BAYESIAN HIERARCHICAL MODEL FOR ESTIMATING GENE ASSOCIATION NETWORK FROM MICROARRAY DATA

Dongxiao Zhu,[a] *and Alfred O Hero* [b]

[a]Bioinformatics Program,[b]Depatments of EECS, Biomedical Engineering and Statistics
University of Michigan,Ann Arbor, MI 48105

1. INTRODUCTION

Estimating gene association networks from gene microarray data is the key to decipher complicated web of functional relationship between genes [1]. However, the process remains to be challenging due to the relatively few independent samples and the large amount of correlation parameters [2]. In a gene association network, vertices represent genes, and edges represent biological association between genes. The network edges are declared to be present if the corresponding correlation parameters are significantly different from a non-zero threshold [3]. The approach has been very useful in inferring gene association networks, and facilitating network based discovery [3]. However, as a Frequentist approach, it often suffers from the "overfitting" problem especially for analyzing small sample size data. Approaches that are able to globally estimate the correlation parameters with variance regularization followed by the seamless correlation thresholding are highly desirable.

The desirable approaches fall naturally into the framework of Bayesian hierarchical models [4]. We assume the correlation parameters are *exchangeable* meaning that the joint distribution (Eq. 1) is invariant to permutations of the indexes. Biologically, this represents a lack of knowledge that could differentiate one pair of biological associations from the others. We then regularize variances of the correlation estimations by specifying a parent normal distribution from which marginal correlation parameters are sampled (Fig. 1). The posterior distributions of correlation parameters provide a seamless combination of the correlation estimation and strength thresholding.

2. METHODS

We use ρ to denote the true strength of association between a pair of gene expression profiles. For G gene expression profiles in a microarray data set, there are $\Lambda = \binom{G}{2}$ correlation parameters ρ that need to be estimated, denoted as $\rho_\lambda, \lambda = 1, \ldots, \Lambda$. We define $\hat{\rho}_\lambda$ as the λth sample correlation coefficient, and $\hat{\Gamma}_\lambda$ as the hyperbolic arc-tangent transformation of $\hat{\rho}_\lambda$. Then the transformed sample correlation coefficients $\hat{\Gamma}_\lambda = \mathrm{atanh}(\hat{\rho}_\lambda)$ are asymptotically Gaussian distributed with means of Γ_λ and stabilized variance approximations of $\sigma_\lambda^2 = 1/(N-3)$. Here N is the sample size.

Simulation studies show that the variance approximation works reasonably well even at a relatively small sample size, e.g. $N \leq 10$. We assume known variances to reduce computational complexity. Furthermore, we don't have *a prior* information about these $\Gamma's$, and assuming independency between them in marginal correlation approaches cause the "overfitting" problem [2]. In the Bayesian hierarchical model, we assume that these parameters are *exchangeable*, and are drawn from a normal distribution with unknown hyperparameters (α, β) (Fig. 1):

$$p(\Gamma_1, \ldots, \Gamma_\Lambda|\alpha, \beta) = \prod_{\lambda=1}^{\Lambda} N(\Gamma_\lambda|\alpha, \beta^2) \qquad (1)$$

In order to generate conditional posterior distributions $p(\Gamma_\lambda|\alpha, \beta, y)$ for each parameter $\Gamma_\lambda, \lambda = 1, \ldots, \Lambda$, where y represents the microarray data, we performed simulation steps as follows [4]:

1. Assign prior distribution for β, i.e. uniform prior distribution $p(\beta) \propto 1$.

2. Draw β from posterior distribution $p(\beta|y)$.

$$p(\beta|y) \propto \frac{p(\beta)\prod_{\lambda=1}^{\Lambda} N(\widehat{\Gamma}_\lambda|\hat{\alpha}, \sigma_\lambda^2 + \beta^2)}{N(\hat{\alpha}|\hat{\alpha}, V_\alpha)} \qquad (2)$$

$$\propto p(\beta)V_\alpha^{1/2}\prod_{\lambda=1}^{\Lambda}(\sigma_\lambda^2 + \beta^2)^{-1/2}\exp(-\frac{(\widehat{\Gamma}_\lambda - \widehat{\alpha})^2}{2(\sigma_\lambda^2 + \beta^2)}) \qquad (3)$$

where $\hat{\alpha}$ and V_α are defined as:

$$\hat{\alpha} = \frac{\sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_\lambda^2 + \beta^2}\hat{\Gamma}_\lambda}{\sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_\lambda^2 + \beta^2}}, \qquad (4)$$

This work was partially supported by grants from the National Institute of Health (EY01115), The Foundation Fighting Blindness, Sramek Foundation and Research to Prevent Blindness.

