Gene Regulatory Network Reconstruction Using an Objective Bayes Approach

Maryam Shahdoust[a], Hossein Mahjub[a], Mehdi Sadeghi[b], Hamid Pezeshk[c,d]*

[a] Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
[b] National Institute of Genetic Engineering and Biotechnology, Tehran, Iran
[c] School of Mathematics, Statistics and Computer Sciences, University of Tehran, Iran
[d] School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Abstract. Although the common estimation approaches to estimate covariance matrix, such as maximum likelihood estimation, provide suitable estimation, but usually they are based on sample information. External information of variables may provide further information about covariance relationships between them. This paper aims to reconstruct the gene regulatory network of normal human bronchial epithelial cells using an objective Bayesian method to estimate the covariance matrix by exploiting the further information about genes (variables). Gene regulatory networks are reconstructed via graphical lasso approach (Glasso) and context likelihood relatedness (CLR) algorithm based on proposed Bayesian estimation of covariance matrix. The dataset is split into two subsets each including 612 genes expressions under 26 stimuli. To estimate the covariance matrix of each set, the information of the other set is applied as external knowledge. The inverse Wishart distribution is applied as informative conjugate prior distribution for covariance matrix. External knowledge is applied in hyperparameter estimation of prior distribution as the Euclidian distances between genes. The results show that applying the further information about genes relationships decreases the correlations between them and resulted networks are sparser than those obtained by approaches which use just the sample information.

Keywords: Gene regulatory network; Covariance matrix; Bayesian inference; Inverse Wishart distribution; Glasso; CLR.

1. Introduction

Multivariate analysis methods are most widely used statistical approaches in systems biology. They are applied in unsupervised methods such as clustering, feature selection and dimension reduction to find regularities and hidden structures in the data [1].

* Corresponding author.
E-mail address: pezeshk@khayam.ut.ac.ir

A covariance matrix plays a main role in almost every multivariate analysis method. In fact, the first step in multivariate analysis is computing the mean and covariance matrix. Therefore, finding an efficient estimator for covariance matrix is an interesting issue in altering of multivariate analysis methods to get better application in different subjects such as high dimensional studies [2-4].

Covariance matrix is a positive semi-definite matrix which its diagonal elements indicate variances and each non-diagonal element in ij position is the covariance between i^{th} and j^{th} variables [5]. Estimation of covariance matrix usually is based on sample data. Although the sample covariance matrix estimation by regular methods such as maximum likelihood estimation, are able to provide suitable estimation, they are just based on sample information. Some studies on multivariate analysis methods showed that using further information about variables could extend the multivariate methods in a way of apply more information [6-8]. For instances, Krupka et al. [6] proposed an approach to use prior knowledge on variables (features) into supervised learning. They showed that the prior information can be incorporated to learning process. They applied an approach to support vector machine classification method. The approach assigns weights to each feature in linear discriminant function and prior information are applied to define a prior weight. These priors are based on a Gaussian process and the weights are assumed to be a smooth function (covariance function) of the prior information [6]. Smart PCA is another extension of the frequently used multivariate analysis, principle component analysis (PCA), which is proposed by Zhang [8]. Smart PCA uses external knowledge through probabilistic interpretation of PCA to extend the regular approach. The inverse Wishart distribution is used as the informative prior for the population covariance and external knowledge is applied by reparameterization of hyperparameters.

Exploiting of further information also was considered in some systems biology studies in order to inferring gene networks. Most of these studies tried to use further information of gene regulatory relationship as prior

information in Bayesian framework of network inferring [9-11].

For instance, prior Lasso (Plasso) proposed by Wang et al. [9] incorporates prior knowledge into gene network studies. The prior pathways and gene network information are considered as prior information. Gene interactions in prior should be confirmed at higher frequency compared with undocumented interactions in real networks. Plasso partitions edges in a graphical model into two subsets based on known gene interactions; a known gene interactions group and an unknown gene interactions group. Then the approach, assigns the former group with a smaller regularization parameter in Lasso regression compared with the other group. Implementation of Plasso in simulated and real data showed better performance compared to the regular Lasso. In other study, Isci et al. [10] proposed an automatic method to incorporate multiple sources of prior knowledge into Bayesian network (BN) learning. The proposed method does not use likelihood approximation to find the optimal network. The approach assumes that infrastructure of BN yields gene interaction information for pairs of genes, which can be applied as informative priors to calculate the probability of a candidate graph. Therefore, the approach uses BN infrastructure itself to apply external biological knowledge when learning network. An introduced approach by Kpogbezan et al. [7] used Bayesian simultaneous equation model for incorporating prior information in undirected network reconstruction. The results of the study showed that the presented approach is able to improve the reconstruction of network using accurate prior data.

