Reconstruction of Metabolic Association Networks Using High-throughput Mass Spectrometry Data*

Imhoi Koo[1,2], Xiang Zhang[2], and Seongho Kim[1]

[1] Department of Bioinformatics and Biostatistics,
University of Louisville, Louisville KY, 40292, USA
[2] Department of Chemistry, University of Louisville, Louisville KY, 40292, USA
{imhoi.koo,xiang.zhang,s0kim023}@louisville.edu

Abstract. Graphical Gaussian model (GGM) has been widely used in genomics and proteomics to infer biological association networks, but the relative performances of various GGM-based methods are still unclear in metabolomics. The association between two nodes of GGM is calculated by partial correlation as a measure of conditional independence. To estimate the partial correlations with small sample size and large variables, two approaches have been introduced, which are arithmetic mean-based and geometric mean-based methods. In this study, we investigated the effects of these two approaches on constructing association metabolite networks and then compared their performances using partial least squares regression and principal component regression along with shrinkage covariance estimate as a reference. These approaches then are applied to simulated data and real metabolomics data.

Keywords: metabolomics, graphical Gaussian model, partial correlation, partial least squares regression, principal component regression, false discovery rate.

1 Introduction

Metabolomics is a rapidly emerging field to systemically analyze small-molecule metabolites in a biological organism [1]. It is equally important in systems biology as other "-omics" such as genomics, transcriptomics, and proteomics. One of the important approaches to integrating the individual "-omics" data for system level analysis is the reconstruction of cellular networks, which is collection and visualization of all physiologically relevant cellular processes.

In metabolomics, a relatively small number of studies have been reported for metabolic network construction. For instance, Arkin et al. [2] predicted interactions within reaction networks over time for the glycolytic pathway, where Pearson's correlation coefficient was used to construct the interaction networks. A major drawback of Pearson's correlation-based networks is unable to distinguish between

* This work was also partially supported by National Institute of Health (NIH) grant 1RO1GM087735 through the National Institute of General Medical Sciences (NIGMS).

D.-S. Huang et al. (Eds.): ICIC 2012, LNCS 7389, pp. 160–167, 2012.

direct and indirect associations. On the other hand, graphical Gaussian models (GGMs) reveal direct associations with conditional independences/dependences among variables, using partial correlation coefficients that are calculated by the correlation of two variables after removing affection of other variables [3]. GGMs have been employed in metabolomics for several studies [4, 5]. Note that the size of samples (experiments) was larger than the number of variables (metabolites) for these studies so that network construction was straightforward.

If the number of samples is much smaller than number of variables, it is difficult to directly estimate partial correlation due to singularity. To overcome this difficulty, several methods have been developed by either reducing the number of given variables or a regularized estimation [6, 7]. Another alternative is to use dimension-reduced regression such as partial least squares regression (PLSR) and principal component regression (PCR) approaches. When calculating the partial correlations using regression coefficients, arithmetic and geometric means of regression coefficients were employed in Kramer et al. [8] and Pihur et al. [9], respectively. The partial correlation coefficients estimated by these two methods are not the same to each other and, it is important to investigate the effects of the different calculation methods on network reconstruction. Therefore, we evaluated the performance of PLSR and PCR using shrinkage covariance estimate as a reference in terms of network construction.

2 Methods and Materials

The graphical Gaussian model (GGM) is a statistical multivariate analysis to infer the direct relationship among variables using nodes and edges [3], where the nodes correspond to the variables under consideration, and the edges represent the conditional independence between two variables as measured by partial correlation coefficient.

Suppose a data matrix X consists of n observed samples and p metabolites with a mean of zero. Then the partial correlation coefficient matrix $P = (\pi_{ij})$ is calculated by the inverse of the covariance matrix $\Sigma = \frac{1}{n}X^\top X$ as follows:

$$\pi_{ij} = -\frac{w_{ij}}{\sqrt{w_{ii}w_{jj}}}, \tag{1}$$

where $\Sigma^{-1} = (w_{ij})$.

The covariance matrix Σ becomes singular when the sample size n is smaller than the number p of variables. To deal with singularity, several methods have been introduced [8]. In this study, the following three methods are considered.

