Communications for Statistical Applications and Methods
2022, Vol. 29, No. 3, 397–402

https://doi.org/10.29220/CSAM.2022.29.3.397
Print ISSN 2287-7843 / Online ISSN 2383-4757

P-value calculation methods for semi-partial correlation coefficients

Seongho Kim[1,a]

[a]Biostatistics and Bioinformatics Core, Karmanos Cancer Institute, Wayne State University, USA

Abstract

The mathematical expression of the p-value calculation for the semi-partial correlation coefficient differs between Kim (2015) and Cohen *et al.* (2003). These two expressions were compared and the advantages of Kim (2015)'s approach over Cohen *et al.* (2003) were discussed.

Keywords: correlation, partial correlation, part correlation, ppcor, semi-partial correlation

1. Introduction

Suppose the random vector $X = (x_1, x_2, \ldots, x_i, \ldots, x_n)'$ where $|X| = n$. The variance of a random variable x_i and the covariance between two random variables x_i and x_j are denoted as v_i and c_{ij}, respectively. The correlation between two random variables x_i and x_j is denoted by $r_{ij} = c_{ij}/(\sqrt{v_i}\sqrt{v_j})$. Then the partial correlation of x_i and x_j given x_k and the semi-partial correlation of x_i with x_j given x_k are defined, respectively, by

$$r_{ij|k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{1 - r_{ik}^2}\sqrt{1 - r_{jk}^2}}, \quad r_{i(j|k)} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{1 - r_{jk}^2}}. \tag{1.1}$$

The partial and semi-partial correlations can be also defined, using multiple correlations, by

$$r_{ij|k} = \frac{\sqrt{R_{i.jk}^2 - R_{i.k}^2}}{\sqrt{1 - R_{i.k}^2}}, \quad r_{i(j|k)} = \sqrt{R_{i.jk}^2 - R_{i.k}^2}. \tag{1.2}$$

where $R_{i.k}^2$ is the squared multiple correlation of x_i on x_k, which is equal to r_{ik}^2, and $R_{i.jk}^2$ is the squared multiple correlation of x_i on x_j and x_k. Note that the superscripts k and c will be used to denote expressions of Kim (2015) and Cohen *et al.* (2003).

Kim (2015) used the statistics $t_{ij|S}^k$ and $t_{i(j|S)}^k$, respectively, in order to calculate the p-values of the partial and semi-partial correlations of x_i and (with) x_j given x_S $(= x_{(i,j)})$, which are

$$t_{ij|S}^k = r_{ij|S}\sqrt{\frac{N - 2 - g}{1 - r_{ij|S}^2}}, \quad t_{i(j|S)}^k = r_{i(j|S)}\sqrt{\frac{N - 2 - g}{1 - r_{i(j|S)}^2}}, \tag{1.3}$$

This work is partially supported by NIH/NCI P30 CA022453 and NIH/NIGMS R21GM140352.

[1] Biostatistics and Bioinformatics Core, Karmanos Cancer Institute, Department of Oncology, School of Medicine, Wayne State University, 87 E. Canfield St., Detroit, MI 48201, USA. E-mail: kimse@karmanos.org

Published 31 May 2022 / journal homepage: http://csam.or.kr

where N is the sample size and g is the total number of given (or controlled) variables, and x_S (= $x_{(i,j)}$) is the random sub-vector of X after removing the random variables x_i and x_j and its size is $|S|$ (= $|X| - 2$). The corresponding p-values are then calculated by

$$p^k_{i,j|S} = 2\Phi_t\left(-\left|t^k_{i,j|S}\right|, N-2-g\right), \quad p^k_{i(j|S)} = 2\Phi_t\left(-\left|t^k_{i(j|S)}\right|, N-2-g\right), \tag{1.4}$$

where $\Phi_t(\cdot)$ is the cumulative density function of a Student's t distribution with the degree of freedom $N - 2 - g$.

We can readily observe that the statistic $t^k_{i(j|S)}$ can equal zero only when the semi-partial correlation $r_{i(j|S)}$ is zero in equation (1.3). This demonstrates that the statistic $t^k_{i(j|S)}$ in equation (1.3) is suitable for testing the deviation of the semi-partial correlation coefficient from zero. For the same reason, the statistic $t^k_{i,j|S}$ is also sufficient to test the null hypothesis $H_0 : r_{i,j|S} = 0$.

