

# Partial correlation in linear models

Brice Ozenne

September 2, 2022

## 1 Summary

This document starts by presenting how to extract from a (univariate) linear regression model partial correlation coefficients. It also precise what type of "partial" (i.e. adjusted on which covariate) we get. When having multiple measurements of pairs of variables, various technics to estimate (partial) correlations are being compared.

## 2 Example

For illustration we will use the following packages:

```
library(LMMstar)
library(ggplot2)
library(lme4)
library(lmerTest)
library(Matrix)
library(data.table)
```

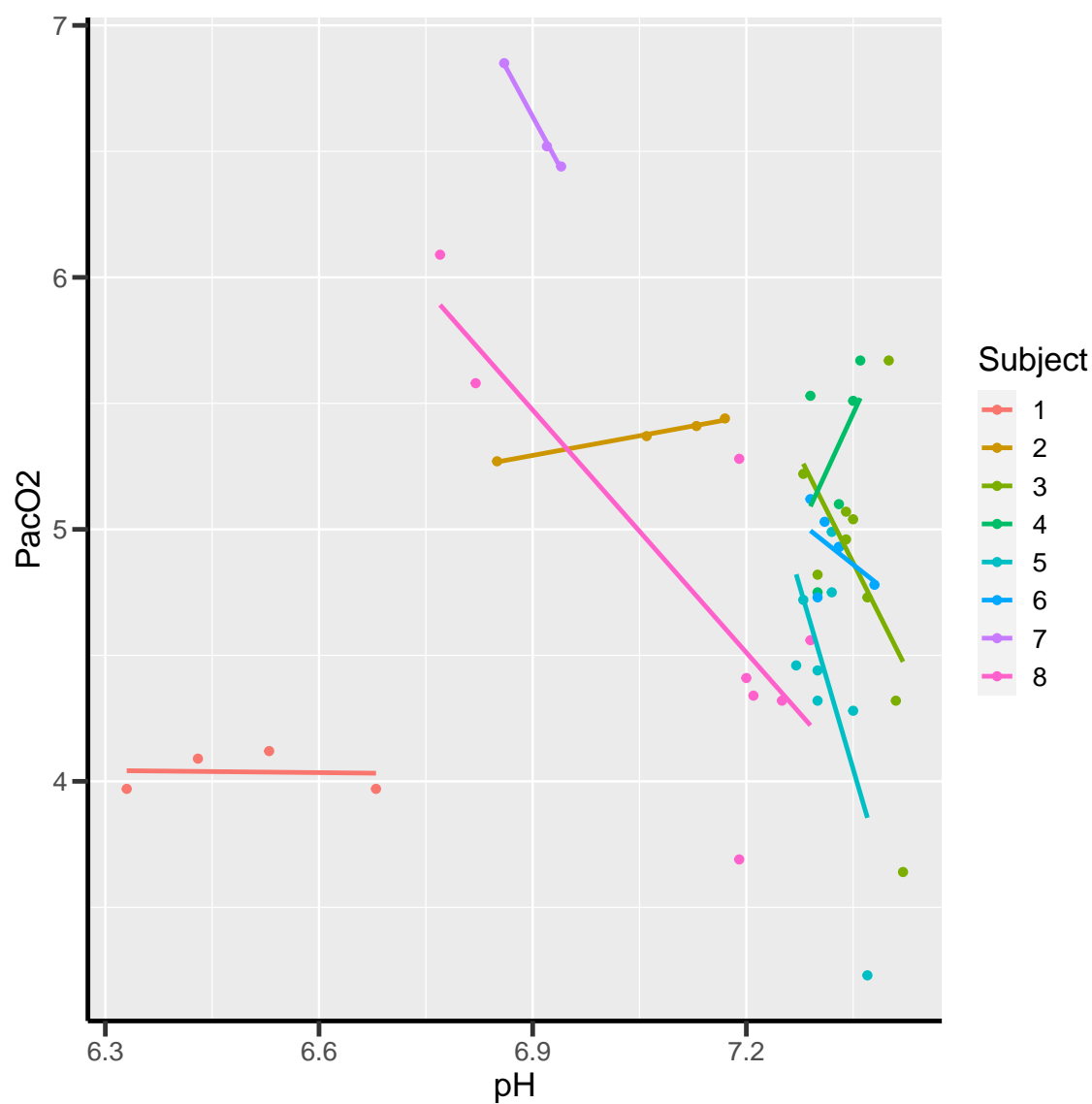
and dataset ([Bland and Altman, 1995](#)):

```
data("bland1995", package = "rmcorr")
bland1995$Subject <- as.factor(bland1995$Subject)
bland1995$time <- unlist(tapply(bland1995$Subject, bland1995$Subject,
  function(x){1:length(x)}))
head(bland1995)
```