Using external knowledge, Safari et al. [11] presented a theoretical model to Bayesian inferring of genes covariance matrix inspired of smart PCA. The simulation study showed that the incorporation of external knowledge about genes relationship could improve the estimation of covariance matrix providing the availability of confident external information.

In this paper we aim to reconstruct the gene regulatory network using one objective Bayes approach. To reconstruct gene network, two undirected gene association network algorithms; graphical Lasso method (Glasso) and context likelihood relatedness (CLR) are applied based on Bayesian estimation of covariance matrix with external knowledge about genes relationships. To estimate the covariance matrix, the method of genes covariance estimation presented by Safari et al. [11] is implemented. The proposed approach is applied on normal human bronchial epithelial cells genes expression in order to reconstruct the corresponding gene regulatory network. A Bayesian approach with inverse Wishart prior distribution for covariance matrix is applied to estimate the covariance matrix. In order to use external information, the dataset is partitioned into two sub-datasets and the information of each group is applied as external knowledge for the other group. The gene network in each group is reconstructed by Glasso and CLR.

2. Methods

2.1. Bayesian inference of covariance matrix

To apply a Bayesian approach to estimate the population covariance matrix (Σ), it is assumed that n observations $Y=\{y_1, \dots, y_n\}$ are distributed as $N(0, \Sigma)$ where Σ follows an inverse Wishart distribution $IW(\upsilon, G)$ [12]:

$$P(\sum|G,\upsilon) = \frac{c_0}{\frac{\upsilon}{\Sigma^2}}|G|^{\frac{\upsilon-p-1}{2}}\exp\left\{-\frac{tr(\Sigma^{-1}G)}{2}\right\}, \quad (1)$$

where scale matrix G is a p×p positive definite matrix. υ is the degrees of freedom parameter which should exceed of 2p and c_0 is a normalizing constant.

Using the mode of inverse Wishart distribution (G/υ), the parameter G could be reparameterized as $G = \upsilon\Omega$ where Ω is prespecified structural information about Σ in that mode of Σ is ($\upsilon\Omega/\upsilon = \Omega$) [13].

Since the Inverse Wishart distribution is the conjugate prior, the posterior distribution of Σ also follows an inverse Wishart distribution where its parameters are:

$$(\sum | G, \upsilon, Y) \sim IW(\upsilon^*, \upsilon^*\Omega^*)$$
$$\upsilon^* = \upsilon + n \ , \ \Omega^* = \left(\frac{n}{n+\upsilon}\right)S + \left(\frac{\upsilon}{n+\upsilon}\right)\Omega . \quad (2)$$

The mean of posterior distribution could be a posterior estimation of Σ:

$$E(\sum|\Omega^*,\upsilon^*Y) = \frac{\upsilon^*\Omega^*}{\upsilon^*-p-1}. \quad (3)$$

2.2. Estimation of hyperparameters of prior Wishart distribution by external knowledge

2.2.1. Prior degrees of freedom

The Ω^* is a weighted combination of the sample covariance S and the hyperparameter Ω. The prior degrees of freedom (υ) controls the strength of this combination and it can be set empirically as any non-negative real number which exceeds 2p [12].

2.2.2. Hyperparameter Ω

Inspired by decomposition of covariance matrix into diagonal matrix of standard deviations and correlation matrix, Ω could be decomposed as VCV where V is a diagonal matrix of standard deviations which could easily be estimated by sample standard deviations. The p×p matrix C is a positive definite matrix represents the prespecified correlation structure. In this paper, it is assumed that the matrix C could be defined from the external information about genes relationships. Regarding to covariance function forms [14], C could be defined as a function of genes distances (4). Thus, the external information about gene relationships could be used as gene distances:

$$C_{ij} = exp(-\frac{d(g_i^T, g_j^T)}{\alpha}), \quad (4)$$

where g^r_i, g^r_j are i^{th} and j^{th} genes and $d(g^r_i, g^r_j)$ is the Euclidean distances between them. Parameter α should be estimated according to the scale of function d. One reliable choice for alpha is the one that satisfies the following criteria; (5), where ρ_{med} is the median element in the sample correlation matrix and d_{med} is the median element of gene distance matrix [8].

$$\rho_{med} = exp\left(-\frac{d_{med}}{\alpha}\right). \quad (5)$$

2.2.3. Reconstructing the gene regulatory network

To construct the gene regulatory network two frequent methods were considered; graphical lasso approach (Glasso) [4] and context-likelihood relatedness (CLR) with Pearson correlation as a similarity measure between genes [15].

