A COMPARATIVE STUDY ON DYNAMIC BAYESIAN NETWORKS BASED ON STATE SPACE MODELS FOR GENE NETWORK RECONSTRUCTIONS

Y.C. Liu and C.M. Chen

Institution of Biomedical Engineering, NTU, Taipei, Taiwan

F90548051@ntu.edu.tw

Abstract: The capabilities of three dynamic Bayesian network methods to infer latent biological actions have been evaluated in this work by biologically reasonable synthesised data. The results show that performances of DBNs methods are proportional to sample size and VBEM method is more stable than MAPEM and SO-LDS methods. Furthermore, five replicates would be sufficient to achieve a stable performance.

Introduction

There has been growing interest in inferring genetic regulatory networks based on microarray data recently. One of the most promising methodologies is Dynamic Bayesian network (DBN), which has been adopted by several previous works [1, 2, 3, 4, 6, 7] for reconstruction of genetic networks from time-series data. DBNs have the great capability of taking into account the influence of latent variables while estimating the gene-gene interactions. The latent variables are any unobservable factors contributing to the observed gene expressions. Nevertheless, since only a small number of replications (e.g., less than 4) and a limited amount of time points are available in practice, estimation of model parameters in a DBN may not be robust and reliable statistically.

This study aims to evaluate the effect of sample size and normality assumption on the performance of three DBNs. The sample size is defined as the product of the number of replications and the number of time points in each replication in a microarray study. Two DBNs considered in this study are proposed by Rangel et al. [2] and Beal et al. [3], using the same dynamic system structure to infer genetic behavior but different learning strategies. While the former is based on Expectation-Maximization (EM) algorithm plus bootstrap approach, which is called MAPEM for convenience, the latter is rooted on Variational Bayesian Expectation-Maximization (VBEM) algorithm. The third DBN evaluated is proposed by Perrin et al. [6], called SO-LDS thereinafter, which describes genetic regulation by second order differential model. The model combines two genetic behavior parameters, which are absorption coefficients and natural frequency, to describe genetic temporal activities. It may be considered as a dynamic Bayesian network based on non-input state-space model. The common feature of these three methods is that they all can be accomplished by Kalman Filter and Kalman smoother process.

Materials and Methods

Bayesian network is one of the graphical models, which can elucidate both causal and diagnostic reasoning problems. As shown in Figures 1(a) and 1(b), a graphical model is defined by a structure, denoted as *M*, which comprises a set of nodes, $\nu = \{X_1, X_2, ..., X_n\}$, representing different variables, and a set of edges, $\varepsilon = \{(X_1, X_3), (X_2, X_3), (X_2, X_4), (X_3, X_5)\}$, indicating the "cause-effect" relationship of two variables. A Bayesian network is a directed acyclic graph, showed as Figure 1(a).

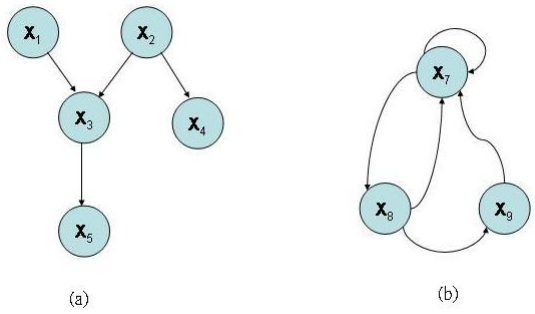

Figure 1: Examples of graphical models

To model biological systems, in a graph, nodes represent biological expressions (i.e., gene expression level), assuming all nodes are continuous variables subject to an underlying distribution and independent of one another. Edges indicate interaction of two biological molecular compounds, which can be assessed by some quantitative technologies, e.g. microarray. The joint probability of a structure under first Markov assumption may be expressed as equation (1),

$$p(X_1, ..., X_n) = \prod_i p(X_i \mid p_i) \qquad (1)$$

where p_i indicate the parent of X_i. Static Bayesian networks can not describe the recurrent or cyclic structure such as Figure 1(b). However, dynamic Bayesian networks can overcome this deficiency and may model causal relation with time lags. Rangel et al.

