The 'HDMT' package: A High-Dimensional Multiple Testing Procedure

Xiaoyu Wang, James Y. Dai

December 4, 2019

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 article 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 two studies. 1) genetic regulation of gene expression in primary prostate cancer (PCa) samples from The Cancer Genome Atlas (TCGA) with risk SNPs as the exposure, and 2) regulation of prostate cancer progression in a Seattle-based cohort of patients diagnosed with clinically localized PCa with excercise as the exposure.

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

[1] 69602     2

> data(exercise_input)
> dim(exercise_input)

[1] 47900     2

> #We only included 10% of the excercise data from the paper
> #due to storage space limit.
>
> #Each matrix contains two columns of p-values for candidate mediators.
> #Column 1 is the p-value of testing if an exposure  is associated with
> #the mediator (alpha!=0).
> #Column 2 is the p-value of testing if a mediator is associated with
> #the outcome adjusted for the exposure (beta!=0)
>
```

2.1 The example of using risk SNPs as the exposure

We read the input first:

```
> input_pvalues <- snp_input
> #To save time for the illustration, we use 10% of rows; to reproduce the
> # figure in the paper, please don't run the following line
> input_pvalues=input_pvalues[sample(1:nrow(input_pvalues),
+                   size=ceiling(nrow(input_pvalues)/10)),]
```

We proceed to estimate the proportion of nulls:

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

We next compute the null distribution of pmax (maximum of studied two input p-values) using either approximation (method=0) or exact method (method=1):

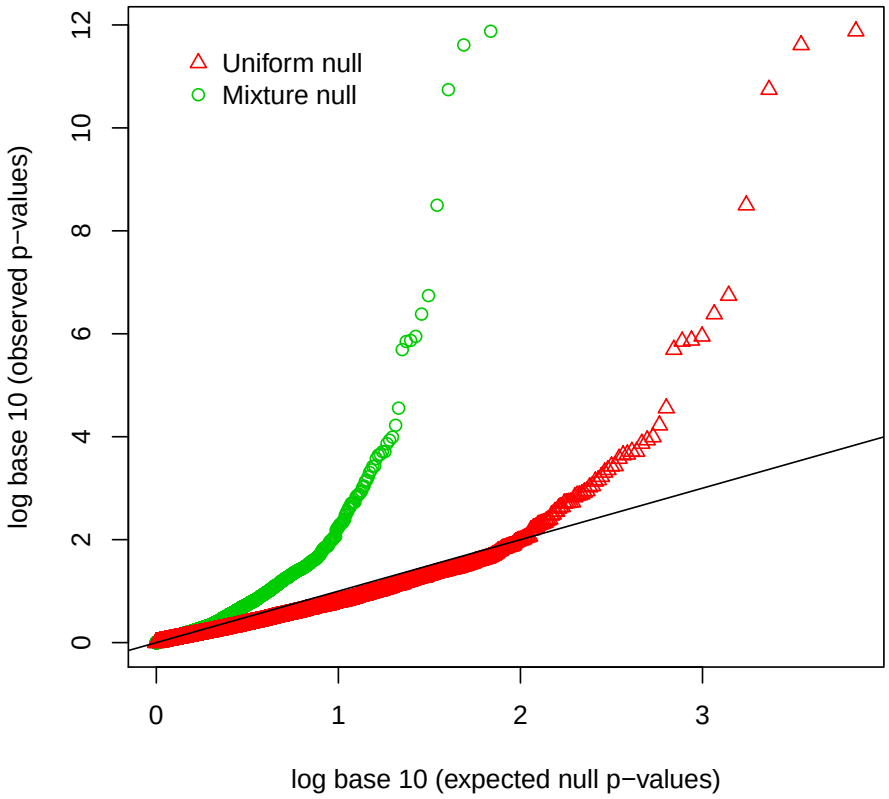
```
> pnull1<-adjust_quantile(nullprop$alpha00,nullprop$alpha01,nullprop$alpha10,
+ nullprop$alpha1,nullprop$alpha2,input_pvalues,method=1)
```

We can compute the pointwise FDR using the approximation method:

```
> fdr <- fdrest(nullprop$alpha00,nullprop$alpha01,nullprop$alpha10,
+ nullprop$alpha1,nullprop$alpha2,input_pvalues,method=0)
```

'HDMT' provides a function correct_qqplot to draw the quantile-quantile plots for pmax, based on null distribution of JS-mixture (green dots) and JS-uniform (red dots)

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



The above figure shows the proposed method JS-mixture provides much more accurate control of the FWER and the FDR compared to the JS-uniform method.

2.2 The example of using exercise as the exposure

We read the input as:

```
> 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 following procedures are identical to the previous example:

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

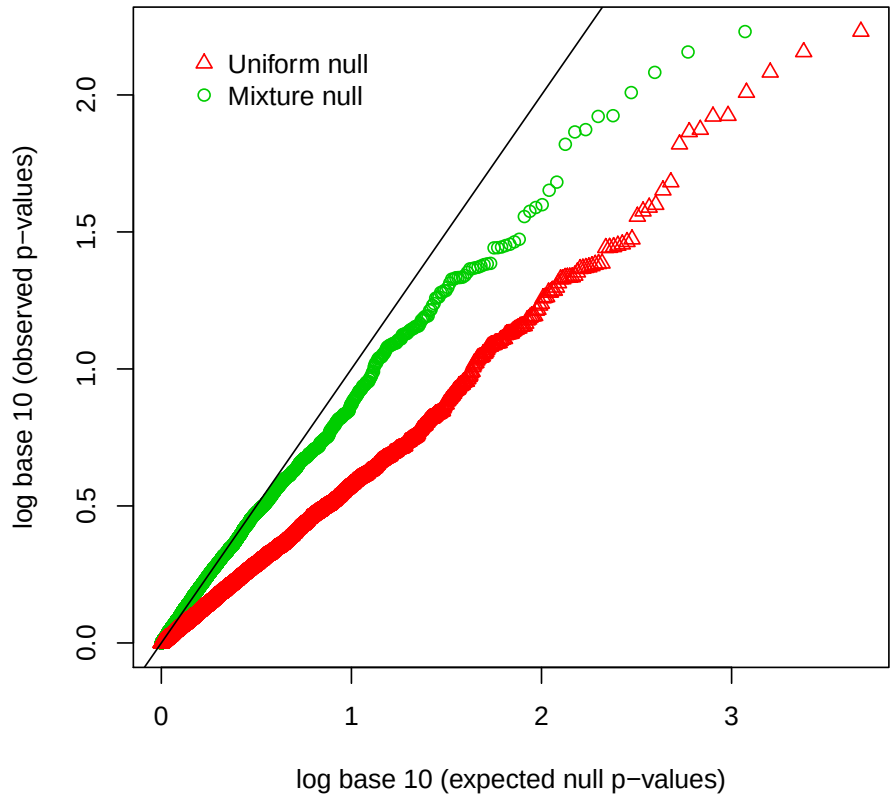
We compute the null distribution of pmax using approximation:

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

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

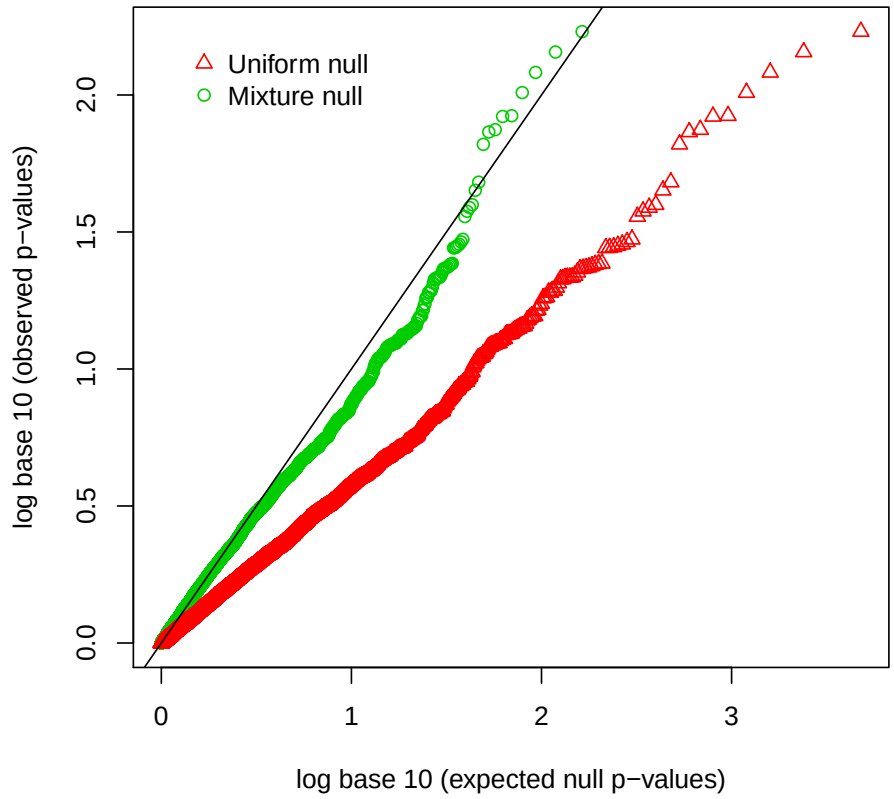
The Q-Q plot based on the approximation method is shown as follows:

```
> 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 3.4.3 (2017-11-30)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.5 LTS

Matrix products: default
BLAS/LAPACK: /app/easybuild/software/OpenBLAS/0.2.18-GCC-5.4.0-2.26-LAPACK-
3.6.1/lib/libopenblas_prescottp-r0.2.18.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_1.0.1

loaded via a namespace (and not attached):
 [1] cp4p_0.3.6          Rcpp_0.12.17        magrittr_1.5
 [4] MASS_7.3-47         BiocGenerics_0.22.1 splines_3.4.3
 [7] fdrtool_1.2.15      munsell_0.4.3       colorspace_1.3-2
[10] lattice_0.20-35     rlang_0.1.4         stringr_1.2.0
[13] plyr_1.8.4          tools_3.4.3         parallel_3.4.3
[16] grid_3.4.3          Biobase_2.36.2      gtable_0.2.0
[19] survival_2.41-3     multtest_2.32.0     lazyeval_0.2.1
[22] tibble_1.3.4        Matrix_1.2-12       reshape2_1.4.2
[25] ggplot2_2.2.1       MESS_0.5.5          qvalue_2.8.0
[28] limma_3.32.1        stringi_1.1.6       compiler_3.4.3
[31] geepack_1.2-1       scales_0.5.0        stats4_3.4.3
[34] geeM_0.10.1
```

References

[1] James Y. Dai, Janet L. Stanford, and Michael LeBlanc. A multiple-testing procedure for high-dimensional mediationhypotheses. *Journal of the American Statistical Association*, 2019, submitted.