|   | Subject | pH   | PacO2 | time |
|---|---------|------|-------|------|
| 1 | 1       | 6.68 | 3.97  | 1    |
| 2 | 1       | 6.53 | 4.12  | 2    |
| 3 | 1       | 6.43 | 4.09  | 3    |
| 4 | 1       | 6.33 | 3.97  | 4    |
| 5 | 2       | 6.85 | 5.27  | 1    |
| 6 | 2       | 7.06 | 5.37  | 2    |

The aim is to relate intramural pH and PaCO<sub>2</sub> using eight subjects:

```
gg <- ggplot(bland1995, aes(x = pH, y = PacO2,  
  group = Subject, color = Subject))  
gg <- gg + geom_point() + geom_smooth(method = "lm", se = FALSE)  
gg
```



### 3 Partial partial in multiple linear regression

Consider the linear model:

```
e.lmm <- lmm(pH ~ Subject + Pac02, data = bland1995)
eTable.lmm <- model.tables(e.lmm)
eTable.lmm
```

|             | estimate   | se         | df | lower      | upper       | p.value      |
|-------------|------------|------------|----|------------|-------------|--------------|
| (Intercept) | 6.9298543  | 0.12946898 | 38 | 6.6677580  | 7.19195056  | 0.000000e+00 |
| Subject2    | 0.7046113  | 0.07735488 | 38 | 0.5480145  | 0.86120804  | 4.277623e-11 |
| Subject3    | 0.9500127  | 0.06109545 | 38 | 0.8263314  | 1.07369394  | 0.000000e+00 |
| Subject4    | 0.9715577  | 0.07350906 | 38 | 0.8227464  | 1.12036905  | 8.881784e-16 |
| Subject5    | 0.8603817  | 0.05839543 | 38 | 0.7421663  | 0.97859708  | 0.000000e+00 |
| Subject6    | 0.9264284  | 0.06599450 | 38 | 0.7928295  | 1.06002730  | 0.000000e+00 |
| Subject7    | 0.6921056  | 0.10490935 | 38 | 0.4797277  | 0.90448342  | 8.670218e-08 |
| Subject8    | 0.7033361  | 0.06157141 | 38 | 0.5786913  | 0.82798087  | 7.460699e-14 |
| Pac02       | -0.1083230 | 0.02989281 | 38 | -0.1688379 | -0.04780822 | 8.471081e-04 |

We claim the partial correlation (adjusting pH and Pac02 for Subject) can be deduced from the Wald statistic and degrees of freedom:

$$\rho = \frac{\frac{\beta}{\sigma_{\beta}}}{\sqrt{\frac{\beta^2}{\sigma_{\beta}^2} + df}} = \frac{\beta}{\sqrt{\beta^2 + df * \sigma_{\beta}^2}} \quad (1)$$

```
Wald <- eTable.lmm["Pac02","statistic"]
Wald/sqrt(Wald^2+eTable.lmm["Pac02","df"])
```

```
[1] -0.5067697
```

The proof can be split in three steps:

1. the F-statistic testing the effect of each factor equals the Wald-statistic squared (divided by 1, the number of parameters)

```
Wald^2
```

```
[1] 13.13132
```

```
anova(e.lmm)
```

## Multivariate Wald test

|               | F-statistic     | df     | p.value |
|---------------|-----------------|--------|---------|
| mean: Subject | 48.247 (7,38.0) | <0.001 | ***     |
| : Pac02       | 13.131 (1,38.0) | <0.001 | ***     |

2. this F-statistic equals  $\frac{MSSR}{MSSE}$  where  $MSSR = SSR/1$  and  $MSSE = SSE/(n-p)$  with  $SSE$  and  $SSR$  being the explained and residual sum of squares. We can check that this extends to multiple regression using the usual anova table:

```
anova(lm(pH ~ Subject + Pac02, data = bland1995))
```