2.1 Shrinkage Covariance Estimation

Schafer and Strimmer [10] introduced shrinkage covariance estimator (SCE) for the partial correlation estimation when the covariance matrix $\Sigma = \frac{1}{n-1}X^\top X$ is singular. Under singularity of covariance matrix, the SCE is to trade off the unbiased sample covariance Σ and low dimensional shrinkage target matrix T:

$$\hat{\Sigma} = \lambda T + (1 - \lambda)\Sigma, \tag{2}$$

where $\lambda \in (0, 1]$ is shrinkage intensity. The optimal value of the tuning parameter λ is analytically determined and estimated from the data.

2.2 Regression with Dimension Reduction

Partial Least Squares Regression and Principal Component Regression. The common property of both partial least squares regression (PLSR) and principal component regression (PCR) is to use dimension reduction method to avoid the singularity for the "small n, large p" paradigm. PLSR finds orthogonal vector $\mathbf{w}$ to maximize the covariance between $\mathbf{t} = X\mathbf{w}$ and dependent (response) variable $\mathbf{y}$, while PCR searches for orthogonal vector $\mathbf{w}$ to maximize the variance of $\mathbf{t} = X\mathbf{w}$.

Consider linear regression of dependent variable $\mathbf{y}$ on data matrix X as follows:

$$\mathbf{y} = X\boldsymbol{\beta} + \epsilon, \tag{3}$$

where β is a vector of regression coefficients and ϵ is error. The estimation of coefficient β using PLSR consists of two steps [11]. The first step is to extract a latent variable set $T = (\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_k)$ of orthogonal components $(k < p)$, which maximizes a covariance with dependent variable $\mathbf{y}$. The second step is to estimate the coefficient of regression of $\mathbf{y}$ on the new latent variable set T and then to transform it into space spanned by data X. The first PLSR component $\mathbf{t}_1 = X\mathbf{w}_1$ is obtained by maximizing the covariance as follows:

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \mathrm{Cov}^2(X\mathbf{w}, \mathbf{y}). \tag{4}$$

The next components $\mathbf{t}_i$, $i = 2, \cdots, k$, are satisfied with maximizing the squared covariance to $\mathbf{y}$ and are mutually orthogonal to each other. Consider the orthogonal part X_k of X on all components $\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_{k-1}$:

$$X_k = X - \mathcal{P}^{\perp}_{\mathbf{t}_1,\cdots,\mathbf{t}_{k-1}}(X), \tag{5}$$

where $\mathcal{P}^{\perp}_{\mathbf{t}_1,\cdots,\mathbf{t}_{k-1}}$ is the projection operator related to $\mathbf{t}_1, \cdots, \mathbf{t}_{k-1}$. The kth latent variable, $\mathbf{t}_k = X_k\mathbf{w}_k$, is then obtained by solving the optimization problem:

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} \mathrm{Cov}^2(X_k\mathbf{w}, \mathbf{y}). \tag{6}$$

Using the following equations, the vector of regression coefficients $\hat{\beta}^{PLSR}(k)$ is determined to predict the output of a model including k components:

$$\hat{y}_k^{PLSR} = X\hat{\beta}^{PLSR}(k) \text{ and } \hat{\beta}^{PLSR}(k) = (\mathbf{w}_1, \cdots, \mathbf{w}_k)T^{\top}\mathbf{y}. \tag{7}$$

For PCR, the equations (4) and (6) are replaced with the following equations, respectively:

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \mathrm{Var}(X\mathbf{w}), \text{ and } \mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} \mathrm{Var}(X_{k-1}\mathbf{w}). \tag{8}$$

Then, the predicted output $\hat{y}^{PCR}$ and regression coefficients $\hat{\beta}^{PCR}$ can be calculated by

$$\hat{y}^{PCR} = X\hat{\beta}^{PCR}(k) \text{ and } \hat{\beta}^{PCR}(k) = (\mathbf{w}_1, \cdots, \mathbf{w}_k)T^{\top}\mathbf{y}, \tag{9}$$

Once the regression coefficients in equations (7) and (9) are computed, the partial correlation coefficients are estimated by using either geometric or arithmetic mean of regression coefficients.

