Bayesian Hierarchical Model for Large-Scale Covariance Matrix Estimation

Dongxiao Zhu[a] and Alfred O Hero, III[b]
Stowers Institute for Medical Research[a],
Departments of EECS, Statistics and Biomedical Engineering,
University of Michigan, Ann Arbor[b]
{doz}@stowers-institute.org {hero}@umich.edu

Abstract

Many bioinformatics problems can implicitly depend on estimating large-scale covariance matrix. The traditional approaches tend to give rise to high variance and low accuracy estimation due to "overfitting", and hence not completely satisfactory. We cast the large-scale covariance matrix estimation problem into the Bayesian hierarchical model framework, and introduce dependency between covariance parameters. We demonstrate the advantages of our approaches over the traditional approaches using simulations and an exemplary omics data analysis.

Estimating covariance matrix from high-throughput "omics" data is indispensable for many tasks, notably for finding clusters in the data, whether of the hierarchical or network flavor. The problem remains to be challenging due to the large number of variables *p* (such as genes or proteins) and the comparatively small number of samples *n* (such as conditions under which gene expression is measured). The existing approaches that rely on the maximum likelihood estimation or the related unbiased empirical covariance matrix suffer from low accuracy and high variance inherent in any "large p, small n" type of data. A regularized and conditioned covariance matrix would be a great improvement over the unconstrained simple estimation of the covariance matrix in the high-throughput omics data setting. Estimation of such a matrix is a difficult problem because relatively few observations do not provide adequate degree of freedom to draw reliable statistical inference on tens of thousands of correlation parameters. Proper constraints need to be imposed on these parameters to overcome this difficulty.

There are two sets of existing approaches. One is based on the pairwise correlation estimation followed by the variance reduction techniques such as bagging (Hastie *et al*. 2001) and bootstrap aggregation (Breiman 1996). Representative work includes the full order (also called Gaussian Graphical Modeling (GGM)) partial correlation estimation approach (Schafer and Strimmer 2005a). It introduced a Bayes model from which all correlations are estimated using an Empirical Bayes method (Schafer and Strimmer 2005a). Another is to obtain improved estimates of the covariance matrix via shrinkage combined with analytic determination of the shrinkage intensity according to the Ledoit-Wolf theorem (Ledoit and Wolf, 2003). The authors showed that the new regularized estimator greatly enhances inferences of gene association networks using synthetic data (Schafer and Strimmer 2005b). Their approach is based on the assumption that the omics data is independently and identically distributed (i.i.d) p-variate observations sampled from a p-variate Gaussian distribution with the $(p \times p)$ covariance matrix of interest. The assumption is plausible only for small sized homogenous data because the underlying statistical distribution of larger sized heterogenous data is often mixed (Yeung *et al*. 2001). In both approaches, dependency was introduced among the correlation parameters but with different ways.

We advocate the framework of the first set of approaches since it does not reply on the stringent assumption, and it's usage has been demonstrated by numerous biological examples. We improve over the existing Empirical Bayes method by providing a full Bayesian treatment of the problem. In the Bayesian framework, we derive the posterior distribution for each correlation parameter based on the observed the $n \times p$ data matrix. The posterior distributions allow the statistical inference of the correlation parameters to be conveniently drawn.

In our previous work (Zhu *et al*. 2005), we described an error control procedure based on correlation statistic that simultaneously controls statistical significance and biological significance of the estimated covariance matrix. The correlation statistic works reasonably well for the data with relatively large sample size. However, it has poor accuracy for data with small sample size due to overfitting (Ledoit and Wolf 2004, Schafer and Strimmer 2005b). Introducing some form of strong dependency among correlation parameters can lead to improved accuracy in this small sample situation. Many approaches to introducing dependency can be adopted. Bayesian hierarchical models accomplish this introduction of dependency in a simple but effective manner.

The remainder of the paper is organized into five parts: Introduction of Bayesian Hierarchical Model for large-scale covariance matrix estimation (Sec. 2); Simulation studies of comparing the Bayesian estimator versus simple estimator (Sec 3); Analyzing the galactose metabolism data using proposed Bayesian approach and compared with the traditional approach (Sec 4); Conclusion and discussion (Sec 5).

1 The Bayesian Hierarchical Model of Covariance Matrix

The framework of Bayesian hierarchical models is a powerful technique that allows for high complexity of modeling structure without a large number of parameters (pairwise correlation parameters in this context) (Gelman *et al*. 2004). We assume the correlation parameters are *exchangeable* meaning that their joint distribution is invariant to permutations of their indexes. It represents a kind of topological invariance that imposes prior assumptions on the location of high correlations in the network. We then regularize variances of the marginal correlation densities by specifying a parent Gaussian distribution from which marginal correlation parameters are sampled. Using a prior population distribution we are able to introduce dependency into the parameters that tends to avoid problems of overfitting. Using quantiles of posterior distributions of the correlation parameters provide a seamless combination of correlation estimation and strength thresholding that can be used as an alternative to FDR-CI methods (Benjamini and Yekutieli 2005, Zhu *et al*. 2005) for small samples.

