Comparison between different local tests: Simes, Simes with Storey and Wilcoxon-Mann-Whitney

2023-07-19

The aim is to compare on real datasets the performance of four closed testing procedures, which respectively use Simes local test with and without Storey estimator for the proportion of true null hypotheses, Wilcoxon-Mann-Whitney local test and the test statistic corresponding to $k = 3$ derived in Theorem 1 of *Testing for outliers with ranks*. Denoting inliers distribution by F, we are going to simulate an outliers distribution corresponding to F^k with $k = 3$ in order to show that closed testing procedure with the local test based on the test statistic corresponding to $k = 3$ is more powerful than closed testing with Simes local test with and without Storey estimator and than closed testing with Wilcoxon-Mann-Whitney local test.

R functions and libraries

```
library(nout)
library(R.matlab)
library(isotree)
library(farff)
library(tictoc)
library(ggplot2)


sim_realdata = function(B, dataset, m1, m, n, l, in_index,
                        out_index=NULL, alpha=m/(l+1), lambda = 0.5){
  k = 3
  m0=m-m1
  if(m1!=0 & is.null(out_index)){
    stop("Error: arg out_index must be initialized.")
  }

  # if(m!=(m1+m0)){
  #   stop("Error: equation m=m1+m0 must be verified.")
  # }

  if(m1!=0){
    tr_ind = sample(in_index, size = n)
    tr = dataset[tr_ind,]
    iso.fo = isolation.forest(tr, ndim=ncol(dataset), ntrees=10, nthreads=1,
                              scoring_metric = "depth", output_score = TRUE)
    in_index2 = setdiff(in_index, tr_ind)

    crit=critWMW(m=m, n=l, alpha=alpha)

    d_WMW = rep(0,B)
    d_WMWk3 = rep(0,B)
    d_Simes = rep(0,B)
    d_StoSimes = rep(0,B)
```

```
  d_BH = rep(0,B)
  d_StoBH = rep(0,B)

  for(b in 1:B){
    cal_ind = sample(in_index2, size = l)
    in_index3 = setdiff(in_index2, cal_ind)
    tein_ind = sample(in_index3, size = m0)
    teoutaug_ind = sample(out_index, size = (k*m1))

    cal = dataset[cal_ind,]

    teout_ind = vector()
    for(j in 1:m1){
      inds = teout_ind[(j+(j-1)*(k-1)):(j+j*(k-1))]
      ind = max(inds)
      teout_ind[j] = ind
    }
    te = dataset[c(tein_ind, teout_ind),]

    S_cal = predict.isolation_forest(iso.fo$model, cal, type = "score")
    S_te = predict.isolation_forest(iso.fo$model, te, type = "score")

    d_WMW[b] = d_MannWhitney(S_X=S_cal, S_Y=S_te, alpha=alpha)
    d_WMWk3[b] = d_MannWhitneyk3(S_X=S_cal, S_Y=S_te, alpha=alpha)
    d_Simes[b] = d_Simes(S_X=S_cal, S_Y=S_te, alpha=alpha)
    d_StoSimes[b] = d_StoreySimes(S_X=S_cal, S_Y=S_te, alpha=alpha)$d
    d_BH[b] = d_benjhoch(S_X=S_cal, S_Y=S_te, alpha=alpha)
    d_StoBH[b] = d_StoreyBH(S_X=S_cal, S_Y=S_te, alpha=alpha)
  }
}

else{
  tr_ind = sample(in_index, size = n)
  tr = dataset[tr_ind,]
  iso.fo = isolation.forest(tr, ndim=ncol(dataset), ntrees=10, nthreads=1,
                            scoring_metric = "depth", output_score = TRUE)
  in_index2 = setdiff(in_index, tr_ind)

  crit=critWMW(m=m, n=l, alpha=alpha)

  d_WMW = rep(0,B)
  d_WMWk3 = rep(0,B)
  d_Simes = rep(0,B)
  d_StoSimes = rep(0,B)
  d_BH = rep(0,B)
  d_StoBH = rep(0,B)

  for(b in 1:B){
    cal_ind = sample(in_index2, size = l)
    in_index3 = setdiff(in_index2, cal_ind)
    te_ind = sample(in_index3, size = m0)

    cal = dataset[cal_ind,]
```

```
      te = dataset[te_ind,]

      S_cal = predict.isolation_forest(iso.fo$model, cal, type = "score")
      S_te = predict.isolation_forest(iso.fo$model, te, type = "score")

      d_WMW[b] = d_MannWhitney(S_X=S_cal, S_Y=S_te, alpha=alpha)
      d_WMWk3[b] = d_MannWhitneyk3(S_X=S_cal, S_Y=S_te, alpha=alpha)
      d_Simes[b] = d_Simes(S_X=S_cal, S_Y=S_te, alpha=alpha)
      d_StoSimes[b] = d_StoreySimes(S_X=S_cal, S_Y=S_te, alpha=alpha)$d
      d_BH[b] = d_benjhoch(S_X=S_cal, S_Y=S_te, alpha=alpha)
      d_StoBH[b] = d_StoreyBH(S_X=S_cal, S_Y=S_te, alpha=alpha)
    }
  }

  discov = as.data.frame(cbind("d_BH"=d_BH, "d_StoBH"=d_StoBH, "d_Simes"=d_Simes,
                               "d_StoSimes"=d_StoSimes, "d_WMW"=d_WMW, "d_WMWk3"=d_WMWk3))
  colnames(discov) = c("BH", "BHSto", "CTSim", "CTSimSto", "CTWMW", "CTWMWk3")
  mean.discov = apply(discov, MARGIN = 2, FUN = mean)

  powerGlobalNull = as.data.frame(cbind("d_BH"=d_BH>0, "d_StoBH"=d_StoBH>0, "d_Simes"=d_Simes>0,
                                        "d_StoSimes"=d_StoSimes>0, "d_WMW"=d_WMW>0,"d_WMWk3"=d_WMWk3>0))
  colnames(powerGlobalNull) = c("BH", "BHSto", "CTSim", "CTSimSto", "CTWMW", "CTWMWk3")
  mean.powerGlobalNull = apply(powerGlobalNull, MARGIN = 2, FUN = mean)

  return(list("discoveries"=discov, "mean.discoveries" = mean.discov,
              "powerGlobalNull"=powerGlobalNull, "mean.powerGlobalNull"=mean.powerGlobalNull,
              "m1"=m1, "alpha"=alpha))

}
```

