Lijun Wang[1] / Md. Asif Ahsan[2] / Ming Chen[2]

A Generalized Approach for Measuring Relationships Among Genes

[1] Department of Statistics, School of Mathematical Sciences, Zhejiang University, Hangzhou 310058, P.R. China
[2] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, P.R. China, E-mail: mchen@zju.edu.cn

Abstract:
Several methods for identifying relationships among pairs of genes have been developed. In this article, we present a generalized approach for measuring relationships between any pairs of genes, which is based on statistical prediction. We derive two particular versions of the generalized approach, least squares estimation (LSE) and nearest neighbors prediction (NNP). According to mathematical proof, LSE is equivalent to the methods based on correlation; and NNP is approximate to one popular method called the maximal information coefficient (MIC) according to the performances in simulations and real dataset. Moreover, the approach based on statistical prediction can be extended from two-genes relationships to multi-genes relationships. This application would help to identify relationships among multi-genes.

Keywords: Generalize, Prediction, Relationship

DOI: 10.1515/jib-2017-0026

Received: April 4, 2017; **Revised:** April 18, 2017; **Accepted:** May 4, 2017

1 Introduction

There are novel, high-throughput technologies, such as transcriptome microarrays, opening global perspectives of living organisms on the molecular level. With a vast experimental literature on biomolecular processes, these data indicated that there are regulatory networks rather than isolated genes [1] . Co-expression network plays an important role in underlying cellular processes and pathways, which is one common form of biological network. Co-expression network is constructed by analyzing the relationships between pairs of genes expression data. So, it is important to identify the pairwise relationships between genes in biological network. Many computational methods for analyzing the relationships have been developed. For example, weighted correlation network analysis (WGCNA) [2] is a method based on correlation, which can find the clusters or modules of highly correlated genes. Besides, there are other methods that based on information-theoretic technique to measure the associations between genes, such as mutual information (MI) [3], maximal information coefficient (MIC) [4], maximal information-based nonparametric exploration (MINE) [4] and maximal information component analysis (MICA) [5]. Here, we will present a measure of dependence for two-variable relationships, and illustrate that this is a generalized form for identifying relationships. In other words, whether it is the method based on correlation, or that methods base on information-theoretic technique, can be attributed to our approach which based on statistical prediction.

We have derived two versions of our generalized approach, and found the equivalences to methods based on correlation and MIC, respectively.

2 The Approach Based on Statistical Prediction

Imagine that we observe each of p genes expression level for N individuals, or at N different time, we will get an expression data matrix, whose dimension is $N \times p$. With statistical language, the p genes can be treated as p random variables, and each of them have N observations. In order to construct the biological network, we need to identify the pairwise relationships between two genes. Let random variables X, Y denote two distinct genes, and a $N \times 1$ vector $\boldsymbol{x}$ denotes gene X's expression data, similarly, $\boldsymbol{y}$ denotes gene Y's expression data. In fact, $\boldsymbol{x}$, $\boldsymbol{y}$ are the two columns of expression data matrix. To measure the relationship or dependence between gene X

Ming Chen is the corresponding author.

and gene Y, it is necessary to judge to what extent gene X's expression decides gene Y's expression. In other words, given the expression level of gene X, how much can we predict gene Y's expression. If the prediction is perfect, then it is obvious that gene Y has a strong dependence on gene X, and vice versa. Thus, this approach is based on statistical prediction, and there are two steps to implement. Firstly, predict Y conditioned on X; then measure the goodness of prediction.

According to statistical decision theory [6], let $f(x)$ denote the prediction of Y using X. We use the most common and convenient squared error loss $L(Y, f(X)) = (Y - f(X))^2$, it leads to the criterion for choosing f

$$EPE(f) = E((Y - f(X))^2) \tag{1}$$

where EPE means the expected (squared) prediction error. To minimize EPE pointwise, we will get the solution

$$f(x) = E(Y \mid X = x) \tag{2}$$

which is known as the regression function [6].

