Microarray Time Series Modeling and Variational Bayesian Method for Reverse Engineering Gene Regulatory Networks

M. Sánchez-Castillo[1], I.M. Tienda Luna[2], D. Blanco-Navarro[1],
and M.C. Carrión-Pérez[1]

[1] Department of Applied Physics, University of Granada, 18071, Spain
[2] Department of Electrical and Computer Engineering, University of Granada, 18071, Spain
{mscastillo,isabelt,dblanco,mcarrion}@ugr.es

Abstract. Gene expression is a complex process controlled by underling biological interactions. One model that tries to explain these relationships at a genetic level is the gene regulatory networks. Uncovering regulatory networks are extremely important for live sciences to understand how genes compete and are associated. Despite measurement methods have been successfully developed within the microarray technique, the analysis of genomic data is difficult due to the vast amount of information considered. We address here the problem of modeling the gene regulatory networks by a novel linear model and we propose a Bayesian approach to learn this structure from microarray time series.

Keywords: microarray, gene regulatory networks, VBEM algorithm.

1 Introduction

Microarray experiments have supposed a breakthrough into genomic research. With this technique, the expression of thousand of genes may be quantified simultaneously. Genomic studies demand help from computer science community to process and analyze such a vast amount of information. One topic of special interest is the study of genetic interactions. Uncovering that kind of relationships is extremely important to understand how genes compete and are associated to produce complex responses and co-operative effects, information which can be used in many fields such as disease treatment and new drug design.

One model that tries to explain genetic interactions is the gene regulatory network (GRN). In a GRN it is considered that the expression of a gene, known as child, depends on others presented in the network, known as parents. We address here the problem of modeling and inferring the GRN from microarray time series. Specifically, this paper revises the linear model presented in [1] and proposes a new one that fits better microarray data. Additionally, a variational Bayesian method based on new model is proposed.

V.V. Das and P.M. El-Kafrawy (Eds.): SPIT 2012, LNICST 117, pp. 70–75, 2014.

2 Gene Regulatory Networks Modeling

Gene regulatory networks are characterized by two important aspects [2]. First is the connectivity, also referred as network topology, which represents the linkage pattern of the network. This logical structure have been modeled in [1] by a set of binary latent variables, denoted by

$$\mathbf{x}_i = [x_i(1),\ldots,x_i(G)]^{\mathrm{T}} \in \{0,1\}^{G\times 1} \tag{1}$$

where $x_i(j)=1$ specifies that the j-th gene is a parent of the i-th gene or $x_i(j)=0$ otherwise. Second, genetic networks also specifies regulatory effects between elements, i.e. strength and type of interaction. This scheme has been described in [1] by an additional set of weights, denoted by

$$\boldsymbol{\omega}_i = [\omega_i(1),\ldots,\omega_i(G)]^{\mathrm{T}} \in \Re^{G\times 1} \tag{2}$$

with $\omega_i(j)>0$ for gene activation and $\omega_i(j)<0$ for gene inhibition.

3 Linear Models for Microarray Time Series Fitting

Consider a microarray data set $\mathbf{Y}\in\Re^{G\times(N+1)}$ with G genes and $N+1$ time samples, such as $[\mathbf{Y}]_{i,n} = y_i(n)$ the observed expression level: relative mRNA abundance of the i-th gene at the n-th time sample. Assuming a Markov process, a first order autoregressive (AR1) model have been proposed in [1]. This approach expressed microarray data as a linear combination of the observations and the variables describing the gene regulatory network, plus independent and identically distributed (IID) Gaussian white noise, as

$$y_i(n) = \sum_{j=1}^{G} y_i(n-1)\omega_i(j)x_i(j) + e_i(n) \tag{3}$$

with

$$p(e_i(n)) = N\left(e_i(n)\middle|0,\sigma_i^2\right), \forall n. \tag{4}$$

However, this model establishes relationships between the observed expression levels, $y_i(n)$, which are supposed to be noisy. It would be much more realistic to establish these relationships between the real expression level, denoted by $z_i(n) = y_i(n) - e_i(n)$. Therefore, we propose a novel approach where genetic relationships are established between the real expression levels instead of its noisy observation, leadding to a first order autoregressive moving-average (AR1MA1) model as

$$y_i(n) = \sum_{j=1}^{G} y_j(n-1)\omega_i(j)x_i(j) - \sum_{j=1}^{G} e_j(n-1)\omega_i(j)x_i(j) + e_i(n). \tag{5}$$