## Analysis of Variance Table

Response: pH

|           | Df | Sum Sq  | Mean Sq | F value | Pr(>F)        |
|-----------|----|---------|---------|---------|---------------|
| Subject   | 7  | 2.86484 | 0.40926 | 46.600  | < 2.2e-16 *** |
| Pac02     | 1  | 0.11532 | 0.11532 | 13.131  | 0.0008471 *** |
| Residuals | 38 | 0.33373 | 0.00878 |         |               |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

which is to be compared to<sup>1</sup>

```
sigma2 <- as.double(sigma(e.lmm))
beta <- eTable.lmm["Pac02","estimate"]
sigma_beta <- eTable.lmm["Pac02","se"]
c(MSSE = sigma2, MSSR = sigma2 * beta^2 /sigma_beta^2)
```

| MSSE        | MSSR        |
|-------------|-------------|
| 0.008782435 | 0.115324959 |

This result can be easily proved when considering a model with a single regressor:

$$Y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where we would have centered the outcome  $Y$ . Here we denote by  $X$  the design matrix,  $n$  the number of observations and  $p = 1$  the number of coefficients,  $H =$

<sup>1</sup> ⚠ Since **R** output type 1 anova only the last and second to last line are relevant. The first line (Subject) is for a model without Pac02 so it should be expected that the F-value does not match with the one of Subject in a model with Pac02.

$X(XX^\top)^{-1}X^\top$  the hat matrix and  $\hat{\beta} = (XX^\top)^{-1}X^\top Y$  the OLS estimator of the regression coefficients.

$$\begin{aligned}\text{Var}(Y) &= YY^\top = YHY^\top + Y(1-H)Y^\top \\ SST &= SSR + SSE \\ &= \hat{\beta}(XX^\top)\hat{\beta}^\top + Y(1-H)Y^\top \\ &= \sigma^2(\hat{\beta}\Sigma_{\hat{\beta}}^{-1}\hat{\beta}^\top + n - p) \\ \frac{MSSR}{MSSE} &= \frac{\hat{\beta}^2}{\Sigma_{\hat{\beta}}} = Wald^2\end{aligned}$$

3. the  $R^2$  is defined as the proportion of variance explained, so using the previous results we get:

$$\begin{aligned}R^2 &= \frac{SSR}{SSR + SSE} \\ &= \frac{1}{1 + SSE/SSR} \\ &= \frac{1}{1 + (n - p)/(\hat{\beta}^2/\sigma_{\hat{\beta}}^2)} \\ &= \frac{Wald^2}{Wald^2 + n - p}\end{aligned}$$

This formula matches exactly the partial correlation coefficient when **both** outcome are adjusted for Subject:

```
e.partialCor <- partialCor(list(pH ~ Subject, Pac02 ~ Subject),
  data = bland1995)
print(e.partialCor, digit = 5)
```

Partial correlation

|               | estimate | se    | df   | lower | upper  | p.value |
|---------------|----------|-------|------|-------|--------|---------|
| rho(pH,Pac02) | -0.507   | 0.125 | 25.7 | -0.71 | -0.225 | 0.00178 |

Note: estimate, standard error, confidence interval have been back-transformed (rho para

Similar values can be obtained using dedicated packages, e.g.:

```
library(rmcorr)
rmcorr(Subject, Pac02, pH, bland1995)$r
```

```
[1] -0.5067697
```

## 4 Partial correlation with repeated measurements

### 4.1 Marginal and conditional correlation

There are several references on the subject (Bland and Altman, 1995; Lipsitz et al., 2001; Bakdash and Marusich, 2017; Shan et al., 2020). We will focus on the mixed model approach. The idea is to jointly model the variance and covariance of all measurements under appropriate constraints. For instance denoting one measurement  $X$  and the other measurement  $Y$ , both indexed by time  $t$ , our target parameter may be  $\rho = \text{Cor}(X(t), Y(t))$  (marginal) assumed independent of  $t$  while  $X$  and  $Y$  may or may not be stationnary. Another target parameter could be the correlation between a de-noised version of  $X$  and  $Y$ , where we have for instance removed individual-specific variations (conditional).