In fact, the $E(Y \mid X)$ is impossible to compute, that means we cannot get the exact $f(X)$ in real problem. Thus, the thing that we only can do is to estimate it. There are many estimations, such as the least squares regression based on Gauss hypothesis. After getting the estimation $\hat{f}(x)$, we can measure the goodness of fitting by using the coefficient of determination

$$R^2 = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}} \tag{3}$$

where $\hat{\boldsymbol{y}}$ represents the prediction of gene Y's expression values computed by the estimated regression function $\hat{f}(x)$, while $\boldsymbol{y}$ represents the real gene expression vector, and we assume both of them have been standardized.

In the approach based on statistical prediction, once choosing an estimation of $f(x)$, we would obtain the coefficient of determination R^2, which is suitable for measuring the dependence of gene Y on gene X. It is worth noting that the relationship in this approach is bilateral, because we only consider predicting gene Y's expression conditioned on gene X's expression. In other words, it is not necessary that the relationship between Y and X equals the relationship between X and Y.

If applying the generalized approach to construct biological network, it turns out to be a bilateral network. In the methods based on correlation, it is obvious that $cor(X, Y) = cor(Y, X)$; and MIC also satisfies symmetry of mutual information. Although the approach based on prediction might not satisfy symmetry in some case, we will show that the coefficients of determination would approximately equal in some cases.

3 Correlation and Least Square Estimation

In this section, we demonstrate the equivalence of correlation and the least squares estimation (LSE), derived from the generalized approach based on statistical prediction. Assume that $f(x)$ is well approximated by a globally linear function,

$$f(\boldsymbol{x}) \approx \boldsymbol{x}^T \beta \tag{4}$$

where $\boldsymbol{x}^{\boldsymbol{T}}$ denotes the transpose of gene expression vector $\boldsymbol{x}$. Then by plugging this linear hypothesis into EPE (1) and differentiating, we can solve for β theoretically, and obtain the prediction of gene expression values

$$\begin{aligned}\hat{f}(\boldsymbol{x}) &= \boldsymbol{x}^{\boldsymbol{T}} \cdot E(\boldsymbol{x}\boldsymbol{x}^T)^{-1} E(\boldsymbol{x}y) \\ &= \boldsymbol{x}^{\boldsymbol{T}} (\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}})^{\boldsymbol{-1}} \boldsymbol{x}y\end{aligned} \tag{5}$$

This prediction is known as least squares estimation. Finally, the coefficient of determination R^2 for measuring the dependence of Y on X can be computed by plugging $\hat{f}(\boldsymbol{x})$ as $\hat{\boldsymbol{y}}$ into Equation (3).

In fact, we can prove that correlation $r = cor(x, y) = \sqrt{R^2}$. Thus, the coefficient of determination R^2 under the linear hypothesis is equivalent to the correlation. Besides, due to the linear hypothesis, we can obtain the dependence of X on Y in linear form. In this case, the coefficient of determination R^2 is symmetry. So, the methods based on correlation can come down to our approach based on statistical prediction. There are some simulations which show LSE is indeed equivalent to the methods based on correlation.

Let R^2_{LSE} denote the coefficient of determination R^2 for least squares estimation. From Figure 1, we can see R^2_{LSE} and $cor(x, y)$ decrease synchronously as the noise increases. And both of them fail to capture the sinusoid relationships, but they both can seizure the linear relationships. Therefore, the R^2_{LSE} is indeed equivalent to the correlation.

