# Network reconstruction with local partial correlation: comparative evaluation

Henrique Bolfarine*, Lina Thomas**, and Anatoly Yambartsev***

Instituto de Matemática e Estatística,
Universidade de São Paulo, São Paulo-SP, Brazil
bolfarin@ime.usp.br

**Abstract.** Over the past decade, various methods have been proposed for the reconstruction of networks modeled as Gaussian Graphical Models. In this work we analyzed three different approaches: the Graphical Lasso (GLasso), Graphical Ridge (GGMridge) and Local Partial Correlation (LPC). For the evaluation of the methods, we used high dimensional data generated from simulated random graphs (Erdös-Rényi, Barabási-Albert, Watts-Strogatz). The performance was assessed through the Receiver Operating Characteristic (ROC) curve. In addition the methods were used for reconstruction of co-expression network, for differentially expressed genes in human cervical cancer data. LPC method outperformed the GLasso in most of the simulation cases, even though GGMridge produced better ROC curves then both other methods. LPC obtained similar outcomes as GGMridge in real data studies.

**Keywords:** network reconstruction, Gaussian Graphical Model, partial correlation, gene co-expression, regularization.

## 1 Introduction

The reconstruction of network structures through estimated associations has become more popular over the past decade, mainly due to the availability of massive data sets. Several methods of network reconstruction have been recently developed. Most of these methods are based on Graphical Models (GM) [1] due to its ability to represent conditional dependencies over graph structures. If the variables in these models are assumed to have Gaussian distribution, then we have a Gaussian Graphical Model (GGM), which facilitates the use of partial correlation to identify conditional dependencies [2]. The main problem that emerges during network reconstruction on large data sets is that in some cases the number of variables is much bigger than the sample size.

To overcome this problem in [3], and [4] we proposed Local Partial Correlation (LPC) method. The method estimates partial correlation between two variables using the neighborhood of the relevance network [5], composed of the

* The corresponding author received support by CNPq grant 133935/2015-9.
** The author received support by FAPESP grant 2013/06223-1.
*** The author thanks CNPq grant 301050/2016-3.

highest Pearson correlations of both variables. Here our objective is to assess its efficiency. We compared the LPC to two other approaches based on regularization methodologies: GLasso and GGMridge. GLasso, defined in [6] is the most popular. It is based on the LASSO [7], a technique widely applied in regression analysis mainly, for variable and model selection. GGMridge [8] based on the work [9] estimates the partial correlation matrix using Ridge penalty and then performs statistical estimation using empirical distributions.

The assessment of the methods was performed using simulations and real-world data. In the simulation studies, we compared the performance of the methods by generating high dimensional data from the following random graph structures: Erdös-Rényi , a well-known uniform model [10], Barabási-Albert or scale free model [11], frequently used to model biological interactions, and Watts-Strogatz or small world [12], a popular model for social interactions. ROC curves were employed to compare the performance of each method, where there are original network structures to liken.

The real-world data came from [13], which is composed of gene expressions from tumorous cervical cancer cells. From this original data set, we selected 1268 genes known as differentially expressed genes (DEGs), identified by [14]. We applied the methods for this data set and compared the results of the reconstructions in terms of the number of nodes and edges that were identified in common by the methods.

Before presenting the different approaches in Section 2, and the results in Section 3 and Section 4, lets recall basic notions of GGM.

### 1.1 Gaussian Graphical Models

Let $X_{\mathcal{V}} = (X_1, \dots X_p) \in \mathbb{R}^p$ be $p$-dimensional random vector, with $|\mathcal{V}| = p$ and covariance matrix $\mathbf{\Sigma}_{\mathcal{V}}$, and let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a finite graph with set of vertices $\mathcal{V}$ and set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. If $X_{\mathcal{V}}$ has Gaussian distribution and $\mathbf{\Sigma}_{\mathcal{V}}$ is positive definite, then $\mathbf{\Omega} \equiv (\omega_{ij})_{i,j\in\mathcal{V}} = \mathbf{\Sigma}_{\mathcal{V}}^{-1}$ is the precision matrix. Under multivariate Gaussian distribution assumption [2], if $\omega_{ij} = 0$ then the partial correlation

