Special issue of the 3rd International Conference on Computational and Experimental Science and Engineering (ICCESEN 2016)

Comprehensive Analyses of Gaussian Graphical Model under Different Biological Networks

D. Dokuzoğlu, V. Purutçuoğlu*
Middle East Technical University, Department of Statistics, Ankara, Turkey

Naturally, genes interact with each other by forming a complicated network and the relationship between groups of genes can be shown by different functions as gene networks. Recently, there has been a growing concern in uncovering these complex structures from gene expression data by modeling them mathematically. The Gaussian graphical model is one of the very popular parametric approaches for modelling the underlying types of biochemical systems. In this study, we evaluate the performance of this probabilistic model via different criteria, from the change in dimension of the systems to the change in the distribution of the data. Hereby, we generate high dimensional simulated datasets via copulas and apply them in Gaussian graphical model to compare sensitivity, specificity, F-measure and various other accuracy measures. We also assess its performance under real datasets. We consider that such comprehensive analyses can be helpful for assessing the limitation of this common model and for developing alternative approaches, to overcome its disadvantages.

DOI: 10.12693/APhysPolA.132.1106
PACS/topics: 02.50.–r, 02.50.Ng, 02.70.Rr

1. Introduction

The biologists routinely use high-throughput technologies such as microarrays to measure the expressions of genes. Accordingly, it is usual to apply multivariate methodologies in order to analyze these large datasets and to disclose various interactions among genes, that cannot be established from individual gene-based approaches.

Thereby, the inference of gene networks plays an important role in enlightening the underlying interactions among genes, that may lead to a better understanding of biological activations in organisms. In this study, we focus on a graphical modeling approach, that purposes at finding relationships in a group of genes, where a graph is used for encoding relationships among multiple variables.

When a graph is used for a gene network, the nodes represent genes and the edges indicate interactions between the linked genes. In other words, if any two genes are connected to each other by an edge, indirectly, they can be affected by other genes. Therefore, the appearance profiles of two genes are correlated, as long as they are both regulated by some other genes. Hence, the large datasets allow us to infer the relationships among these genes and the Gaussian graphical model (GGM) is one of the well-known alternative approaches to get these findings.

Indeed, this model is also suggested as the alternative of differential equations (DE) modelling in the description of the steady-state activation of the biological systems. But the DE models are deterministic [1–7], whereas, GGM is probabilistic. GGM is simply dependent on the estimated partial correlation matrix, interpretation of which is straightforward under the normality assumption of the data [8]. Here, the zero entry implies no relation between the associated pair of genes, due to the feature of the conditional independence under the multivariate normal distribution.

Thus, under the GGM assumption, the graph structure can be estimated using the sparsity pattern of the inverse covariance matrix Σ, also called the precision or concentration matrix $\Sigma^{-1} = \Theta$ [8]. There are different approaches to infer Θ in GGM. Among alternatives, the neighborhood solution method [9] and the graphical lasso or glasso approach [10], are the most commonly used ones.

In this study, we choose the glasso approach in our analyses in the estimation of Θ. This method is originated from the lasso regression with the ℓ_1-norm via $Y^{(p)} = \beta Y^{(-p)} + \varepsilon$, where the node $Y^{(p)}$ depends on the rest of the nodes $Y^{(-p)}$ and β denotes the p-dimensional regression coefficients [8].

Here, it is assumed that ε has a p-dimension and it is a multivariate normally distributed random error with the zero mean vector and the covariance matrix $\sum_{p\times p}$. We can present the objective function that is maximized with respect to Θ under the ℓ_1-penalized log-likelihood function as $\log\det(\Theta) - \mathrm{tr}(S\Theta) - \lambda|\Theta|_1$, in which n is the number of observations, p denotes the number of nodes and S shows the empirical covariance matrix. Furthermore, λ is the non-negative Lagrange multiplier. When λ gets larger, the biological network becomes sparser [10]. Finally, tr(.) and det(.) describe the trace and determinant, respectively, and $|\Theta|_1$ represents the ℓ_1-norm of Θ.