To be more specific let's consider the following statistical model:

$$\begin{aligned} X_i(t) &= \mu_{X,i}(t) + u_i + \varepsilon_{X,i}(t) \\ Y_i(t) &= \mu_{Y,i}(t) + v_i + \varepsilon_{Y,i}(t) \end{aligned}$$

where 
$$\begin{bmatrix} u \\ v \\ \varepsilon_X(t) \\ \varepsilon_Y(t) \end{bmatrix} = \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_u & \tau_{uv} & 0 & 0 \\ \tau_{uv} & \tau_v & 0 & 0 \\ 0 & 0 & \sigma_X & \sigma_{XY} \\ 0 & 0 & \sigma_{XY} & \sigma_Y \end{bmatrix} \right)$$

It implies the following residual covariance matrix:

$$\begin{aligned} \Omega = \text{Var} \begin{bmatrix} X(1) \\ X(2) \\ X(3) \\ Y(1) \\ Y(2) \\ Y(3) \end{bmatrix} &= \begin{bmatrix} \tau_u + \sigma_X & \tau_u & \tau_u & \tau_{uv} + \sigma_{XY} & \tau_{uv} & \tau_{uv} \\ \tau_u & \tau_u + \sigma_X & \tau_u & \tau_{uv} & \tau_{uv} + \sigma_{XY} & \tau_{uv} \\ \tau_u & \tau_u & \tau_u + \sigma_X & \tau_{uv} & \tau_{uv} & \tau_{uv} + \sigma_{XY} \\ \tau_{uv} + \sigma_{XY} & \tau_{uv} & \tau_{uv} & \tau_v + \sigma_Y & \tau_v & \tau_v \\ \tau_{uv} & \tau_{uv} + \sigma_{XY} & \tau_{uv} & \tau_v & \tau_v + \sigma_Y & \tau_v \\ \tau_{uv} & \tau_{uv} & \tau_{uv} + \sigma_{XY} & \tau_v & \tau_v & \tau_v + \sigma_Y \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 & \sigma_2 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_4 \\ \sigma_2 & \sigma_1 & \sigma_2 & \sigma_4 & \sigma_3 & \sigma_4 \\ \sigma_2 & \sigma_2 & \sigma_1 & \sigma_4 & \sigma_4 & \sigma_3 \\ \sigma_3 & \sigma_4 & \sigma_4 & \sigma_5 & \sigma_6 & \sigma_6 \\ \sigma_4 & \sigma_3 & \sigma_4 & \sigma_6 & \sigma_5 & \sigma_6 \\ \sigma_4 & \sigma_4 & \sigma_3 & \sigma_6 & \sigma_6 & \sigma_5 \end{bmatrix} \end{aligned}$$

and the following residual correlation matrix:

$$R = \mathbb{C}or \begin{bmatrix} X(1) \\ X(2) \\ X(3) \\ Y(1) \\ Y(2) \\ Y(3) \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_1 & \rho_2 & \rho_3 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_3 & \rho_2 & \rho_3 \\ \rho_1 & \rho_1 & 1 & \rho_3 & \rho_3 & \rho_2 \\ \rho_2 & \rho_3 & \rho_3 & 1 & \rho_4 & \rho_4 \\ \rho_3 & \rho_2 & \rho_3 & \rho_4 & 1 & \rho_4 \\ \rho_3 & \rho_3 & \rho_2 & \rho_4 & \rho_4 & 1 \end{bmatrix}$$

The marginal correlation is:

$$\begin{aligned} \rho_M &= \frac{\mathbb{C}ov[u_i + \varepsilon_{X,i}(t), v_i + \varepsilon_{Y,i}(t)]}{\sqrt{\mathbb{V}ar[u_i + \varepsilon_{X,i}(t)] \mathbb{V}ar[v_i + \varepsilon_{Y,i}(t)]}} \\ &= \frac{\tau_{uv} + \sigma_{XY}}{\sqrt{(\tau_u + \sigma_X)(\tau_v + \sigma_Y)}} = \frac{\sigma_3}{\sqrt{\sigma_1 \sigma_5}} = \rho_2 \end{aligned}$$

while the conditional correlation is:

$$\begin{aligned} \rho_C &= \frac{\mathbb{C}ov[\varepsilon_{X,i}(t), \varepsilon_{Y,i}(t)]}{\sqrt{\mathbb{V}ar[\varepsilon_{X,i}(t)] \mathbb{V}ar[\varepsilon_{Y,i}(t)]}} \\ &= \frac{\sigma_{XY}}{\sqrt{\sigma_X \sigma_Y}} = \frac{\sigma_3 - \sigma_4}{\sqrt{(\sigma_1 - \sigma_2)(\sigma_5 - \sigma_6)}} = \frac{\rho_2 - \rho_3}{\sqrt{(1 - \rho_1)(1 - \rho_2)}} \end{aligned}$$