Without loss of generality, we employ marginal correlation coefficient to demonstrate the Bayesian hierarchical model for large-scale (marginal) correlation matrix estimation. The model can be easily extended for large-scale partial correlation matrix estimation, and we will discuss this issue in section 5. We use ρ to denote the true correlation coefficient between a pair of gene expression profiles (Bickel and Doksum 2000). Specifically, let $X_{g_j(n)}$ be the n-th condition index of the i-th gene profile and let $S_{X_{g_i},X_{g_i}}$, $S_{X_{g_j},X_{g_j}}$, and $S_{X_{g_i},X_{g_j}}$ are sample variances and covariance defined as:

$$S_{X_{g_i},X_{g_i}} = (N-1)^{-1}\sum_{n=1}^{N}(X_{g_i(n)} - \overline{X_{g_i}})^2,$$

$$S_{X_{g_j},X_{g_j}} = (N-1)^{-1}\sum_{n=1}^{N}(X_{g_j(n)} - \overline{X_{g_j}})^2,$$

$$S_{X_{g_i},X_{g_j}} = (N-1)^{-1}\sum_{n=1}^{N}(X_{g_i(n)} - \overline{X_{g_i}})(X_{g_j(n)} - \overline{X_{g_j}}).$$

The true correlation coefficient is defined as

$$\rho = \frac{\mathrm{E}[S_{X_{g_i},X_{g_j}}]}{\sqrt{\mathrm{E}[S_{X_{g_i},X_{g_i}}]\mathrm{E}[S_{X_{g_j},X_{g_j}}]}}, \tag{1}$$

where E[.] is statistical expectation. For G gene expression profiles in a gene microarray sequence, there are $\Lambda = \binom{G}{2}$ of these 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 ρ_λ and stabilized variance approximations of $\sigma_\lambda^2 = 1/(N-3)$ (Fisher 1923), here N is the sample size. We define $\Gamma_\lambda = \mathrm{atanh}(\rho_\lambda)$ as the corresponding transformed true correlation coefficients.

Our previous simulation studies showed that this variance approximation works reasonably well even at a relatively small sample size (e.g. $N < 10$) (Zhu *et al*. 2005). In this sequel we assume known variance of the transformed correlation matrix to reduce computational complexity of the full Bayesian correlation matrix estimation. In case of unknown variances, the conditional posterior distribution can not generally be written in closed form, for this reason, Markov Chain Monte Carlo (MCMC) techniques might be applied but at high cost.

From our assumption that the $\{\rho_\lambda\}_{\lambda=1}^{\Lambda}$ are *exchangeable* we model $\{\rho_\lambda\}_{\lambda=1}^{\Lambda}$ as random variables drawn from a Gaussian distribution with unknown hyperparameters (α, β^2) (Fig. 1).

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

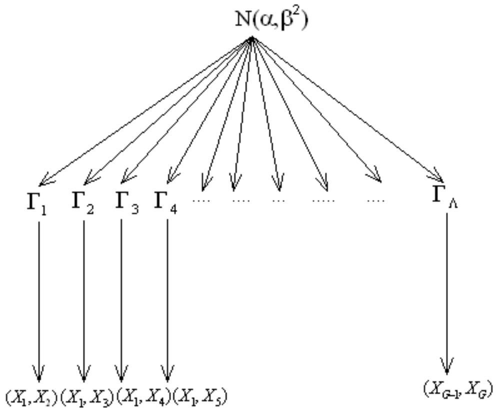

Figure 1: Bayesian hierarchical model structure (Gelman *et al*. 2004, Chapter V).

where $P(\Gamma_\lambda|\alpha,\beta^2)$ is a Gaussian distribution with mean α and variance β^2.

In order to generate conditional posterior distributions $p(\Gamma_\lambda|\alpha,\beta,y)$ for each parameter $\Gamma_\lambda, \lambda = 1,\ldots,\Lambda$, we performed simulation steps as follows: (Gelman *et al*. 2004, Chapter V) (refer to Appendix for details):

- Assign prior distribution for β, e.g. uniform prior distribution $p(\beta) \propto 1$. Note, the choice of uniform prior yields a *proper* posterior density while other *noninformative* prior distributions such as, $p(\beta) \propto \beta^{-1}$ do not. (refer to Appendix for mathematical proof.)

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

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

$$\propto p(\beta)V_\alpha^{1/2}\prod_{\lambda=1}^{\Lambda}(\sigma_\lambda^2+\beta^2)^{-1/2}\exp(-\frac{(\widehat{\Gamma}_\lambda-\widehat{\alpha})^2}{2(\sigma_\lambda^2+\beta^2)}), \tag{4}$$

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

$$\hat{\alpha} = \frac{\sum_{\lambda=1}^{\Lambda}\frac{1}{\sigma_\lambda^2+\beta^2}\hat{\Gamma}_\lambda}{\sum_{\lambda=1}^{\Lambda}\frac{1}{\sigma_\lambda^2+\beta^2}}, \tag{5}$$

and

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

See Appendix for detailed derivation of $p(\beta|y)$.

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

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

  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.

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

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

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

  $$\hat{\Theta}_\lambda = \frac{\frac{1}{\sigma_\lambda^2}\hat{\Gamma}_\lambda + \frac{1}{\beta^2}\alpha}{\frac{1}{\sigma_\lambda^2}+\frac{1}{\beta^2}}, \tag{9}$$

  and

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

  The atanh-transformed posterior mean correlation coefficient $\hat{\Theta}_\lambda$ is a precision-weighted average of the prior population mean α and the λth sample mean $\hat{\Gamma}_\lambda$.

The posterior distribution (Eq. 8) contains all the current information about the atanh-transformed parameter ρ_λ. In particular, the *posterior mean* and *posterior interval* are derived as the following:

$$\begin{aligned} \mathrm{E}[\Gamma_\lambda] &= \mathrm{E}[\operatorname{atanh}(\rho_\lambda)] \\ &= \operatorname{atanh}(\mathrm{E}[\rho_\lambda]) = \hat{\Theta}_\lambda. \end{aligned} \tag{11}$$