In the calculation, the optimal selection of λ can be done by different approaches, such as STAR [11], EBIC [12] and RIC [13]. Here, we select RIC (rotation information criterion) since it is the most common measure of GGM, if the inferences are conducted by the glasso method [13]. Accordingly, in this study, we investigate

*corresponding author; e-mail: vpurutcu@metu.edu.tr

the performance of GGM comprehensively under different distributions, from normal to skew densities, with distinct mixtures of marginals, via copula and various dimensions. We evaluate the outputs based on various measures of accuracy and interpret the results.

2. Methods

2.1. Copula

In our analyses, we perform copulas to generate different joint distributions in modelling via GGM. In general, copulas provide the theoretical framework in which the multivariate associations can be modeled separately from the univariate distributions of the observed variables, based on the Sklar theorem by $H(x, y) = C\{F(x), G(y)\}$ [14]. Here, if x and y represent two continuous random variables, the copula function C, which characterizes the joint dependency of x and y, should be unique.

There are seven major types of copulas, namely, Gaussian, Student-t, Gumbel, FGM, Clayton and Frank copulas. Each of them presents distinct ranges for the random variables and denotes different levels of correlations. Archimedean copula families, which are composed of the Gumbel, Frank and Clayton copulas, are constructed with only a single dependency parameter θ [14]. Furthermore, it is not clear which parameters create a reliable model under which values and which dependence structure can be created by the given copula function. Moreover, in the Archimedean family, the Gumbel and Clayton copulas do not have explicit density expressions, if we infer their copula terms [15].

Hereby, as the Gaussian copula does not have these limitations, is applicable for high dimensional data and covers both low and highly correlated systems, we use it in our analyses with a wide variety of marginals. The Student-t copula also has a similar feature. Whereas, as it is not mathematically convenient and its outcomes are similar to the Gaussian copula, we choose the Gaussian copula due to its computational facilities in calculations.

2.2 Measure of accuracy

In this research, in order to assess the performance of GGM, we apply the well-known accuracy measures, which are the precision, recall, F_1-score, false positive rate (FPR), true positive rate and the accuracy whose mathematical descriptions are given in Table I. In these expressions, TP denotes the true positive value, which is the number of correct predictions of actually positive entry; FP represents the false positive value, that indicates the number of incorrect predictions of actually negative entry; FN refers to the false negative rate, which implies the number of incorrect predictions of actually positive entry and finally, TN shows the true negative rate that is the number of correct predictions for actually negative entries. In these accuracy measures, apart from FPR, the perfection levels are equal to one and for FPR, the best performance is seen under the zero entry.

TABLE I

Formulas of accuracy measures.

Accuracy measure	Formula
Precision	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$
Recall	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$
F_1-score	$2\times\frac{(\mathrm{Precision})(\mathrm{Recall})}{\mathrm{Precision+Recall}}$
Accuracy	$\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$
False positive rate	$\frac{\mathrm{FP}}{\mathrm{TN+FP}}$

3. Application

3.1. Application with simulated data

In the light of the assessment of the limitation of this common model, different runs are completed under various scenarios by performing the Monte Carlo simulation. In each analysis, we use 1000 iterations. To evaluate the adequacy of GGM, we compare the exact graph path, i.e., the population graph path, with the estimated graph path, i.e., sample graph path, under different dimensions, graph structures and copula functions. Moreover, in our simulation, we take the total number of nodes, i.e. genes, that is also named as the dimension of the networks as 20, 50 and 100 nodes. For each dimension, a random sample of size twenty ($n = 20$) is drawn from the simulated multivariate data to be modeled by GGM. The findings are shown in Tables II and III. In these comparative analyses, the simulations are separated into two parts, which are the GGM application in multivariate normal data and multivariate data with the Gaussian copula function.