## 4.2 Approximated conditional correlation

We now show that formula 1 generalizes to mixed models. Consider the following mixed model relating  $\mathbf{Y} = (Y_1, \dots, Y_T)$  and  $\mathbf{X} = (X_1, \dots, X_T)$ :

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

where  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Omega)$ . Introducing the cholesky decomposition  $\Omega = \omega\omega^\top$ , we can equivalently study:

$$\omega^{-1}\mathbf{Y} = \omega^{-1}\mathbf{X} + \boldsymbol{\zeta}$$

where  $\boldsymbol{\zeta}$  follow a standard normal distribution. We are back the univariate case up to a factor  $\omega^{-1}$ .

1. F-statistics are still equal the Wald statistic squared (divided by the number of parameters).
2. F-statistics still equal  $\frac{MSSR}{MSSE}$ . Indeed:

$$\begin{aligned}
SSE &= (\omega^{-1}\mathbf{Y})^\top \left( I - \omega^{-1}\mathbf{X} \left( (\omega^{-1}\mathbf{X})^\top (\omega^{-1}\mathbf{X}) \right)^{-1} (\omega^{-1}\mathbf{X})^\top \right) (\omega^{-1}\mathbf{Y}) \\
&= \mathbf{Y}^\top \Omega^{-1} \mathbf{Y} - \mathbf{Y}^\top \Omega^{-1} \mathbf{X} \left( \mathbf{X}^\top \Omega^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^\top \Omega^{-1} \mathbf{Y} \\
&= \mathbf{Y}^\top (I - H^\top) \Omega^{-1} (I - H^\top) \mathbf{Y}
\end{aligned}$$

where  $H = \mathbf{X} (\mathbf{X}^\top \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Omega^{-1}$ . Indeed:

$$(I - H^\top) \Omega^{-1} (I - H^\top) = \Omega^{-1} - H^\top \Omega^{-1} - \Omega^{-1} H + H^\top \Omega^{-1} H = \Omega^{-1} - H^\top \Omega^{-1}$$

and  $MSSE = \frac{SSE}{n-p} = \sigma^2$  with  $p$  being the rank of  $X$ . Using that  $HH = H$  :

$$\begin{aligned}
SSR &= (\omega^{-1}\mathbf{Y})^\top \left( \omega^{-1}\mathbf{X} \left( (\omega^{-1}\mathbf{X})^\top (\omega^{-1}\mathbf{X}) \right)^{-1} (\omega^{-1}\mathbf{X})^\top \right) (\omega^{-1}\mathbf{Y}) \\
&= \mathbf{Y}^\top \Omega^{-1} \mathbf{X} \left( \mathbf{X}^\top \Omega^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^\top \Omega^{-1} \mathbf{Y} \\
&= \mathbf{Y}^\top H^\top \Omega^{-1} \mathbf{Y} = \mathbf{Y}^\top H^\top H^\top \Omega^{-1} \mathbf{Y} \\
&= \mathbf{Y}^\top H^\top \Omega^{-1} H \mathbf{Y} \\
&= \hat{\beta}^\top X^\top \Omega^{-1} X \hat{\beta} = \hat{\beta}^\top \Sigma_{\hat{\beta}}^{-1} \hat{\beta}
\end{aligned}$$

where  $\hat{\beta} = (\mathbf{X}^\top \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Omega^{-1} \mathbf{Y}$  is the GLS estimator of  $\beta$ . So for a single covariate:

$$F = \frac{MSSR}{MSSE} = \frac{\hat{\beta}^\top \Sigma_{\hat{\beta}}^{-1} \hat{\beta}}{\sigma^2}$$

