The 'HDMT' package: A Multiple Testing Procedure For High-dimensional Mediation hypotheses

Xiaoyu Wang, James Y. Dai

October 6, 2020

1 Introduction

Mediation analysis is of rising interest in clinical trials and epidemiology. The advance of high-throughput technologies has made it possible to interrogate molecular phenotypes such as gene expression and DNA methylation in a genome-wide fashion, some of which may act as intermediaries of treatment, external exposures and life-style risk factors in the etiological pathway to diseases or traits. When testing for mediation in high-dimensional studies like ours [1], properly controlling the type I error rate remains a challenge due to the composite null hypothesis. Among existing methods, the joint significance (JS) test is an intersection-union test using the maximum p-value for testing the two parameters, though a naive significance rule based on the uniform null p-value distribution (JS-uniform) may yield an overly conservative type I error rate and therefore low power. This is particularly a concern for high-dimensional mediation hypotheses for genome-wide molecular intermediaries such as DNA methylation. In this R package we develop a multiple-testing procedure that accurately controls the family-wise error rate (FWER) and the false discovery rate (FDR) for testing high-dimensional mediation composite null hypotheses. The core of our procedure is based on estimating the proportions of three types of component null hypotheses and deriving the corresponding mixture distribution (JS-mixture) of null p-values. Theoretical derivation and extensive simulations show that the proposed procedure provides adequate control of FWER and FDR when the number of mediation hypotheses is large.

2 Examples

We show two examples assessing the mediation role of DNA methylation in prostate cancer studies. The first study is on assessing the mediation potential role of DNA methylation in genetic regulation of gene expression in primary prostate cancer (PCa) samples from The Cancer Genome Atlas (TCGA). The second study investigated the potential mediation of DNA mColumn 1 contains the p-values for testing if an exposure

is associated with the mediator ($\alpha \neq 0$). Column 2 contains the p-value for testing if a mediator is associated with the outcome after adjusted for the exposure ($\beta \neq 0$) methylation in exercise effect on prostate cancer progression in a Seattle-based patient cohort.

```
> data(snp_input)
> dim(snp_input)

[1] 69602     2

> data(exercise_input)
> dim(exercise_input)

[1] 47900     2
```

The input data matrix contains two columns of p-values for candidate mediators. See notation in [1].

2.1 DNAMethylation in Genetic Regulation of Gene Expression Among 147 Prostate Cancer Risk SNPs

We first read the input data matrix first. To save time for compiling this vignettes file, we use 10% of p-values in the input data matrix. To reproduce the figure in the paper, one need to run the code on the entire data matrix.

```
> input_pvalues <- snp_input
> input_pvalues=input_pvalues[sample(1:nrow(input_pvalues),
+                   size=ceiling(nrow(input_pvalues)/10)),]
```

The first step of the procedure is to estimate the proportions of the three type of null hypotheses

```
> nullprop <- null_estimation(input_pvalues,lambda=0.5)
> nullprop

$alpha10
[1] 0.03476512

$alpha01
[1] 0.4189053

$alpha00
[1] 0.4987789
```

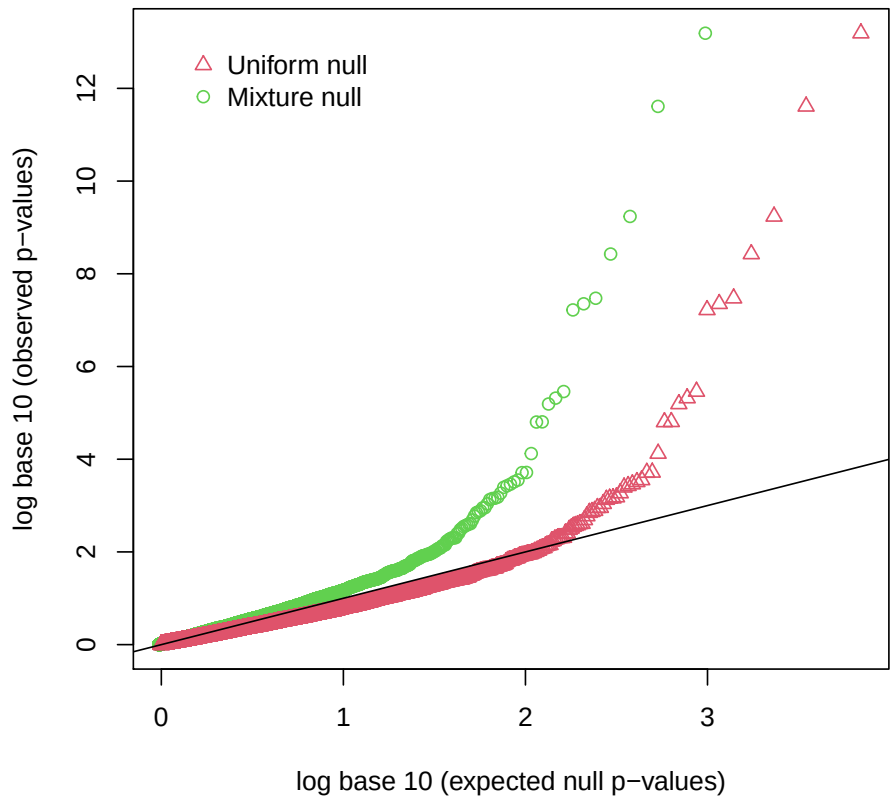
```
$alpha1
[1] 0.9176842

$alpha2
[1] 0.533544
```