1-4244-0385-5/06/$20.00 ©2006 IEEE

and

$$V_\alpha^{-1} = \sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_\lambda^2 + \beta^2}. \tag{5}$$

3. Draw α from $p(\alpha|\beta, y)$. Combining the data with the uniform prior density $p(\alpha|\beta)$ yields,

   $$p(\alpha|\beta, y) \sim N(\hat{\alpha}, V_\alpha). \tag{6}$$

   where $\hat{\alpha}$ is a precision-weighted average of the $\hat{\Gamma}$'s and V_α is the total precision. Note, we define precision as inverse of variance.

4. Draw Γ_λ from $p(\Gamma_\lambda|\alpha, \beta, y)$

   $$p(\Gamma_\lambda|\alpha, \beta, y) \sim N(\hat{\Theta}_\lambda, V_\lambda), \tag{7}$$

   where $\hat{\Theta}_\lambda, V_\lambda$ are defined as:

   $$\hat{\Theta}_\lambda = \frac{\frac{1}{\sigma_\lambda^2}\hat{\Gamma}_\lambda + \frac{1}{\beta^2}\alpha}{\frac{1}{\sigma_\lambda^2} + \frac{1}{\beta^2}}, \tag{8}$$

   and

   $$V_\lambda = \frac{1}{\frac{1}{\sigma_\lambda^2} + \frac{1}{\beta^2}}. \tag{9}$$

5. Take hyperbolic tangent transformation of $\hat{\Gamma}_\lambda$, i.e.

   $$\hat{\rho}_\lambda = \tanh(\hat{\Gamma}_\lambda), \tag{10}$$

   and the sampling distribution $\hat{\rho}_\lambda$ is the desired posterior distribution.

3. SIMULATIONS

We evaluated the performance of the full Bayesian estimation by comparing with the marginal estimation in terms of Mean Squared Error (MSE) and variance. Fig. 2 plots MSE's and variances of the Bayesian correlation estimation and the marginal correlation estimation over 500 simulations. It is evident in Fig. 2 that the MSE of Bayesian estimation is about three-fold smaller than that of the marginal estimation. A much dramatic contrast was observed at the small sample size that clearly shows the advantages of the Bayesian estimations for the small sample size problem. Comparison of variances follows the same trend (Fig. 2).

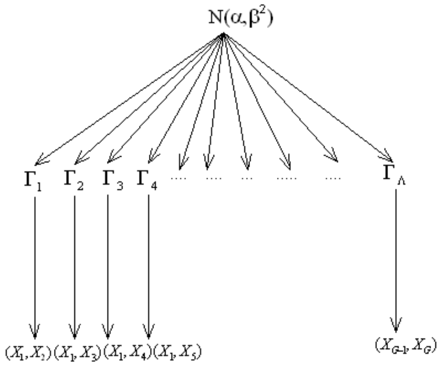

Fig. 1. *Bayesian hierarchical model structure.*

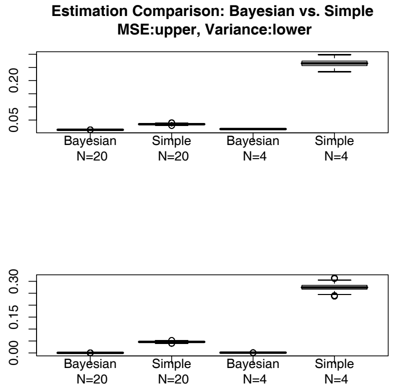

Fig. 2. *Mean Squared Errors (MSE's) and Variances of the Bayesian estimations versus the simple estimations over* 500 *runs of simulations.*

4. ESTIMATING CO-EXPRESSION NETWORK FROM GALACTOSE METABOLISM DATA

We also demonstrated the application of our Bayesian approach and compared it with the previous Frequentist approach [3] using a yeast galactose metabolism two-color microarray data [5]. Following the procedure in method section, we simulated the empirical posterior distribution for each correlation parameter Γ. Similar to the previous analysis, we used 0.6 as the correlation cutoff value, and declared the statistical association to be biologically relevant when their $(1 - q) \times 100\%$ *posterior confidence intervals* do not intersect with [-0.6, 0.6] at the significant level q. Comparison of the networks inferred from Bayesian hierarchical model and from the previous approach in terms of top hub nodes shows much agreement with certain discrepancies.

5. REFERENCES

[1] Butte,A., Tamayo,P. Slonim,D., Golub,T.R. and Kohane,I.S. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci USA*, **97**, 12182-6.

[2] Schfer, J. and Strimmer, K. (2005) An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**(6), 754-764.

[3] Zhu, D., Hero, A.O., Qin, Z.S., Swaroop, A. (2005) High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS). *J Comput Biol*, **12**(7), 1029-1045.

[4] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) Bayesian Data Analysis. *Chapmann & Hall/CRC*, Boca Raton, FL, USA.

[5] Ideker,T., Thorsson,V. et al. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol*, 7(6): 805-17.