**3.** Defining  $R^2$  as the proportion of variance explained, we get back

$$R^2 = \frac{\beta^2}{\beta^2 + df \sigma_\beta^2}$$

where  $df = n - p$ . A corresponding correlation coefficient can be computed as:

$$\rho = \frac{\beta}{\sqrt{\beta^2 + df \sigma_\beta^2}}$$



### 4.3 Back to the example

In the example, we see a very small marginal correlation and a large conditional one:

```
e.pcor <- partialCor(c(pH,Pac02)~1, repetition = ~time|Subject, data =  
  bland1995, heterogeneous = 0.5)  
e.pcor
```

Partial correlation

```
              estimate    se  df  lower    upper p.value  
rho(1.pH,1.Pac02) -1.63e-05 0.313 1.24 -0.988  0.98791  1.0000  
r(1.pH,1.Pac02)   -5.09e-01 0.125 2.63 -0.806 -0.00546  0.0489  
-----  
rho: marginal correlation  
r  : correlation conditional on the individual  
estimates, standard errors, confidence intervals have been back-transformed (tanh).
```

This matches the estimate (but not the uncertainty) of another software:

```
c(r = rmcorr(Subject, pH, Pac02, bland1995)$r,  
  p = rmcorr(Subject, pH, Pac02, bland1995)$p)
```

```
          r          p  
-0.5067697422  0.0008471081
```

We can also extract the underlying correlation coefficients:

```
round(coef(attr(e.pcor,"lmm"), effects = "correlation"),5)
```

```
rho(1.pH,1.Pac02)    rho(1.pH,2.Pac02) rho(1.Pac02,2.Pac02)    rho(1.pH,2.pH)  
      -0.00002              0.10168              0.66317              0.88129
```

that reveal a very strong within pH correlation (almost 0.9) and a rather strong within Pac02 correlation (about 0.65). The instantaneous correlation is nearly 0 but the lag correlation is about 0.1 leading to the observed conditional correlation.

An alternative approach is to fit a mixed model on only one outcome, regressing out the other:

```
e.CS <- lmm(pH ~ Pac02, repetition = ~time|Subject, data = bland1995,  
  structure = "CS")
```

Then estimate the partial correlation formula:

```
e.CSaov <- anova(e.CS, effects = "Pac02=0")  
confint(e.CSaov, columns = c("estimate","se","df","partial.r"))
```

```
      estimate    se  df partial.r  
Pac02   -0.103 0.0295 39.6   -0.486
```

Here approximate degrees of freedom are used, i.e. 39.6 instead of:

```
NROW(bland1995)-2
```

```
[1] 45
```

which would lead to a correlation of:

```
e.CSaov$univariate$statistic/sqrt(e.CSaov$univariate$statistic^2+45)
```

```
[1] -0.4627676
```

## 4.4 Simulation study (compound symmetry model)

We'll compare  $\rho$  and  $r$  in the case of 3 timepoints,  $r = 0.8$ , and 250 individuals:

```
n.time <- 3
n.id <- 250
Sigma <- matrix(c(1,0.8,0.8,1),2,2)
Sigma
```

```
      [,1] [,2]
[1,]  1.0  0.8
[2,]  0.8  1.0
```

```
set.seed(11)
df.W <- data.frame(id = unlist(lapply(1:n.id, rep, n.time)),
  time = rep(1:n.time,n.id),
  rmvnorm(n.time*n.id, mean = c(3,3), sigma = Sigma)
)
head(df.W)
```

```
   id time      X1      X2
1  1    1 2.483259 2.759470
2  1    2 1.034157 1.102983
3  1    3 3.636308 2.691506
4  2    1 4.463341 4.150878
5  2    2 2.510048 2.081439
6  2    3 2.103239 2.317938
```

We use random effects to obtain a constant correlation within  $X$  and within  $Y$ :

```
sd.id <- 1.5
df.W$X1 <- df.W$X1 + rnorm(n.id, sd = sd.id/4)[df.W$id]
df.W$X2 <- df.W$X2 + rnorm(n.id, sd = sd.id)[df.W$id]
df.W$id <- as.factor(df.W$id)
df.L <- reshape2::melt(df.W, id.vars = c("id","time"))
df.L$time2 <- as.factor(as.numeric(as.factor(paste(df.L$variable,df.L$time
,sep="."))))
```