ISSN: 1727-1983

[2] and Beal et al. [3] expressed DBN models by input dependent State-Space model (2), abbreviated as SS model thereinafter, and the system dynamic structure is illustrated in Figure 2. The 'gene-gene' interactions matrix, that is $CB + D$ matrix, is then found by MAPEM and VBEM algorithm, respectively.

$$\begin{aligned} x_t &= Ax_{t-1} + Bu_t + w_t \\ y_t &= Cx_t + Du_t + v_t \end{aligned} \tag{2}$$

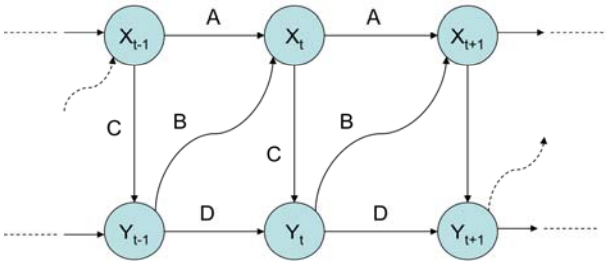

Figure 2: input-dependent SS model.

In Equation (2), x_t is a hidden state vector, e.g. unobserved molecular expressions, t time index, y_t an observation vector, e.g. observed gene expressions, and u_t an input vector. The random vector w_t represents zero-mean Gaussian noise with covariance Q. The random vector v_t stands for zero-mean Gaussian noise with covariance R of observation. The parameters A, B, C, D, are referred as state dynamic, input-to-state, observation and input-to-observation matrices, respectively. All parameters mentioned above are collected by a parameter vector $\theta = \{A, B, C, D, Q, R\}$. By first-order Markov assumption the complete data joint likelihood of $\{Y_{1:T}\}$ based on state-space model (2) is given by

$$\begin{aligned} &p(x_{1:T}, y_{1:T}) \\ &= p(x_1 \mid u_1) p(y_1 \mid u_1) \prod_{t=2}^{T} p(x_t \mid x_{t-1}) p(y_t \mid u_t) p(y_t \mid x_t) p(x_t \mid u_t) \end{aligned} \tag{3}$$

where $p(x_t \mid x_{t-1})$ denotes the probability of dynamic state transitions , $p(y_t \mid x_t)$ the probability of hidden state x_t generating observation y_t at time t , $p(x_t \mid u_t)$ and $p(y_t \mid u_t)$ probability of hidden state and observation conditioned on input u_t , respectively. The joint log-likelihood of Equation (3) can express as Equation (4).

$$\begin{aligned} &\ln p(x_{1:T}, y_{1:T}) \\ &= \sum_{t=1}^{T} \ln[p(x_t \mid x_{t-1}) p(y_t \mid u_t) p(y_t \mid x_t) p(x_t \mid u_t)] \end{aligned} \tag{4}$$

The parameters in SS model (2) may be estimated by the forward-backward recursive algorithm, i.e. Kalman Filter and Kalman smoother coupled-process, presented in [5]. The former computes expectation of x_t given $\{Y_{1:t}\}$, i.e., $E[x_t \mid \{Y_{1t}\}]$ and the latter calculates expectation of the state posterior given whole observations, i.e. $E[x_t \mid \{Y_{1:T}\}]$. The EM algorithm iterates E step and M step to find the most fit for the data parameters of Equation (2). Rangel et al. [2] applied boostrap analysis to find the confidence intervals of θ and determined the state dimension by cross-validation to aviod the over-fitting and under-fitting problems. Instead of maximizing the expectation of log-likehood equation (3), Beal et al. [3] maximized the marginal likelihood defined in Equation (5).

$$p(y \mid m) = \int p(y, x, \theta \mid m) d\theta dx \tag{5}$$

In E step and M step of VBEM algorithm, it finds the state distribution $q_x(x)$ and parameters distribution $q(\theta)$, as shown in Equations (6) and (7), respectively.

