Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Bayesian Dynamic Multivariate Models for Inferring Gene Interaction Networks

Yulan Liang[1], Arpad Kelemen[1,2]

[1]Department of Biostatistics
The State University of New York at Buffalo, Buffalo, NY 14214, USA
[2]Department of Computer and Information Sciences
Niagara University, Niagara University, NY 14109, USA
akelemen@buffalo.edu

Abstract - **Constructions of gene and protein dynamic network is a challenging and important problem in genomic research while estimating the temporal correlations and non-stationarity are the keys in this process. In this paper, we develop Bayesian dynamic multivariate models to tackle this challenge for inferring the gene network profiles associated with diseases and treatments. We treat both the stochastic transition matrix and the observation matrix time-variant and include temporal correlation structures in the covariance matrix estimations in the multivariate Bayesian setting. The unevenly spaced short time courses with unseen time points are treated as hidden state variables. Bayesian approaches with various prior and hyper-prior models with MCMC algorithms are used to estimate the model parameters. We apply our models to multiple tissue polygenetic affymetrix data sets. Preliminary results show that the genomic dynamic behavior can be well captured by the proposed model.**

KEYWORDS: Bayesian approach, Dynamic linear model, Multivariate time series, Temporal gene expression, Deviance Information Criterion, Affymetrix data

I. INTRODUCTION

After the completion of the genome sequencing project, new computational challenges arouse in functional genomics, which include gene/protein interaction network modeling, pathway discovery and function prediction. However, complex phenotypes such as diseases typically involve multiple inter-correlated genetic and environmental factors that interact in a hierarchical fashion, microarrays hold tremendous latent information that require more sophisticated computational tools to tackle the hidden information.

Time-course gene expression data are often measured to study dynamic biological systems since knowing when or whether a gene is expressed and how one interacts with others can provide a strong clue of its biological roles. One of the goals of modeling time course microarray data is to infer and predict the genetic networks and gene-gene interactions from expression data. A study with genetic network approach using gene expression data was discussed and developed by D'haeseleer, Liang, and Somogyi [1]. Friedman and co-workers have used static Bayesian networks, which are graph-based models of joint multivariate probability distributions that assess conditional independence between variables, to obtain simpler sub-models to describe gene interactions from array data [2, 3]. Probabilistic Boolean Networks were recently developed as models of gene regulatory networks which are able to cope with uncertainty and discover the relative sensitivity of genes in their interactions with other genes. Chen et al. developed a stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae [4].

Dynamic linear models have greater flexibility in modeling non-stationary and nonlinear short time course microarray data. However, current existing methods were based on standard Kalman filter methods that rely on the linear state transitions and Gaussian errors. Perrin et al. used a penalized likelihood maximization (MAP) implemented through an extended version of EM algorithm to learn the parameters of the model [5]. Rangel, et al. used classical cross-validations and Bootstrap techniques and Beal et al. used variation approximations with linear time invariant Gaussian setting for constructions of the regulatory network [6, 7]. Kim et al. developed an algorithm to identify interaction network and coupled it with non-parametric regression methods [8].

In this paper we utilize the merits of Bayesian flexibility of estimation procedures and the stochastic state space process of modeling the temporal dynamics and develop Dynamic multivariate model in the fully Bayesian setting for inferring the interaction networks associated with diseases. Monte Carlo Markov Chain (MCMC) algorithms are used to sample the posterior distribution of the hidden variables and the model parameters [9, 10]. Various prior models with different hyper-prior distributions are simulated and compared, and Deviance Information Criterion (DIC) is used for model checking and selections [11]. The developed models were applied to common gene expressions data derived from multiple tissues polygenic phenomena in complex biological systems [12].