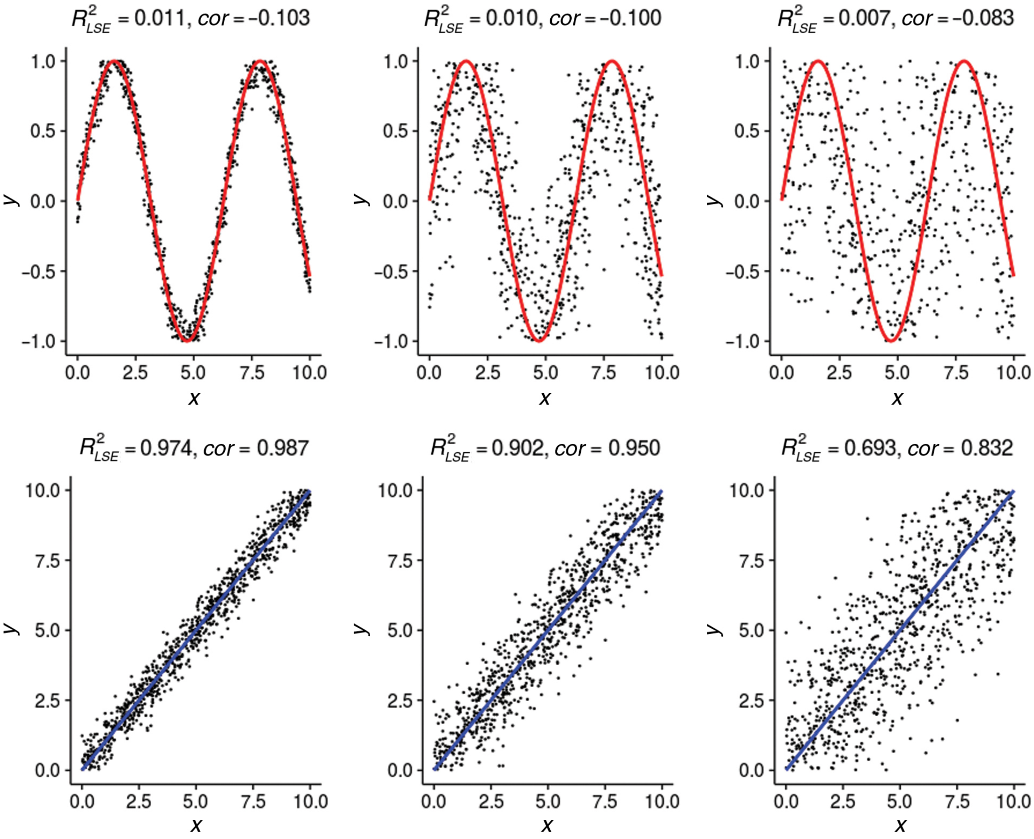

Figure 1: Comparison of R^2_{LSE} and correlation. The first row indicates that Y and X have the sinusoid relationship. As increasing noise from left to right, the corresponding R^2_{LSE} and $cor(x, y)$ decrease synchronously. And these values are too small to capture the sinusoid relationships. The second row indicates that Y and X have the linear relationships. Similarly, as increasing noise from left to right, the corresponding R^2 and $cor(x, y)$ decrease synchronously. But these values are relatively large, which indicates that both R^2_{LSE} and $cor(x, y)$ can capture the linear relationships.

4 Nearest Neighbors Prediction and MIC

The maximal information coefficient (MIC) is a measure of dependence for two-variable relationships, which based on information-theoretic technique, particularly the mutual information. MIC falls between 0 and 1, and it assigns scores that tend to 1 for noiseless relationships and assigns scores that tend to 0 to statistically independent variables [4]. In other words, higher MIC means that the paired relationship is more strong. In this section, we will present another version of prediction method called nearest neighbors prediction, termed as NNP. We establish the equivalence of MIC and NNP through the performances of simulations and real dataset. In LSE, we estimate $E(Y \mid X)$ based on the linear hypothesis, whereas in NNP, we make full use of the raw data, and use as less hypothesis as possible. At each expression level x of gene X, we need to figure out the expression level y of gene Y. NNP uses the average of expression y_is in the neighborhood of x to estimate y (see Figure 2), that means