$$\rho_{ij.\mathcal{Z}} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \quad \text{for} \quad i,j \in \mathcal{V}, \quad \text{and} \quad \mathcal{Z} = \mathcal{V} \setminus \{i,j\}, \tag{1}$$

between $X_i$ and $X_j$ is also zero, which implies conditional independence of these two variables given the rest $X_{\mathcal{Z}}$ [1]. In the GGM context this is equivalent to $\{i,j\} \notin \mathcal{E}$. Therefore, we can reconstruct a gene co-expression network by identifying if the elements of the precision and partial correlation matrices are different from zero. In summary, we have that

$$\omega_{ij} = 0 \Leftrightarrow \rho_{ij.\mathcal{Z}} = 0 \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_{\mathcal{Z}} \Leftrightarrow \{i,j\} \notin \mathcal{E}, \quad \text{for} \quad i,j \in \mathcal{V}. \tag{2}$$

## 2 Methods

### 2.1 Local Partial Correlation

Let $\mathbf{X}$ be a data matrix with $p$ variables and $n$ sample size, and assume $p \gg n$ (high dimensional problem). For any fixed $i \in \mathcal{V}$ and p-value threshold $\alpha \in (0,1)$

define $\mathcal{Z}_\alpha(i) \subset \mathcal{V} \setminus \{i\}$ the set of indices's of the variables that are significantly non-zero Pearson correlated with variable $X_i$. For any $i, j \in \mathcal{V}$ we call neighborhood of $\{i, j\}$ the set $\mathcal{Z}_\alpha(i) \cup \mathcal{Z}_\alpha(j)$. If the neighborhood has more variables than samples, we select $\lfloor n/2 \rfloor$ [3] variables from $\mathcal{Z}_\alpha(i) \cup \mathcal{Z}_\alpha(j)$ with highest Pearson correlations. Denote $\mathcal{Z}_\alpha(i,j) \equiv \mathcal{Z}_\alpha(i) \cup \mathcal{Z}_\alpha(j)$, and $\mathcal{J} \equiv \{i, j\} \cup \mathcal{Z}_\alpha(i, j)$.

We build an empirical covariance matrix $\hat{\mathbf{\Sigma}}_\mathcal{J}$. If inverted, $\hat{\mathbf{\Sigma}}_\mathcal{J}^{-1}$ provides the estimation $\hat{\rho}_{ij.\mathcal{Z}}$ [2], where for lighten the notation $\mathcal{Z} \equiv \mathcal{Z}_\alpha(i, j)$. Finally, we test $H_0 : \rho_{ij.\mathcal{Z}} = 0$ versus $H_a : \rho_{ij.\mathcal{Z}} \neq 0$, using the *z-transfomation*,

$$\psi(\hat{\rho}_{ij.\mathcal{Z}}) = \frac{1}{2} \log\{(1 + \hat{\rho}_{ij.\mathcal{Z}})/(1 - \hat{\rho}_{ij.\mathcal{Z}})\}.$$

For a significance level $\alpha_{LPC} \in (0, 1)$, the null hypothesis is rejected if

$$\sqrt{n - |\mathcal{Z}| - 3} \times |\psi(\hat{\rho}_{ij.\mathcal{Z}})| > \Phi^{-1}(1 - \alpha_{LPC}/2), \tag{3}$$

where $\Phi(x)$ is the cumulative Gaussian distribution $\mathcal{N}(0, 1)$, and $|\mathcal{Z}|$ is the size of the neighborhood. If (3) is true, then from (2), $\{i, j\} \in \mathcal{E}$.