II. METHODS

II.A. Bayesian Dynamic Multivariate Model Formulation and Prior Model Specification

Gene and protein expression measurements (observations) are contaminated by noise. A dynamic state space model will decompose the signal and noise processes into two model equations: the stochastic (evolution) equation and the observation equation. In a dynamic linear model, a sequence of P-dimensional real valued observation vector X_t of gene or protein expression is modeled by assuming that each time

step, X_t are generated from a K-dimensional unobserved or hidden state variable S_t, and the sequence of S_t's define a Markov process. The joint probability of $\{X_t, S_t\}$:

$$P(S_t, X_t) = P(S_1)P(X_1 \mid S_1)\prod_{t=2}^{T} P(S_t \mid S_{t-1})P(X_t \mid S_t) \quad (1)$$

where $P(S_1)$ is the first unobserved state and is assumed to be generated from conjugate distributions such as Gaussian, or student t-distributions. $P(S_t \mid S_{t-1})$ is the transition density or probability of hidden states (such as genes that are not included in the study) and it can be defined in the stochastic evolution equations as:

$$S_t = g_t(S_{t-1}) + V_t \quad (2)$$

where g_t is the deterministic transition function determining the mean of S_t given S_{t-1}. $P(X_t \mid S_t)$ is the observation density or probability that can be defined in the observation equations as:

$$X_t = f_t(S_t) + W_t \quad (3)$$

where f_t is the statistical transition function of the observation processes. The observation vector X_t are conditionally independent given S_t and is independent of S_z; $z \neq t$. W_t, V_t are assumed to follow Gaussian or non-Gaussian distributions with means zeros of both population processes. g_t, f_t follow either linear or nonlinear settings.

We start with simple dynamic linear models in univariate case. Here, both stochastic and observation equations take linear forms and the distributions of the states and the observation variables are assumed to follow Gaussian distributions. The model is as follows:

$$S_{g,t} = AS_{g,t-1} + \varpi_t,\ X_{g,t} = CS_{g,t} + \nu_{,t} \quad (4)$$

where g=1,…,P (number of genes), t=1,…,T (number of time points). ω_t, ν_t are noise variables, A is state transition matrix. C is the observation matrix. The observations often can be divided into a set of input (or exogenous) variables and a set of output (response) variables. Including inputs in both the state equation and observation equation, the model (4) can be modified as follows:

$$S_{g,t} = A_t S_{g,t-1} + B_t U_t + \varpi_t,\ X_{g,t} = C_t S_{g,t} + D_t U_{g,t} + \nu_t \quad (5)$$

where U_t is the input (covariate) observation vector. B is input to the state matrix and D is input to the observation matrix. Here we are particularly interested in modeling the effects of the influence of the expression of one gene at a previous time point on another gene, its associated hidden variables and the gene-gene interactions. Therefore we take $U_t = X_{t-1}$, in which the input is replaced by the pervious time step for modeling the gene-gene interaction. Thus our model is further simplified into:

$$X_{g,t} = (AC)_t S_{g,t-1} + (CB + D)_t X_{g,t-1} + \varepsilon_t,\ \varepsilon_t = C\varpi_t + \nu_{t+1} \quad (6)$$

where AC is the hidden state dynamic matrix with the influence of hidden state variables on gene expression level at each time point, CB+D contains both the gene-to-gene interaction and the gene-to-gene interactions 'through' the hidden state over time and ε_t is the noise.

Rangel, et al. and Beal, et al. assumed that A, B, C, D are linear and time-invariant matrices in the dynamic linear model setting [6, 7]. Here in our model (6), we initialized our model with a time varying matrix. The motivation of the above dynamic model with time varying coefficient comes from our major prediction goal, which requires a model not only to have good fit in the sampling period, but also a good generalization performance. Since microarray experiments are more concerned about the short term prediction given the short term time course data, as Congdon suggested, the introduction of time variability is advisable [9]. In this way, the underlying parameters to be estimated evolve through time with continuous measure instead of discrete.

One weakness of the above model is that it is restricted to univariate multiple time courses measured simultaneously on a common system. Multivariate models that can describe patterns of dependency among multiple series (genes across time) may be helpful to discover the gene dependences of the underlying processes. We extend the above model from univariate to multivariate dynamic model via covariance structure for learning gene correlations and their temporal behavior for constructing the gene networks. Here, each series depends on both its own past and the past values of the other series, therefore the variations in expression for a given gene can be predicted by a small set of other genes. One advantage of simultaneously modeling several series is the possibility of pooling information from related genes to improve the precision and out of sample forecasts [9].