Glasso is a popular graphical Gaussian model to infer a gene regulatory network [16, 17]. In graphical Gaussian models the estimated precision matrix can be seen as direct correlation between gene `pairs in the gene association network. Assuming a multivariate Gaussian distribution for gene expression data, Glasso estimates the sparse precision matrix (Θ) by penalize the maximum likelihood estimate of Θ using an L1 penalty function:

$$log\ det\ \Theta - tr(S\Theta) - \rho\|\Theta\|_1, \quad (6)$$

where S is emperical covariance matrix, $\|\Theta\|_1$ denotes the sum of absolute value of Θ and ρ is regulazation prameter of the algorithm.

In this work, the estimated covariance matrix by Bayesian method is used as Glasso entry instead of using sample covariance matrix.

The CLR algorithm is an extension of the relevance network algorithms [15]. Originally, CLR calculates the statistical likelihood of each mutual information values, but it could use other kinds of similarity measures such as Pearson correlation [18]. The correlation matrix is calculated from the proposed Bayesian estimation. To estimate the correlation matrix the bellow algebra equation (7) is applied [5]. In this equation R is a correlation matrix, D is a diagonal matrix of diagonal elements of covariance matrix and S is considered as the covariance matrix.

$$R = D^{-\frac{1}{2}} S\, D^{-\frac{1}{2}} \quad (7)$$

3. Results

The dataset includes gene expressions of normal human lung epithelial cells exposed to 52 different stimuli deposited in ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) with the accession number E-MTAB-2091. All genes expressions are normalized and values are corresponded to log2 expression.

The dataset is split into two sub-datasets by random choose of the stimuli. 26 stimuli are in each sub-dataset. Each sub-dataset included 612 genes which their standard deviations are greater than one.

To estimate the covariance matrix by Bayesian approach in each group, the Euclidean distances matrix of the other set is applied as external knowledge. The distance matrices are normalized due to the minimum and maximum distances of each matrix.

The estimation of prior degrees of freedom is totally empirical. The first prior degrees of freedom (υ) is chosen in order to satisfy the condition of (υ>2p) as 53 and then it is increased. Increasing the υ showed approximately similar results. Therefore, it is set to 53 in rest of the estimation process. After estimation of covariance matrix then the correlation matrix is constructed by the use of expression (7) in each sub-dataset. Table 1 shows some percentiles for Bayesian correlation matrices and their corresponding sample matrices in two sub-datasets.

Table 1. percentiles for sample correlation matrix and estimated correlation matrix in each sub-dataset.

		Percentile			
		25	50	75	90
Sub-data1	Sample correlation	0.24	0.5	0.73	0.84
	Bayesian estimation	0.01	0.03	0.08	0.17
Sub-data2	Sample correlation	0.23	0.47	0.71	0.84
	Bayesian estimation	0.01	0.04	0.11	0.21

Table 2. the number edges, common and different edges for CLR nets based on sample and estimated correlation matrices in each sub-data.

		Zero cells	Edges	Common information	Different information
Sub-data1	Sample correlation	96789	90177	137860	49106
	Bayesian estimation	119651	67315		
Sub-data2	Sample correlation	94095	92871	130019	56947
	Bayesian estimation	116010	70956		

The number of potential edges in a complete graph of 612 genes is 186966. The median of each weighted adjacency CLR network was chosen as threshold in each network.

Gene regulatory networks are constructed by Glasso method and CLR algorithm based on sample genes correlations and Bayesian genes correlations.

The Glasso approach is applied for several parameters for Lasso (ρ) from 0.1 to 0.9. Fig. 1. and Fig. 2. illustrate the constructed networks based on sample covariance matrices and Bayesian correlations for different ρ in sub-data1 and sub-data2, respectively. Fig. 1. shows the networks only for ρ in the range of 0.1 to 0.7 because none edges are estimated based on Bayesian correlation matrix in sub-data1.