### 2.2 GLasso and GMMridge

The constrained or penalized Maximum Likelihood Estimate (MLE) is often used for high dimensional problems when $p$ is lager than $n$. It is known that the MLE of the precision matrix $\mathbf{\Omega}$ is $\mathbf{S}^{-1}$, where $\mathbf{S} = \mathbf{X}\mathbf{X}'/n$ [1], with $\mathbf{X}$ the data matrix. GLasso, uses complex optimization tools [6] to maximize the following penalized log likelihood

$$\mathcal{L}(\mathbf{\Omega}) = \log\det(\mathbf{\Omega}) - tr(\mathbf{S}\mathbf{\Omega}) - \lambda_L ||\mathbf{\Omega}||_1, \quad \lambda_L > 0,$$

where $\|\cdot\|_1$ is the $\ell_1$ norm, $||\mathbf{A}||_1 = \sum_{ij} |a_{ij}|$, with $a_{ij} \in \mathbf{A}$, and $\lambda_L$ the tunning parameter. The procedure results in a sparse precision matrix rather than precise partial correlation estimation [15]. The GGMridge, on the other hand, uses a "ridge inverse" $(\mathbf{S} + \lambda_R \mathbf{I}_p)^{-1}$ in the analogy to ridge regression [16], which generates estimates for the partial correlation matrix

$$\hat{\mathbf{P}} = -\,\mathbf{scale}((\mathbf{S} + \lambda_R \mathbf{I}_p)^{-1}), \quad \lambda_R > 0,$$

where $\mathbf{scale}(\mathbf{A}) = \mathrm{diag}(\mathbf{A})^{-1/2}\mathbf{A}\,\mathrm{diag}(\mathbf{A})^{-1/2}$, and $\lambda_R$ is the restriction factor. The elements of $\hat{\mathbf{P}}$, which are in the same form as (1), are tested with $\alpha_R \in (0, 1)$, using empirical distributions [8], providing significant estimates for the partial correlation, and reconstructing the underlying graph structure.

## 3 Simulation Studies

One of the goals of this study was to evaluate if network structure affects the overall performance of the methods. We have chosen three graph models with different topologies: Erdös-Rényi [10], Barabási-Albert [11], and Watts-Strogatz [12]. In the Erdös-Rényi network we used a uniform distribution for the edges. The Barabási-Albert graph was generated through a preferential attachment algorithm [11], and for the Watts-Strogatz network we used the algorithm described in [12]. Parameters where chosen such that the generated graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, or its adjacency matrix $\mathbf{A} = (a_{ij})_{i,j \in \mathcal{V}}$ is sparse.

### 3.1 From adjacency to covariance matrix

In this section we explain how for given adjacency matrix we construct a covariance matrix following the procedure described in [17]. For each generated adjacency matrix $\mathbf{A}$ we attribute random values to the non-zero elements of $\mathbf{A}$, transforming it into a positive definite covariance matrix. First define the matrix

$$\mathbf{\Omega}_1 = \begin{cases} \omega_{ij} = \omega_{ji} = u_{ij} \cdot \delta_{ij}, & \text{if } a_{ij} = 1,\ i < j, \\ \omega_{ij} = \omega_{ji} = 0, & \text{otherwise,} \end{cases}$$

where $u_{ij}$ is a uniform random variable in the interval $(0.4, 0.8)$, and $\delta_{ij}$ has discrete uniform distribution with values in $\{-1, 1\}$. Next, define $\mathbf{\Omega}_2 = \mathbf{\Omega}_1 + (|\lambda_{min}(\mathbf{\Omega}_1)| + 0.05)\mathbf{I}_p$, where $\lambda_{min}(\mathbf{\Omega}_1)$ is the minimum eigenvalue of $\mathbf{\Omega}_1$, and $\mathbf{I}_p$ is a $p \times p$ identity matrix. The inverse of the precision matrix, or covariance matrix, is obtained from the transformation; $\mathbf{\Omega}^{-1} = \text{diag}(\mathbf{u})\mathbf{\Omega}_2^{-1}\,\text{diag}(\mathbf{u})$, where $\text{diag}(\mathbf{u})$ is a diagonal matrix formed by the $p$-dimensional vector $\mathbf{u}$, that is uniformly distributed in the interval $(1, 5)$. Finally, we have a multivariate Gaussian distribution $X_{\mathcal{V}} \sim \mathcal{N}_{\mathcal{V}}(\mathbf{0}, \mathbf{\Omega}^{-1})$ from which we obtain $n$ sample.