In order to incorporate the fully hierarchical Bayesian setting into the multivariate model for learning the model parameters and model structures, and using probability density functions, we formulate Bayesian Dynamic multivariate model (BDM) as follows by modifying (6):

$$X_t \sim \text{MVN}(u_t, \varepsilon_t), \text{ (observation equation)} \quad (7)$$

$$u_t = (AC)_t S_{t-1} + (CB + D)_t X_{t-1}, \text{ (systematic equation)} \quad (8)$$

$$AC_t \sim \text{MVN}(\underline{\beta_1}, \Omega_1),\ (CB+D)_g \sim \text{MVN}(\underline{\beta_2}, \Omega_2) \quad (9)$$

$$\varepsilon_t \sim \text{MVN}(\underline{0}, V) \quad (10)$$

where Ω_1, Ω_2, V are inverse positive definite correlated covariance matrices (or precision matrices) with particular structures and they are generated from an inverse Wishart distribution [9]. $\underline{\beta_1}$, $\underline{\beta_2}$ are generated from t-distributions or Gaussian distributions with vague hyper inverse-Gamma distributions. The biological merit of this model compared to the previous one is that the correlated gene structure via the estimated correlation-covariance matrix can be used to infer the gene-gene interaction networks. Here we start with a model with Multivariate Gaussian distribution with mean zero and correlated covariance matrix while the hyper-parameters of the correlated covariance matrix will be generated from a multivariate distribution, such as an inverse Wishart distribution.

The estimations of the parameters and hyper-parameters in the covariance matrix are conducted by MCMC algorithms. The sensitivity to prior specifications are tested and Deviance Information Criterion is used for model selection and comparison.

II.B. Monte Carlo Markov Chain for Posterior Inference

Since temporal gene array data can be considered to be generated from a continuous dynamical system, it is reasonable to assume that the values of the hidden states $S_{(t)}$ do not change much from one time index to the next. Therefore, the dynamic model representing function g is used to model only the change. Let us denote the observed gene expression data with $X = (x(1), x(2), ..., x(T))$, the hidden state values with $S = (s(1), s(2), ..., s(T))$ and all the other model parameters with θ. Here, the presence of unobserved components (hidden variables) makes the maximum likelihood inference more difficult to apply. Furthermore, since the temporal expression data includes a series of highly correlated variables, which may have varying degrees of relevance to the outcome, conventional statistical models with Maximized Likelihood approaches can not set the coefficients of irrelevant variables (genes) to zero (multi-collinearty). These variables with nonzero values reduce the model's generalization performance and cause overfitting.

The fully Bayesian approach is preferable, since it allows treatment of general dynamic multivariate models with shrinkage and regularization, can deal with small size problems and make the simulation based approach easier, such as MCMC to parameter estimation and latent factor filtering. In a fully Bayesian approach not only the uncertainty resulting from the error is accounted for in the estimation of the dynamic parameters via credible intervals, but the hyper-parameters are also estimated through the hierarchically constructed priors and the integration of all the parameters. Monte Carlo methods can be used to approximate this integration by simulation. Monte Carlo Markov Chain methods can be used to sample from the fully Bayesian posterior distributions of all unknown quantities. This approach also allows us to examine the robustness of the inference with respect to choices for the priors and hyper-priors. For the Monte Carlo method Gaussian approximations centered on the posterior models was implemented in WinBugs using Gibbs sampling [11]. We utilized this in our proposed models. Moreover, we used over-relaxation methods to aid the convergence and reduce the chance of local maxima.

II.C. Model Comparisons, Selection Criteria and Validations

In order to choose the best model for prediction, the model selection criteria, such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) or Bayes factors can be considered. Although AIC is useful for non-nested models, it works poorly in the case of multi-collinearty, which is typical for gene expression data. It has drawbacks of tending to be biased for complicated models due to the fact that log-likelihood increases faster than the model complexity component. Deviance Information Criterion is a new measure proposed by Spiegelhalter, et al. for model complexity and goodness of fit under the Bayesian setting and it's more appropriate when comparing complex hierarchical models in the Bayesian setting, where the number of parameters is not clearly defined [11]. One advantage is its inclusion of a prior distribution, which induces a dependency between parameters that is likely to reduce the effective dimensionality. Furthermore, it helps the prior models' identifications. DIC can be summarized by the posterior expectation of the deviance and complexity (effective number of parameters) as the expected deviance minus deviance at the posterior expectation of the parameters. We used DIC for within sample fit measure for model selections.

