PPCOR Documentation

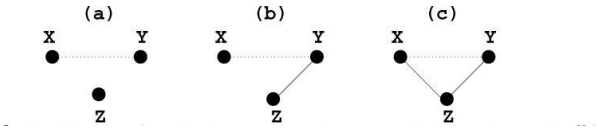
About the algorithm

In this package, a general matrix formula is derived for the semi-partial correlation calculation. It provides a means for fast computing partial and semi-partial correlation as well as the level of statistical significance. The package includes the calculation of statistics and p-values of each correlation coefficient for both higher-order partial and semi-partial correlations. It also provides users with nonparametric partial and semi-partial correlation coefficients based on Kendall's and Spearman's rank correlations.

The ppcor package provides users with 4 functions: pcor(), pcor.test(), spcor(), and spcor.test(). The function pcor() and spcor() calculates the partial/semi-partial correlations of all pairs of two random variables of a matrix or data frame and provides the matrices of statistics and p-values of each pairwise partial/semi-partial correlation. In order to compute the pairwise partial/semi-partial correlation coefficient of a pair of two random variables given one or more random variables, pcor.test()/spcor.test() can be also used instead.

What is partial and semi-partial correlation

The partial and semi-partial (also known as part) correlations are used to express the specific portion of variance explained by eliminating the effect of other variables when assessing the correlation between two variables. The partial correlation can be explained as the association between two random variables after eliminating the effect of all other random variables, while the semi-partial correlation eliminates the effect of a fraction of other random variables. The rationale for the partial and semi-partial correlations is to estimate a direct relationship or association between two random variables. Suppose there are three random variables, X, Y and Z, and we are interested in the relationship between X and Y. In case a, Z is correlated with none of X and Y. In case b, only Y is correlated with Z. In case c, Z is correlated with both X and Y. Because Z is independent of both X and Y in case a, the correlation, partial correlation, and semi-partial correlation all should theoretically have the identical value. When only Y is correlated with Z in case b, the partial correlation is exactly the same as the semi-partial correlation but is different from the correlation. In case c, all three correlations are different from each other.



In more familiar terms, partial correlation is a measure of the strength and direction of association between two variables while controlling for the effect of one or more other variables (also known as covariates or control variables). Hence, it is used to find out the strength of the unique portion of association. Semi-partial correlation is the correlation of two variables with variation from a third or more other variables removed only from the second variable. Partial correlation holds constant one variable when computing the relations between two others. Suppose we want to know the correlation between X and Y holding Z constant for both X and Y. That would be the partial correlation between X

and Y controlling for Z. Semi-partial correlation holds Z constant for either X or Y, but not both, so if we wanted to control X for Z, we could compute the semi-partial correlation between X and Y holding constant for X.

The three correlation methods

1. A Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related and is the default method. A few assumptions should be made before using Pearson's correlation. First, the variables should be continuous. If one or both of the variables are ordinal in measurement, then a Spearman rank correlation should be conducted. Next, the variables need to be linearly related. If they are not, the data could be transformed or a non-parametric approach such as the Spearman's or Kendall's rank correlation tests could be used. Also, homoscedasticity should exist in the dataset. If the variance between the two variables is not constant, then r will not provide a good measure of association. And when the variables consist of high levels of skewness or contain significant outliers, it is recommended to use Spearman's correlation.
2. Spearman's correlation is a non-parametric test that measures the degree of association between two variables based on using a monotonic function. The Spearman approach does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on an ordinal scale. The first assumption when using Spearman's approach is that one or both of the variables are ordinal in measurement. The approach is appropriate to use if both variables are continuous but are heavily skewed or contain sizable outliers. A linear relationship is not required, the only requirement is that one variable is monotonically related to the other variable.
3. Kendall's correlation is also a non-parametric approach that assesses statistical associations based on the ranks of the data. Kendall's correlation is less sensitive to error and the p-values are more accurate with smaller sample sizes.

Findings about the parameter:

1. When a pcor.test is performed on the same three genes with the three different correlation methods, Pearson's method will generate the highest correlation coefficient while Kendall's method will generate the lowest correlation coefficient. (not sure if it applies for all combinations of three genes).
2. When a spcor.test is performed on the same three genes with the three different correlation methods, Spearman's method will generate the highest correlation coefficient while Kendall's method will generate the lowest correlation coefficient. (not sure if it applies to all combinations of three genes).
3. When a pcor.test and a spcor.test is performed on the same three genes over the three correlation methods, the correlation coefficients would all be slightly smaller in the spcor.test. (not sure if it applies to all combinations of three genes).
4. When the pcor.test is performed on three genes, the same correlation coefficient could be derived by running two linear regressions over the genes and calculating the correlation between the residuals, as showed in the code.
5. When the pcor() function and the spcor() function is applied to the entire dataset, correlation coefficients generated by the pcor() function will generate NaN values and correlation

coefficients generated by the spcor() function will generate NaN and infinity values. These may occur when the column sum of the corresponding gene is equal to zero.

Things to note:

1. The datasets used for the ppcor() function must not contain any missing values.

References:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4681537/