### 3.2 Results

The area under the ROC curve was used as a measure to compare the methods. Usually applied in binary classifiers, it describes the trade off between the false positives fraction, which is the probability of the method misclassify (specificity), and the true positives, which is the probability of a correct classification (sensitivity). Therefore, the methods with larger areas under the ROC curve are considered more efficient classifiers. The values for parameters used to plot the curves varies from the less regularized to the point where the graph is almost empty (full regularization). For the GLasso we used $\lambda_L \in \{0.001, 0.006 \ldots, 1; \text{by } 0.005\}$, in the GGMridge we used $\alpha_R \in \{0.0001, 0.0011 \ldots, 40; \text{by } 0.001\}$, and for the LPC $\alpha = \alpha_{LPC} \in \{10^{-4}, \ldots, 0.4; \text{by } 0.01\} \cup \{0.6, 0.7, 0.8, 0.9, 1\}$ varying at the same rate. In total we ran 300 simulations for each method with sizes: $(p = 50, n = 20)$, $(p = 100, n = 50)$, and $(p = 200, n = 50)$. Next we took the average of the produced specificity and sensitivity, generating the curves in Figure 1, and the areas in Table 1. We can observe that, in most cases, the GGMridge performed better than Glasso and LPC, with LPC in the Scale-free, and Small-world having a better performance than GLasso.

## 4 Gene Co-expession Network

### 4.1 Motivation

Gene co-expression analysis aims to identify genes whose expression differ in healthy cells in comparison with those with abnormal behavior, like tumorous cancer cells. Since most gene co-expression data is high dimensional, with the number of genes in the thousands and just a small number of samples, performing this kind of analysis can be very challenging. This situation inspires most of the network reconstruction methods currently being developed.

