MERIT R Package

Yunxiao Li and Yi-Juan Hu

2022-02-22

1 Overview

The MERIT (controlling Monte-carlo Error Rate In large-scale monte-carlo hypothesis Testing) package implements our method MERIT (Li, Hu, and Satten 2022) for large-scale Monte-Carlo (MC) hypothesis testing that controls the Monte-Carlo error rate (MCER). MCER is the probability that any decisions on accepting or rejecting a hypothesis based on MC p-values (i.e., resampling-based p-values such as permutation or bootstrap p-values) are different from decisions based on ideal p-values (based on analytical methods or exhaustive resampling of replicates). MC errors can easily occur in large-scale hypothesis testing because there are at least some (ideal) p-values near the (FDR-based) threshold of significance (Benjamini and Hochberg 1995), which require a larger number of MC replicates than p-values that are far from the threshold. With a finite number of MC replicates, the list of detections can vary when different sets of MC replicates (e.g., generated using different seeds) are used, resulting in lack of reproducibility.

This package contains one function **merit** that implements the MERIT method and one dataset **gene.expression** that is used to illustrate the use of the function.

2 Getting Started

The package should be downloaded from GitHub at **https://github.com/yijuanhu/MERIT** to a local hard drive, installed, and loaded:

```
install.packages("MERIT_1.0.tar.gz", repos=NULL)
library(MERIT)
```

This package is compatible with Windows, Mac, and Linux systems. Note that the R package **doParallel** should be installed as a prerequisite.

3 Input and Output of merit()

The input parameters are:

- **exceedance.matrix**: The $m \times n$ "exceedance" matrix in which the (i, j)th entry has 1 or 0 value indicating, for testing the ith hypothesis, whether the test statistic based on the jth MC replicate exceeds the observed test statistic, where m is the total number of hypotheses and n is the total number of MC replicates. To improve computational efficiency, this matrix can be replaced by the "collapsed" matrix such that each column of the collapsed matrix is the sum of a consecutive number of (e.g., 1-1000, 1001-2000, etc.) columns in the original matrix and the number of columns is thus reduced (e.g., $n/1000$). We recommend at least 100 columns in the collapsed matrix for a good performance of the bootstrap procedure.
- **n.MCreplicate**: The total number of MC replicates, which may differ from the column number of **exceedance.matrix** if the matrix has been collapsed.
- **MCER.type**: A capital letter among 'I,' 'II,' and 'O' (default) corresponding to the type-I, type-II, and overall MCER to be controlled for. The type-I MCER is the probability of rejecting any hypotheses

based on MC p-values which should be accepted based on the ideal p-values, the type-II MCER is the probability of accepting any hypotheses which should be rejected, and the overall MCER is the probability that any decisions on accepting or rejecting a hypothesis based on MC p-values are different from decisions based on ideal p-values.
- **MCER**: The nominal level for the type of MCER specified in **MCER.type**.
- **fdr**: The nominal FDR level with the default 10%.
- **n.cores**: The number of cores to use in parallel computing. The default is 4.
- **seed**: A user-supplied integer seed for the random number generator in the bootstrap procedure. The default is NULL; with the default value, an integer seed will be generated internally.

The output consists of:

- **rejected**: A vector of indices of rejected hypotheses.
- **accepted**: A vector of indices of accepted hypotheses.
- **undecided**: A vector of indices of hypotheses that are neither rejected nor accepted.
- **seed**: The seed that is user supplied or internally generated, stored in case the user wants to reproduce the permutation replicates

4 Analysis of Gene Expression Data from A Prostate Cancer Study

4.1 Data description and preprocessing

In the prostate cancer study published in (Singh et al. 2002), the gene expression data for 6033 genes were generated for 102 subjects comprised of 52 prostate cancer patients and 50 healthy controls. To detect differentially expressed (DE) genes, the p-value on each gene was calculated based on the t-statistic (assuming equal variances in the case and control groups) and 1 million permutation replicates by shuffling the case-control labels. The **gene.expression** object is a list that contains the exceedance matrix (collapsed into 500 columns) and the total number of permutation replicates; no individual-level gene expression data were provided here.

```
data(gene.expression)
dim(gene.expression$exceedance.matrix)
```

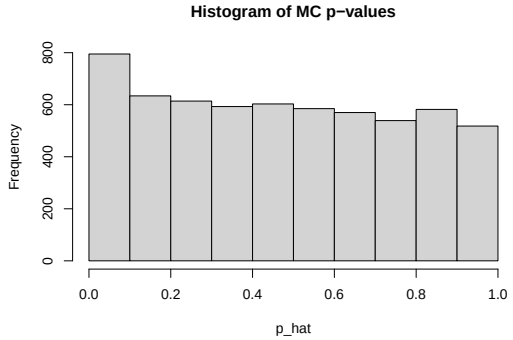
```
## [1] 6033  500
```

```
gene.expression$n.MCreplicate
```

```
## [1] 1e+06
```

Below we calculate the 6033 MC (i.e., permutation) p-values and display the histogram. The enrichment at the left side suggests that a proportion of genes are DE genes.

```
p_hat = (rowSums(gene.expression$exceedance.matrix) + 1) / (gene.expression$n.MCreplicate + 1)
hist(p_hat, main = 'Histogram of MC p-values')
```



4.2 Running merit()

Using these MC p-values, we call **merit()** as follows to detect DE genes by controlling the FDR at 10% and the overall MCER at 10%:

```
system.time({
    res.merit = merit(
        exceedance.matrix = gene.expression$exceedance.matrix,
        n.MCreplicate = gene.expression$n.MCreplicate,
        MCER.type = 'O',
        MCER = 0.1,
        fdr = 0.1,
        seed = 123)
})
user    system  elapsed
144.722  12.876  40.194
```

We take a quick look at the results:

```
length(res.merit$rejected)
[1] 57
length(res.merit$accepted)
[1] 5971
length(res.merit$undecided)
[1] 5
```

MERIT rejected 57 hypotheses, concluding that they are DE genes and ensuring that the probability that any of these hypotheses would be accepted by the ideal p-values is less than 5% (half of the overall MCER). MERIT accepted 5971 hypotheses, concluding that they are non-DE genes and ensuring that the probability that any of these hypotheses would be rejected by the ideal p-values is less than 5% (half of the overall MCER). MERIT made no decisions for 5 hypotheses, meaning that their ideal p-values are too close to the threshold of significance based on 10% FDR and require more permutation replicates in order to make a decision between "rejected" and "accepted."

5 References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Li, Yunxiao, Yijuan Hu, and Glen A Satten. 2022. "MERIT: Controlling Monte-Carlo Error Rate in Large-Scale Monte-Carlo Hypothesis Testing." *bioRxiv*, doi: https://doi.org/10.1101/2022.01.15.476485.

Singh, Dinesh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, et al. 2002. "Gene Expression Correlates of Clinical Prostate Cancer Behavior." *Cancer Cell* 1 (2): 203–9.