In the first class, the multivariate normally distributed data are created under different biologic networks and dimensions. Four types of biologic networks, which are scale-free, cluster, random and hubs are used to assess their differences in the implementation of GGM. However, it is known that the scale-free networks are the most common types for the biological systems [16]. Thereby, after generating 20 observations for each node in the system, the graphical lasso (glasso) method is implemented to infer the graph path.

In the application of the glasso method, the estimation of the penalty constant λ is performed via the RIC criterion, as stated beforehand. On the other hand, in the second stage of the study, the multivariate data are simulated via the Gaussian copula functions and applied in GGM. In this assessment, we use the normal, Student-t, log-normal [17] and the exponential marginals within the copula function, with their suitable parameters. In the selection of these marginals, we consider that the Student-t is one of the close alternatives of normal distributions. The log-normal distribution is particularly preferable for the data with extreme positive values, as observed in most of the biological mechanisms (e.g., exponential growth) and chemical phenomena (e.g., the velocity of a simple reaction) [15].

In Table II, we represent the accuracy measures of different network types under multivariate normal data. From the results, it is seen that the accuracy measures

TABLE II

Accuracy measures of GGM for different types of networks under multivariate normality.

Graph type	Number of nodes	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Scale-free	20	0.467	0.902	0.207	0.0262	0.320
	50	0.390	0.960	0.016	0.001	0.030
	100	0.249	0.980	0.001	0.000	0.001
Random	20	0.540	0.862	0.192	0.028	0.303
	50	0.497	0.942	0.018	0.001	0.035
	100	0.492	0.970	0.001	0.000	0.002
Cluster	20	0.573	0.738	0.118	0.035	0.198
	50	0.540	0.888	0.015	0.002	0.030
	100	0.485	0.943	0.000	0.000	0.002
Hubs	20	0.487	0.909	0.296	0.033	0.428
	50	0.456	0.962	0.033	0.002	0.061
	100	0.466	0.926	0.002	0.000	0.004

of GGM decrease, while the network becomes larger for all network types.

Moreover, the best performance is observed under scale-free and hubs networks, as they are very close to each other in terms of the sparsity performance of the graphs [16]. Whereas, the accuracy measures decrease significantly if the systems are random or cluster types.

Furthermore, in Table II, it is found that even under multivariate normality, apart from accuracy and FPR, none of other measures indicate good performance, although GGM is originally designed for the data from this sort of distribution.

On the other hand, in order to observe the accuracy of GGM under non-normal data, we initially apply the Student-t margins with the degrees of freedom 10, since the higher degrees of freedom bring us the wider data distribution, similar to the normal distribution.

From the results in Table III, it is seen that all entries considerably decrease, even though the Student-t distribution is one of the close alternatives of the multivariate normal distribution. If we repeat the analyses under log-normal marginals, we detect that the performance of GGM becomes worse. Here, we take two choices of standard deviations, as the higher deviation implies more skewed structure and captures more extreme values.

On the contrary, when the standard deviation decreases, the shape of the distribution looks like a more symmetric, similar to the normal distribution. Then, to detect the GGM performance in skew data, we take the exponential distribution margins with rate $\lambda = 4$. From Table III, it is found that GGM can capture some direct edges between nodes and its accuracy is slightly higher with respect to the results of symmetric data.

Finally, besides the simulation of the Gaussian copula with single marginal types, we also model the copula function with mixed marginals, to observe the performance of this well-known model under mixed densities. Hereby, we apply a join distribution, whose half of the data is from the exponential marginals with rate 4 and other half is from the normal marginal with mean zero and variance 4. As seen from the Monte Carlo results, GGM cannot capture any direct edges between nodes and the model can merely assign zeros.

3.2 Application with real data

In this part, we apply GGM to two bench-mark real biological datasets. As the first dataset, we use the cell signaling data, which contain information about 11 phosphoproteins and some phospholipids [18]. These 11 proteins are called as praf, pmek, plcg, PIPP2, PIP3, p44.42,

TABLE III

The accuracy table of GGM for different marginal distributions under scale-free networks (Not Comp. stands for not computable and stnd. dev. stands for standard deviation).