This will lead to the following correlation structure:

```
Sigma.GS <- as.matrix(bdiag(Sigma,Sigma,Sigma))[c(1,3,5,2,4,6),c
(1,3,5,2,4,6)]
Sigma.GS[1:3,1:3] <- Sigma.GS[1:3,1:3] + (sd.id/4)^2
Sigma.GS[4:6,4:6] <- Sigma.GS[4:6,4:6] + sd.id^2
cov2cor(Sigma.GS)
```

```
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 1.0000000 0.1232877 0.1232877 0.4155056 0.0000000 0.0000000
[2,] 0.1232877 1.0000000 0.1232877 0.0000000 0.4155056 0.0000000
[3,] 0.1232877 0.1232877 1.0000000 0.0000000 0.0000000 0.4155056
[4,] 0.4155056 0.0000000 0.0000000 1.0000000 0.6923077 0.6923077
[5,] 0.0000000 0.4155056 0.0000000 0.6923077 1.0000000 0.6923077
[6,] 0.0000000 0.0000000 0.4155056 0.6923077 0.6923077 1.0000000
```

We can now estimate two types of correlation: marginal and conditional

```
e.LMMstar <- partialCor(c(X1,X2) ~ 1, repetition = ~ time|id, data = df.W
, heterogeneous = 0.5)
e.LMMstar
```

#### Partial correlation

```
      estimate      se    df lower upper  p.value
rho(1.X1,1.X2)    0.427 0.0346 34.7 0.356 0.493 6.76e-13
r(1.X1,1.X2)      0.798 0.0251 58.9 0.764 0.829 0.00e+00
```

-----  
rho: marginal correlation

r : correlation conditional on the individual

estimates, standard errors, confidence intervals have been back-transformed (tanh).

The conditional coefficient is identical to what other packages output:

```
rmcorr:::rmcorr(id, X1, X2, df.W)$r
```

```
[1] 0.7983617
```

Here the modeled correlation matrix is:

```
Omega <- sigma(attr(e.LMMstar,"lmm"))
Rho <- cov2cor(Omega)
Rho
```

```
      1.X1      2.X1      3.X1      1.X2      2.X2      3.X2
1.X1 1.00000000 0.06545230 0.06545230 0.42652595 -0.00432106 -0.00432106
2.X1 0.06545230 1.00000000 0.06545230 -0.00432106 0.42652595 -0.00432106
3.X1 0.06545230 0.06545230 1.00000000 -0.00432106 -0.00432106 0.42652595
1.X2 0.42652595 -0.00432106 -0.00432106 1.00000000 0.68836567 0.68836567
2.X2 -0.00432106 0.42652595 -0.00432106 0.68836567 1.00000000 0.68836567
3.X2 -0.00432106 -0.00432106 0.42652595 0.68836567 0.68836567 1.00000000
```

From which the conditional correlation can be deduced:

```
(Rho[1,4]-Rho[1,5])/sqrt((1-Rho[1,2])*(1-Rho[4,5]))
```

```
[1] 0.7983617
```

or equivalently:

```
(Omega[1,4]-Omega[1,5])/sqrt((Omega[1,1]-Omega[1,2])*(Omega[4,4]-Omega[4,5]))
```

```
[1] 0.7983617
```

Replicating this a thousand times:

```
n.id <- 100
n.sim <- 1000
n.cpus <- 25 ## run on the server
warper <- function(n){
  df.W <- data.frame(id = unlist(lapply(1:n, rep, n.time)),
    time = rep(1:n.time,n),
    rmvnorm(n.time*n, mean = c(3,3), sigma = Sigma)
  )
  df.W$X1 <- df.W$X1 + rnorm(n, sd = sd.id/4)[df.W$id]
  df.W$X2 <- df.W$X2 + rnorm(n, sd = sd.id)[df.W$id]
  df.W$id <- as.factor(df.W$id)

  res1 <- setNames(c(rmcorr(id, X1, X2, df.W)$r, rmcorr(id, X1, X2, df.W)$
    CI), c("estimate","lower","upper"))
  res2 <- partialCor(c(X1,X2) ~ 1, repetition = ~ time|id, data = df.W,
    heterogeneous = 0.5)
  return(rbind(cbind(as.data.frame(as.list(res1))), se = NA, method = "
    rmcorr"),
```

```

      cbind(res2[2,c("estimate", "lower", "upper", "se")], method="lmm")))
}

ls.res <- pbapply::pblapply(1:n.sim, function(iSim){
  cbind(sim = iSim, warper(n.id))
}, cl = n.cpus)
dt.res <- as.data.table(do.call(rbind, ls.res))