$$\hat{f}(x) = \text{Ave}(y_i \mid x_i \in N(x)) \tag{6}$$

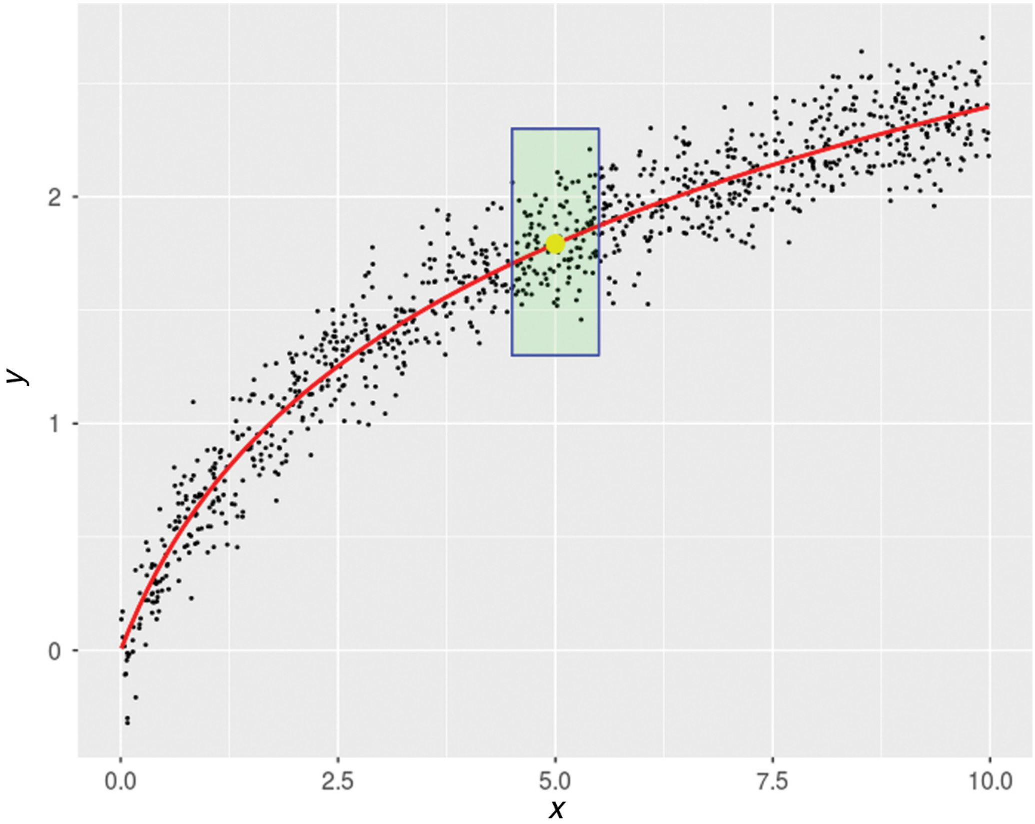

Figure 2: Nearest Neighbors Prediction. The red curve represents the true relationship between Y and X; the blue box indicates the yellow point's neighborhood. Here, we set $\delta = 0.5$.

where "Ave" denotes average, and $N(x)$ means the neighborhood of x, which we have two choices. One is to choose $N_k(x)$, which means the neighborhood contain k points that closest to x, another is to use $x \pm \delta$ to measure x's neighborhood, where δ represents the region width. And we adopt the second way to implement our method NNP.

After getting the prediction of Y at the corresponding X, we use the coefficient of determination R^2 to measure the dependence of Y on X.

We choose some function types to do simulations, and obtain the results in Figure 3, it is obvious that both R^2 and MIC are good at capturing the relationship between Y and X, no matter what the form of functions (or relationships) are, and they change synchronously as the noise increases. More importantly, these values are very close. So we can assume that there is a coefficient of determination R^2, corresponding to one estimation of $E(Y \mid X)$, is exactly the statistic MIC. If it does exist the MIC version of prediction method (need further theoretical proof), we can use the prediction method to interpret the MIC. Even if there is not such estimation of $E(Y \mid X)$, we found that NNP's performance is very close to MIC's, so it can be regarded as the approximate version of MIC.

