

# Vignette for `MultiMed` package

Simina M. Boca  
Innovation Center for Biomedical Informatics and  
Department of Oncology, Georgetown University Medical Center  
email: `smb310@georgetown.edu`,

Joshua N. Sampson  
Biostatistics Branch, Division of Cancer Epidemiology and Genetics,  
National Cancer Institute  
email: `joshua.sampson@nih.gov`

April 27, 2020

## 1 Overview

The `MultiMed` package implements a permutation method which adjusts for “multiple comparisons” when testing whether multiple biomarkers are mediators between a known risk factor and a disease. The approach is described in the companion paper [Boca et al., 2014], “Testing multiple biological mediators simultaneously.” This method can significantly improve the power to detect mediators over the standard Bonferroni correction.

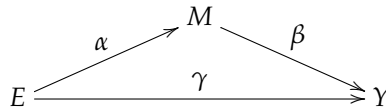
We first need to load the package:

```
> library(MultiMed)
```

## 2 Performing the test of mediation

The scenarios which can be considered are shown in Figure 1 for the single mediator case and Figure 2 (also shown in the [Boca et al., 2014] paper) for the multiple mediator case. Here, we consider simulating data where the exposure  $E$ , the mediator(s)  $M$  (or  $M_i, i = 1, \dots, K$ ), and the outcome  $Y$  are normally distributed. We denote by  $\sigma_E^2$  the variance of  $E$ , by  $\sigma_M^2$  ( $\sigma_{M_i}^2$ ) the variance of  $M$  ( $M_i$ ) conditional on  $E$ , and by  $\sigma_Y^2$  the variance of  $Y$  conditional on  $E$  and  $M$  ( $M_i$ ).

Figure 1: A scenario with a single possible mediator between exposure and outcome.



### 2.1 The `medTest` function

The function used to perform the test of mediation is `medTest`. It has seven arguments: `E`, `M`, `Y`, `Z`, `nperm`, `w`, and `useWeightsZ`. `E`, `M`, and `Y` represent matrices of size  $n \times 1$ ,  $n \times K$ , and  $n \times 1$ , respectively, giving the exposure, mediator, and outcome values, where  $n$  is the sample size and  $K$  is the number of mediators. `E` and `Y` can also be inputted as vectors. The `Z` argument is either `NULL` or a numerical matrix having  $n$  rows. If it is not `NULL`, then the exposure, mediators, and outcome will all be initially regressed on `Z`, with the

residuals being used in the mediation analysis. The `nperm` argument gives the number of permutations used to estimate the null distribution, the default being 100. The `w` argument specifies whether any weighting should be done for the  $E$ - $M$  association, as would be needed, for instance, in a scenario which considers a case-control study. The default is `w=1`, which means that all the study participants are equally weighted; `w` may also be given as a vector of length  $n$ , in which case it is first standardized to sum to 1. The `useWeightsZ` argument can be `TRUE`, in which case the weights in `w` are used for the initial regression on  $Z$ , or `FALSE`, in which case equal weights are used for this initial step.

## 2.2 Simulated example: Single mediator case

For a sample size of  $n = 100$ , we can simulate a dataset with a single mediator in the following way:

```
> set.seed(20183)
> alpha <- 0.2
> beta <- 0.2
> gamma <- 0.4
> n <- 100
> sigma2E <- 1
> sigma2M <- 1 - alpha^2
> sigma2Y <- 1 - beta^2 * (1 - alpha^2) - (alpha * beta + gamma)^2
> ## exposure:
> E <- rnorm(n, 0, sd = sqrt(sigma2E))
> ## mediator:
> M <- matrix(0, nrow = n, ncol = 1)
> M[, 1] <- rnorm(n, alpha * E, sd = sqrt(sigma2M))
> ## outcome:
> Y <- rep(0, n)
> for (subj in 1:n) Y[subj] <- rnorm(1, beta * M[subj, ], sd = sqrt(sigma2Y))
```

Note that the values of  $\sigma_E^2$ ,  $\sigma_M^2$ , and  $\sigma_Y^2$  were chosen so that the marginal variances of  $E$ ,  $M$ , and  $Y$  are 1.

