Supplementary Materials for "PCM: A Pairwise Correlation Mining Package for Biological Network Inference"

Zenyou He, Feiyang Gu, Xiaoqing Liu, Qiong Duan, Bo Tian, and Can Zhao

School of Software, Dalian University of Technology, China
zyhe@dlut.edu.cn, gufeiyang1000@gmail.com, eileenwelldone@gmail.com

1 Supplementary Methods

1.1 Marginal Correlation Measures

In the current version of the PCM package, 9 marginal correlation measures for continuous variables and 26 marginal correlation measures for binary variables are provided. The names, calculation functions and other supplementary details are presented in Supplementary Table 1 and Table 2, respectively.

Table 1: The correlation measures for continuous variables implemented in the PCM package. More details and properties about these measures can be found in [1–3]. In this table, $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a set of joint observations from two univariate continuous variables X and Y, $\overline{x}$ and $\overline{y}$ denote the mean value of X and Y, respectively.

ID	Measure	Definition
1	Pearson	$\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$
2	Cosine	$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}}$
3	Jaccard	$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}+\sqrt{\sum_{i=1}^{n} {y_i}^2}-\sum_{i=1}^{n} x_i y_i}$
4	Overlap	$\frac{\sum_{i=1}^{n} x_i y_i}{min(\sqrt{\sum_{i=1}^{n} {x_i}^2}, \sqrt{\sum_{i=1}^{n} {y_i}^2})}$
5	Dice	$\frac{2\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}+\sqrt{\sum_{i=1}^{n} {y_i}^2}}$
6	Spearman	$1 - \frac{6\sum_{i=1}^{n}(r_i-s_i)^2}{n(n-1)(n+1)}$
7	Dot Product	$\sum_{i=1}^{n} x_i y_i$
8	Kendall Rank	$\frac{n_c-n_d}{\sqrt{(n_0-n_1)(n_0-n_2)}}$
9	Hoeffding's D measure	$\frac{(n-2)(n-3)D_1+D_2-2(n-2)D_3}{n(n-1)(n-2)(n-3)}$

Spearman. The Spearman's correlation coefficient is defined as the Pearson's correlation coefficient between two variables that have been transformed into ranks, where r_i and s_i are the ranks of x_i and y_i, respectively.

Kendall Rank. The variables appeared in the Kendall's Rank correlation coefficient are defined as:

$n_c =$ The number of concordant pairs, where two pairs are concordant if $(\mathrm{x_i} - \mathrm{x_j})(\mathrm{y_i} > \mathrm{y_j}) > 0$,

$n_d =$ The number of discordant pairs, where two pairs are disconcordant if $(\mathrm{x_i} - \mathrm{x_j})(\mathrm{y_i} > \mathrm{y_j}) < 0$,

$n_0 = n(n-1)/2$,

$n_1 = \sum_i t_i(t_i - 1)/2$,

$n_2 = \sum_j u_j(u_j - 1)/2$,

$t_i =$ The number of tied values in the $\mathrm{i^{th}}$ group of ties for X,

$u_j =$ The number of tied values in the $\mathrm{j^{th}}$ group of ties for Y.

Hoeffding's D measure. In Hoeffding's D measure, r_i and s_i are the ranks of x_i and y_i, respectively. And $q_i = \sum_{j=1}^{n} \phi(x_j, x_i)\phi(y_j, y_i)$, where $\phi(a, b) = 1$ if $a < b$ and $\phi(a, b) = 0$ otherwise.

D_1, D_2 and D_3 are defined as:

$$
\begin{aligned}
D_1 &= \sum\nolimits_{i=1}^{n} q_i(q_i - 1), \\
D_2 &= \sum\nolimits_{i=1}^{n} (r_i - 1)(r_i - 2)(s_j - 1)(s_j - 2), \\
D_3 &= \sum\nolimits_{i=1}^{n} (r_i - 2)(s_i - 2)q_i.
\end{aligned}
$$

Table 2: The correlation measures for binary variables provided in the PCM package. Further details and discussions on these measures can be found in [4, 5]. To simplify the notations, we use X (Y) to denote the event that $X = 1$ ($Y = 1$) and $\overline{X}$ ($\overline{Y}$) to denote the event that $X = 0$ ($\overline{Y} = 0$), respectively. $P(\cdot)$ represents the probability that one event will occur and N is the number of samples in the data set.

