# PLACO Software Documentation

August 30, 2020

**Version** 0.1.1
**Date** August 30, 2020
**Title** Pleiotropic analysis under composite null hypothesis
**Correspondence** Debashree Ray, Ph.D. <dray@jhu.edu>
**Description** PLACO implements a variant-level formal statistical test of pleiotropy of two traits using summary-level GWAS data, and can account for potential correlation across traits, such as that arising due to shared controls in case-control studies.
**Depends** R (>= 3.0.1)

---

`cor.pearson` *To estimate correlation between Z-scores from two traits*

---

## Description

The R function `cor.pearson` allows the user to estimate the correlation matrix between Z-scores from two traits. This matrix is required to decorrelate Z-scores coming from traits that share samples either partially or fully. It is estimated only once irrespective of the number $p$ of genetic variants to be tested for pleiotropic association.

## Usage

```
cor.pearson(Z.matrix, P.matrix, p.threshold=1e-4)
```

## Arguments

`Z.matrix` It is the $p \times 2$ matrix of Z-scores where $p$ is total number of variants in the GWAS dataset. The 2 columns correspond to the 2 traits. For a trait, Z-score is the ratio of estimated genetic effect to its standard error ($Z = \hat{\beta}/\hat{se}$).

`P.matrix` It is the $p \times 2$ matrix of p-values of the 2 traits where $p$ is total number of variants in the GWAS dataset. Like the Z-scores (summary statistics) in `Z.matrix`, the p-values of `P.matrix` correspond to individual test of each trait against each genetic variant. The order in which traits and the genetic variants are arranged in `Z.matrix` and `P.matrix` must be same.

`p.threshold` The p-value threshold used to determine which genetic variants are likely not associated. A liberal threshold needs to be used to screen out any signal that may affect the estimate of the correlation matrix `R`. Genetic variants (here, rows) with p-values smaller than this threshold for any trait are removed before estimating `R`. Default value is $10^{-4}$.

**Value**

`R` The estimated $2 \times 2$ correlation matrix of the GWAS summary statistics of the 2 traits under the complete null hypothesis of no association.

---

## `var.placo` *To estimate the variance parameters needed to implement PLACO*

---

**Description**

The R function `var.placo` estimates the variances of the $Z$-scores of the two traits under the composite null hypothesis of no pleiotropy. Output from this function goes as input for R function `placo`. This estimation procedure is done only once for a given study using the single-trait Z-scores and p-values (or GWAS summary statistics) that are usually publicly available.

**Usage**

```
var.placo(Z.matrix, P.matrix, p.threshold=1e-4)
```

### Arguments

Arguments for `var.placo` are the same as those for `cor.pearson`.

### Value

| | |
|---|---|
| `VarZ` | A vector of estimated variances for the Z-scores of 2 traits. |

---

## `placo` *Pleiotropic association test of two traits using GWAS summary statistics*

---

### Description

PLACO uses genome-wide summary statistics (Z-scores and p-values) on two traits to test for variant-level pleiotropic association between a genetic marker and two traits. It can be applied on independent traits, or moderately correlated traits after decorrelating the Z-scores as described in the manuscript.

### Usage

```
placo(Z, VarZ, AbsTol=.Machine$double.eps^0.8)
```

### Arguments

| | |
|---|---|
| `Z` | The vector of Z-scores of 2 traits for a given genetic variant. |
| `VarZ` | The vector of estimated variances for the Z-scores of 2 traits as estimated by the function `var.placo()`. |
| `AbsTol` | The user can specify the absolute tolerance value used in the numerical integration for evaluating PLACO p-value. Default value is $3 \times 10^{-13}$. Function `integrate()` is used for numerical integration. |

### Details

Consider two genome-wide studies of traits $Y_1$ and $Y_2$ on $n_1$ and $n_2$ individuals respectively who were genotyped and/or imputed or sequenced at $p$ genetic variants. Let $\boldsymbol{Y}_k$ and $\boldsymbol{X}_k$ be the vectors of $k$-th trait values and genotypes at a given genetic variant respectively on all $n_k$ individuals ($k = 1, 2$). For the $k$-th trait, suppose $\beta_k$ is the genetic effect and the corresponding summary statistic for testing no genetic association of the trait is $Z_k = \hat{\beta}_k / \operatorname{se}(\hat{\beta}_k)$, where $\hat{\beta}_k$ is the maximum likelihood estimate (MLE) of $\beta_k$ and $\operatorname{se}(\hat{\beta}_k)$ is its standard error. Publicly available GWAS data usually have information on $\hat{\beta}_k$ and $\operatorname{se}(\hat{\beta}_k)$, and/or $Z_k$ and the corresponding p-value $p_k$, $k = 1, 2$.