Applying function tanh to both sides of the Eq. 11, we have,

$$\mathrm{E}[\rho_\lambda] = \tanh(\hat{\Theta}_\lambda). \tag{12}$$

For deriving the posterior interval of the ρ_λ, we used the fact that the cumulative density function (cdf) of $\Gamma_\lambda{}' = \frac{\Gamma_\lambda - \hat{\Theta}_\lambda}{\sqrt{V_\lambda}}$ is Φ, the cdf of standard Gaussian random variable. Hence, we define its quantile function as Φ^{-1}, and write down the $(1-q)\times 100\%$ posterior interval of the parameter $\Gamma_\lambda{}'$:

$$I^{\Gamma_\lambda{}'}(q) : [\Phi^{-1}(q/2), \Phi^{-1}(1-q/2)]. \tag{13}$$

After some algebraic derivation and based on the fact that tanh is a monotonically increasing function, we have a $(1-q)\times 100\%$ posterior interval for the parameter ρ_λ:

$$I^{\rho_\lambda}(q) : [\tanh(\sqrt{V_\lambda}(\Phi^{-1}(q/2)) + \hat{\Theta}_\lambda), \tanh(\sqrt{V_\lambda}(\Phi^{-1}(1-q/2)) + \hat{\Theta}_\lambda)]. \tag{14}$$

2 Simulation Studies

2.1 Comparisons in terms of Confidence Interval, Mean Squared Error, and Variance

We evaluated the performance of full Bayesian hierarchical model estimation of correlations and compared with the frequentist method in Zhu *et al*. 2005. We define the frequentist CI as follows: If L and U are statistics (i.e., observable random variables) whose probability distribution depends on some unobservable parameter θ, and

$$Pr(L \leq \theta \leq U) = q, q \in (0,1),$$

then the random interval [L,U] is a $(1-q) \times 100\%$ *confidence interval* for θ. A frequentist interval may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated trials, while a *Bayesian (confidence) interval* for an unknown quantity of interest can be directly regarded as having a high probability of containing the unknown quantity. Therefore, Bayesian approach where a reliable prior is available, facilitates a common-sense interpretation of statistical conclusions (Gelman *et al*. 2004).

We first compared two point estimators of correlations in terms of the average width of the individual frequentist (Pearson) CI's for the correlation parameters versus that of the posterior CI's for the same set of correlation parameters at the corresponding significance levels. Obviously, more concentrated (narrower) CI's, at the given significance level, are superior to less concentrated CI's. It is clear from Fig. 2 and Fig. 3 that the average Bayesian posterior CI's are uniformly narrower than the average freqentist CI's in both small ($N = 4$) and larger sample data ($N = 20$). This dramatic contrast indicates the advantages of Bayesian approach for small sample size problems (Fig. 3). From Eqs. 22 and 3, the posterior distributions of the mean $p(\alpha|\beta,y)$ and of the variance $p(\beta|y)$ are decreasing functions of Λ, i.e., the number of correlation parameter $\Gamma' s$. Therefore, narrower posterior CI's are expected for larger Λ. On the other hand, wider CI's are expected when transforming individual frequentist CI's into simultaneous FDR-CI's.

We also compared these two correlation estimators in terms of Mean Squared Error (MSE) and variance criteria. Similar to the definition in Zhu *et al*. 2005, the MSE is defined as:

$$MSE = \frac{1}{\Lambda}\sum_{\lambda=1}^{\Lambda}(\hat{\rho}_\lambda - \rho_\lambda)^2, \tag{15}$$

where ρ_λ is the true population correlation, and $\hat{\rho}_\lambda$ is the sample correlation estimator, λ is the parameter index, and Λ is the total number of parameters.

The simulation steps proceed as follows:

- Draw Λ population correlations from a normal distribution with known mean (α) and variance (β) (hyperparameters) as defined in Eq. 2.

- Re-estimate the Λ parameters either separately using the frequentist (Pearson) correlation estimator or using Bayesian hierarchical model. For the Bayesian approach, the correlation estimator is the posterior mean (Eq. 11).

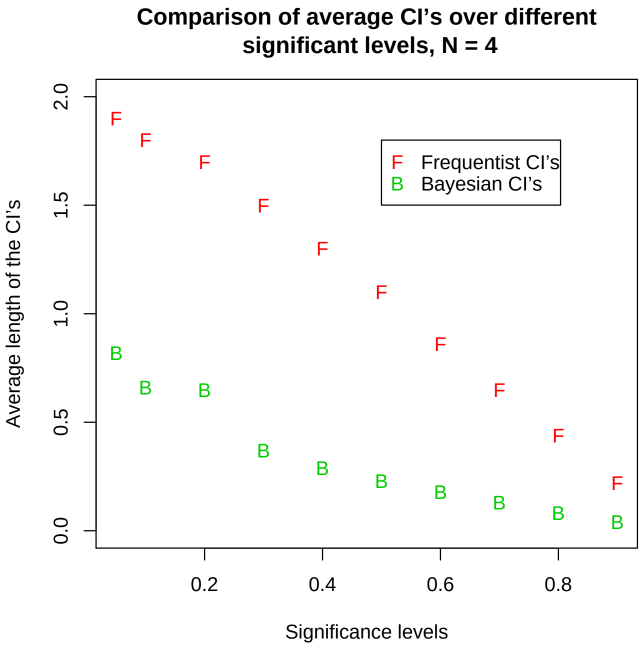

Figure 2: Comparison of average posterior CI's versus average individual frequentist CI's over a wide range of significance levels at a small sample size ($N = 4$).

- Compare the two estimators in terms of both MSE and variance. An estimator with low MSE and variance are considered to be superior.

Fig. 4 plots MSE's (upper panel) and variances (lower panels) of Bayesian correlation estimators and frequentist (Pearson) correlation estimators at a small sample size (e.g. $N = 4$) and a larger sample size (e.g. $N = 20$) over 500 runs of simulations. It is evident in upper panel of the Fig. 4 that the MSE of Bayesian estimators is about threefold smaller than the frequentist estimators for larger sample size. Similarly to the CI's comparisons, this indicates the advantages of the Bayesian correlation estimator for the small sample size problems (Fig. 4). The lower panel of the Fig. 4 plots variances of the Bayesian correlation estimator and the frequentist correlation estimator. Again, the comparison of variances follow the same trend as that of the MSE's (Fig. 4).