ID	Measure	Definition
1	Odds ratio	$\frac{P(X,Y)P(\overline{X},\overline{Y})}{P(X,\overline{Y})P(\overline{X},Y)}$
2	Cosine	$\frac{P(X,Y)}{\sqrt{P(X)P(Y)}}$
3	Support	$P(X,Y)$
4	Yule's Q	$\frac{P(X,Y)P(\overline{X},\overline{Y})-P(X,\overline{Y})P(\overline{X},Y)}{P(X,Y)P(\overline{X},\overline{Y})+P(X,\overline{Y})P(\overline{X},Y)}$
5	Yule's Y	$\frac{\sqrt{P(X,Y)P(\overline{X},\overline{Y})}-\sqrt{P(X,\overline{Y})P(\overline{X},Y)}}{\sqrt{P(X,Y)P(\overline{X},\overline{Y})}+\sqrt{P(X,\overline{Y})P(\overline{X},Y)}}$
6	Kappa	$\frac{P(X,Y)+P(\overline{X},\overline{Y})-P(X)P(Y)-P(\overline{X})P(\overline{Y})}{1-P(X)P(Y)-P(\overline{X})P(\overline{Y})}$
7	J-Measure	$max\{P(X,Y)log(\frac{P(Y\|X)}{P(Y)}) + P(X,\overline{Y})log(\frac{P(\overline{Y}\|X)}{P(\overline{Y})}),$ $P(X,Y)log(\frac{P(X\|Y)}{P(X)}) + P(\overline{X},Y)log(\frac{P(\overline{X}\|Y)}{P(\overline{X})})\}$
8	Gini index	$max\{P(X)[P(Y\|X)^2 + P(\overline{Y}\|X)^2] + P(\overline{X})[P(Y\|\overline{X})^2 + P(\overline{Y}\|\overline{X})^2]$ $- P(Y)^2 - P(\overline{Y}^2),$ $P(Y)[P(X\|Y)^2 + P(\overline{X}\|Y)^2] + P(\overline{Y})[P(X\|\overline{Y})^2 + P(\overline{X}\|\overline{Y})^2]$ $- P(X)^2 - P(\overline{X}^2)\}$
9	Confidence	$max\{P(Y\|X), P(X\|Y)\}$
10	Laplace	$max\{\frac{N\times P(X,Y)+1}{N\times P(X)+2}, \frac{N\times P(X,Y)+1}{N\times P(Y)+2}\}$
11	Conviction	$max\{\frac{P(X)P(\overline{Y})}{P(X,\overline{Y})}, \frac{P(Y)P(\overline{X})}{P(Y,\overline{X})}\}$
12	Interest	$\frac{P(X,Y)}{P(X)P(Y)}$
13	Piatetsky-Shapiro's	$P(X,Y) - P(X)P(Y)$
14	Certainty factor	$max\{\frac{P(Y\|X)-P(Y)}{1-P(Y)}, \frac{P(X\|Y)-P(X)}{P(X)}\}$
15	Added value	$max\{P(Y\|X) - P(Y), P(X\|Y) - P(X)\}$
16	Collective strength	$\frac{P(X,Y)+P(\overline{X},\overline{Y})}{P(X)P(Y)+P(\overline{X})P(\overline{Y})} \times \frac{1-P(X)P(Y)-P(\overline{X})P(\overline{Y})}{1-P(X,Y)-P(\overline{X})}$
17	Jaccard	$\frac{P(X,Y)}{P(X)+P(Y)-P(X,Y)}$
18	Klosgen	$\sqrt{P(X,Y)}max\{P(Y\|X) - P(Y), P(X\|Y) - P(X)\}$
19	ϕ-coefficient	$\frac{P(X,Y)-P(X)P(Y)}{\sqrt{P(X)P(Y)(1-P(X))(1-P(Y))}}$
20	Probability Ratio	$log\frac{P(X,Y)}{P(X)P(Y)}$
21	BCPNN	$log(\frac{P(X,Y)+cc}{P(X)P(Y)+cc})$
22	SCWCC	$N\frac{(P(X,Y)-P(X)P(Y))^2}{P(X)P(Y)+cc}$
23	Two-way Support	$P(X,Y)log(\frac{P(X,Y)}{P(X)P(Y)})$
24	SCWS	$N \times P(X,Y)\frac{(P(X,Y)-P(X)P(Y))^2}{P(X)P(Y)}$
25	Simplified χ^2-statistic	$N\frac{(P(X,Y)-P(X)P(Y))^2}{P(X)P(Y)}$
26	Likelihood Ratio	$N[P(X,Y)log(\frac{P(X,Y)}{P(X)P(Y)}) + (1 - P(X,Y))log(\frac{1-P(X,Y)}{1-P(X)P(Y)})]$

Odds ratio. It is defined as the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. Here one group corresponds to $Y = 1$ and another group corresponds to $Y = 0$.