The conventional cross-phenotype association methods test the global null hypothesis that none of the traits is associated with the given genetic variant (i.e., $\beta_1 = \beta_2 = 0$). Rejection of this global null can be due to one associated trait ($\beta_1 \neq 0, \beta_2 = 0$ or $\beta_1 = 0, \beta_2 \neq 0$). Here, we are interested in identifying the genetic variants that are associated with both the traits or outcomes (i.e., $\beta_1 \neq 0, \beta_2 \neq 0$). The effects of such a genetic variant on the traits may or may not be equal. Formally, our null hypothesis of no pleiotropy is $H_0$ : at most 1 trait is associated with the genetic variant while the alternative hypothesis is $H_a$ : both traits are associated. Mathematically, our null hypothesis of no pleiotropy is a composite null hypothesis, and can simply be written as $H_0 : \beta_1\beta_2 = 0$ vs. the alternative hypothesis $H_a : \beta_1\beta_2 \neq 0$.

The PLACO test statistic and approximate asymptotic p-value for testing the composite null hypothesis $H_0$, assuming the two traits are independent, are

$$
\begin{aligned}
T_{\text{PLACO}} &= Z_1 Z_2 \\
p_{\text{PLACO}} &= \mathbb{F}\left(z_1 z_2 / \sqrt{\text{Var}(Z_1)}\right) + \mathbb{F}\left(z_1 z_2 / \sqrt{\text{Var}(Z_2)}\right) - \mathbb{F}\left(z_1 z_2\right)
\end{aligned}
$$

where $z_1$ and $z_2$ are the observed $Z$-scores for the two traits at a given genetic variant; $\text{Var}(Z_1)$ and $\text{Var}(Z_2)$ are the estimated marginal variances of the $Z$-scores (as estimated by the function `var.placo()`); and $\mathbb{F}(u) = 2\int_{|u|}^{\infty} \mathbb{f}(x)dx$ is the two-sided tail probability of a normal product distribution at value $u$.

If the two traits come from studies with overlapping samples, either partially (e.g. case-control traits with shared controls) or completely, then the $Z$-scores will be correlated and may lead to inflated p-values if the correlation is not accounted for in the pleiotropic analysis.

If $\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ be the vector of $Z$-scores for a given genetic variant and $\hat{\boldsymbol{R}} = \begin{pmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{pmatrix}$ be the estimated correlation matrix (as obtained from function `cor.pearson()`), one needs to de-correlate the $Z$-scores as $\boldsymbol{Z}^{\text{decor}} = \boldsymbol{R}^{-1/2}\boldsymbol{Z} = \begin{pmatrix} Z_1^{\text{decor}} \\ Z_2^{\text{decor}} \end{pmatrix}$ so that $Z_1^{\text{decor}}$ and $Z_2^{\text{decor}}$ are uncorrelated. PLACO, as described before, can now be applied on these de-correlated $Z$-scores to test for pleiotropy of two correlated traits. However, we advocate that PLACO be applied to uncorrelated or moderately correlated traits. PLACO is only applicable genome-wide and not to a selected set of genetic variants.

For more details on how PLACO may be used, please refer Ray and Chatterjee (2020). We request that the reference for Ray and Chatterjee (2020) be cited if this software is used in any publication.

## Value

| | |
|---|---|
| `T.placo` | The PLACO statistic for the test of pleiotropic association of a single variant and two traits. |
| `p.placo` | The approximate asymptotic p-value of PLACO. |

## Reference

Ray, D. and Chatterjee, N. A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between type 2 diabetes and prostate cancer. *bioRxiv*, https://doi.org/10.1101/2020.04.11.037630.

## Example

```
#-------- Download or directly source PLACO
# require(devtools)
# source_url("https://github.com/RayDebashree/PLACO/blob/master/
PLACO_v0.1.1.R?raw=TRUE")
#--------

source("PLACO_v0.1.1.R")
set.seed(1)
## For an example, let's first simulate a toy set of GWAS summary
## statistics on 2 uncorrelated traits and 1000 variants
require(MASS)
```

```
k <- 2
p <- 1000
Z.matrix <- mvrnorm(n=p, mu=rep(0,k), Sigma=diag(1,k))
P.matrix <- matrix(NA, nrow=p, ncol=k)
for(j in 1:k){
    P.matrix[,j] <- sapply(1:nrow(Z.matrix),
        function(i) pchisq(Z.matrix[i,j]^2,df=1,ncp=0,lower.tail=F))
}
colnames(Z.matrix) <- paste("Z",1:k,sep="")
colnames(P.matrix) <- paste("P",1:k,sep="")

## Steps to implementing PLACO
# Step 1: Obtain the variance parameter estimates (only once)
VarZ <- var.placo(Z.matrix, P.matrix, p.threshold=1e-4)
# Step 2: Apply test of pleiotropy for each variant
out <- sapply(1:p, function(i) placo(Z=Z.matrix[i,], VarZ=VarZ))
# Check the output for say variant 100
dim(out)
out[,100]$T.placo
out[,100]$p.placo

## If the traits are dependent or correlated, we suggest
## decorrelating the Z-scores (only once), then apply Steps 1 and 2

# Step 0a: Obtain the correlation matrix of Z-scores
R <- cor.pearson(Z.matrix, P.matrix, p.threshold=1e-4)
# Step 0b: Decorrelate the matrix of Z-scores
    # function for raising matrix to any power
    "%^%" <- function(x, pow)
        with(eigen(x), vectors %*% (values^pow * t(vectors)))
Z.matrix.decor <- Z.matrix %*% (R %^% (-0.5))
```