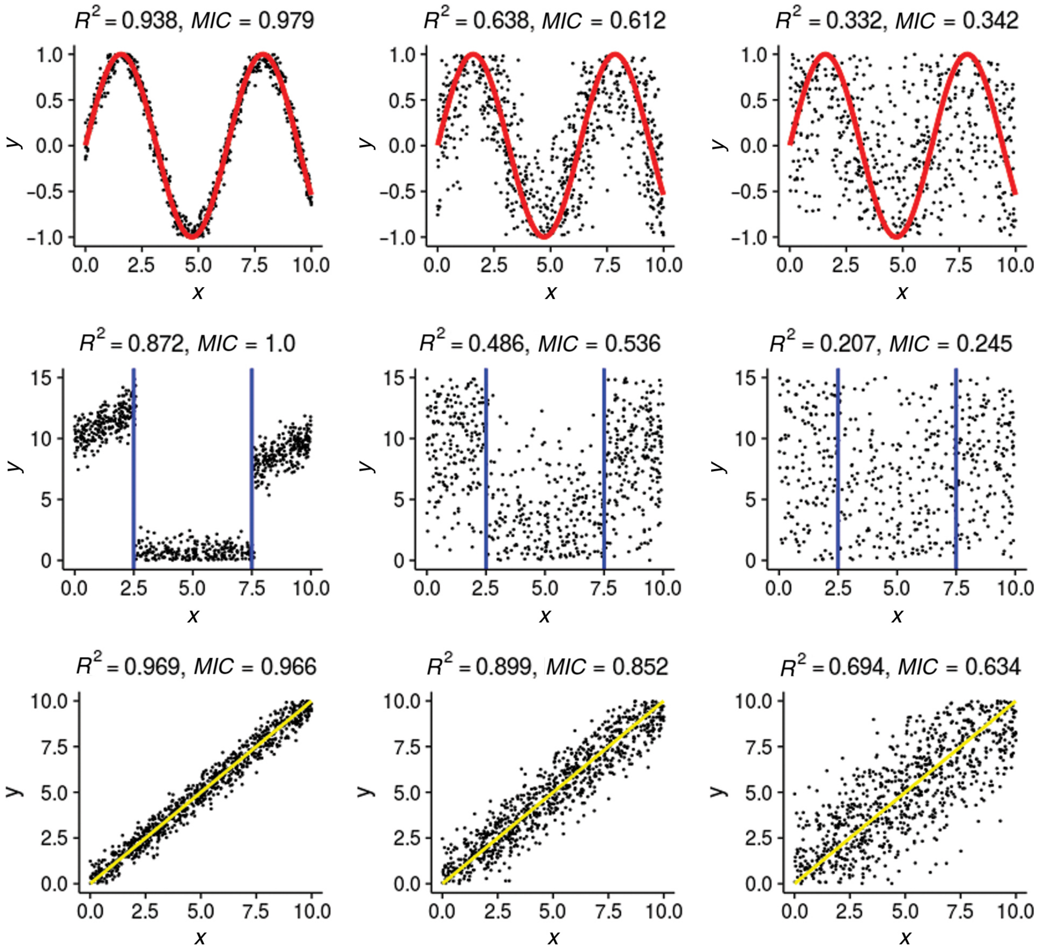

Figure 3: Comparison of MIC and NNP. At the first row, Y and X have the sinusoid relationship, and as the noise increases from left to right, the corresponding R^2 and $MIC(x, y)$ decrease synchronously. More importantly, these values are relatively large, which indicates that both of them capture the sinusoid relationship perfectly. At the second row, Y and X have the segmentation relationship. Similarly, as the noise increases from left to right, the corresponding R^2 and $MIC(x, y)$ decrease synchronously, and these values are large enough to capture the dependence of Y on X. The third row shows the linear relationship, both R^2 and $MIC(x, y)$ can perfectly capture the linear relationship. And these values decrease synchronously as the noise increases from left to right.

To further examine the approximate equivalence of NNP and MIC, we explored a yeast expression dataset [7]. MIC has used this yeast expression dataset to compute the relationships between genes expression levels and time. We have tested the performance of MIC and NNP, and found both of them exhibit similar results.

From Figure 4, it is clearly that the pattern of MIC is the same as NNP, especially when the relationships computed by these two approaches are higher than 0.5. In general, the value that is higher than 0.5, indicates that the corresponding relationships are strong enough. Although there are some differences between MIC and NNP where these values are less than 0.5, which means that the NNP would overestimate the relationship which is weak indicated by the MIC, it does not affect the main point because these trends are consistent. In other words, the paired genes have high MIC would have high NNP, and vice versa.