Method 1: Geometric mean approach. In this approach, the partial correlation coefficient π_{ij} of X in the equation (1) is estimated by

$$\pi_{ij} = \mathrm{sign}(\hat{\beta}_{ij})\sqrt{\hat{\beta}_{ij}\hat{\beta}_{ji}} \tag{10}$$

Method 2: Arithmetic mean approach. Arithmetic mean approach of the association/interaction scores was introduced by Pihur et al. [9]. The partial correlation is calculated by

$$\hat{\pi}_{ij} = \frac{\sum_{l=1}^{k} \hat{\beta}_{il} c_{jl}^{(i)} + \sum_{l=1}^{k} \hat{\beta}_{jl} c_{il}^{(j)}}{2}. \tag{11}$$

In this equation (11), the coefficients are obtained from

$$x_i = \sum_{j=1}^{k} \beta_{ij} t_j^{(i)} + \epsilon, \text{ and } t_j^{(i)} = \sum_{l \neq i}^{p} c_{jl}^{(i)} X_l^{(i)}, \tag{12}$$

where $t_j^{(i)}$ is a jth latent variable of PLSR and PCR , and k is the number of latent variables which is pre-determined by user.

2.3 False Discovery Rate

After estimating partial correlation coefficients, statistical hypothesis test is performed to select the significant edges indicating strong association between two variables. To do this, false discovery rate (FDR) is applied to control the expected proportion of incorrectly rejected null hypotheses by using the q-value method in R software package *fdrtool* [12].

2.4 Data

Simulation Data. The simulated data were generated using two conditions, sample size and network complexity. The number of variables p was always set to 100. We used three different densities, 5%, 15%, and 25%, to describe the complexity of the network. Given each density, we considered five different sample sizes, 25, 50, 100, 150, and 200, to generate simulated data. For each case, we generated 100 data sets and then compared the performance of each method with their averages. The R software package *GeneNet* was used to generate the simulated data [13].

Experimental Data. We also investigated the performance of each method using experimental data of metabolites extracted from mouse liver. The experimental data consist of all compounds detected from mouse samples on a linear trap quadruple-Fourier transform ion cyclotron resonance mass spectrometer (LTQ-FTICR MS) via direct infusion electrospray ionization (DI-ESI)-mass spectrometry. For the association network study, we used 99 compound peaks that were detected in all 40 samples by *MetSign* software [14].

2.5 Performance Evaluation

We evaluated the five estimation methods, shrinkage covariance estimation (SCE), geometric/arithmetic mean-based partial least squares regression (PLSR.G and PLSR.A, respectively), and geometric/arithmetic mean-based principle component regression (PCR.G and PCR.A, respectively), in this study. In order to evaluate their performance, the following three criteria were considered:

1. The true positive rate is the proportion of true positives which are correctly predicted; $TPR = \frac{TP}{TP+FN}$,
2. The positive predictive value is the proportion of subjects with positive output results which are correctly predicted; $PPV = \frac{TP}{TP+FP}$,
3. F1 score is a measure of accuracy, which is the harmonic average of TPR and PPV; $F1 = 2 \cdot \frac{TPR \cdot PPV}{TPR+PPV}$.

3 Results and Discussion

Fig. 1 (a)-(c) show the F1 scores of each method in terms of network construction based on simulated data. It can be seen that the performance of geometric mean-based and arithmetic mean-based approaches relies on the estimation methods (PLSR and PCR). As for PCR, geometric mean-based approach performs better than arithmetic mean-based approach when the network is complex regardless of the sample size. However, arithmetic mean-based approach has the larger F1 score than geometric mean-based approach with PLSR (PLSR.G) when the sample size is large. In particular, when the sample size is less than 50, PLSR.G performs the best with density of 5%, while PCR.G is the best method if the density is 15% or 25% based on F1 score, as shown in the figure.

Interestingly, in case of PPV as shown in Fig. 1 (d)-(f), arithmetic mean-based approach with PCR outperforms geometric mean-based approach regardless of sample size and network complexity. On the other hand, as for TPR in Fig. 1 (g)-(i), geometric mean-based approach performs better than arithmetic mean-based approach when PCR is applied, while arithmetic mean-based approach with PLSR (PLSR.A) is better when the density is 15% or 20%.