We next compute the expected quantiles of the mixture null distribution of p_{max} (maximum of two p-values) using either the approximation (`exact`=0) method or the exact method (`exact`=1). This set of quantiles can be used to draw the corrected q-q plot.

```
> pnull1<-adjust_quantile(nullprop$alpha00,nullprop$alpha01,nullprop$alpha10,
+                         nullprop$alpha1,nullprop$alpha2,input_pvalues,exact=1)
```

'HDMT' provides a function correct_qqplot to draw the corrected quantile-quantile plot for p_{max}, based on the estimated quantile of the mixture null distribution (green dots) and compared to the standard q-q plot based on the uniform distribution (red dots).

```
> pmax <- apply(input_pvalues,1,max)
> correct_qqplot(pmax, pnull1)
```



We can also compute the pointwise estimated FDR for p_{max} using either the approximation method or the exact method.

```
> efdr <- fdr_est(nullprop$alpha00,nullprop$alpha01,nullprop$alpha10,
+                 nullprop$alpha1,nullprop$alpha2,input_pvalues,exact=0)
> plot(pmax[order(pmax)],efdr[order(pmax)],type="l",ylim=c(0,1),
+      xlab="p-max",ylab="Estimated FDR")
```

2.2 DNA Methylation and the Association of Vigorous Physical Activity With Lower Risk of Metastatic Progression

For this example, we only included 10% of p-values from the genome-wide testing as shown in the paper, due to data storage space limit. We read in the input data matrix with two columns of p-values as follows.

```
> input_pvalues <- exercise_input
> #To save time, we use 10% of rows
> input_pvalues=input_pvalues[sample(1:nrow(input_pvalues),
+                   size=ceiling(nrow(input_pvalues)/10)),]
```

The estimation procedures are identical to the previous example:

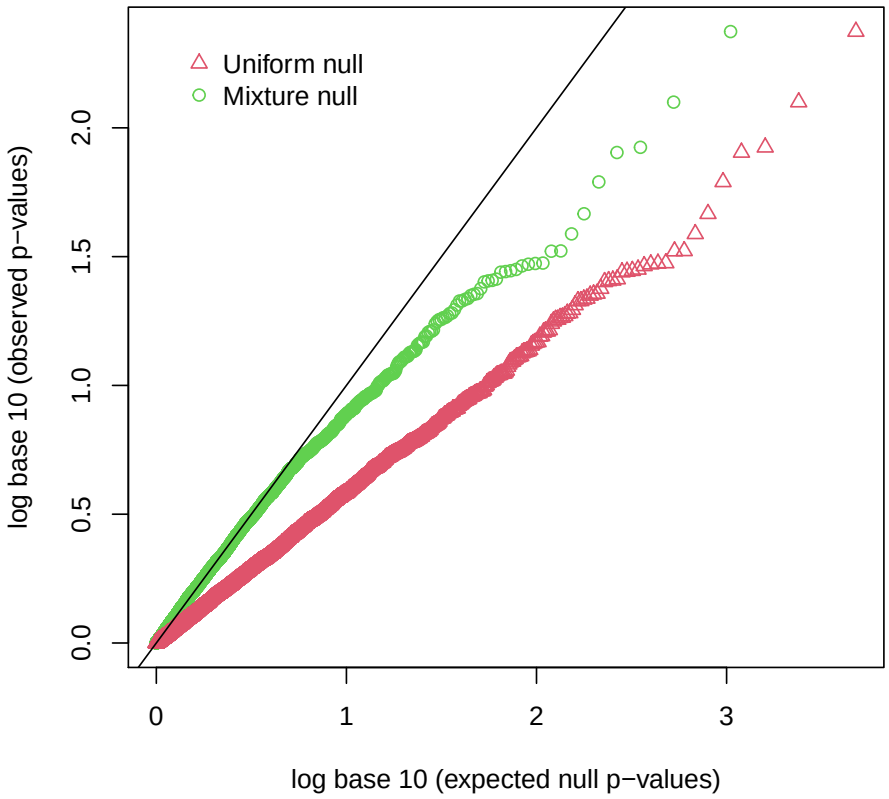
```
> nullprop <- null_estimation(input_pvalues,lambda=0.5)
```