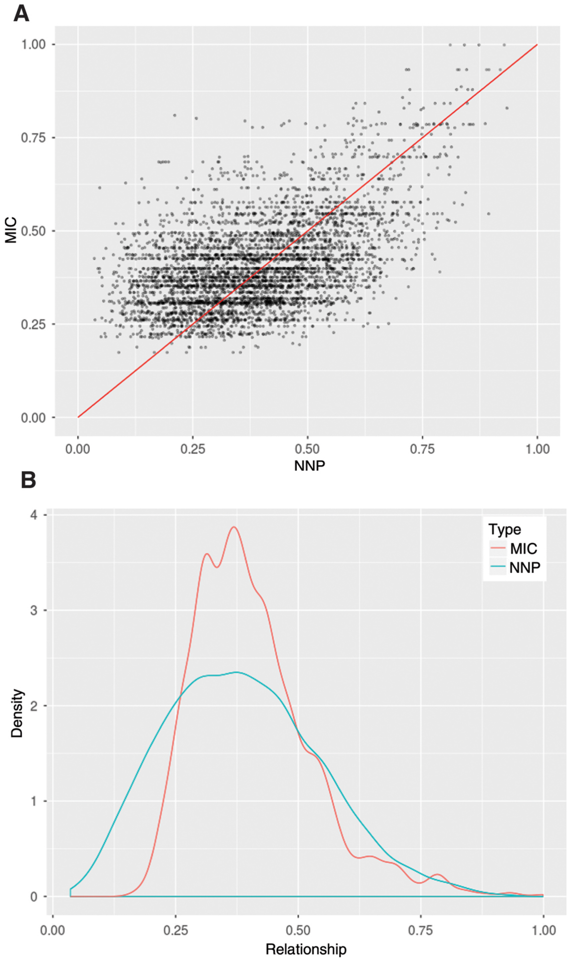

Figure 4: Synchronization of NNP and MIC. Panel A shows how the MIC change with NNP. The red line is angle bisector. It indicated that they change synchronously. Panel B shows the density of MIC and NNP when measuring the relationships between all genes expression levels and time.

From the simulations and real dataset analysis, NNP has approximate performances with MIC, so we conclude that MIC could nearly come down to one statistical prediction, NNP.

5 The Generalized Form

For the methods based on correlation, such as WGCNA, we have figure out the equivalence to least squares estimation under linear hypothesis based on mathematical proof. Hence, we conclude that these methods based on correlation can come down to the generalized statistical prediction method.

For MIC, which based on information-theoretic technique, we assume that there is a version of prediction method whose coefficient of determination exactly equals to MIC. Although this assumption needs further research to validate, we have found an approximate prediction method NNP, which has a good approximation to MIC.

As a result, we obtain the generalized form for measuring the relationship among genes. The generalized approach is superior to the methods based on correlation and information-theoretic technique. We will discuss the superiorities of this generalized approach as follows.

As we can see, either MIC or correlation can only compute the relationships between paired genes. It is urgent need to compute the relationships for more than two genes. Here we have extended the generalized form to measure the relationships for multi-genes. Suppose there are multi-genes, $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$, where each component represents a gene, and another particular gene Y. With the expression data of these genes, we can measure the relationship between gene Y and genes $\boldsymbol{X}$ through our statistical prediction method. Firstly, use the conditional expectation $E(Y \mid \boldsymbol{X})$ to predict gene Y's expression, then use the coefficient of determination to measure the relationship between gene Y and genes $\boldsymbol{X}$. If p is very large, it becomes to be a high dimension problem. For the method NNP, we need to find a fairly large neighborhood of observations and average them, which will be difficult in high dimensions. This phenomenon is commonly referred to as the curse of dimensionality. Besides, there are other concerns in high-dimensional spaces, such as high variance and overfitting. Fortunately, some methods that overcome these problems have developed, such as highly regularized approaches [6].

6 Multi-Genes Relationships

In this section, we will demonstrate how to measure the relationship between a particular gene with multi-genes. For simplicity, here we just discuss the situation of $p < N$ and use the most common linear regression form of $E(Y \mid \boldsymbol{X})$,

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j \tag{7}$$

To estimate $f(X)$, the most popular estimation method is least squares, in which we pick up the coefficients $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ to minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2 \tag{8}$$

There might be some components of X which contribute little to Y, that means these corresponding coefficients are nearly zero. Therefore, it is necessary to estimate $E(Y \mid \boldsymbol{X})$ by using the subset of $\boldsymbol{X}$ that exhibit the strongest effects. There are a number of different strategies for choosing the subset, such as best-subset selection, forward-stepwise and backward-stepwise selection [6].

Best subset regression finds for each $k \in \{0, 1, 2, ..., p\}$, find the subset size that gives smallest residual sum of squares Equation (8). We use this approach to explore the yeast gene expression dataset [7].