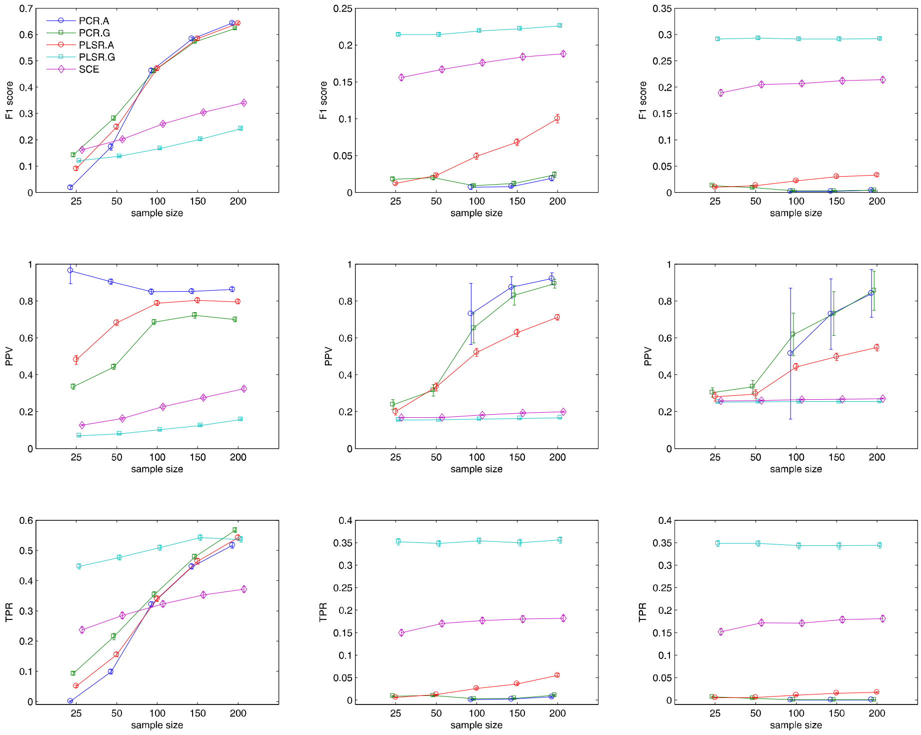

Fig. 1. Performance plots. (a), (b), and (c) show the F1 scores. (d), (e), and (f) show the positive predictive value. (g), (h), and (i) show the true positive rate. (a), (d), and (g) correspond to density 5%. (b), (e), and (h) correspond to density 15%. (c), (f), and (i) correspond to density 25%. SCE, PLSR.G, PLSR.A, PCR.G and PCR.A stand for shrinkage covariance estimate, geometric mean-based partial least squared regression, arithmetic mean-based partial least squared regression, geometric mean-based principle component regression and arithmetic mean-based principle component regression, respectively. Error bar stands for average value and 95% confidence interval of F1 score, PPV and TPR over 100 runs.

Table 1 shows the numbers of empty network estimated by SCE, PLSR.G and PLSR.A out of 100 independent runs. Note that PCR methods generated no empty network. When the true network becomes more complex, those methods generated more estimated empty network. Furthermore, the number of estimated empty networks is decreased as the sample size goes to 200. However, the trend of PLSR.G is different with other methods. For example, the empty network for PLSR.G with density of 25% is increased as the sample size is increased.

The results of network construction using real experimental data are shown in Table 2. The number of significant edges and the number of intersection of edges of pair of two methods are reported. The geometric mean-based approaches, PLSR.G and PCR.G, generate larger significant edges than arithmetic mean-based approaches. Namely, PLSR.A and PCR.A selected at least 5.4 times more edges. Most edges (90% and 89%) of PLSR.A and PCR.A are overlapped with these of PLSR.G and PCR.G, respectively.

The reason for the difference of the F1 score between two mean-based approaches for PLSR and PCR in complex density is likely due to the different statistical property of latent variables from them. The difficulty of regression using PLSR and PCR under complex network can also be another reason for disagreement of performance pattern of them. Furthermore, since the output of arithmetic mean is larger than that of geometric mean for the same input, discrimination power of arithmetic mean-based approach to estimating significant edges combined with FDR method increases when n = 100, 150, 200 and the density is 5%. This makes the trend of PPV and TPR for the final approaches consistent in density of 5%. For the real experimental data, the condition seems similar to the case that density is 5% and sample size is 50 or 100 in terms of the number of significant edges.