```

lead to the same estimate for the two implementations:

```

range(dt.res[method=="rmcorr", estimate] - dt.res[method=="lmm", estimate], na.rm=TRUE)

```

```

[1] -8.572216e-10  2.108167e-09

```

and lead to a reasonable coverage:

```

dt.res[,.(missing = mean(is.na(estimate)), coverage = mean((0.8>=lower)*
  (0.8<=upper), na.rm=TRUE)), by = "method"]

```

```

  method missing coverage
1: rmcorr    0.000 0.941000
2:   lmm     0.026 0.949692

```

## 4.5 Simulation study (crossed random effect model)

We will modify the previous simulation setting by introducing more structure on the correlation. More precisely, observations will be correlated within individual (biological variation) and within timepoint (batch effect). This violates the compound symmetry structure and therefore we expect `rmcorr` to give biased estimates. We will use `lmer` instead of `lmm` as a reference since `lmer` is very convenient to use and fast when dealing with crossed random effects. Note that, however, it is not straightforward to have a measure of uncertainty.

```

n.time <- 4
n.id <- 100
warper <- function(n){
  df.W <- data.frame(id = unlist(lapply(1:n.id, rep, n.time)),
    time = rep(1:n.time, n.id),
    rmvnorm(n.time*n.id, mean = c(3,3), sigma = Sigma)
  )
  df.W$X1 <- df.W$X1 + rnorm(n.id, sd = sd.id/4)[df.W$id]
  df.W$X2 <- df.W$X2 + rnorm(n.id, sd = sd.id)[df.W$id]
  df.W$X1 <- df.W$X1 + rnorm(n.time, sd = sd.id/3)[df.W$time]
  df.W$X2 <- df.W$X2 + rnorm(n.time, sd = sd.id/2)[df.W$time]
}

```

```

df.W$id <- as.factor(df.W$id)

e.lmer <- lmer(X2 ~ X1 + (1|time) + (1|id), data = df.W)
e.Slmer <- summary(e.lmer)$coefficient

res1 <- setNames(c(rmcorr(id, X1, X2, df.W)$r, rmcorr(id, X1, X2, df.W)
)$CI), c("estimate", "lower", "upper"))
res2 <- c(estimate = e.Slmer["X1", "t value"]/sqrt(e.Slmer["X1", "t
value"]^2+e.Slmer["X1", "df"]), lower = NA, upper = NA)

return(rbind(cbind(as.data.frame(as.list(res1)), method = "rmcorr"),
cbind(as.data.frame(as.list(res2)), method= "lmer"))))
}

ls.res <- pbapply::pblapply(1:101,function(iSim){
  cbind(sim = iSim, warper(100))
})
dt.res <- as.data.table(do.call(rbind, ls.res))

```

We can clearly see that the `rmcorr` estimator is biased and very variable while the `lmer`-based estimator (i.e. using Equation 1) gives reasonable results:

```

rbind(rmcorr = quantile(dt.res[method=="rmcorr",estimate]),
      lmer = quantile(dt.res[method=="lmer",estimate]))

```

|                     | 0%          | 25%       | 50%       | 75%       | 100%      |
|---------------------|-------------|-----------|-----------|-----------|-----------|
| <code>rmcorr</code> | 0.001331043 | 0.4887141 | 0.6207839 | 0.7273354 | 0.8670439 |
| <code>lmer</code>   | 0.732343195 | 0.7738880 | 0.7900921 | 0.8094797 | 0.8472541 |

## 5 Reference

- Bakdash, J. Z. and Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in psychology*, 8:456.
- Bland, J. M. and Altman, D. G. (1995). Calculating correlation coefficients with repeated observations: Part 2—correlation between subjects. *Bmj*, 310(6980):633.
- Lipsitz, S. R., Leong, T., Ibrahim, J., and Lipshultz, S. (2001). A partial correlation coefficient and coefficient of determination for multivariate normal repeated measures data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(1):87–95.
- Shan, G., Zhang, H., and Jiang, T. (2020). Correlation coefficients for a study with repeated measures. *Computational and mathematical methods in medicine*, 2020.