In the yeast gene expression dataset, there are 4381 genes. To illustrate the relationship between a particular gene and multi-genes, we choose one from 4381 genes as the particular gene, termed as Y and 10 from 4381 as the multi-genes, termed as $X_1, X_2, ..., X_{10}$. To begin with, we can plot all possible subset models' residual sum-of-squares for this dataset (see Figure 5). The lower boundary represents the models that are eligible for selection by the best-subsets approach.

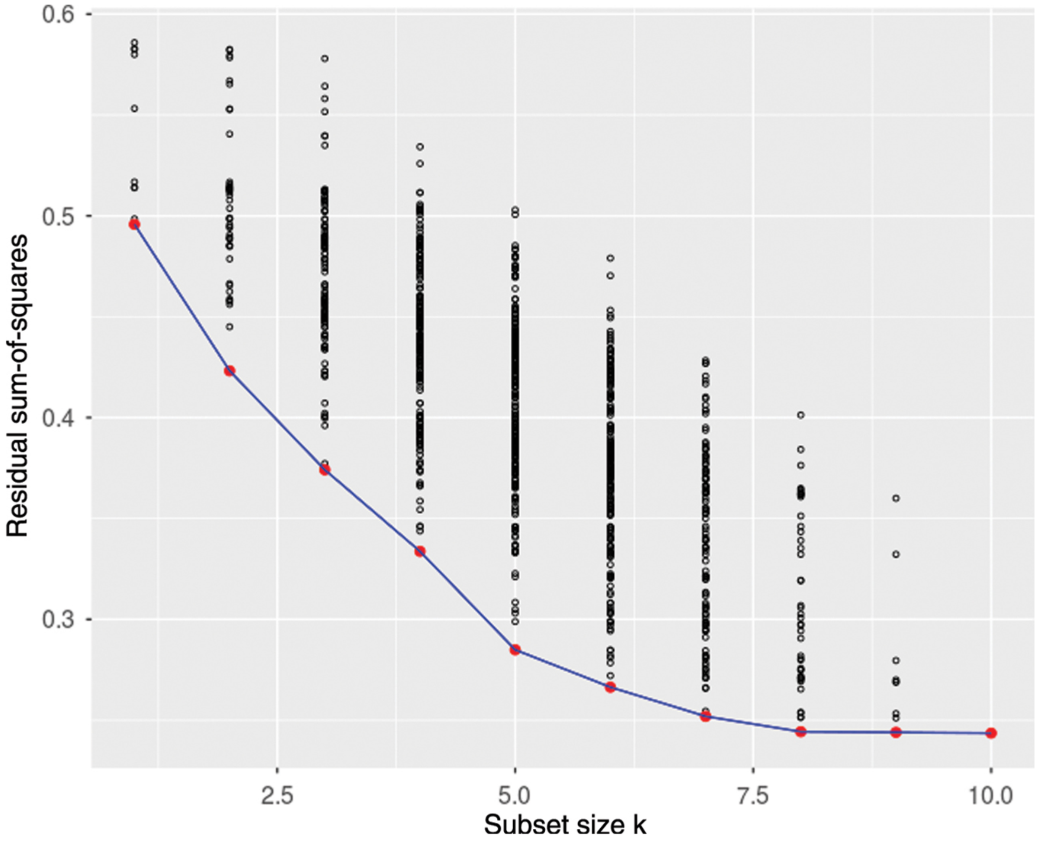

Figure 5: All possible subset models for the yeast gene expression dataset. At each subset size is shown the residual sum-of-squares for each model of that size.

To measure the relationship between particular gene and multi-genes, it is necessary to select an approximate model to capture their relationship. However, the best-subset curve is necessarily decreasing, so it cannot used to select the subset size k [6]. Fortunately, there are a number of criteria for choosing the appropriate k, such as Mallows's C_p which is used to assess the fit of a regression model that has been estimated using ordinary least squares [8]. We adopt the C_p statistic to pick up the best subset size and figure out which gene should be put into the regression model. A small value of C_p means that the model is relatively precise, but at the same time need to let the (p, C_p) be close to the line $C_p = p$ as much as possible (see the left panel in Figure 6). Thus, we select the subset model which presented by the red point in Figure 6. Finally, we obtain the subset $\{X_5, X_9, X_{10}\}$, and use the variables in this subset to measure the relationship between gene Y and genes $\boldsymbol{X}$.