It is worth mentioning that the above simulations were biased towards the assumptions of Bayesian hierarchical model. In order to test robustness of our algorithm to model mismatch, we also generated data using the uniform distribution but implemented with Pearson CI's and Bayesian CI's that assume mismatched Gaussian and hierarchical models, respectively. In Fig. 5, we compared the average width of individual Pearson CI's with that of individual Bayesian intervals. The superior perfor-

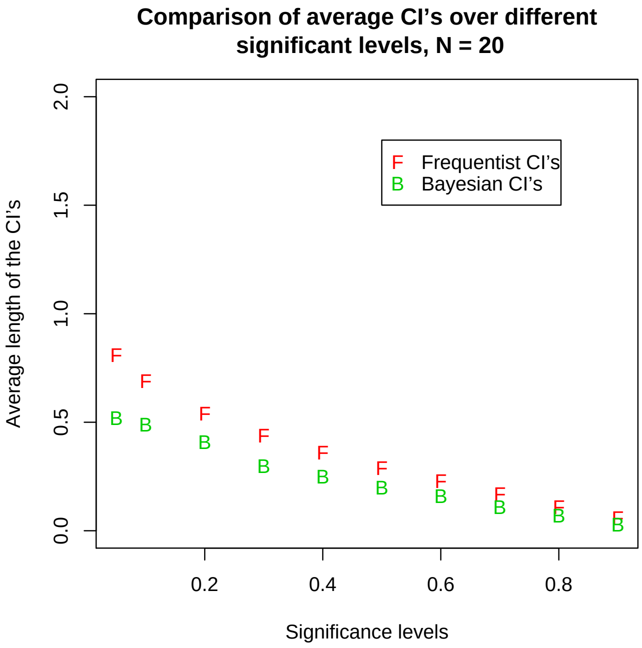

Figure 3: Comparison of average posterior CI's versus average individual frequentist CI's over a wide range of significance levels at a larger sample size ($N = 20$).

mance of hierarchical Bayesian estimator (Fig. 2, Fig. 3) is clearly offset by the invalid model assumption in that average Bayesian CI's are uniformly wider than average frequentist CI's (Fig. 5). This simulation results highlight the importance of Fisher transformation.

2.2 Evaluation of the Bayesian Hierarchical Model

In order to evaluate our Bayesian approach in terms of error control and compare with the frequentist counterpart, we simulated pairwise gene expression data based on known population covariances, and then simulated Bayesian intervals for each parameter from the hierarchical model. The actual False Positive (FP) at a given MAS level is calculated as a ratio of the number of screened gene pairs whose corresponding population correlation parameters $\rho_{i,j}$ are less than the MAS level specified, divided by the total number of gene pairs. The actual MAS is the minimum true discovery of population correlation $\rho_{i,j}$ among the screened pairs. We specified 16 pairs of (FP,MAS) criteria (Four FP levels: 0.2, 0.4, 0.6, 0.8; Four MAS levels: 0.2, 0.4, 0.6, 0.8), and each is plotted as a different upper case Roman alphabet (Red) in Fig. 6. The 16

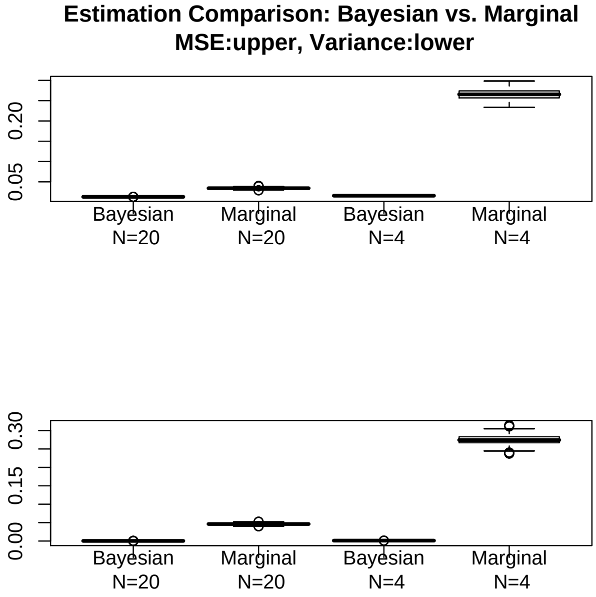

Figure 4: Mean Squared Errors (MSE's) and Variances of the Bayesian estimations versus the simple estimations over 500 runs of simulations.

corresponding pairs of actual (FP,MAS) criteria are also shown in Fig. 6 using the same set of lower case Roman alphabets (Blue). It can be observed that generally the actual FP's (lower case) fall further below the specified constraint (upper case) than those did in Fig. 4 of Zhu *et al*. 2005 (Fig. 6), and the actual MAS's (lower case) fall above the specified constraints (upper case). The more dramatic deviations of actual FP's from their specified levels are due to multiple factors, such as, lack of multiplicity adjustment and the conservative asymptotic approximation. Simulations using some other combinations of N and Λ, as compared with the FDR-CI approach, give rise to the similar results. We conclude that Bayesian hierarchical model yields better correlations estimates. However, the false positive rate is overestimated by the Bayesian procedure and this leads to overly stringent error control.