Cosine. The cosine of two variables can be considered as the normalized inner product, which ranges from -1 to +1.

Support. It is originally proposed for association rule mining in the data mining society, which

has been widely used in different applications.

Yule's Q and *Yule's Y*. These two measures are normalized variants of the odds ratio, whose values fall into the interval [-1, 1].

Kappa. This measure captures the degree of consistency between two binary variables. If these two variables are highly correlated with each other, then the values for $P(X, Y)$ and $P(\overline{X}, \overline{Y})$ will be large, which in turn, leads to a higher correlation value.

J-measure and *Gini*. The entropy is the key concept in information theory, which is indeed related to the variance of a probability distribution. Intuitively, the entropy of a uniform distribution is larger than that of a skewed distribution. The mutual information evaluates the dependencies among variables by the amount of reduction in the entropy of one variable when the value of another variable is given. The measures such as J-Measure and Gini index are defined according to the same principle.

Confidence, *Laplace* and *Conviction*. The confidence measure is originally used in association rule mining for measuring the accuracy of a given rule. The Laplace function and the conviction function are two variants of the confidence measure.

Interest, *Piatetsky-Shapiro's*, *Certainty factor*, *Collective strength* and *Added value*. The Interest measure is widely used in data mining for measuring the deviation from statistical independence. Piatetsky-Shapiro's measure, Certainty factor, Collective strength and Added value are four variants that extend the Interest measure from different angles.

Jaccard. The Jaccard measure has been used extensively in information retrieval to measure the similarity between two documents.

ϕ-coefficient. This measure is the variant of Pearson's correlation coefficient for binary variables, which is closely related to the χ^2 statistic since $\phi^2 = \chi^2/N$.

BCPNN and *SCWCC*. In these two measures, *cc* is a user-specified parameter, which is usually assigned a larger value if the data set is noisy and a smaller value otherwise.

In addition, the marginal mutual information between two discrete variables is also implemented in PCM as a correlation measure. Formally, the mutual information between two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log(\frac{p(x,y)}{p(x)p(y)}).$$

1.2 Clustering-based Approximate Correlation Mining

To conduct an exhaustive pairwise correlation mining, the time complexity is at least $O(n^2)$, where n is the number of variables. To handle the data sets with large number of variables, an alternative approach is to generate an approximate set of correlated pairs so as to reduce the time complexity.

In the PCM package, we implement a clustering-based method for this purpose. This method has two steps: clustering and correlation mining. In the first step, it uses clustering algorithms to partition the variables into different groups. For continuous variables, the clustering algorithm employed is the well-known k-means method, which requires the number of clusters as an input parameter. For binary variables, the CLOPE algorithm [6] is utilized, which also has an user-specified parameter called repulsion for controlling the level of intra-cluster similarity. It is expected that highly correlated variables will be put into the same cluster so that it is sufficient to perform correlation mining within each cluster. Therefore, in the second step, the correlation coefficients of all candidate pairs within each group are calculated. And the mining results across all clusters are gathered to generate the final set of correlated pairs.

Suppose the number of generated clusters is k (all clusters has the same size) and the clustering algorithm has a linear time complexity, then this method will reduce the time complexity of pairwise correlation mining from $O(n^2)$ to $O(n^2/k)$. Apparently, such performance gain in running efficiency is at the cost of missing some really correlated pairs of variables if these variables are not assigned into the same cluster. For instance, if we handle the example data set used in the main paper when the number of cluster is set to be 2 and the threshold for the Pearson's correlation coefficient is 0.5, 9 of the 49 correlated pairs that are above the threshold will be missed.

1.3 Algorithms for Conditional Correlation Mining

In this section, two algorithms that consider the conditional correlation between two variables are introduced briefly [7, 8].

1.3.1 CMI2NI

The basic procedure of CMI2NI [7] is described in Algorithm 1. This algorithm takes a matrix and a user-specified threshold as the input and outputs a set of variable pairs (i.e., a network with a set of detected edges), which are correlated even on condition the existence of other variables.

In this algorithm, the correlation measure CMI2 is defined based on the conditional mutual information. To make the computation fast and exact, the authors further assume a Gaussian distribution on the target data. This makes it possible to calculate the conditional mutual information with a concise formula involving the covariance matrix of the data set.

To handle data sets with binary variables, we also provide an implementation that follows the same procedure. In this implementation named CMI2NIB, we can perform the conditional pairwise correlation mining from the binary data set.