Each CLR networks, based on sample correlation matrix and Bayesian correlation, is constructed due to median of their corresponding weighted adjacency matrix (WAJ). The WAJ cells which are less than median replaced by zero and non-zero cell indicated an edge in the network. According to CLR algorithm some cells in weighted adjacency matrix

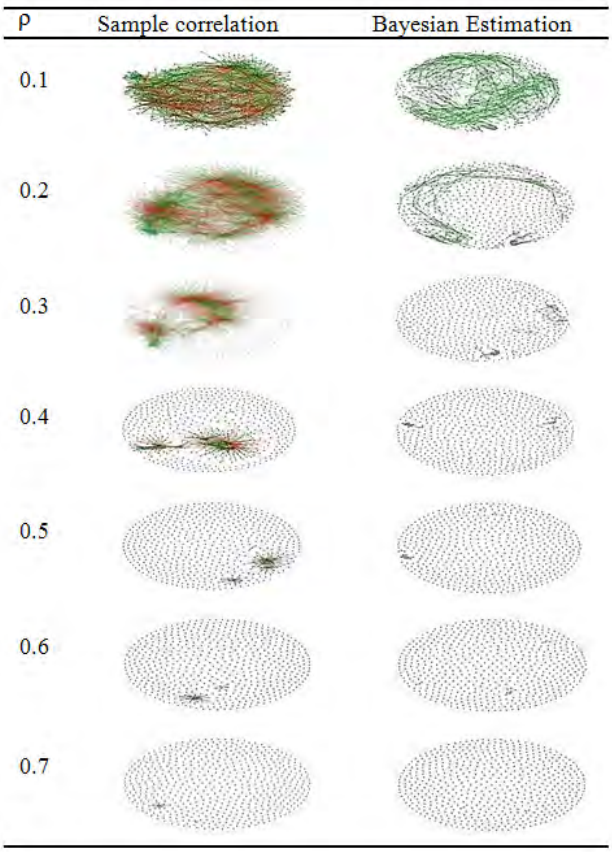

Fig. 1. Estimated networks by Glasso for different regularization parameter (ρ) based on sample correlation matrix and estimated correlation matrix in sub-data1.

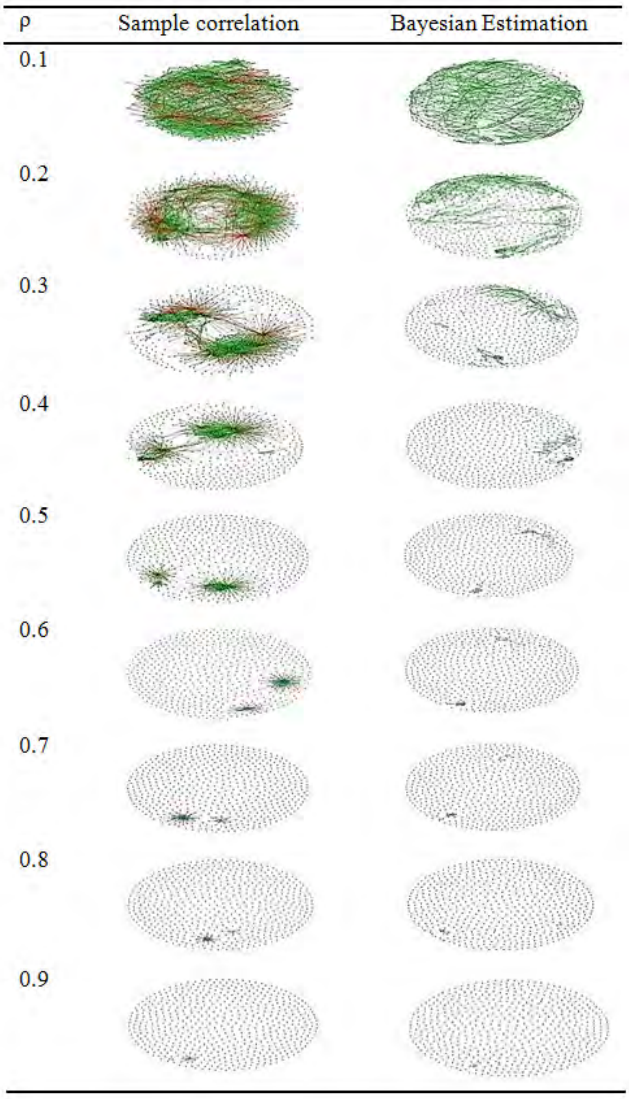

Fig. 2. Estimated networks by Glasso for different regularization parameter (ρ) based on sample correlation matrix and estimated correlation matrix in sub-data2.

could be zero, thus the number of zero and nonzero cells are not equal regarding the median. The number of edges in CLR networks could be found in Table 2. Also, the number of common and different edges in CLR networks based on two correlation matrices in each sub-group is shown in Table 1. In each sub-data, the column of "common information" is the sum of common zero and non-zero cells in adjacency CLR matrix in both networks. Also, the column of "different information" denotes the number of different cells in two adjacency CLR matrices.