4 Variational Bayesian Expectation-Maximization Framework

Consider $\mathbf{y}_i$ the set of observations, $\mathbf{x}_i$ the set of latent or hidden variables and $\boldsymbol{\theta}_i$ the set of unknowns parameters for the i -th variable. The posterior distribution could be derived from the priors and the likelihood as

$$p(\mathbf{x}_i,\boldsymbol{\theta}_i|\mathbf{y}_i)=\frac{p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)p(\mathbf{x}_i,\boldsymbol{\theta}_i)}{p(\mathbf{y}_i)} \tag{6}$$

with $p(\mathbf{y}_i)$ the marginal likelihood obtained by marginalization as

$$p(\mathbf{y}_i)=\int p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)p(\mathbf{x}_i,\boldsymbol{\theta}_i)d\mathbf{x}_i d\boldsymbol{\theta}_i. \tag{7}$$

Finding an analytical solution for the marginal likelihood and posterior distributions usually is a difficult task. An alternative to compute the posterior distribution by marginalization have been presented by Beal et al. in [3]. Instead of integrating out the unknowns, variational Bayes computes a lower bound of the logarithm of the marginal likelihood. In virtue of Jensen's inequality, lower bound can be expressed by a functional depending on a free distribution as,

$$\log p(\mathbf{y}_i)\geq F[q(\mathbf{x}_i,\boldsymbol{\theta}_i)]=\int q(\mathbf{x}_i,\boldsymbol{\theta}_i)\log\frac{p(\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\theta}_i)}{q(\mathbf{x}_i,\boldsymbol{\theta}_i)}d\mathbf{x}_i d\boldsymbol{\theta}_i. \tag{8}$$

Optimization of (8) is a problem that may be solved by variational calculus. Alternatively, based on a mathematical convenience, variational Bayesian choose a free distribution that factorizes into conjugate families as

$$q(\mathbf{x}_i,\boldsymbol{\theta}_i|\xi)\approx q(\mathbf{x}_i|\xi_{\mathbf{x}_i})q(\boldsymbol{\theta}_i|\xi_{\boldsymbol{\theta}_i}) \tag{9}$$

with $\xi=\{\xi_{\mathbf{x}_i},\xi_{\boldsymbol{\theta}_i}\}$ hyperparameters that characterizes the conjugate families.

For conjugate models, the computation of the posterior becomes into a set of posterior hyperparameters learning rules. Therefore, variational Bayesian Expectation-Maximization (VBEM) methods consist of the following two steps, in which one of the free distributions is optimized whilst the other one is fixed as

$$q\left(\mathbf{x}_i|\hat{\xi}_{\mathbf{x}_i}^{(t+1)}\right)\propto \mathbf{e}^{\left\langle \log p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)\right\rangle_{q\left(\boldsymbol{\theta}_i|\hat{\xi}_{\boldsymbol{\theta}_i}^{(t)}\right)}+p(\mathbf{x}_i)} \tag{10}$$

$$q\left(\theta_i|\hat{\xi}_{\boldsymbol{\theta}_i}^{(t+1)}\right)\propto \mathbf{e}^{\left\langle \log p(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\theta}_i)\right\rangle_{q\left(\mathbf{x}_i|\hat{\xi}_{\mathbf{x}_i}^{(t+1)}\right)}+p(\boldsymbol{\theta}_i)} \tag{11}$$

Subsequetnly, the lower bound is updated and VBEM algorithm iterates until the difference after two consecutive steps satisfies a convergence criterion as

$$\left| \frac{F\left[q\left(\mathbf{x}_i, \boldsymbol{\theta}_i \middle| \hat{\boldsymbol{\xi}}^{(t+1)}\right)\right] - F\left[q\left(\mathbf{x}_i, \boldsymbol{\theta}_i \middle| \hat{\boldsymbol{\xi}}^{(t)}\right)\right]}{F\left[q\left(\mathbf{x}_i, \boldsymbol{\theta}_i \middle| \hat{\boldsymbol{\xi}}^{(t+1)}\right)\right]} \right| \le \varepsilon. \tag{12}$$