To perform a test of mediation, we use the `medTest` function. The output is a matrix with two columns: `S`, the test statistic used (the absolute value of the product of the correlations between  $E$  and  $M$  and between  $r_{M|E}$  and  $r_{Y|E}$ , where  $r_{Z_1|Z_2}$  represents the residual obtained from regressing  $Z_1$  on  $Z_2$ ) and `p`, the p-value:

```
> medTest(E, M, Y, nperm = 500)
```

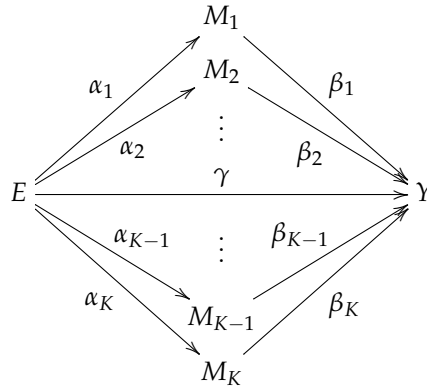
```
      S      p
[1,] 0.01322964 0.53
```

## 2.3 Simulated example: Multiple mediator case

Now consider a scenario with  $K = 10$  mediators and a sample size of  $n = 100$ .

```
> set.seed(380184)
> alpha <- c(rep(0, 6), rep(0.3, 2), rep(0, 2))
> beta <- c(rep(0, 6), rep(0, 2), rep(0.3, 2))
> gamma <- 0.6
> alpha
[1] 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.3 0.3 0.0 0.0
> beta
[1] 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.3 0.3
```

Figure 2: A scenario with  $K$  possible mediators between exposure and outcome.



```
> n <- 100
> sigma2E <- 1
> sigma2M <- 1-alpha^2
> sigma2Y <- 1-sum(beta^2*sigma2M)-(sum(alpha*beta)+gamma)^2
> sigma2M
```

```
[1] 1.00 1.00 1.00 1.00 1.00 1.00 0.91 0.91 1.00 1.00
```

```
> sigma2Y
```

```
[1] 0.46
```

Note that in this case `alpha` and `beta` are vectors having the  $i^{\text{th}}$  elements be  $\alpha_i$ , respectively  $\beta_i$ , where  $i = 1, \dots, 10$  indexes the mediators. Similarly, `sigma2M` is a vector, with the  $i^{\text{th}}$  element being  $\sigma_{M_i}^2$ . The values of  $\sigma_E^2$ ,  $\sigma_{M_i}^2$ , and  $\sigma_Y^2$  were chosen so that the marginal variances of  $E$ ,  $M_i$ ,  $Y$  are 1.

We first simulate the data:

```
> K <- length(alpha)
> ## exposure:
> E <- rnorm(n, 0, sd = sqrt(sigma2E))
> ## mediator:
> M <- matrix(0, nrow = n, ncol = K)
> for (i in 1:K) {
+   M[, i] <- rnorm(n, alpha[i] * E, sd = sqrt(sigma2M[i]))
+ }
> ## outcome:
> Y <- rep(0, n)
> for (subj in 1:n)
+   Y[subj] <- rnorm(1, sum(beta*M[subj,])+gamma*E[subj], sd=sqrt(sigma2Y))
```