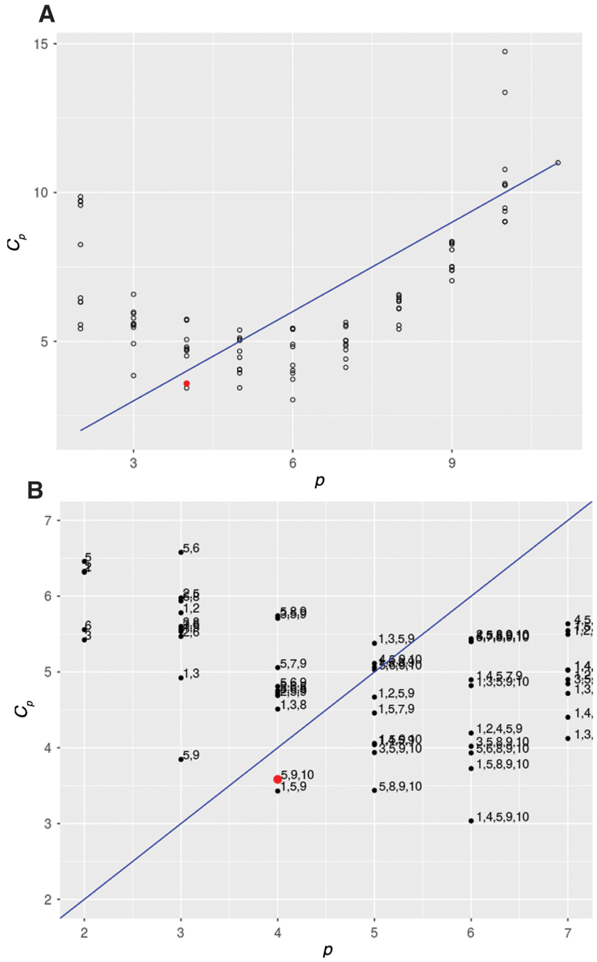

Figure 6: C_p plot. Plot of C_p versus p. In left panel, the dots are the C_p for the subset models for each p. The curve is the line $C_p = p$. The red dot means the subset size we choose. In right panel, the text near one dot represents the corresponding subset variables. It's worth noting that subset size p consider the intercept term, so the number of index in the text equals $p - 1$.

Then we build the regression model using these genes in the subset by least squares estimation. And we get the coefficients of determination in Table 1.

Table 1: The relationships measured by R^2 and MIC.

Formula	R^2	MIC
$Y \sim X_5$	0.1178	0.3070
$Y \sim X_9$	0.0057	0.3676
$Y \sim X_{10}$	0.0049	0.3725
$Y \sim X_5 + X_9 + X_{10}$	0.3561	None

The relationships between Y and $X_i, i = \{5, 9, 10\}$ measured by R^2 and MIC.

From the data in the table, we can see that although single gene might present low relationship with a particular gene, the relationship would be strengthened by integrating other single genes.

As a result, we can use the prediction method to measure the relationship between particular gene and the multi-genes. It is impossible for MIC to measure this kind of relationship. While this application is important, it is helpful when we study the dependence of a particular gene on multi-genes.

7 Conclusion

Our approach based on statistical prediction provides a general framework to identify the relationships among genes. On the basis of simulations and real dataset analysis, the method NNP derived from this generalized approach is useful for identifying the relationships among genes, having fairly similar performance with MIC. We also have extended the traditional relationships between paired genes (or variables) to multi-genes (or variables) by using this generalized approach.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

[1] D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics. 2000;16:707–26.
[2] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
[3] Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci. 2000;97:12182–6.
[4] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011;334:1518–24.
[5] Rau CD, Wisniewski N, Orozco LD, Bennett B, Weiss JN, Lusis AJ. Maximal information component analysis: a novel non-linear network analysis method. Front Genet. 2013;4:28.
[6] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Springer series in statistics. Berlin: Springer, 2001.
[7] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998;9:3273–97.
[8] Mallows CL. Some comments on Cp. Technometrics. 1973;15:661–75.