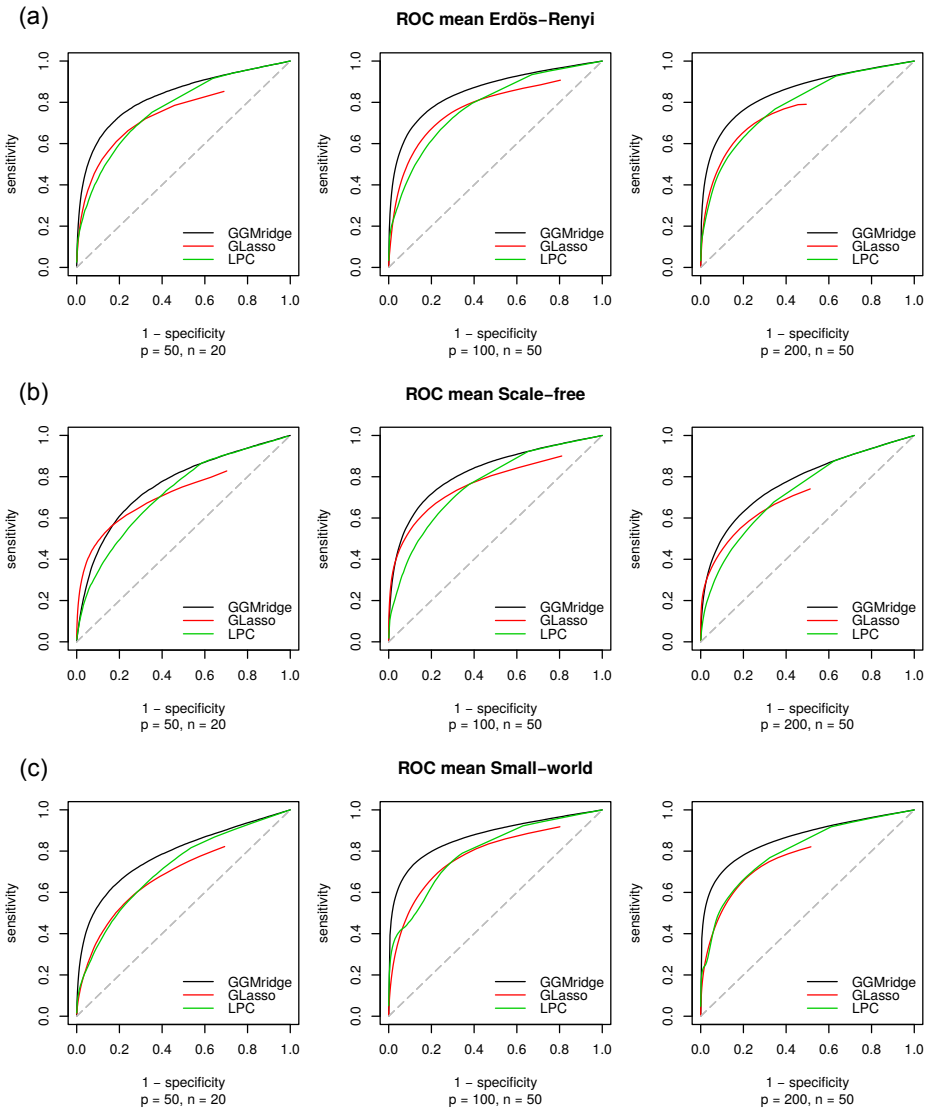


**Fig. 1.** In the Erdös-Rényi structure (a), we can observe that the GGMridge method had a better performance than GLasso and LPC. In this case GLasso performed better, by a small margin, than LPC in the network with $p = 50$, with LPC obtaining better results with $p = 100$, and $p = 200$. In the Scale-free structure (b), the GGMridge outperformed the GLasso and LPC, with LPC having better results than the GLasso in every case, as seen in Table 1. In the Small-world structure (c), LPC outperforms GLasso in every case, fact confirmed in Table 1. Again, the GGMridge method obtained curves with larger areas.

**Table 1.** ROC curve area by method, size, and methods.

| (variables, samples) | Graph type | GGMridge | GLasso | LPC |
|---|---|---|---|---|
| $(p = 50, n = 20)$ | Erdös-Rényi | .83022 | .75222 | .77078 |
| | Watts-Strogatz | .78478 | .69262 | .71712 |
| | Barabási-Albert | .76098 | .72620 | .72061 |
| $(p = 100, n = 50)$ | Erdös-Rényi | .85598 | .77944 | .78206 |
| | Watts-Strogatz | .87118 | .78364 | .79717 |
| | Barabási-Albert | .85133 | .76657 | .78932 |
| $(p = 200, n = 50)$ | Erdös-Rényi | .85133 | .76657 | .78932 |
| | Watts-Strogatz | .86213 | .77483 | .80106 |
| | Barabási-Albert | .77424 | .71715 | .72723 |

Motivated by this problem, the data used in the application comes from [13], which contains the expression of 25,387 genes extracted from 21 tumor tissue samples from human cervical cancer [18]. From this data, we selected a subset of 1,268 differently expressed genes (DEGs), identified by [14]. . Since networks can not be compared using the same threshold we followed heuristics provided by the biological community [19], [4], and decided to reconstruct the networks with 3 times more edges then nodes. We used specific thresholds and regularization coefficients. In GGMridge, the values where $\lambda_R = 1$, and $\alpha_R = 0.01$, for GLasso the value was $\lambda_L = 0.6$, and for the LPC, $\alpha = 0.1$, and $\alpha_{LPC} = 0.02$.