We then use the `medTest` once again to perform the test of mediation. The output is now a matrix with 10 rows, each row giving the test statistic `S` and the p-value `p` for each mediator. Note that the p-values are already implicitly considering the multiple tests being performed, so no further adjustment is necessary:

```
> medTest(E, M, Y, nperm = 500)
```

```
      S      p
[1,] 0.0115085655 1.000
[2,] 0.0008037094 1.000
```

```
[3,] 0.0009221887 1.000
[4,] 0.0161794377 0.998
[5,] 0.0016529532 1.000
[6,] 0.0001764986 1.000
[7,] 0.0343911724 0.774
[8,] 0.0554955400 0.298
[9,] 0.0031333508 1.000
[10,] 0.0447346023 0.534
```

## 2.4 Data analysis: Metabolites as mediators

We consider a data example from the [Boca et al., 2014] paper, using the Navy Colorectal Adenoma case-control study [Sinha et al., 1999], with daily fish intake as the exposure of interest  $E$  and colorectal adenoma status as the outcome  $Y$ . The possible mediators are 149 serum metabolites, whose values were previously batch normalized and log transformed.

We first load the dataset:

```
> data(NavyAdenoma)
```

The first 5 columns of the `NavyAdenoma` object represent: daily fish intake, BMI, gender (coded as 0 for male, 1 for female), age, and current smoking status (coded as 0 for non-smoker, 1 for current smoker):

```
> colnames(NavyAdenoma)[1:5]
```

```
[1] "Fish"      "BMI"      "Female"   "Age"      "Smoking"
```

The next 149 columns represent the metabolite values, while the last column represents the case-control status:

```
> colnames(NavyAdenoma)[c(6:9,154)]
```

```
[1] "glycine"   "serine"   "betaine"  "alanine"  "erythritol"
```

```
> colnames(NavyAdenoma)[155]
```

```
[1] "Adenoma"
```

```
> table(NavyAdenoma$Adenoma)
```

```
 0  1
129 129
```

Due to the retrospective sampling, we consider weights incorporating the prevalence of adenoma in this age category (approximately 0.228) and the fraction of cases in the dataset for the E-M associations:

```
> prev <- 0.228
```

```
> p <- sum(NavyAdenoma$Adenoma==1)/nrow(NavyAdenoma)
```

```
> p
```

```
[1] 0.5
```

```
> w <- rep(NA, nrow(NavyAdenoma))
```

```
> w[NavyAdenoma$Adenoma == 1] <- prev/p
```

```
> w[NavyAdenoma$Adenoma == 0] <- (1-prev)/(1-p)
```

```
> table(w)
```

```
w
0.456 1.544
129 129
```

We use `medTest` to perform the test of mediation, adjusting for the covariates BMI, gender, age, and current smoking status. As in the Boca et al. [2014] paper, we perform this adjustment using equal weights, rather than using the weights in `w`, but users can consider using the weights in `w` both here and downstream:

```
> set.seed(840218)
> medsFish <- medTest(E=NavyAdenoma$Fish,
+                   M=NavyAdenoma[, 6:154],
+                   Y=NavyAdenoma$Adenoma,
+                   Z=NavyAdenoma[, 2:5],
+                   nperm=1000, w=w,
+                   useWeightsZ=FALSE)
```

Now find metabolite which has the lowest p-values:

```
> rownames(medsFish) <- colnames(NavyAdenoma[, -c(1:5, 154)])
> medsFish[which.min(medsFish[, "p"]), , drop=FALSE]
```

```

                S      p
docosahexaenoate (DHA; 22:6n3) 0.04989712 0.051
```

Thus, we conclude that DHA (fish oil) is a possible mediator of the association between fish intake and colorectal adenoma.

## References

- S. M. Boca, R. Sinha, A. J. Cross, S. C. Moore, and J. N. Sampson. Testing multiple biological mediators simultaneously. *Bioinformatics*, 30(2):214–220, 2014.
- R. Sinha, W. H. Chow, M. Kulldorff, J. Denobile, J. Butler, M. Garcia-Closas, R. Weil, R. N. Hoover, and N. Rothman. Well-done, grilled red meat increases the risk of colorectal adenomas. *Cancer Research*, 59(17):4320–4324, 1999.