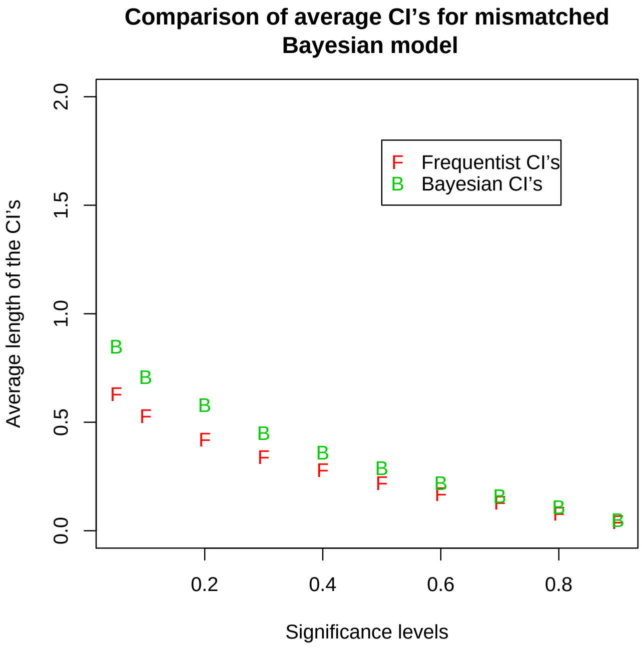

Figure 5: Comparison of average CI's when the Bayesian model is unsustained.

3 Applications to Network Construction and Seeded Clustering

3.1 Constructing Relevance Networks

We applied the Bayesian hierarchical model to high-throughput data and compared it with the frequentist approach using the same subset of yeast galactose catabolism two-color microarray data that was described in Zhu *et al*. 2005. The data contains 997 gene expression profiles across 20 genetic/physilogical conditions that was identified by Ideker et al using the generalized likelihood ratio test (Ideker *et al*. 2000).

Following the procedure described in section 1, we simulated the empirical posterior distribution for each of the $\binom{997}{2} = 496,506$ correlation parameters ρ_λ. The $(1-q) \times 100\%$ *posterior interval* for each 'parameter' was obtained by thresholding $q/2 \times 100\%$ and $(1-q/2) \times 100\%$ of it's quantile function (Eq. 14). Analogous to the FDR-CI screening procedure described in
citealtZhu05a, a network edge is declared to be present at the significance level q and the MAS level *cormin* if it's posterior CI does not intersect with $[-cormin, cormin]$. We sought to compare the two approaches in terms of network topological properties

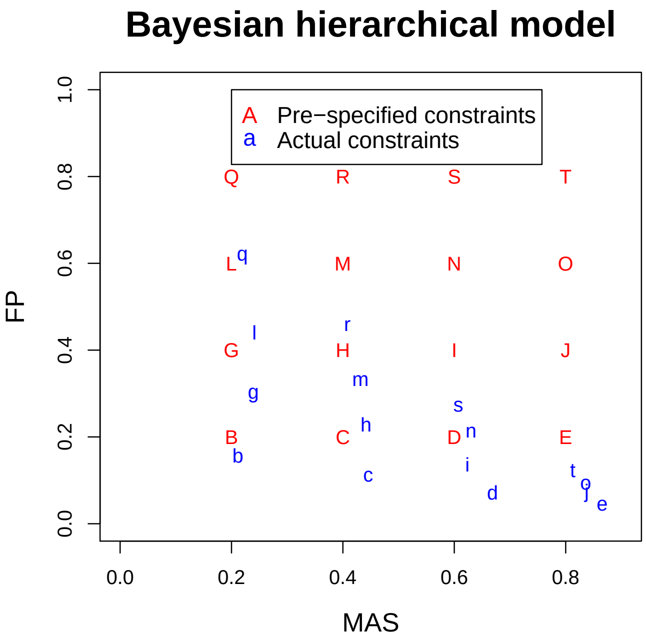

Figure 6: Evaluation of error control of the Bayesian hierarchical model. Sample size $N = 20$, and $\Lambda = 1000$ correlation coefficients were simulated. Simulations using smaller sample size data yield more stringent error control.

that are interesting to the biologists. In particular, we compared the biological functional annotations of the top hub genes of the two networks. In Zhu *et al*. 2005, we controlled FDR at 5%, and constructed networks at five MAS levels, i.e. 0.5, 0.6, 0.7, 0.8, 0.9. Correspondingly, 18135, 9337, 4151, 1346, 133 edges were declared to be present using Pearson correlation statistic alone. Controlling the significance level at 5%, we screened the same set of numbers of edges using Bayesian hierarchal model to construct the five networks that are more comparable to those in Zhu *et al*. 2005. A list of stable hub genes were obtained by calculating and sorting the average rank of each vertex (gene) degree over five networks (Table 1).

Comparing the Table 1 with the that was reported in Zhu *et al*. 2005, note that the GO biological process annotation “protein biosynthesis[GO:0006412]” and/or it’s children annotations “hypusine biosynthesis[GO:0046515]”, “branched chain family amino acid biosynthesis[GO:0009082]”, and “tryptophan biosynthesis[GO:0000162]” are significantly enriched in both tables. This is consistent with the established fact that protein biosynthesis plays a key role in galactose metabolism (Berg *et al*. 2006). The underlying biological mechanism is that many types of proteins need to be syn-

Table 1: Top twenty "hub genes" from Bayesian hierarchical model applied to the galactose metabolism data (Ideker *et al*. 2000). The rank of each gene is the average rank over five different networks with the same set of edge numbers as in Table 1 of Zhu *et al*. 2005. The highest ranked gene is the most connected and stable gene under varying constraints of (FP,MAS).

Gene Name	Average Rank	GO Annotation
YJR070C	4	hypusine biosynthesis[GO:0046515]
YBR043C	4.4	multidrug transport[GO:0006855]
AGA2	4.4	agglutination[GO:0000771]
RPP0	4.6	protein biosynthesis[GO:0006412]
RPL26A	4.6	protein biosynthesis[GO:0006412]
YOR263C	5	biological process unknown
TRP2	5.4	tryptophan biosynthesis[GO:0000162]
ASC1	5.6	regulation of protein biosynthesis[GO:0006417]
YIL064W	5.6	biological process unknown
BOP2	5.6	biological process unknown
GAP1	5.8	amino acid transport[GO:0006865]
RPS2	6	protein biosynthesis[GO:0006412]
RPL11A	6.2	protein biosynthesis[GO:0006412]
SSF2	6.2	ribosomal subunit assembly[GO:0042257]
ILV5	6.2	branched chain family amino acid biosynthesis[GO:0009082]
YPL185W	6.2	biological process unknown
PCK1	6.4	hexose biosynthesis[GO:0019319]
YDR100W	6.4	biological process unknown
YMR291W	6.6	biological process unknown
ATC1	6.6	bipolar bud site selection[GO:0007121]

thesized upon switching from primary carbon source (glucose) to secondary carbon source (galactose) or the other way around (Wieczorke *et al*. 1999).