### 4.2 Reconstructed Networks

The thresholds and regularization parameters defined in 4.1, results in three different networks. GGMridge generated a network with 558 nodes and 1876 edges, GLasso, 630 nodes and 1942 edges, and LPC identified 658 nodes, with 1774 edges. In Figure 2, generated by Cytoscape [20], we have the reconstructed gene co-expression network of all nodes and edges obtained in common. Figure 3, is the Venn diagram representation of Figure 2. It provides valuable information of the results. We can observe that LPC and GGMridge share a large number of nodes, 343 in total. If we include the nodes shared with the GLasso we have 570 common identifications, almost 60% of the total. GLasso and LPC share 281 nodes, or 27%, and GLasso with GGMridge share 250 nodes, 26% of the total. On the other hand, the edges are not shared proportionally as the nodes, only 0.2% of the edges were identified in common between all methods. With LPC and GGMridge sharing considerably more, 4%. LPC and GLasso 0.06%, and GGMridge and GLasso 1.4%. The degree of the nodes, which is the number of edges incident to the node, where also different. The node with maximum degree was identified by Glasso, with size 188. In GGMridge, it was 35, and in LPC, 14. An interesting fact occurs when GGMridge and Glasso are observed together generating a maximum degree of 188. LPC and Glasso give 165, and LPC and GGMridge 35. The three methods identified a maximum degree of 177, which suggests that GLasso favors nodes with high degrees, or "hubs"[11].

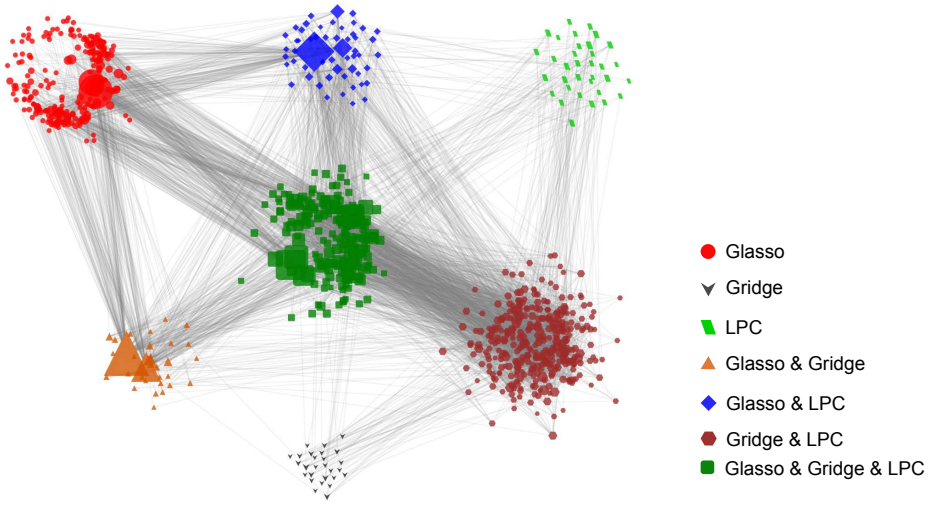


**Fig. 2.** The 244 red circles are the nodes identified only by GLasso, the 27 gray inverted triangles are the nodes identified by GGMridge, and the 34 green parallelogram where detected by LPC. The 33 orange triangles are the nodes identified by both GLasso and GGMridge. The 343 blue hexagons are the nodes identified by both the GGMridge and LPC. The 54 yellow diamonds are the nodes detected by both the GLasso and LPC, and The 227 light blue squares are the nodes identified by the three methods.The gray lines between the nodes, are the edges that where identified by the methods.

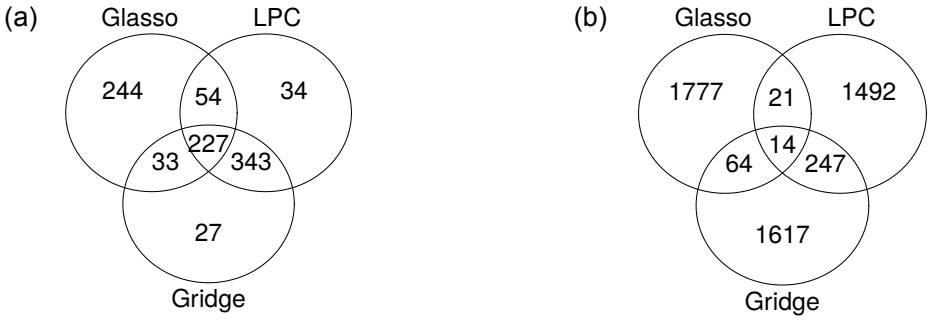


**Fig. 3.** In (a) we have the Venn diagram with the detected nodes by method, where in (b) we have the detected edges by method.