4. Discussion

Reconstructed gene regulatory network was done by using external information about genes relationships in a Bayesian framework. In order to use the external information about genes, the gene regulatory network was reconstructed based on the covariance matrix estimated by the Bayesian method presented in Safari et al. study. The proposed Bayesian estimation for genes covariance matrix prepares the possibility of applying the prior information about genes relatedness as a function of their corresponding distances. Therefore, in this study, the reconstruction of gene networks due to the external hints about genes relationship was applied by utilizing Safari et al. presented method to estimate the genes covariance matrix. The results of our study showed that applying common network construction methods (such as Glasso and CLR algorithm) based on the Bayesian estimation of covariance matrix could profit the external information about genes relationships in a simple way.

Comparing the Bayesian estimated correlation matrix with sample correlations showed that the Bayesian correlations were smaller than the sample ones. Table1 shows that the percentiles of Bayesian estimations are less than corresponding sample percentiles in each sub-dataset. Smaller correlations in Bayesian correlation matrix suggested the sparser networks and this is consistent with networks obtained from Glasso and also CLR algorithms based on Bayesian estimated covariance matrix. The reconstructed networks using

external knowledge about genes relationships are sparser than reconstructed networks just based on sample information. As long as theses sparse networks could conserve the valuable information about genes regulatory interactions, applied Bayesian approach would be useful to gene regulatory networks reconstruction. Further studies are still in progress.

Acknowledgment

Maryam Shahdoust and Hossein Mahjub would like to thank Biostatistics department of Hamadan University of Medical sciences for their supports.

References

1. Prokop, Ales, and Béla Csukás, eds. Systems Biology: Integrative Biology and Simulation Tools. Vol. 1. Springer Science & Business Media, 2013.
2. Kuismin, Markku, and Mikko J. Sillanpää. "Use of Wishart prior and simple extensions for sparse precision matrix estimation." PloS one 11.2 (2016): e0148171.
3. Pourahmadi, Mohsen. High-dimensional covariance estimation: with high-dimensional data. John Wiley & Sons, 2013.
4. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." Biostatistics 9.3 (2008): 432-441.
5. Rencher, Alvin C. Methods of multivariate analysis. Vol. 492. John Wiley & Sons, 2003.
6. Krupka, Eyal, and Naftali Tishby. "Incorporating Prior Knowledge on Features into Learning." AISTATS. Vol. 2. 2007.
7. Kpogbezan, Gino B., et al. "An empirical Bayes approach to network recovery using external knowledge." Biometrical Journal (2017).
8. Zhang Y. Smart PCA. Proceedings of the 21st international jont conference on Artifical intelligence; Pasadena, California, USA. 1661662: Morgan Kaufmann Publishers Inc.; 2009. p. 1351-6.
9. Wang, Zixing, et al. "Incorporating prior knowledge into gene network study." Bioinformatics 29.20 (2013): 2633-2640.
10. Isci, Senol, et al. "Bayesian network prior: network analysis of biological data using external knowledge." Bioinformatics 30.6 (2014): 860-867.
11. Safari, Abdollah, et al. "A Multivariate Bayesian Model for Gene Networks." Journal of Statistical Sciences 6.2 (2013): 187-200.
12. Gelman, Andrew, et al. "Bayesian data analysis, (chapman & hall/CRC texts in statistical science)." (2003).
13. Chen, Chan-Fu. "Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis." Journal of the Royal Statistical Society. Series B (Methodological) (1979): 235-248.
14. Yaglom, Akira Moiseevich. Correlation theory of stationary and related random functions: Supplementary notes and references. Springer Science & Business Media, 2012.
15. Faith, Jeremiah J., et al. "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles." PLoS biol 5.1 (2007): e8.
16. Li, Yupeng, and Scott A. Jackson. "Gene network reconstruction by integration of prior biological knowledge." G3: Genes| Genomes| Genetics 5.6 (2015): 1075-1079.
17. Wang, YX Rachel, and Haiyan Huang. "Review on statistical methods for gene network reconstruction using expression data." Journal of theoretical biology 362 (2014): 53-61.
18. Joshi, Anagha, Yvonne Beck, and Tom Michoel. "Multi-species network inference improves gene regulatory network reconstruction for early embryonic development in Drosophila." Journal of Computational Biology 22.4 (2015): 253-265.