A salient feature in Table 1 that is not possessed in that of Zhu *et al*. 2005 is that it includes several transporters and regulators such as GAP1[GO:0006865], YBR043C[GO:0006855], and ASC1[GO:0006417] etc. These proteins are essential for a smooth transition from glucose to galactose (Berg *et al*. 2006, Wieczorke *et al*. 1999). In addition, Table 1 also includes several uncharacterized genes that are hypothesized to be important for galactose metabolism. In general, the Bayesian data analysis results not only conform to the previous frequentist data analysis results, but also provide additional justification for the biological mechanism and motivation for illustrating new gene functions.

3.2 Seeded Clustering

In parallel with the application of the two-stage algorithm to rediscover the galactose metabolic pathway reported in Zhu *et al*. 2005, we also applied the Bayesian hierarchical model to perform the seeded (one-to-all) clustering. Performance was evaluated according to the relative ranks of a handful of known members of the galactose metabolic pathway. The gene ranks were used instead of p-values due to substantial differences of the two statistical frameworks.

We selected gene "GAL10" as the "seed gene" in order to compare the results with those reported in Zhu *et al*. 2005. The comparison was made at a large sample size $N = 20$ and a smaller sample size $N = 4$ respectively aiming to examine the performance of the two methods as a function of the sample size. In the former, we used all the 20 genetic/physiological conditions under which gene expression levels were measured (Table 2); In the later, we sampled a small subset (e.g. $N = 4$) of these 20 conditions each time without replication and repeated a number of times to obtain a "bagged" (stable) estimation of gene ranks in the seeded clusters (Table 2).

When all the 20 observations were used, the two approaches give rise to very similar seeded clusters indicating that the Bayesian hierarchial model approach is as powerful as the frequentist approach for relatively large sample size problems. As shown in Table 2, all of the top 20 seeded gene pairs have the identical rank across two methods. When multiple random subset data were used, many genes have dissimilar average ranks across the two approaches. Among the top five genes (GAL10, GAL7, GCY1 GAL1, GAL2) screened by the seeded clustering using "GAL10" as the seed gene (see Zhu *et al*. 2005 and Table 3), 4 out of 5 (GAL10, GAL7, GAL1, GAL2) genes rank higher in Bayesian estimation than those in marginal estimation, and the remaining "GCY1" gene receives tie ranks. In addition, our results provide strong experimental motivation for examining the genes that received higher ranks in the Bayesian analysis, for example, gene YEL057C. The evaluation using "GAL7" as the "seed gene" gave the similar results.

4 Discussion

Numerous previous studies have demonstrated the suitability of using gene co-expression networks for functional discoveries (e.g. Butte and Kohane 2000, Zhou *et al*. 2002). There are different approaches to building the co-expression network - in particular, different ways of estimating correlation matrix, of testing significance of these correlations, and of controlling the error rate have been proposed. We emphasize that our goal is to estimate correlation matrix with reduced variance and improved accuracy.

Towards this goal, the major improvement that we have made is that we provided a full Bayesian treatment that combines the correlation estimation and testing seamless. For the estimation, we improve over existing approaches by providing a regularized full Bayesian estimation. For the hypothesis test, the main improvement over the existing approaches is that we test whether the magnitude correlation is different from a non-zero positive number instead of 0. This allows for more stringent control of biological significance. For example, in small-sample data the traditional test declares many small but statistically significant correlations to be biologically relevant. However, these may be caused by non-biological effects such as spatial and positional effects of genes along the chromosome (Kluger *et al*. 2003).

Our framework is sufficiently general to be extended to many different correlation measures, such as full order (Schafer and Strimmer 2005a) and limited order (Fuente

Table 2: Comparison of Bayesian estimations versus Marginal estimations using "seeded" clustering at a small and a larger sample sizes. In the former, the ranks were averaged over 100 estimations, in each of which a subset data of sample size $N = 4$ was randomly sampled from the whole data of sample size $N = 20$. In the later, the ranks were obtained using the whole data of sample size $N = 20$.

	$N = 4$						$N = 20$	
Gene1	Gene2	Bayesian	Frequentist		Gene1	Gene2	Bayesian	Frequentist
GAL10	**GAL1**	5.25	5.35		**GAL10**	**GAL7**	1	1
GAL10	**GAL2**	6.65	7.4		**GAL10**	**GCY1**	2	2
GAL10	**GAL7**	6.7	6.85		**GAL10**	**GAL1**	3	3
GAL10	**GCY1**	7.7	7.7		**GAL10**	**GAL2**	4	4
GAL10	YOR121C	8.05	7.8		GAL10	YOR121C	5	5
GAL10	YEL057C	8.55	10.6		GAL10	YEL057C	6	6
GAL10	SSU1	8.6	7.65		GAL10	YDR010C	7	7
GAL10	FKS1	8.75	8.25		GAL10	SSU1	8	8
GAL10	PCL10	9.95	7.85		GAL10	PCL10	9	9
GAL10	YJL212C	11	8.85		GAL10	YJL212C	10	10
GAL10	MET14	11.1	10.4		GAL10	FKS1	11	11
GAL10	YDR010C	11.3	10.9		GAL10	MET14	12	12
GAL10	MCM1	11.35	12.3		GAL10	MCM1	13	13
GAL10	EXG1	11.85	13.1		GAL10	EXG1	14	14
GAL10	CRH1	12.05	12.95		GAL10	ARG1	15	15
GAL10	ARG7	12.8	12.3		GAL10	CRH1	16	16
GAL10	YPR157W	13.2	15.35		GAL10	PRY2	17	17
GAL10	PRY2	14.4	13.3		GAL10	YPR157W	18	18
GAL10	YKR012C	14.6	16.25		GAL10	YKR012C	19	19
GAL10	CPA2	16.15	14.85		GAL10	CPA2	20	20

Table 3: Clustering co-expressed genes with Bayesian hierarchical model at the significance level 5% using “GAL10” as the “seed gene”. Known genes in the pathway are in bold face ($N = 20$).