VB-E step

$$\begin{aligned} &q_X^{(\ell+1)}(x) \\ &\propto \exp[\int \ln p(x, y \mid \theta, m) q_\theta^{(\ell)}(\theta) d\theta] \end{aligned} \tag{6}$$

VB-M step

$$\begin{aligned} &q_\theta^{(\ell+1)}(\theta) \\ &\propto p(\theta \mid m) \exp[\int \ln p(x, y \mid \theta, m) q_X^{(\ell+1)}(x) dx] \end{aligned} \tag{7}$$

Perrin et al. [6] described gene regulation model based on a deterministic inertial model (8),

$$\begin{aligned} &\frac{d^2 E_i(t)}{dt^2} + 2\lambda_i \omega_i \frac{dE_i(t)}{dt} + \omega_i^2 E_i(t) \\ &= \sum_j w_{ij} E_j(t) \end{aligned} \tag{8}$$

where i, j are gene indices. $E_i(t)$ is the gene expression level of gene i. λ_i indicates an absorption coefficient specific to gene i while ω_i acts as a natural frequency of gene i. The model is basically a linear dynamic system and can be viewed as a dynamic Bayesian network based on non-input dependent state-space model and can be solved by an EM algorithm.

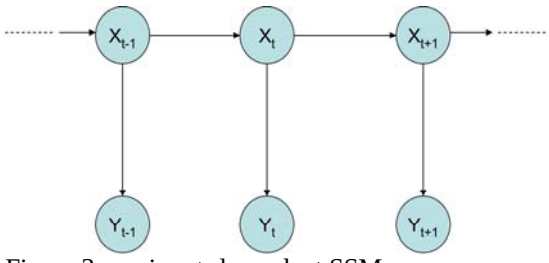

Figure 3: noninput-dependent SSM

ISSN: 1727-1983

Based on the structure defined in figure 4, we have synthesized biologically reasonable data with different numbers of time points and data sets by Monte Carlo simulations. In figure 4, F's represent unobserved factors that affect the other genes, which are considered as latent variables (also known as hidden variables in other studies). The ability of gene reconstructions from limited data is rarely discussed in previous studies [[1, 2, 3, 4, 6], which will be evaluated in this work.

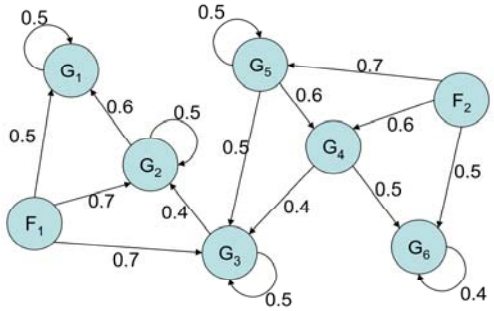

Figure 4: Functional structure of simulation data

Results

Tables 1 and 2 display evaluation results of the three DBN methods with various sample sizes, for which the data are assumed to normal-distributed and Gamma-distributed, respectively. The results show that the performances of three DBNs methods are proportional to the sample size, which is defined as the product of the number of time points (T) and the number of data sets (S). All three methods have poor performances with small sample sizes. To investigate the effects of the data sets and time points on DBNs methods, Figures 5 to 7 illustrate the performances of fixing time points to T=20 while varying S from 1 to 25. The results show that just five data sets are almost enough to achieve the average performance, the accuracies of which are 0.82 and 0.61 for VBEM and MAPEM, respectively. Instead of fixing the number of time points, Figures 8 to 10 illustrate the performance for a fixed S (=5) with T=10 to 200. The results show that average accuracy of VBEM is about 0.85 and that of MAPEM is about 0.6. Generally speaking, VBEM is more stable than MAPEM as either T or S varies.

Discussion

While it has been claimed that the latent variables can be estimated or at least taken into account by the DBNs in reconstructing gene networks, this ability has never been studied extensively before. For different sample sizes, the simulation results show that the VBEM performs best among the three DBNs for both Normal-distributed and Gamma-distributed data. When either the number of time points or the number of data sets is fixed, the VBEM is still more robust than the MAPEM. The SO-LDS seems to be the worst among the three DBNs. One of the possible reasons for this poor performance may be because the second-order model does not match the simulated data.

Table 1: Performances for the three DBNs with Normal-distributed data.

Sample size		250	200	100	50	25	20
VBEM	Sen.	0.876	0.870	0.752	0.711	0.594	0.594
	Spe.	0.912	0.874	0.841	0.796	0.767	0.767
	Acc.	0.9	0.872	0.810	0.766	0.706	0.706
MAPEM	Sen.	0.647	0.352	0.117	1	0.058	1
	Spe.	0.483	0.871	0.903	0	1	0
	Acc.	0.541	0.687	0.625	0.354	0.666	0.354
SO-LDS	Sen.	1	1	1	1	1	1
	Spe.	0	0	0	0	0	0
	Acc.	0.354	0.354	0.354	0.354	0.354	0.354

Table 2: Performances for the three DBNs with Gamma-distributed data.