III. RESULTS

The multiple tissue data applied here are the results of three experiments to characterize the response of a single bolus dose of methylprednisolone (MPL) in rats [12]. Affymetrix GeneChips ® Rat Genome (R_U34A) microarray chip, which contained 8799 probe sets, was used in this study. Data were obtained at the following 17 time points: 0, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5, 5.5, 6, 8, 12, 18, 30, 48, 72 hours after treatment. Gene expression levels from 0.25 hour to 72 hours were converted to ratios via a simple calculation that involved dividing the gene expression at time t_i by the gene expression at time t_0, where *i* represents the specific post-dose time-point and t_0 represents baseline at time = 0 hours, i.e. the control group levels at t_0. These ratios were subsequently natural-log transformed to produce approximately normally distributed gene expression levels at each sampling time point. After pre-processing and gene filtering using Bayesian finite Markov mixture model and meta analysis we developed earlier [13], 6 genes that differentially expressed commonly in three organs (kidney, liver and muscle) and 44 genes differentially expressed commonly in two organs (liver and muscle) were selected in the preliminary study for constructing gene interaction networks based on the proposed Bayesian dynamic multivariate models.

The prior models and hyper-priors were initialized with various forms and sensitivity analysis was conducted from various initials. 2000 samples after 6000 burn-ins were used for computation. We were able to achieve the smallest DIC value of 195.1 with one of the proposed model. The estimated covariance matrix was converted into correlation coefficients for constructions of the interaction networks based on the estimated correlation coefficients from the model. All the correlation coefficients that were larger than 0.10 (or smaller than -0.10) were then highlighted and the gene-gene interactions were based on the correlation coefficients. Fig. 1 displays the results of the constructed gene-time-gene interaction network for 6 selected genes that are significantly differentially expressed at 5 time points. Fig. 2 shows the constructed interaction network based on 44 commonly differentially expressed genes in both liver and muscle.

IV. DISCUSSION

Inferring networks of gene and proteins from biological data is a central issue of computational biology. To delineate the possible interactions of all genes in a genome is a task for which conventional experimental techniques are ill-suited. Sorely needed are rapid and inexpensive computational methods that identify candidates for interacting genes. Our proposed model can explicitly learn gene-gene interaction and gene-time-gene interaction networks by its model specification (e.g. estimation of scale matrix) in the Bayesian dynamic multivariate setting, which could be useful for further pathway discovery. The validation analysis of the developed model for the interaction networks and pathways will be further compared to the results from Ingenuity Pathway software (3.0 version) and other existing pathway databases and literature. Future work will include the development of a learning algorithm in the above models that can decompose the latent component of correlated genes matrix and also capture the key component in the multivariate time series. Another extension of the current work is overcome the slow convergence problems of MCMC algorithms by using jump reversible MCMC algorithm and variational method. In this way, we can also inferring the medium size interaction networks rather than small size network.

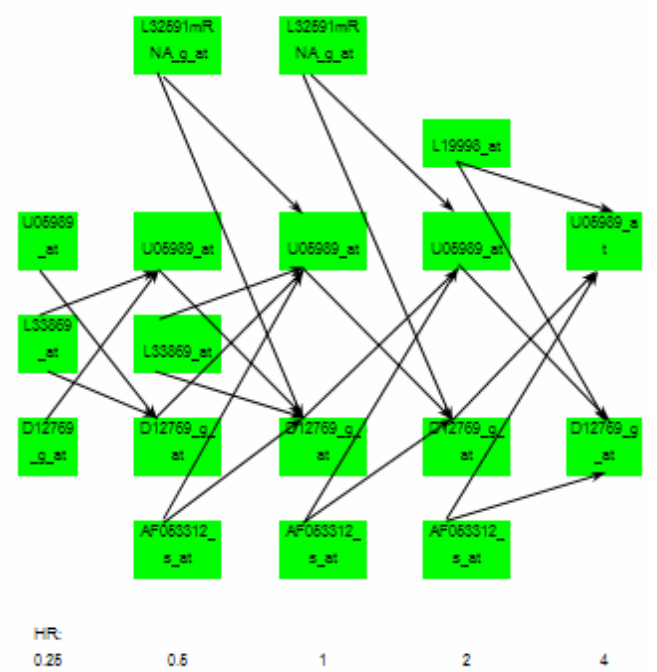

Figure 1: gene-time-gene dependence interaction network derived from six genes that differentially expressed commonly in three organs (kidney, liver and muscle) and five significant time points (x-axis: 5 significant time points (0.25, 0.5, 1, 2, 4 hrs); y-axis: six selected genes) using BDM and multiple tissues data.