On the other hand, Cohen *et al.* (2003) suggested the single statistic for calculating the p-values of both partial and semi-partial correlations of x_i and (with) x_j given x_S (= $x_{(i,j)}$), which is

$$t^c_{i,j|S} = t^c_{i(j|S)} = r_{i(j|S)}\sqrt{\frac{N-2-g}{1-R^2_{i.jS}}}, \tag{1.5}$$

where $R^2_{i.jS}$ is the squared multiple correlation of x_i on x_j and x_S. The corresponding p-values are

$$p^c_{i,j|S} = p^c_{i(j|S)} = 2\Phi_t\left(-\left|t^c_{i(j|S)}\right|, N-2-g\right). \tag{1.6}$$

In equation (1.1), we can notice that the partial and semi-partial correlations have the identical numerator and differ only by their denominators. Indeed, $r_{i(j|k)} = r_{i,j|k} \cdot \sqrt{1-r^2_{ik}}$ and so $r_{i(j|k)} \leq r_{i,j|k}$ because of $1 - r^2_{ik} \leq 1$. This relationship confirms that if one of the correlations is zero, then the other correlation is also zero. Besides, the statistic $t^c_{i(j|S)}$ (or $t^c_{i,j|S}$) cannot equal zero unless the semi-partial correlation $r_{i(j|S)}$ is zero in equation (1.5). Thus, equation (1.5) is appropriate for the hypothesis test to decide whether both partial and semi-partial correlation coefficients are significantly different from zero, which is the rationale for the use of the same statistic for both partial and semi-partial correlations in Cohen *et al.* (2003).

We can also see that, using equation (1.2),

$$r_{i(j|S)}\sqrt{\frac{N-2-g}{1-R^2_{i.jS}}} = r_{i,j|S}\sqrt{\frac{N-2-g}{1-r^2_{i,j|S}}}, \tag{1.7}$$

which implies that $t^k_{i,j|S} = t^c_{i,j|S} = t^c_{i(j|S)}$ and so $p^k_{i,j|S} = p^c_{i,j|S} = p^c_{i(j|S)}$. That is, Cohen *et al.* (2003) uses Kim (2015)'s statistic $t^k_{i,j|S}$ to test whether both partial and semi-partial correlations are significantly different from zero.

However, the partial correlation is not a one-to-one correspondence to the semi-partial correlation. For example, when (r_{ij}, r_{ik}, r_{jk}) is either $(0.816, 0.8, 0.8)$ or $(0.892, 0.8, 0.8)$, the partial correlation $r_{i,j|k}$ is equal to 0.70 but the semi-partial correlation $r_{i(j|k)}$ is 0.56 and 0.42, respectively. This means that there can be two or more corresponding semi-partial correlation coefficients for a partial correlation coefficient, resulting that two different semi-partial correlation coefficients can share the identical statistical significance in case of Cohen *et al.* (2003)'s approach. Moreover, the monotone ordering is not

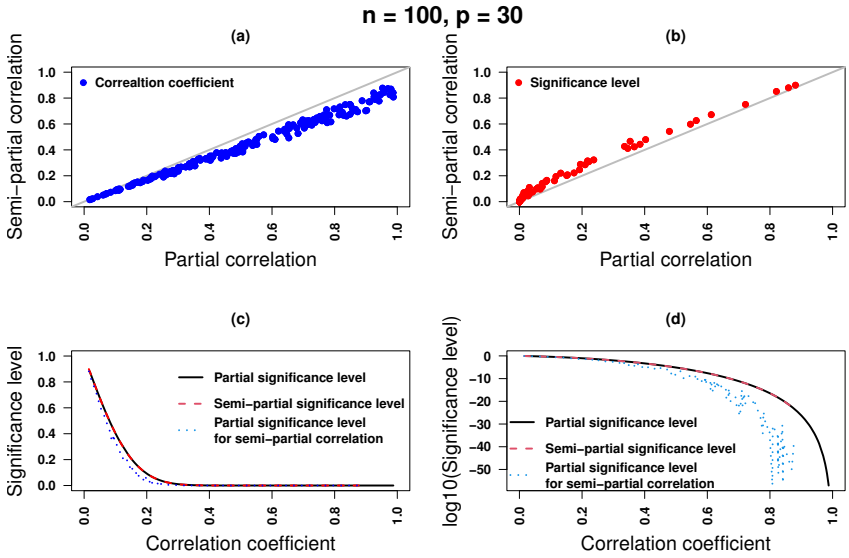

Figure 1: *Comparison between Kim (2015) and Cohen et al. (2003) when the number of variables is 30 (i.e., $p = 30$) and the number of samples is 100 (i.e., $n = 100$). The panels (a) and (b) are the scatter plots of the correlation coefficients and the significance levels (i.e., p-value), respectively, between the partial correlation (i.e., $r_{12|S}$) and the semi-partial correlation (i.e., $r_{1(2|S)}$), where $S = \{3, 4, 5, \ldots, 30\}$. In (b), the p-values were calculated using $p^k_{12|S}$ and $p^k_{1(2|S)}$ for the partial and semi-partial correlations, respectively. The panels (c) and (d) are the relationships between the correlation coefficients and the corresponding significance levels (i.e., p-value). In (d), the significance levels were log-transformed.*