5 VBEM Method Applied to the AR1MA1 Model

Given a generative model as the AR1MA1 one in (5), we are going to consider the binary variables describing the topology of the network $\mathbf{x}_i$ as latent variables whilst the weights and noise variance $\boldsymbol{\theta}_i = \left\{\boldsymbol{\omega}_i, \sigma_i^2\right\}$ are interpreted as model parameters. On the other hand, data will be a microarray time series for the i-th gene as $\mathbf{y}_i = \left[y_i(1), \ldots, y_i(N)\right]^{\mathrm{T}}$ .Taking into account (4) and (5), the likelihood function may be expressed as

$$p\left(\mathbf{y}_i \middle| \mathbf{x}_i, \boldsymbol{\theta}_i\right) = N\left(\mathbf{y}_i \middle| \mathbf{R}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{x}_i, \frac{\sigma_i^2}{\gamma_i}\mathbf{1}^N\right) \tag{13}$$

with $\mathbf{D}_{\boldsymbol{\omega}_i}$ a diagonal matrix with vector $\boldsymbol{\omega}_i$, $\mathbf{R} = \mathbf{T}\mathbf{Y}^{\mathrm{T}}$, $\mathbf{T} = \left[\mathbf{1}^N \middle| \mathbf{0}\right] \in \Re^{G \times (N+1)}$ and

$$\gamma_i = \gamma_i\left(\mathbf{x}_i, \boldsymbol{\omega}_i\right) = \left(1 + \boldsymbol{\omega}_i^{\mathrm{T}}\mathbf{D}_{\boldsymbol{\omega}_i}\mathbf{D}_{\mathbf{x}_i}\mathbf{x}_i\right)^{-1}. \tag{14}$$

According to (9), probability distributions must be chosen from families that factorizes into hidden variables and parameters. We are going to choose priors from the same families as in method proposed in [1] as

$$q\left(\mathbf{x}_i\right) = N\left(\mathbf{x}_i \middle| \boldsymbol{\mu}_{\mathbf{x}_i}, \boldsymbol{\Sigma}_{\mathbf{x}_i}\right) \tag{15}$$

$$q\left(\boldsymbol{\omega}_i, \sigma_i^2\right) = N\left(\boldsymbol{\omega}_i \middle| \boldsymbol{\mu}_{\boldsymbol{\omega}_i}, \sigma_i^2\boldsymbol{\Sigma}_{\boldsymbol{\omega}_i}\right) IG\left(\sigma_i^2 \middle| \alpha_i, \beta_i\right) \tag{16}$$

a Gaussian and Normal scaled Inverse Gaussian distribution with $\boldsymbol{\xi}_{\mathbf{x}_i} = \left\{\boldsymbol{\mu}_{\mathbf{x}_i}, \boldsymbol{\Sigma}_{\mathbf{x}_i}\right\}$ and $\boldsymbol{\xi}_{\boldsymbol{\omega}_i} = \left\{\boldsymbol{\mu}_{\boldsymbol{\omega}_i}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}_i}, \alpha_i, \beta_i\right\}$ the hyperparameters to be learned from data.

Likelihood function in (13) does not satisfies the requirements for the conjugate model. Specifically, dependence of variance scale (14) on the unknowns does not allows to define conjugate priors. As a suboptimal solution, we propose a fixed point approach where scale effect of γ_i is approximated according to the most probable of $\mathbf{x}_i$ and $\boldsymbol{\omega}_i$, given by its means as

$$\bar{\gamma}_i \approx \gamma_i\left(\boldsymbol{\mu}_{\mathbf{x}_i}, \boldsymbol{\mu}_{\boldsymbol{\omega}_i}\right) = \left(1 + \boldsymbol{\mu}_{\boldsymbol{\omega}_i}^{\mathrm{T}}\mathbf{D}_{\boldsymbol{\mu}_{\boldsymbol{\omega}_i}}\mathbf{D}_{\boldsymbol{\mu}_{\mathbf{x}_i}}\boldsymbol{\mu}_{\mathbf{x}_i}\right)^{-1}. \tag{17}$$

6 Results and Discussion with Synthetic Data

We have applied the proposed VBEM algorithm to synthetic data sets. Specifically two VBEM methods were considered: (i) based on the AR1 model proposed in [1], refereed as AR1-VBEM method and (ii) based on the new AR1MA1 proposed model, refereed as AR1MA1-VBEM method. To compare the performance, various sets have been generated with $G = 50$ genes, $N = 50$ time samples and different levels of noise with a signal-to-noise ratio SNR $\in (1,80)$. Each data set have been generated by simulation using the priors and likelihood as in section 5 with subjective priors. According to biological knowledge, sugessting that in a real regulatory network each gene has a limited number of parents, we have set up the netkork topology for having 15 parents (about the 30% of the total number of genes). The inference procedure has been repeated one hundred times for having a satatistically significant result.