Marginals	Number of nodes	Precision	Accuracy	Recall (TPR)	FPR	F_1-score
Student-t	20	0.575	0.906	0.053	0.004	0.092
(degree of freedom=10)	50	0.533	0.961	0.000	0.000	0.000
	100	Not Comp.	0.980	0.000	0.000	0.000
Log-normal	20	Not Comp.	0.905	0.000	0.000	0.000
(mean=10, stnd. dev.=8)	50	Not Comp.	0.9608	0.000	0.000	0.000
	100	Not Comp.	0.9802	0.000	0.000	0.000
Log-normal	20	Not Comp.	0.905	0.000	0.000	0.000
(mean=10, stnd. dev.=0.5)	50	Not Comp.	0.9608	0.000	0.000	0.000
	100	Not Comp.	0.9802	0.000	0.000	0.000
Exponential	20	0.138	0.545	0.723	0.501	0.648
(rate=4)	50	0.058	0.650	0.515	0.352	0.550
	100	0.028	0.748	0.347	0.247	0.434
Semi-exponential (rate=4)	20	Not Comp.	0.905	0.000	0.000	0.000
Semi-normal	50	Not Comp.	1	0.000	0.000	0.000
(mean=0, stnd. dev.=2)	100	Not Comp.	1	0.000	0.000	0.000

pakts473, PKA, PKC, P38 and pjnk, where each of them has 1000 observations, resulting in 11000 measurements and whose true network structure is represented in Fig. 1.

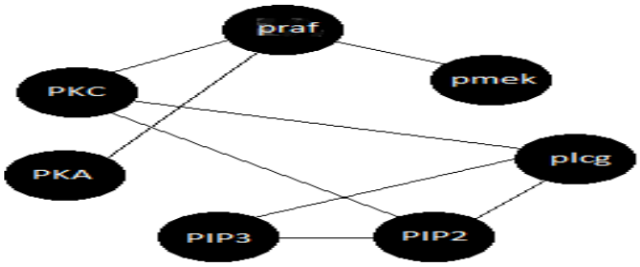

Fig. 1. The true network of the cell signaling proteins.

From modeling of this dataset via GGM, we find none of the underlying true links and GGM can merely assign zero entries in the precision matrix for all estimated interactions. In order to critic the plausible reason behind this estimation, we check the QQ-plots of each protein (see Supplementary material, Fig. S1) and we find that the distributions of each marginal protein are far from the normal density, although the structure of the system is suitable for the GGM-type of the mathematical modelling.

As the second real data application, we apply the human gene expression data which contain 100 transcripts, measured on 60 unrelated individuals. The data are collected by Stranger et al. [19] and are defined by Bhadra and Mallick [20] and Chen et al. [21]. The purpose of the data is to understand the gene expression in the B-lymphocyte cells from the Northern and Western European ancestry from Utah. The main focus of these studies is the 3125 single nucleotide polymorphisms which are found in the 5 UTR (untranslated region).

Hereby, from modelling via GGM, similar to the previous results, we cannot discover any of the validated links. Then, to check the normality of this dataset as the plausible source of deficiencies in the model, we compute the Shapiro-Wilk test for the multivariate normality and we take the significance level of 0.05. The results show the departure from the normality with a p-value $<2.2\times10^{-16}$. When we draw the QQ-plots of the genes, similar to the findings of the first dataset, we find non-normal distribution of the gene expressions. The univariate plots of selected 15 genes are presented in Supplementary material (Fig. S2), as examples of this analysis.

4. Conclusions

In this study, we have considered to comprehensively evaluate the performance of Gaussian graphical model (GGM), which is one of the common modelling approaches for the description of the steady-state behaviors of biological systems. For this purpose, we have assessed the findings of GGM, first of all, under different dimensions and then the topology of the networks and under various distributions.