Algorithm 1 CMI2NI

Input:
A matrix of continuous variables A,
A parameter for dependence threshold θ.
Output:
The set of correlated pairs of variables G.

Step-1. Initialization. Let G_0 be the set of all candidate variable pairs. Set $L = -1$.
Step-2. $L = L + 1$; For each pair $(i, j) \in G_0$, let T be the number of variables that are correlated with both i and j, i.e. $T = |\{h|(i, h) \in G_0, (j, h) \in G_0\}|$.
Step-3. Set $G = G_0$. If $T < L$, stop. If $T \geq L$, select out L variables from these T variables. For each set K from the C_T^L selections, compute the Lth-order conditional mutual information $\mathrm{CMI2}(i, j|K)$. Let $\mathrm{CMI2}_{max}(i, j|K)$ be maximal value among all C_T^L calculated values. If $\mathrm{CMI2}_{max}(i, j|K) < \theta$, set $G = G - (i, j)$.
Step-4. If $G = G_0$, stop; If $G \neq G_0$, set $G_0 = G$ and return to Step-2.

1.4 LOPC

The LOPC [8] algorithm is described in Algorithm 2, whose procedure is similar to that of CMI2NI. In this method, the zero-th, first and second order partial correlation coefficients for a candidate variable pair are calculated.

Algorithm 2 LOPC

Input:
A matrix of continuous variables A,
Output:
The set of correlated pairs of variables G.

Let V be the set of all variables
Zero-th order partial correlation:
for each pair (i, j) **do**
Calculate the zero-th order partial correlation coefficient r_{ij};
Construct the test statistic for r_{ij} and calculate the corresponding p-value $p(r_{ij})$;
end for
Compute the multiple testing adjusted p-value for the zero-th order partial correlation coefficient $\overline{p}(r_{ij})$ across all pairs.

First order partial correlation:
for each pair (i, j) **do**
Calculate the first order partial correlation coefficients $r_{ij.k}$ for all possible $k \in V/\{i, j\}$;
Select the maximum in terms of absolute value as $\widehat{r}_{ij.k}$;
Construct test statistics for $\widehat{r}_{ij.k}$ and compute the corresponding p-value $p(\widehat{r}_{ij.k})$;
end for
Compute the multiple test adjusted p-values for the first order partial correlation coefficient $\overline{p}(\widehat{r}_{ij.k})$ across all pairs.

Second order partial correlation:
for each pair (i, j) **do**
if $\max\{\overline{p}(r_{ij}), \overline{p}(\widehat{r}_{ij.k})\} < 0.05$ **then**
Calculate the second order partial correlation coefficients $r_{ij.kq}$ for all possible $k, q \in V/\{i, j\}$;
Select the maximum in terms of absolute value as $\widehat{r}_{ij.kq}$;
Construct test statistics for $\widehat{r}_{ij.kq}$ and compute the corresponding p-value $p(\widehat{r}_{ij.kq})$;
else
Let $\overline{p}(\widehat{r}_{ij.kq}) = 1$.
end if
end for
Compute the multiple test adjusted p-values for the second order partial correlation coefficient $\overline{p}(\widehat{r}_{ij.kq})$ across all pairs.

References

[1] Y. Zhao, Q. Zou, Y. Jiang, and G. Wang, "A graphic processing unit web server for computing correlation coefficients for gene expression data," *Journal of Computational and Theoretical Nanoscience*, vol. 12, no. 4, pp. 582–584, 2015.

[2] S. de Siqueira Santos, D. Y. Takahashi, A. Nakata, and A. Fujita, "A comparative study of statistical methods used to identify dependencies between gene expression signals," *Briefings in bioinformatics*, p. bbt051, 2013.

[3] R. Deshpande, B. VanderSluis, and C. L. Myers, "Comparison of profile similarity measures for genetic interaction networks," *PloS one*, vol. 8, no. 7, p. e68664, 2013.

[4] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004.

[5] L. Duan, W. N. Street, Y. Liu, S. Xu, and B. Wu, "Selecting the right correlation measure for binary data," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 9, no. 2, p. 13, 2014.

[6] Y. Yang, X. Guan, and J. You, "CLOPE: a fast and effective clustering algorithm for transactional data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 682–687.

[7] X. Zhang, J. Zhao, J.-K. Hao, X.-M. Zhao, and L. Chen, "Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks," *Nucleic acids research*, vol. 43, no. 5, pp. e31–e31, 2015.

[8] Y. Zuo, G. Yu, M. G. Tadesse, and H. W. Ressom, "Biological network inference using low order partial correlation," *Methods*, vol. 69, no. 3, pp. 266–273, 2014.