In Figure 1 we have plotted the performance of each method as an error percentage versus the noise level. The most undesireable performance would correspond to a random assignment with constant error rate around the 50%. We have considered as a satisfactory result an error rate lower than percentile 5%. It can be noticed that AR1MA1-VBEM method outperforms the AR1-VBEM one, producing satisfactory error rates at lower levels of noise.

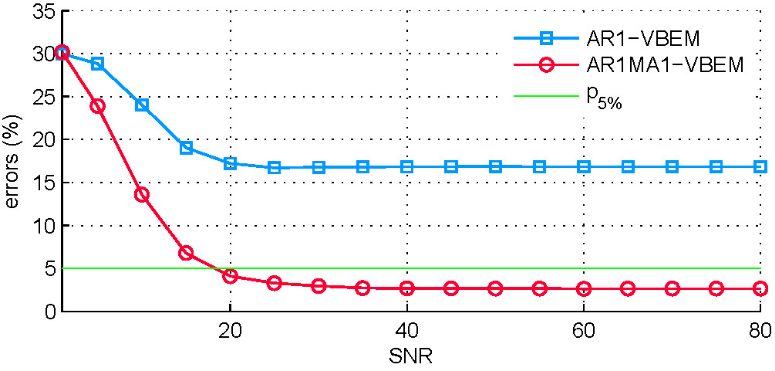

Fig. 1. Performance of AR1-VBEM method (stroke with box tokens) and the AR1MA1-VBEM one (stroke with circle tokens). AR1MA1-VBEM method outperforms the AR1-VBEM one with an error rate under percentile 5% for SNR > 20.

In binary decission, however, another kind of statistics are more suitable for analyzing these results [4]. We are going to consider the receiver operating characteristic (ROC) curve that represents the hits or true postive rate (TPR) versus the false postive or error rate (FPR). Random performance would correspond to a line through the origin with unitary slope, referred as the no-discrimination (ND) line. The area under the ROC curve (AUROCc) summarizes this analysis, with values between 0.5 for the ND line and a maximum value equal to 1.0 corresponding to the best performance.

In Figure 2 we have plotted the AUROCc for both VBEM methods at different levels of noise. Results show that AR1MA1-VBEM method outperforms the AR1-VBEM one, with values closer to one at for higher SNR.

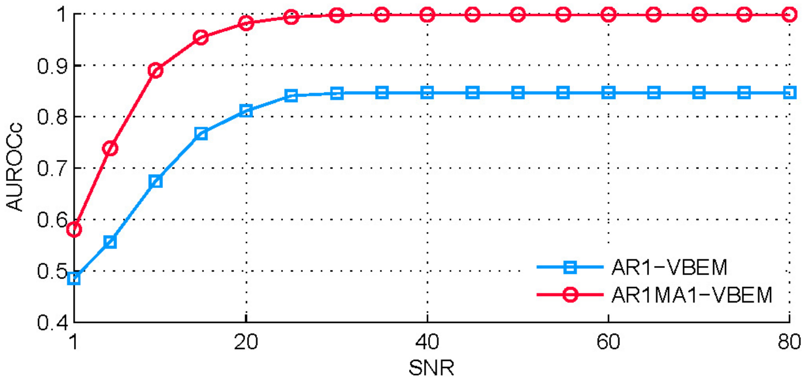

Fig. 2. AUROCc versus the level of noise for the AR1-VBEM method (stroke with box tokens) and the AR1MA1-VBEM one (stroke with circle tokens). AR1MA1-VBEM outperforms AR1-VBEM with higher AUROCc at any level of noise and values closer to one for higher SNR.

References

1. Tienda-Luna, I.M.: Constructing Gene Networks Using Variational Bayesian Variable Selection. Articial Life 14(1), 65–79 (2008)
2. Ribeiro, A.: A General Modeling Strategy for Gene Regulatory Networks with Stochastic Dynamics. J. Comput. Biol. 13(9), 1630–1639 (2006)
3. Beal, M.J., Ghaharamani, Z.: The Variational Bayesian EM Algorithm for Incomplete Data with Application to Scoring Graphical Model Structures. Bayesian Statistics 7 (2003)
4. Huang, Y.: Reverse engineering gene regulatory network, a survey of statistical models. SPMAG 26(1), 76–97 (2009)