Gene1	Gene2	2.5%	50%	97.5%
GAL10	**GAL7**	0.699967273	0.843269806	0.919377659
GAL10	**GCY1**	0.695895931	0.83904824	0.917448689
GAL10	**GAL1**	0.685628575	0.824914454	0.906837751
GAL10	**GAL2**	0.664031223	0.817631953	0.903466008
GAL10	YOR121C	0.652511568	0.814118521	0.901500909
GAL10	YDR010C	0.574348042	0.77081336	0.875409524
GAL10	YEL057C	0.582835775	0.769743768	0.880618535
GAL10	SSU1	0.584487078	0.769335123	0.879019784
GAL10	PCL10	0.552529392	0.751817344	0.871763977
GAL10	YJL212C	0.543601479	0.747480187	0.862433646
GAL10	MET14	0.525320838	0.723128249	0.852859396
GAL10	FKS1	0.515021843	0.719874179	0.854759107
GAL10	MCM1	0.474061933	0.697313988	0.834101087
GAL10	EXG1	0.446476056	0.666889754	0.818233838
GAL10	ARG1	0.382292245	0.63708452	0.807736956
GAL10	CRH1	0.344971636	0.594425382	0.773435199
GAL10	PRY1	0.299057555	0.588919717	0.774038296
GAL10	YPR157W	0.29645952	0.576125639	0.765975044
GAL10	CPA2	0.303356019	0.571475575	0.745218878
GAL10	YKR012C	0.262900828	0.566724743	0.748081117

et al. 2004) partial correlation statistics. The rational is that these correlation statistics are asymptotically normal distributed through transformations (Hotelling 1953). Our approach is also not computational cumbersome. In deriving the posterior distributions of the correlation 'parameters', the conjugate prior and likelihood (i.e. Gaussian parental distribution) were assumed in order to keep the posterior distributions in a closed form. The computational load is thus greatly reduced and we avoided MCMC techniques, making the application to the larger networks become more feasible.

As discussed in Zhu *et al*. 2005, one should seek a good combination of level of significance and correlation strength. The Bayesian approach prescribed here imposes a model of the parameters as random variables sampled from a parental population distribution. This model structure allows the regularization of variances by introducing dependency between the parameters. Using simulations, we have shown the superior performance of Bayesian hierarchical model approach to marginal estimation approach, in terms of width of the CI's, MSE and variance, especially for small sample size. The posterior distribution provides a natural way of correlation thresholding that bridges between statistical correlation and biological relevancy.

Appendix

Selecting Prior Distribution

Here we present the mathematical details of choosing a prior as described in section 3.1. They were adapted from the solution to exercises 2.8 in Gelman *et al*. 2004.

We need to show the joint posterior density $p(\Gamma,\alpha,\beta|y)$ is improper if we select the hyperprior distribution $p(\beta) \propto \beta^{-1}$, while $p(\Gamma,\alpha,\beta|y)$ is proper if we select the hyperprior distribution $p(\beta) \propto 1$.

We first factor the joint posterior distribution $p(\Gamma,\alpha,\beta|y) \propto p(\beta|y)p(\alpha|\beta,y)p(\Gamma|\alpha,\beta,y)$. Note that $p(\alpha|\beta,y)$ and $p(\Gamma|\alpha,\beta,y)$ have proper densities. The joint posterior density $p(\Gamma,\alpha,\beta|y)$ is proper if and only if the marginal density $p(\beta|y)$ is proper, i.e. has a finite integral for β from 0 to ∞.

In Eq. 3.3, as β approaches 0, everything multiplying $p(\beta)$ approaches a nonzero constant limit $C(y)$. Thus the behavior of $p(\beta|y)$ near 0 is determined by the prior density $p(\beta)$. It is easy to show that the function $p(\beta) \propto 1/\beta$ is not integrable for any small interval around 0, and so it leads to a nonintegrable posterior density.

If prior density $p(\beta) \propto 1$, then the posterior density is integrable near zero. We need to examine the behavior as $\beta \to \infty$ and find an upper bound that is integrable. The exponential term is clearly less than or equal to 1. We can rewrite the remaining terms as $(\sum_{j=1}^{J}[\prod_{k\neq j}(\sigma_k^2+\beta^2)])^{-1/2}$. For $\beta > 1$ we make this quantity bigger by dropping all of the σ^2 to yield $\left(J\beta^{2(J-1)}\right)^{-1/2}$. An upper bound on $p(\beta|y)$ for β large is $p(\beta)J^{-1/2}/\beta^{J-1}$. When $p(\beta) \propto 1$, this upper bound is integrable if $J > 2$, and so $p(\beta|y)$ is integrable if $J > 2$.

Deriving Posterior Distribution $p(\beta|y)$

Here we present the mathematical details of deriving posterior distribution $p(\beta|y)$ as described in section 3.1. They were adapted from Chapter V of Gelman *et al*. 2004.