In all these calculations, we have computed the accuracy of the estimates based on various measures. From Monte Carlo studies, under multivariate normal and different marginals bounded by the Gaussian copula, as well as real data analyses, we have detected that the accuracy of GGM is very limited and its performance is good only under the very strict normality assumption and under the scale-free type of networks.

Except for these special conditions, GGM cannot be successful in modeling biochemical networks. Therefore, in order to unravel this challenge, we consider to implement non-parametric alternatives of GGM. For this purpose, we have been working on MARS (multivariate adaptive regression splines) and random forest models [22]. Our current outputs have shown promising results and this topic is still an ongoing study of our group.

Acknowledgments

The authors would like to thank to TÜBİTAK (Project no: 114E636) for their support.

References

[1] Z. Akhmetova, S. Zhuzbaev, S. Boranbayev, *Acta Phys. Pol. A* **130**, 352 (2016).

[2] B. Gürbüz, M. Sezer, *Acta Phys. Pol. A* **130**, 194 (2016).

[3] E. Boutalbi, A. Gougam, F. Mekideche-Chafa, *Acta Phys. Pol. A* **128**, B-271 (2015).

[4] K. Ergen, A. Çıllı, N. Yahnıoğlu, *Acta Phys. Pol. A* **128**, B-273 (2015).

[5] M.J. Pazdanowski, *Acta Phys. Pol. A* **128**, B-213 (2015).

[6] A. Recioui, *Acta Phys. Pol. A* **128**, B-7 (2015).

[7] İ.S. Üncü, A. Arisoy, B. Büyükarikan, *Acta Phys. Pol. A* **128**, B-474 (2015).

[8] J. Whittaker, *Graphical models in Applied Multivariate Statistics*, John Wiley and Sons, Chichester 1990.

[9] N. Meinshausen, P. Bühlmann, *Ann. Stat.* **34**, 1436 (2006).

[10] J.H. Friedman, T. Hastie, R. Tibshirani, *Biostatistics* **9**, 432 (2008).

[11] H. Liu, K. Roeder, L. Wasserman, in: *Advances in Neural Information Processing Systems (NIPS)*, 2010, p. 1.

[12] J. Chen, Z. Chen, *Biometrika* **95**, 759 (2008).

[13] T. Zhao, H. Luin, N. Simon, *J. Mach. Learn. Res.* **13**, 1059 (2012).

[14] R.B. Nelsen, *An Introduction to Copulas*, 2nd ed., Springer Science and Business Media, Portland 2006.

[15] M.M. Wawrzyniak, D. Kurowicka, *Dependence concepts*, Delft University of Technology, Netherlands 2006.

[16] A.L. Barabási, Z.N. Oltvai, *Nature Rev. Genetics* **5**, 101 (2004).

[17] E. Limpert, W.A. Stahel, M. Abbt, *BioScience* **51**, 341 (2001).

[18] K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, G. Nolan, *Science* **308**, 523 (2005).

[19] B. Stranger, A. Nica, M. Forrest, A. Dimas, C. Bird, C. Beazley, C. Ingle, M. Dunning, P. Flicek, S. Montgomery, S. Tavare, P. Deloukas, E. Dermitzakis, *Nature Genetics* **39**, 1217 (2007).

[20] A. Bhadra, B.K. Mallick, *Biometrics* **69**, 447 (2013).

[21] L. Chen, F. Emmert-Streib, J. Storey, *Genome Biol.* **8**, R219 (2007).

[22] T. Hastie, R. Tibshirani, J.H. Friedman, *The Elements of Statistical Learning*, Springer Verlag, New York 2001.