## 5 Conclusion

The simulation study provided an interesting glimpse of the three approaches. We observed that GGMridge outperformed GLasso and LPC in all cases. But as a relatively new method, GGMridge needs more tests in order to prove its attributes. LPC, although having an heuristic approach, performed better than GLasso in most cases, proving to be a viable and reachable alternative for GGM inference, without losing the partial correlation estimates. In the application, the

reconstructed networks differed from each other in significant ways. Specially, in relation to the detected edges, with GGMridge and LPC being the methods with more common identifications, in both edges and nodes.

## References


1. Lauritzen, S. L.: Graphical models. Vol. 17, Clarendon Press, Oxford (1996)
2. Mardia, K. V., Bibby, J. M., Kent, J. T.: Multivariate Analysis. Tenth printing, Academic Press, London (1995)
3. Thomas, L. D., Fossaluza, V., Yambartsev, A.: Building complex networks through classical and Bayesian statistics - A comparison. XI Brazilian Meeting on Bayesian Statistics: EBEB 2012, AIP Conference Proceedings, **1490**, 323–331 (2012)
4. Thomas, L. D., Vyshenska, D., Shulzhenko, N., Yambartsev, A., Morgun, A.: Differentially correlated genes in co-expression networks control phenotype transitions. F1000Research, **5**, 2740 (2016)
5. Schfer, J., Strimmer, K: Learning LargeScale Graphical Gaussian Models from Genomic Data. In: AIP Conference Proceedings, pp. 263–276, **776**(1), AIP (2005)
6. Friedman, J., Hastie T., Tibshirani R.: Sparse inverse covariance estimation with the graphical lasso, Biostatistics **9**(3), 432–441 (2008)
7. Tibshirani, R.: Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B **58**(1), 267–288 (1996)
8. Ha, M. J., Sun, W.: Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation, Biometrics **70**(3), 762–770 (2014)
9. Schäfer, J., Strimmer, K.: An empirical Bayes approach to inferring large-scale gene association networks, Bioinformatics. **21**(6), 754–764 (2005)
10. Erdos, P., Rényi, A.: On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. **5**(1), 762–770 (1960)
11. Barabási, A. L., Albert, R.: Emergence of scaling in random networks, Science, **286**(5439), 509–512 (1999)
12. Watts, D. J., Strogatz, S. H.: Collective dynamics of Small World Networks, Nature, **393**(6684), 440–442 (1998)
13. Zhai, Y., et al: Gene Expression Analysis of Preinvasive and Invasive Cervical Squamous Cell Carcinomas Identifies HOXC10 as a Key Mediator of Invasion, Cancer Research, **67**(21) 10163–10172 (2007)
14. Mine, K., et al: Gene network reconstruction reveals cell cycle and antiviral genes as major drivers of cervical cancer. Nature Communications, **4**(1) (2013)
15. Meinshausen, N.,Bhlmann, P.: High-dimensional graphs and variable selection with the lasso.The Annals of Statistics, **34**(3), 1436–1462 (2006)
16. Hoerl, A., Kennard, R.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Nature Communications, **12**(1), 55–67 (1970)
17. Cai, T., Liu, W., Zhou, H.: Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. The Annals of Statistics, **44**(2), 455–488. (2016)
18. Gene Expression Omnibus Repository, `https://www.ncbi.nlm.nih.gov/geo/`. Access number GSE7803, Affymetrix U133A platform. Last accessed 11 Jun 2018
19. Dong, X., Yambartsev, A., Ramsey, S. A., Thomas, L. D., Shulzhenko, N., Morgun, A.: Reverse enGENEering of Regulatory Networks from Big Data: A Roadmap for Biologists. Bioinformatics and Biology Insights, **9**, 61–74 (2015)
20. Shannon, P.: Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research, **13**(11), 2498–2504 (2003)