Sample size		250	200	100	50	25	20
VBEM	Sen.	0.941	0.941	0.823	0.705	0.647	0.470
	Spe.	0.903	0.741	0.838	0.935	0.677	0.774
	Acc.	0.916	0.812	0.833	0.854	0.666	0.666
MAPEM	Sen.	0.352	0.294	0.176	0.117	0	1.000
	Spe.	0.838	0.806	0.774	0.903	1	0
	Acc.	0.666	0.625	0.562	0.625	0.645	0.354
SO-LDS	Sen.	0.235	0.529	0.235	0.117	1	1
	Spe.	0.580	0.387	0.516	0.709	0	0
	Acc.	0.458	0.437	0.416	0.500	0.354	0.352

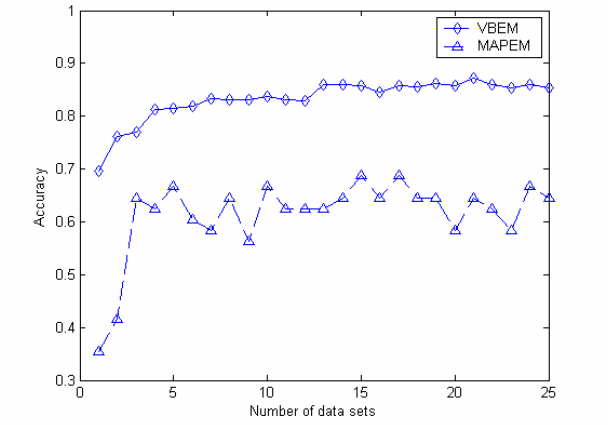

Figure 5: Accuracy with different numbers of data sets for VBEM and MAPEM

ISSN: 1727-1983

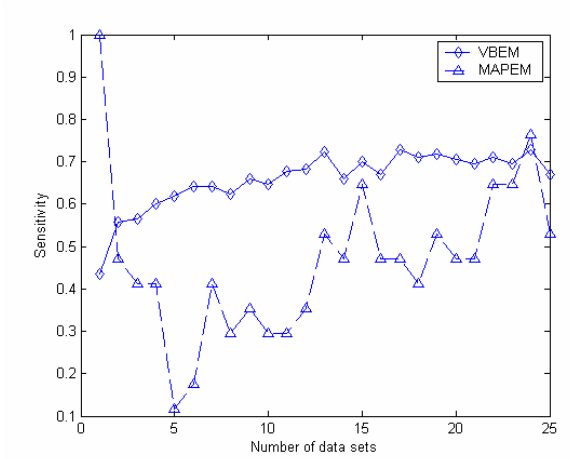

Figure 6: Sensitivity with different numbers of data sets for VBEM and MAPEM

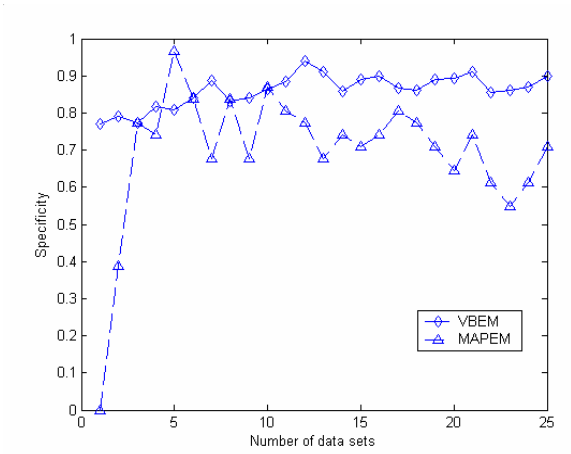

Figure 7: Specificity with different numbers of data sets for VBEM and MAPEM

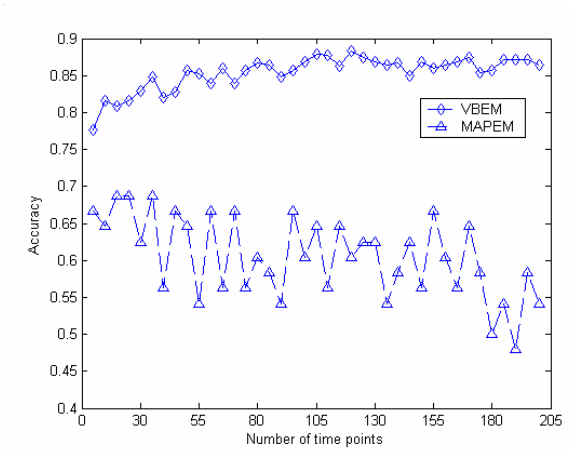

Figure 8: Accuracy with different numbers of time points for VBEM and MAPEM

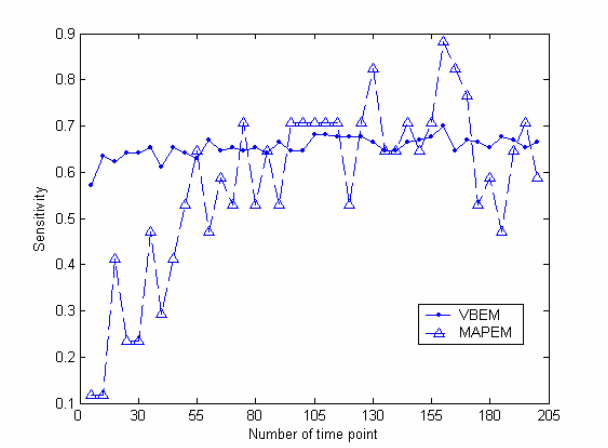

Figure 9: Sensitivity with different numbers of time points for VBEM and MAPEM

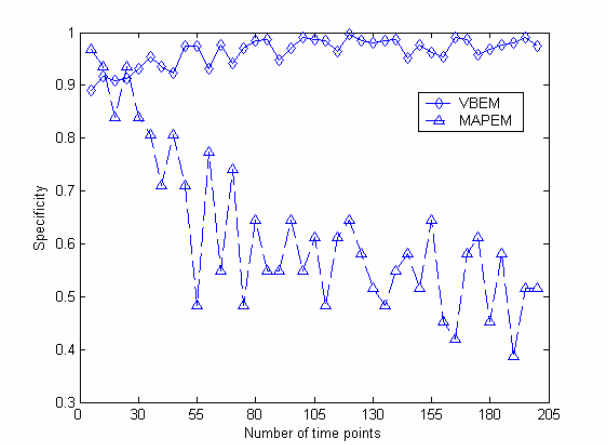

Figure 10: Specificity with different numbers of time points for VBEM and MAPEM

Conclusions

In this study, we have evaluated three DBNs based on simulation data. Different sample sizes and different data distributions have been considered. Although it is not shown in the result section, incorporating a correct number of latent variables is indeed a crucial step