preserved between the partial and semi-partial correlation coefficients. For instance, when (r_{ij}, r_{ik}, r_{jk}) is $(0.892, 0.8, 0.8)$ and $(0.512, 0.80, 0.28)$, the partial correlation $r_{ij|k}$ is equal to 0.7 and 0.5, respectively, but the semi-partial correlation $r_{i(j|k)}$ is 0.42 and 0.48, respectively. In fact, these properties cause the relationship between $p^c_{i(j|S)}$ and $r_{i(j|S)}$ to be non-monotonic, while $p^k_{i(j|S)}$ is monotonic with regard to $r_{i(j|S)}$. This difference between $p^c_{i(j|S)}$ and $p^k_{i(j|S)}$ is evident in the following simulation study.

A simulation study was performed to compare the p-value calculation methods between Kim (2015) and Cohen *et al.* (2003). The expressions $p^k_{i(j|S)}$ and $p^k_{ij|S}$ were used for the p-value calculations for the semi-partial correlation coefficients, which are corresponding to Kim (2015) and Cohen *et al.* (2003)'s approaches, respectively, because $p^c_{i(j|S)} = p^k_{ij|S}$. The number of variables and samples was set to 30 and 100, respectively. For the partial correlation coefficient between the first and the second variables, 200 values were selected between 0.01 and 0.99 with the equal-width binning. For each of 200 partial correlation coefficients, a data set was simulated such that the calculated partial correlation coefficient from the simulated data set was similar to the selected coefficient, resulting in 200 simulated data sets. For each simulated data set, the partial and semi-partial correlation coefficients (i.e., $r_{12|S}$ and $r_{1(2|S)}$) were calculated along with the corresponding p-values $p^k_{12|S}$ and $p^k_{1(2|S)}$, where $S = \{3, 4, 5, \ldots, 30\}$. The simulated outcomes are depicted in Figure 1, and the R codes, which were used to perform the simulation study, are available in Appendix.

Figure 1(a) shows that the partial correlation coefficients are always greater than or equal to the semi-partial correlation coefficients (i.e., $r_{12|S} \geq r_{1(2|S)}$) as expected. Consequently, the corresponding p-values for the partial correlation coefficients are always less than or equal to those for the semi-partial correlation coefficients (i.e., $p^k_{12|S} \leq p^k_{1(2|S)}$) as can be seen in Figure 1(b). In Figure 1(c), the

p-values (the blue dotted line) for the semi-partial correlation coefficients by Cohen *et al.* (2003) are less than or equal to those (the red dashed line) by Kim (2015). This happens because $r_{12|S} \geq r_{1(2|S)}$ and Cohen *et al.* (2003) uses $p^k_{12|S}$ for the corresponding p-values. Furthermore, Figure 1(d) confirms that the p-values ($p^k_{12|S}$, the blue dotted line) for the semi-partial correlation coefficients by Cohen *et al.* (2003) are not monotonic with regard to $r_{1(2|S)}$. On the other hand, the semi-partial correlation coefficients ($r_{1(2|S)}$) by Kim (2015) are monotonic to the p-values ($p^k_{1(2|S)}$, the red dashed line).

In conclusion, both statistics $t^k_{i(j|S)}$ and $t^c_{i(j|S)}$ are appropriate for testing the null hypothesis $H_0 : r_{i(j|S)} = 0$. However, as for comparisons between semi-partial correlations, Kim (2015)'s statistic $t^k_{i(j|S)}$ will be more suitable because it maintains the monotonicity with respect to $r_{i(j|S)}$.