We factor the marginal posterior density of the hyperparameters as follows:

$$p(\alpha,\beta|y) = p(\alpha|\beta,y)p(\beta|y), \tag{16}$$

which is equivalent to:

$$p(\beta|y) = \frac{p(\alpha,\beta|y)}{p(\alpha|\beta,y)}. \tag{17}$$

We then derive $p(\alpha,\beta|y)$ and $p(\alpha|\beta,y)$ respectively as the following. For hierarchical model, we can simply consider the information supplied by data about the hyperparameters directly:

$$p(\alpha,\beta|y) \propto p(\alpha,\beta)p(y|\alpha,\beta). \tag{18}$$

For many problems, decomposition in Eq. 18 is of no help since $p(y|\alpha,\beta)$ cannot generally be written in closed form. For the Gaussian distribution, the marginal likelihood has a particularly simple form. The marginal distributions of the sample correlation $\widehat{\Gamma}_\lambda$ are independent (but not identically distributed) Gaussian:

$$p(\widehat{\Gamma}_\lambda|\alpha,\beta) \propto N(\alpha,\sigma_\lambda^2+\beta^2). \tag{19}$$

Thus we can write the marginal posterior density as

$$p(\alpha,\beta|y) \propto p(\alpha,\beta)\prod_{\lambda=1}^{\Lambda} N(\widehat{\Gamma}_\lambda|\alpha,\sigma_\lambda^2+\beta^2). \tag{20}$$

From inspection of Eq. 20 with β assumed known, and with a uniform conditional prior density $p(\alpha|\beta)$, where $p(\alpha|\beta,y)$ is also Gaussian, i.e.

$$p(\alpha|\beta,y) \propto N(\hat{\alpha},V_\alpha), \tag{21}$$

where

$$\hat{\alpha} = \frac{\sum_{\lambda=1}^{\Lambda}\frac{1}{\sigma_\lambda^2+\beta^2}\widehat{\Gamma}_\lambda}{\sum_{\lambda=1}^{\Lambda}\frac{1}{\sigma_\lambda^2+\beta^2}}, \tag{22}$$

and

$$V_\alpha^{-1} = \sum_{\lambda=1}^{\Lambda}\frac{1}{\sigma_\lambda^2+\beta^2}. \tag{23}$$

$\hat{\alpha}$ is a precision-weighted average of Γ's and V_α is the total precision. We define precision as inverse of variance. From Eqs. 17, 20 and 21,

$$p(\beta|y) = \frac{p(\alpha,\beta|y)}{p(\alpha|\beta,y)} \tag{24}$$

$$\propto \quad \frac{p(\beta)\prod_{\lambda=1}^{\Lambda}N(\Gamma_\lambda|\alpha,\sigma_\lambda^2+\beta^2)}{N(\alpha|\hat{\alpha},V_\alpha)} \tag{25}$$

This identity holds for any value of α, in particular, it holds if we set α to $\hat{\alpha}$, which makes evaluation of the expression quite simple.

$$p(\beta|y) \propto \frac{p(\beta)\prod_{\lambda=1}^{\Lambda} N(\widehat{\Gamma}_{\lambda}|\hat{\alpha},\sigma_{\lambda}^2+\beta^2)}{N(\hat{\alpha}|\hat{\alpha},V_{\alpha})} \tag{26}$$

$$\propto p(\beta)V_{\alpha}^{1/2}\prod_{\lambda=1}^{\Lambda}(\sigma_{\lambda}^2+\beta^2)^{-1/2}\exp(-\frac{(\widehat{\Gamma}_{\lambda}-\widehat{\alpha})^2}{2(\sigma_{\lambda}^2+\beta^2)}), \tag{27}$$

where $\hat{\alpha}$ and V_{α} are defined in Eqs. 22 and 23. Both expressions are functions of β, which means that $p(\beta|y)$ is a complicated function of β.

References

Benjamini,Y. and Yekutieli,D. 2005. False discovery rate adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, **100**,71-80.

Berg, J.M., Tymoczko, J.L. and Stryer, L. 2006. Biochemistry. W. H. Freeman, New York, USA.

Bickel,P.J. and Doksum,K.A. 2000. Mathematical statistics: basic ideas and selected topics. 2nd Edition. *Prentice Hall*, Upper Saddle River, NJ, USA.

Breiman,L. 2000. Bagging predictors. *Machine Learn*, **24**, 123-140.

Butte,A. and Kohane,I.S. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, **5**, 415-426.

Fisher, R.A. 1923. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 1-32.

Fuente, A., Bing, N., Hoeschele, I. and Mendes, P. 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, **20**, 3565-3574.

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

Hastie,T., Tibshirani,R. and Friedman,J. 2001. The Elements of Statistical Learning. Springer, New York, USA.

Hotelling,H. 1953. New light on the correlation coefficient and its transforms. *J. R. Statist. Soc. B*, **15**, 193232.

Ideker, T., Thorsson, V., Siegel, A.F. and Hood, L.E. 2000. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol*, **7**, 805-817.

Kluger, Y., Yu, H., Qian, J. and Gerstein., M. 2003. Relationship between gene co-expression and probe localization on microarray slides. *BMC Genomics*, **4**, 49.

Ledoit, O. and Wolf, M. 2004. A well conditioned estimator for largedimensional co-variance matrices. *J. Multiv. Anal.*, **88**, 365-411.

Schafer, J. and Strimmer, K. 2005. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754-764.

Schafer, J. and Strimmer, K. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, **4**, 32.

Wieczorke,R., Krampe,S., Weierstall,T., Freidel,K., Hollenberg,C.P. and Boles,E. 1999. Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae*. *FEBS Lett*, **464**, 123-128.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L. 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **10**, 977-987.

Zhou,X.J., Kao,M. et al. 2002. Transitive functional annotation by shortest path analysis of gene expression data. *Proc Natl Acad Sci USA*, **99**, 12783-12788.

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**, 1029-1045.