Table 1. Number of empty networks for SCE and PLSR with geometric (.G) and arithmetic (.A) approaches out of 100 independent simulations

		5%			15%			25%	
n	SCE	PLSR.G	PLSR.A	SCE	PLSR.G	PLSR.A	SCE	PLSR.G	PLSR.A
25	92	0	0	100	6	8	100	5	7
50	0	0	0	97	8	1	100	12	2
100	0	0	0	85	39	0	93	59	1
150	0	0	0	41	26	0	86	69	0
200	0	0	0	24	11	0	82	72	0

Table 2. Number of significant edges for SCE, PLSR and PCR with geometric (.G) and arithmetic (.A) approaches for real experimental results and number of intersection of two methods

	SCE	PLSR.G	PLSR.A	PCR.G	PCR.A
SCE	259	150	131	137	85
PLSR.G		1228	172	608	136
PLSR.A			191	133	79
PCR.G				1292	188
PCR.A					211

4 Conclusion

We evaluated the performance of two estimation methods, arithmetic mean-based and geometric mean-based approaches, using regression coefficients to construct association networks. We observed that the performances of geometric mean-based and arithmetic mean-based approaches are dependent on the dimension-reduced regression methods (PLSR and PCR) and simulation settings such as sample size and density. Arithmetic estimation outperforms geometric mean when it is incorporated with PLSR and the sample size is larger, while the geometric mean-based approach performs better when the true network is complex and it is used with PCR in terms of F1 score.

References

1. Watkins, S.M., German, J.B., Hammock, B.D.: Metabolomics: Building on a Century of Biochemistry to Guide Human Health. Metabolomics 1(1), 3–9 (2005)
2. Arkin, A., Shen, P.D., Ross, J.: A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. Science 277(5330), 1275–1279 (1997)

3. Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley series in probability and mathematical statistics. Wiley, Chichester (1990)
4. Theis, F.J., Krumsiek, J., Suhre, K., Illig, T., Adamski, J.: Gaussian Graphical Modeling Reconstructs Pathway Reactions from High-throughput Metabolomics data. BMC Systems Biology 5 (2011)
5. Chan, E., Rowe, H., Hansen, B., Kliebenstein, D.: The Complex Genetic Architecture of the Metabolome. PLoS Genetics 6(11), e1001198 (2010)
6. Dobra, A., Hans, C., Jones, B., Nevins, J.R., Yao, G., West, M.: Sparse Graphical Models for Exploring Gene Expression Data. Journal of Multivariate Analysis 90(1), 196–212 (2004)
7. de la Fuente, A., Bing, N., Hoeschele, I., Mendes, P.: Discovery of Meaningful Associations in Genomic Data using Partial Correlation Coefficients. Bioinformatics 20(18), 3565–3574 (2004)
8. Kramer, N., Schafer, J., Boulesteix, A.L.: Regularized Estimation of Large-scale Gene association Networks Using Graphical Gaussian Models. BMC Bioinformatics 10 (2009)
9. Pihur, V., Datta, S., Datta, S.: Reconstruction of Genetic Association Networks from Microarray Data: a Partial Least Squares Approach. Bioinformatics 24(4), 561–568 (2008)
10. Schafer, J., Strimmer, K.: A Shrinkage Approach to Large-scale Covariance Matrix Estimation and Implications for Functional Genomics. Statistical Applications in Genetics and Molecular Biology 4 (2005)
11. Houskuldsson, A.: Pls Regression Methods. Journal of Chemometrics 2(3), 211–228 (1988)
12. Strimmer, K.: fdrtool: a Versatile r Package for Estimating Local and Tail Area-based False Discovery Rates. Bioinformatics 24(12), 1461–1462 (2008)
13. Schafer, J., Strimmer, K.: An Empirical Bayes Approach to Inferring Large-scale Gene Association Aetworks. Bioinformatics 21(6), 754–764 (2005)
14. Wei, X., Sun, W., Shi, X., Koo, I., Wang, B., Zhang, J., Yin, X., Tang, Y., Bogdanov, B., Kim, S., Zhou, Z., McClain, C., Zhang, X.: Metsign: A Computational Platform for High-resolution Mass Spectrometry-based Metabolomics. Analytical Chemistry 83(20), 7668–7675 (2011)