We can compute the null distribution of pmax using the approximation method in Section 2.2, and we can also compute the null distribution of pmax using the exact method in Section 2.4.

```
> pnull<-adjust_quantile(nullprop$alpha00,nullprop$alpha01,nullprop$alpha10,
+ nullprop$alpha1,nullprop$alpha2,input_pvalues,exact=0)

> pnull1<-adjust_quantile(nullprop$alpha00,nullprop$alpha01,nullprop$alpha10,
+ nullprop$alpha1,nullprop$alpha2,input_pvalues,exact=1)
```

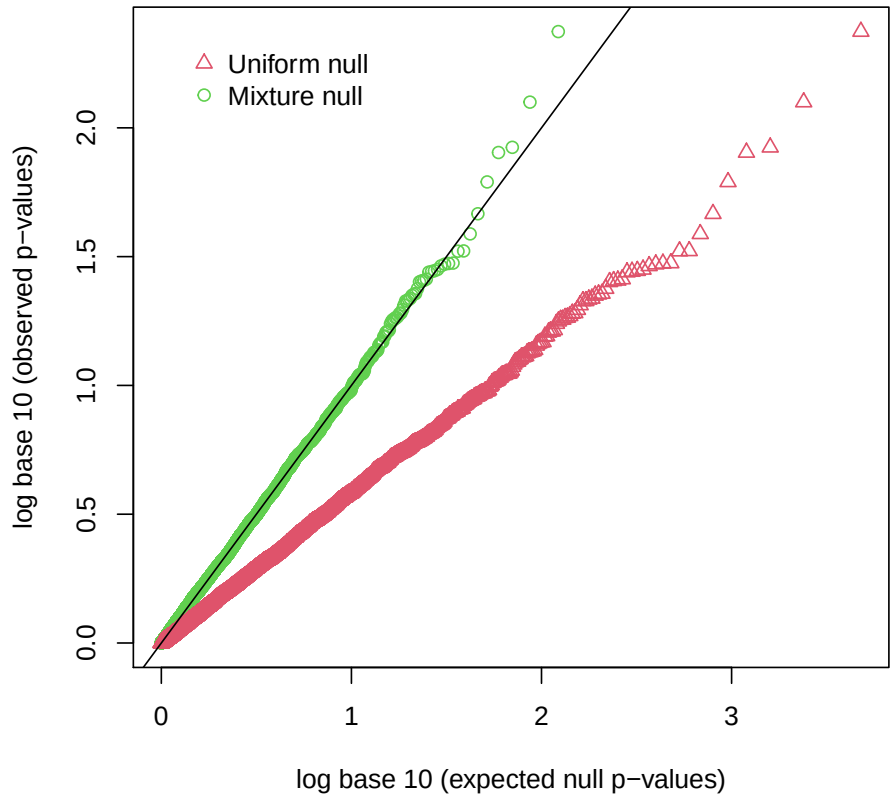
The Q-Q plot based on the approximation method is shown as follows:

```
> pmax <- apply(input_pvalues,1,max)
> correct_qqplot(pmax, pnull)
```



The Q-Q plot based on the exact method is shown as follows:

```
> correct_qqplot(pmax, pnull1)
```



3 session information

The version number of R and packages loaded for generating the vignette were:

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.04.4 LTS

Matrix products: default
BLAS/LAPACK: /app/software/OpenBLAS/0.3.7-GCC-8.3.0/lib/libopenblas_haswellp-
r0.3.7.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
```

```
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] HDMT_1.0.2

loaded via a namespace (and not attached):
[1] compiler_4.0.2 tools_4.0.2    fdrtool_1.2.15
```

References

[1] James Y. Dai, Janet L. Stanford, and Michael LeBlanc. A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, 2020.