

BAYESIAN HIERARCHICAL MODEL FOR ESTIMATING GENE ASSOCIATION NETWORK FROM MICROARRAY DATA

Dongxiao Zhu,^a and Alfred O Hero^b

^aBioinformatics Program, ^bDepartments 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 “overfitting” problem especially for analyzing small sample size data. Approaches that are able to globally estimate the correlation parameters with variance regularization followed by seamless correlation thresholding are highly desirable.

The desirable approaches fall naturally into the framework of Bayesian hierarchical model [4]. By assuming all correlation parameters are *exchangeable* and sampled from a common population distribution, we regularize variances of the marginal correlations by specifying a parent normal distribution from which marginal correlation parameters are sampled from. 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 needs to be estimated, denoted as ρ_λ , $\lambda = 1, \dots, \Lambda$. Let Γ denote the hyperbolic arc-tangent transformation of ρ , that is,

$$\Gamma_\lambda = \text{atanh}(\rho_\lambda), \quad (1)$$

Then parameters Γ_λ are asymptotically normal distributed with stabilized variances, i.e. $\sigma_\lambda^2 = 1/(N - 3)$, N is the sample size. Simulation studies show that the variance approximation works reasonably well even at a relatively small sample size, e.g. $N = 20$. We assume known variances due to the consideration of computational complexity. Furthermore, we don't have *a priori* information about these Γ 's, and assuming independency between them in marginal correlation approaches cause “overfitting” problem [2]. In the Bayesian hierarchical model framework, we assume that these parameters are *exchangeable*, and are drawn from a normal distribution with unknown hyperparameters (α, β) (Fig. 1a):

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

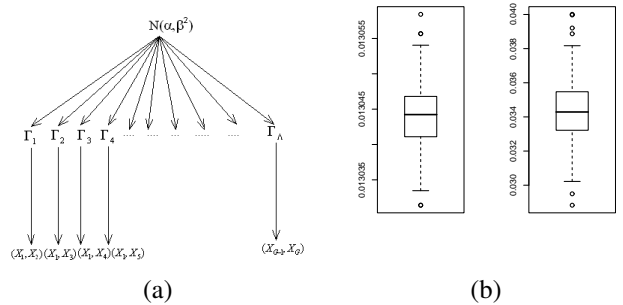


Fig. 1. (a) Bayesian hierarchical model structure. (b) Mean Squared Errors of Bayesian hierarchical model estimation (left) and Marginal correlation estimation (right) over 500 runs of simulations.

In order to generate conditional posterior distributions $p(\Gamma_\lambda | \alpha, \beta, y)$ for each parameter Γ_λ , $\lambda = 1, \dots, \Lambda$, we performed simulation steps as follows [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. Assign prior distribution for β , i.e. uniform prior distribution $p(\beta) \propto 1$.
2. Draw β from posterior distribution $p(\beta|y)$.

$$p(\beta|y) \propto \frac{p(\beta) \prod_{\lambda=1}^{\Lambda} N(\hat{\Gamma}_{\lambda}|\hat{\alpha}, \sigma_{\lambda}^2 + \beta^2)}{N(\hat{\alpha}|\hat{\alpha}, V_{\alpha})} \quad (3)$$

$$\propto p(\beta) V_{\alpha}^{1/2} \prod_{\lambda=1}^{\Lambda} (\sigma_{\lambda}^2 + \beta^2)^{-1/2} \exp\left(-\frac{(\hat{\Gamma}_{\lambda} - \hat{\alpha})^2}{2(\sigma_{\lambda}^2 + \beta^2)}\right) \quad (4)$$

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

$$\hat{\alpha} = \frac{\sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_{\lambda}^2 + \beta^2} \Gamma_{\lambda}}{\sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_{\lambda}^2 + \beta^2}}, \quad (5)$$

and

$$V_{\alpha}^{-1} = \sum_{\lambda=1}^{\Lambda} \frac{1}{\sigma_{\lambda}^2 + \beta^2}. \quad (6)$$

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

$$p(\alpha|\beta, y) \sim N(\hat{\alpha}, V_{\alpha}). \quad (7)$$

where $\hat{\alpha}$ is precision-weighted average of the Γ and V_{α} is the total precision. Note, we define precision as inverse of variances.

4. Draw Γ_{λ} from $p(\Gamma_{\lambda}|\alpha, \beta, y)$

$$p(\Gamma_{\lambda}|\alpha, \beta, y) \sim N(\hat{\Gamma}_{\lambda}, V_{\lambda}), \quad (8)$$

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

$$\hat{\Gamma}_{\lambda} = \frac{\frac{1}{\sigma_{\lambda}^2} \Gamma_{\lambda} + \frac{1}{\beta^2} \alpha}{\frac{1}{\sigma_{\lambda}^2} + \frac{1}{\beta^2}}, \quad (9)$$

and

$$V_{\lambda} = \frac{1}{\frac{1}{\sigma_{\lambda}^2} + \frac{1}{\beta^2}}. \quad (10)$$

5. Take hyperbolic tangent transformation of Γ_{λ} .

$$\rho_{\lambda} = \tanh(\Gamma_{\lambda}), \quad (11)$$

the sampling distribution of ρ_{λ} 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. 1b plots MSE's of Bayesian hierarchical model estimation (left) and marginal correlation estimation (right) over 500 simulations. It is evident in Fig. 1b that the MSE of Bayesian estimation is about three-fold smaller than that of the marginal estimation. Comparison of variances follows the same trend.

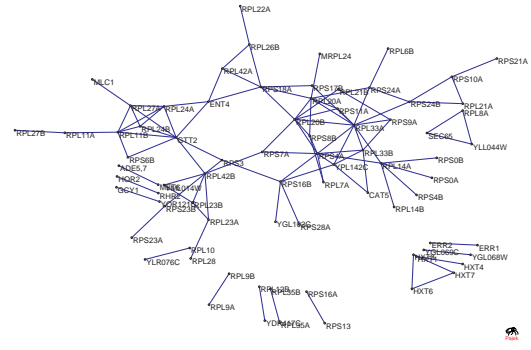


Fig. 2. The network visualization. The network is assembled by screening edges whose posterior intervals do not intersect with [-0.6, 0.6].

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%) posterior confidence intervals do not intersect with [-0.6, 0.6] at the significant level q . Fig. 2 presents a network assembled from screened edges in which all 95% posterior confidence intervals do not intersect with [-0.6, 0.6]. 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. *Chapman & 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.