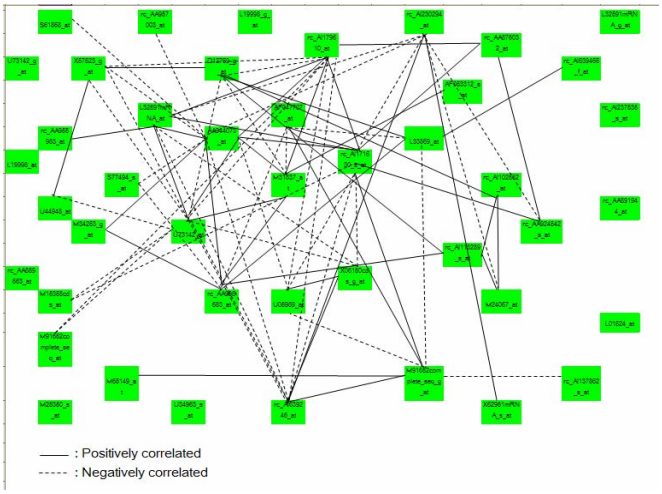

Figure 2: Constructed gene-gene interaction network using BDM and multiple tissues data for the selected 44 differentially expressed genes in two organs (liver and muscle).

REFERENCES

1. P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co expression clustering to reverse engineering", *Bioinformatics* 16:707-726, 2000.
2. N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data", *Journal of Computational Biology* 7:601-620, 2000.
3. N. Friedman, "Inferring cellular networks using probabilistic graphical models", *Science*; 303(5659):799-805, 2004.
4. K. C. Chen, T. Y. Wang, H. H. Tseng, C. Y. Huang, C. Y. Kao, "A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae", *Bioinformatics*. 21(12):2883-90.
5. B. E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, F. D'Alche-Buc, "Gene networks inference using dynamic Bayesian networksa", *Bioinformatics* 19 Suppl 2:II138-II148, 2003.
6. C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. A. Sotheran, A. Gaiba, D. L. Wild, and F. Falciani, "Modeling T-cell activation using gene expression profiling and state space models", *Bioinformatics*, 20(9):1361-1372, 2004.
7. M. J. Beal, F. L. Falciani, Z. Ghahramani, C. Rangel, and D. Wild, "A Bayesian Approach to Reconstructing Genetic Regulatory Networks with Hidden Factors", *Bioinformatics*. Vol. 21, pp. 349-356, 2005.
8. S. Y. Kim, S. Imoto, S. Miyano, "Dynamic bayesian network and nonparametric regression model for inferring gene networks", *Genome Informatics*, 13: 371-372, 2002.
9. P. Congdon, *Applied Bayesian Modeling*. John Wiley & Sons, Ltd., 2002.
10. B. P. Carlin, and S. Chib, "Bayesian model choice via Markov chain Monte Carlo methods", *Journal of the Royal Statistical Society,* Series B, 57, 473-84, 1995.
11. D. Spiegelhalter, N. Best, B. Carlin, and A. Linde, "Bayesian measures of model complexity and fit", *Journal of Royal Statistical Society*, Series B. 64, part 4, pp. 583-639, 2002.
12. R. Almon, J. Chen, J., D. DuBois, W. J. Jusko, and E. P. Hoffman, "In vivo Multi-Tissue Corticosteroid Microarray Time Series Available Online at Public Expression Profile Resource", *Pharmacogenomics*, 4: 791-799, 2003.
13. Y., Liang, A., Kelemen, "Bayesian Finite Markov Mixture Models and Meta-analysis for Temporal Multi-Tissue Polygenic Patterns Following Corticosteroid Administration", Statistics in Medicine, in press.