Appendix

The following R codes were used to perform the simulation study and to generate Figure 1, which were written using R 4.1.0.

```
library(GeneNet)
library(ppcor)

# This is the function to perform a simulation study and generate a plot
# corresponding to a given configuration
#
# Input parameters
# p: the number of variables
# n: the number of samples
# pvmin: the minimum of the partial correlation coefficient;
#            it must be positive
# pvmax: the maximum of the partial correlation coefficient;
#              it must be positive
# pvnum: the number of cases of the partial correlation coefficient
#              between pvmin and pvmax
# nseed: a seed number

simul.ppcor <- function(p=30,n=100,pvmin=0.05,pvmax=0.95,pvnum=100,nseed=153){

  # set a seed number for random generation
  set.seed(nseed)
  # select the partial correlation coefficients to examine
  pvs = seq(pvmin,pvmax,length=pvnum)
  # output
  out0 = c()
  for(i in 1:length(pvs)){
    true.pcor = matrix(0,p,p)
    true.pcor[1,2] = pvs[i]
    true.pcor[2,1] = pvs[i]
    diag(true.pcor) = 1
    # generation of a simulated data set with the partial correlation
```

```
  # coefficient of the true.pcor
  m.sim = ggm.simulate.data(n,true.pcor)
  # calculation of the partial correlation
  epcor = pcor(m.sim)
  # calculation of the semi-partial correlation
  espcor = spcor(m.sim)
  # output
  out0 = rbind(out0,c(pvs[i],epcor$estimate[1,2],espcor$estimate[1,2]
              ,epcor$p.value[1,2],espcor$p.value[1,2]
              ,0,epcor$estimate[1,3],espcor$estimate[1,3]
              ,epcor$p.value[1,3],espcor$p.value[1,3]))
}

# select the output only with the positive correlation coefficients
out = out0[which(out0[,2]>=0 & out0[,3]>=0),]

# generate plots
par(mfrow=c(2,2))
plot(out[,2],out[,3],xlim=c(0,1),ylim=c(0,1),pch=19,col="blue",log=""
       ,xlab="Partial correlation",ylab="Semi-partial correlation"
       ,cex.lab=1.5,font=2,cex.axis=1.,las=2)
abline(0,1,lwd=2,col="grey")
points(out[,2],out[,3],pch=19,col="blue")
legend("topleft","Correaltion coefficient",pch=19,col="blue"
           ,text.font=2,bty="n")
title("(a)",font=2,cex=1.2)

plot(out[,4],out[,5],xlim=c(0,1),ylim=c(0,1),pch=19,col="red",log=""
       ,xlab="Partial correlation",ylab="Semi-partial correlation"
       ,cex.lab=1.5,font=2,cex.axis=1.,las=2)
abline(0,1,lwd=2,col="grey")
points(out[,4],out[,5],pch=19,col="red")
legend("topleft","Significance level",pch=19,col="red"
            ,text.font=2,bty="n")
title("(b)",font=2,cex=1.2)

vlwd = 2
out2 = out[order(out[,2]),]
out3 = out[order(out[,3]),]
plot(out2[,2],out2[,4],lwd=vlwd,lty=1,col="black",type="l"
       ,xlim=c(0,1),ylim=c(0,1),log=""
       ,xlab="Correlation coefficient"
       ,ylab="Significance level",cex.lab=1.5
       ,font=2,cex.axis=1.,las=2)
points(out3[,3],out3[,5],lwd=vlwd,type="l",col="red",lty=2)
points(out3[,3],out3[,4],lwd=vlwd,type="l",col="blue",lty=3)
```

```
  legend("topright",c("Partial significance level"
              ,"Semi-partial significance level"
              ,"Partial significance level \nfor semi-partial correlation")
              ,lty=1:3
              ,col=c(1,2,4),text.font=2,lwd=vlwd,bty="n")
  title("(c)",font=2,cex=1.2)

  plot(out2[,2],log10(out2[,4]),lwd=vlwd,lty=1,col="black",type="l"
         ,xlim=c(0,1)
         ,log=""
         ,xlab="Correlation coefficient"
         ,ylab="log10(Significance level)",cex.lab=1.5
         ,font=2,cex.axis=1.,las=2)
  points(out3[,3],log10(out3[,5]),lwd=vlwd,type="l",col=2
             ,lty=2,pch=17,cex=.5)
  points(out3[,3],log10(out3[,4]),lwd=vlwd,type="l",col=4
             ,lty=3,pch=19,cex=.5)
  legend("bottomleft",c("Partial significance level"
               ,"Semi-partial significance level"
               ,"Partial significance level \nfor semi-partial correlation")
               ,lty=c(1,2,3)
               ,pch=c(-1,-1,-1),col=c(1,2,4),text.font=2,lwd=vlwd,bty="n")
  title("(d)",font=2,cex=1.2)

  mtext(paste0("n = ",n,", p = ",p), side = 3, line = -1.5
             ,outer = TRUE,font=2,cex=1.5)

  # output
  out
}

# Perform a simulation and generate a plot
sout = simul.ppcor(p=30,n=100,pvmin=0.01,pvmax=0.99,pvnum=200,nseed=153)
```

References

Cohen J, Cohen P, West SG, and Aiken LS (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.)*, Lawrence Erlbaum Associates Publishers.

Kim S (2015). ppcor: An R Package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, **22**, 665–674.

Received December 1, 2021; Revised January 12, 2022; Accepted January 13, 2022