Supplementary material

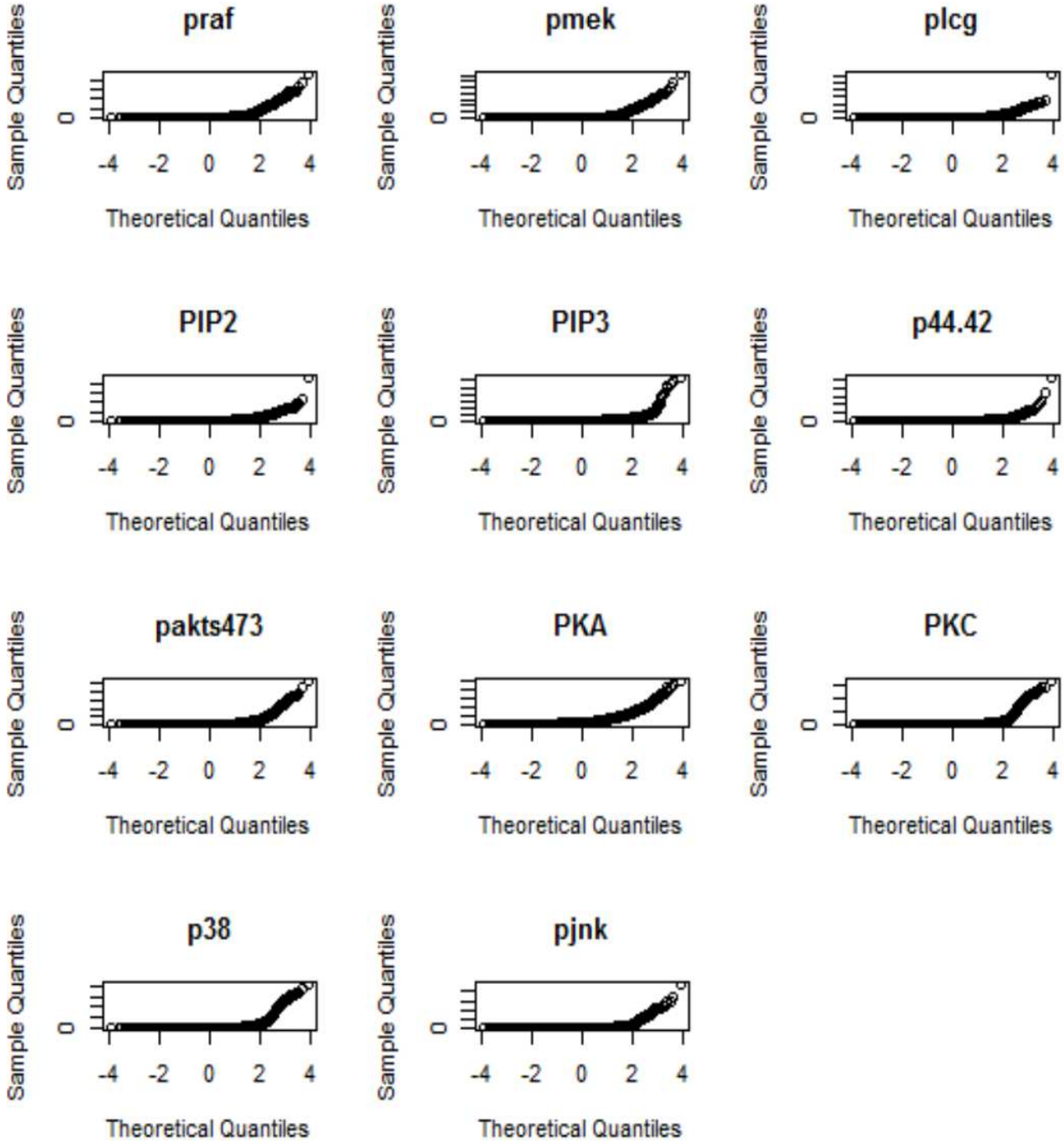

Fig. S1. The QQ-plots of the cell signaling data by comparing the normal density.

