineering Informatics, vol. 1, 2010, pp. 104-124.

[6] J. Gu, W. Ching and T. Siu, "A Markovian Infectious Model for Dependent Default Risk", International Journal of Intelligent Engineering Informatics, vol, 2011, pp. 174-195.

[7] M. Davis and V. Lo, "Infectious Defaults", Quantitative Finance, vol,1, 2001, pp. 382-387.

[8] M. Davis and V. Lo, "Modeling Default Correlation in Bond Portfolio", In C. Alescander (ed.) Mastering Risk Volume 2: Applications. Financial Times, Prentice Hall, 2001, pp. 141-51.

[9] H. Jiang, X. Chen, Y. Qiu and W. Ching, "On Generating Optimal Probabilistic Boolean Networks from a Set of Sparse Matrices", East Asian Journal of Applied Mathematic, vol. 2, 2012, pp.353-372.

[10] G. Giampieri and M. Davis and M. Crowder, "Analysis of Default Data Using Hidden Markov Models", Quantitative Finance, vol. 5, 2005, pp. 27-34.

[11] J. Gu, W. Ching, T. Siu and H. Zheng, "On Modeling Credit Defaults: A Probabilistic Boolean Network Approach", Risk and Decision Analysis, vol. 4, 2013, pp. 119-129.

[12] S. Kauffman, "Metabolic Stability and Epigenesis in Randomly Constructed Gene Nets", Journal of Theoretical Biology, vol. 22, 1969, pp. 437-467.

[13] S. Kauffman, "Homeostasis and Differentiation in Random Genetic Control Networks", Nature, vol. 224, 1969, pp. 177-178.

[14] I. Shmulevich, E. Dougherty, S. Kim and W. Zhang, "Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks", Bioinformatics, vol. 18, 2002, pp. 261-274.

[15] I. Shmulevich, E. Dougherty, S. Kim and W. Zhang, "From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks", Proceedings of the IEEE, vol. 90, 2002, pp. 1778-1792.

[16] I. Shmulevich and E. Dougherty, "Probabilistic Boolean Networks: The Modeling and Control of Gene Regulatory Networks", SIAM, US, 2010.

[17] T. Siu, W. Ching, M. Ng and E. Fung, "On a Multivariate Markov Chain Model for Credit Risk Measurement", Quantitative Finance, vol. 5, 2005, pp. 543-556.

[18] S. Zhang, W. Ching, X. Chen and N. Tsing, "Generating Probabilistic Boolean Networks from a Prescribed Stationary Distribution", Information Sciences, vol.180, 2010, pp. 2560-2570.