toward a better estimation of gene-gene interaction. The simulation results suggest that VBEM is the best among these three DBNs. Nevertheless, this evaluation is by no means complete because only one network has been simulated. Moreover, only linear gene-gene relation has been simulated, which is not necessarily the case in true biological systems.

References

[1] MURPHY,K. AND MIAN,S., Modeling gene expression data using dynamic Bayesian networks. (1999) Technical report, University of California, Berkeley, CA.

ISSN: 1727-1983

[2] RANGEL,C., ANGUS,J., GHAHRAMANI,Z., et al. (2004) Modeling T-cell activation using gene expression profiling and state-space models. Bioinformatics, 20, pp. 1361-1372.

[3] BEAL,M.J., FALCIANI,F., GHAHRAMANI,Z., RANGEL,C. AND WILD,D.L. (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics, 21, pp.349-356.

[4] HUSMEIER, D., (2003) Sensitivity and Specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformaics, 19, pp.2271-2282.

[5] ROSTI,A.-V.I. AND GALES,M.J.F. (2001) Generalised linear Gaussian models. Cambridge University Engineering Department.

[6] PERRIN, B.-E., RALAIVOLA, L., MAZURIE, A., BOTTANI, S. MALLET, J. AND D'ALCHE BUC, F. (2003) Gene networks inference using dynamic Bayesian networks. Bioinformatics, 19, S138-S148.

[7] KIM,S.Y., IMOTO,S., MIYANO,S. (2004) Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. BioSystems, 75, pp.57-6.

ISSN: 1727-1983