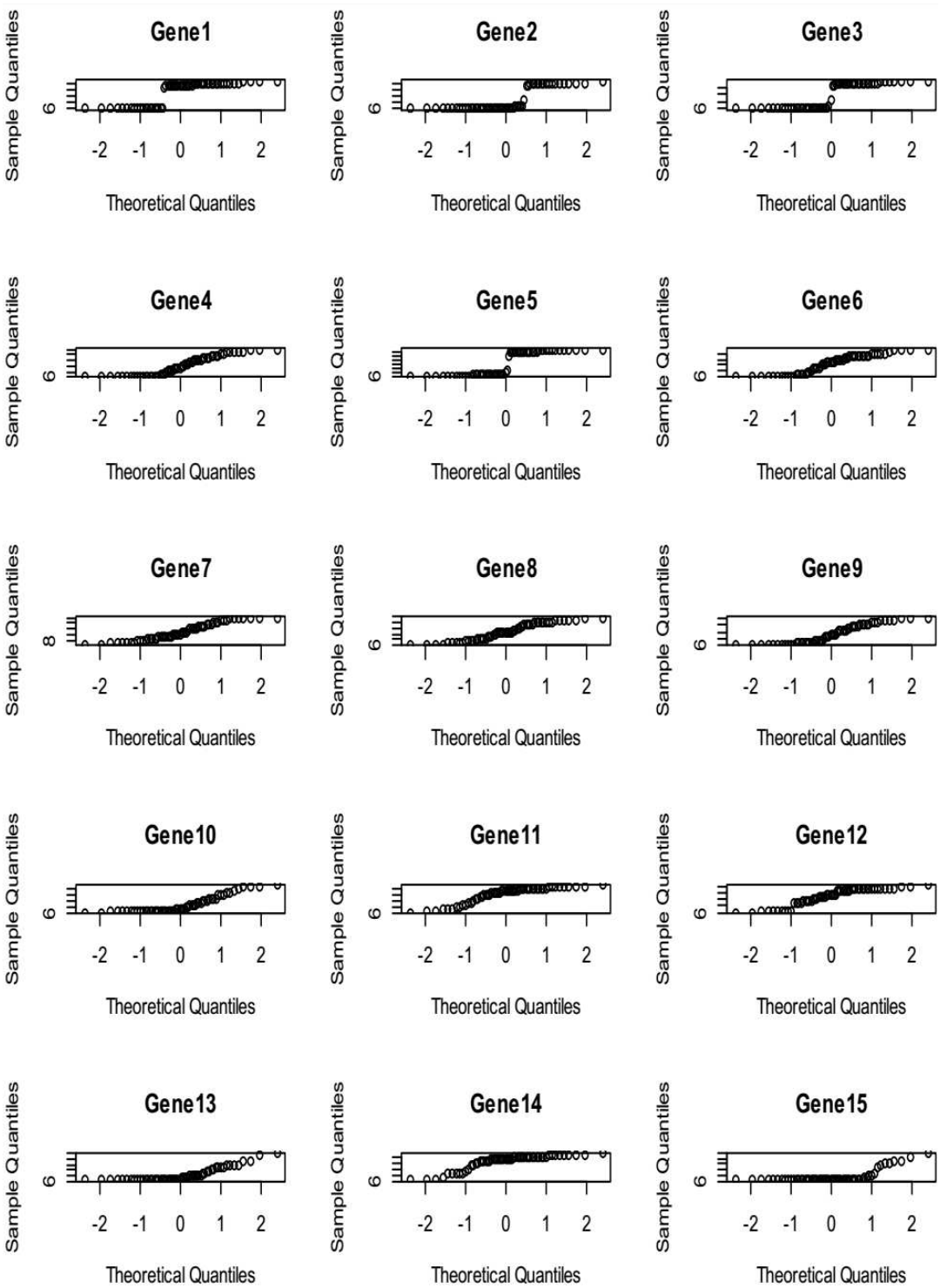

Fig. S2. The QQ-plots of the human gene expression data by comparing the normal density.