In the following we set the calibration set and the test set size, respectively l and m, so that the nominal level α is proportional to $\frac{m}{l+1}$. The train set size is equal to n and the number of iterations is $B = 10^5$.

Statlog (Shuttle) dataset

The dataset is available at http://odds.cs.stonybrook.edu/shuttle-dataset

```
set.seed(321)

# Initializing parameters
B = 10
n = 199
l = 199
m = 20
alpha = m/(l+1)
m1s = seq(from=0, to=m, by=1)

data = readMat("~/nout/trials/RealData/Datasets/Dataset shuttle/shuttle.mat")
dataset = cbind(data$X, data$y); colnames(dataset)[ncol(dataset)] = "y"
in_ind = which(dataset[,ncol(dataset)]==0)
out_ind = which(dataset[,ncol(dataset)]==1)

tic()
res = lapply(m1s,
```

```
            function(m1) sim_realdata(B=B, in_index=in_ind, out_index=out_ind,
                                      dataset=dataset,
                                      alpha=alpha,l=l, n=n, m=m, m1=m1))
toc()
```

```
## 18.69 sec elapsed
```

```
# Storing results
store_res = list("mean.discov" = matrix(nrow=length(m1s), ncol = 6),
                 "mean.powerGlobalNull" = matrix(nrow=length(m1s), ncol = 6))
row.names = rep(NA, times=length(m1s))
for(i in 1:length(m1s)){
  row.names[i] = paste("m1 =",m1s[i])
}
rownames(store_res$mean.discov) = row.names
colnames(store_res$mean.discov) = c("BH", "StoBH", "Simes", "StoSimes", "WMW", "WMWk3")
rownames(store_res$mean.powerGlobalNull) = row.names
colnames(store_res$mean.powerGlobalNull) = c("BH", "StoBH", "Simes", "StoSimes", "WMW", "WMWk3")


for(i in 1:length(res)){
  store_res$mean.discov[i,] = res[[i]]$mean.discov
  store_res$mean.powerGlobalNull[i,] = res[[i]]$mean.powerGlobalNull
}

store_res$mean.discov
```

```
##          BH StoBH Simes StoSimes WMW WMWk3
## m1 = 0  0.3   0.2   0.3      0.2 0.5  16.2
## m1 = 1  0.0   0.0   0.0      0.0 0.0  14.9
## m1 = 2  0.0   0.0   0.0      0.0 0.2  13.6
## m1 = 3  0.1   0.0   0.1      0.0 0.0  12.7
## m1 = 4  0.0   0.0   0.0      0.0 0.0  12.2
## m1 = 5  0.1   0.0   0.1      0.0 0.0  11.7
## m1 = 6  0.2   0.0   0.2      0.0 0.0  10.0
## m1 = 7  0.2   0.0   0.1      0.0 0.0   9.3
## m1 = 8  0.0   0.0   0.0      0.0 0.0   8.7
## m1 = 9  0.1   0.0   0.1      0.0 0.0   7.7
## m1 = 10 0.0   0.0   0.0      0.0 0.0   6.9
## m1 = 11 0.0   0.0   0.0      0.0 0.0   6.8
## m1 = 12 0.0   0.0   0.0      0.0 0.0   5.2
## m1 = 13 0.0   0.0   0.0      0.0 0.0   4.6
## m1 = 14 0.0   0.0   0.0      0.0 0.0   4.4
## m1 = 15 0.0   0.0   0.0      0.0 0.0   2.9
## m1 = 16 0.0   0.0   0.0      0.0 0.0   2.1
## m1 = 17 0.1   0.0   0.1      0.0 0.0   1.7
## m1 = 18 0.0   0.0   0.0      0.0 0.0   0.4
## m1 = 19 0.0   0.0   0.0      0.0 0.0   0.3
## m1 = 20 0.0   0.0   0.0      0.0 0.0   0.0
```

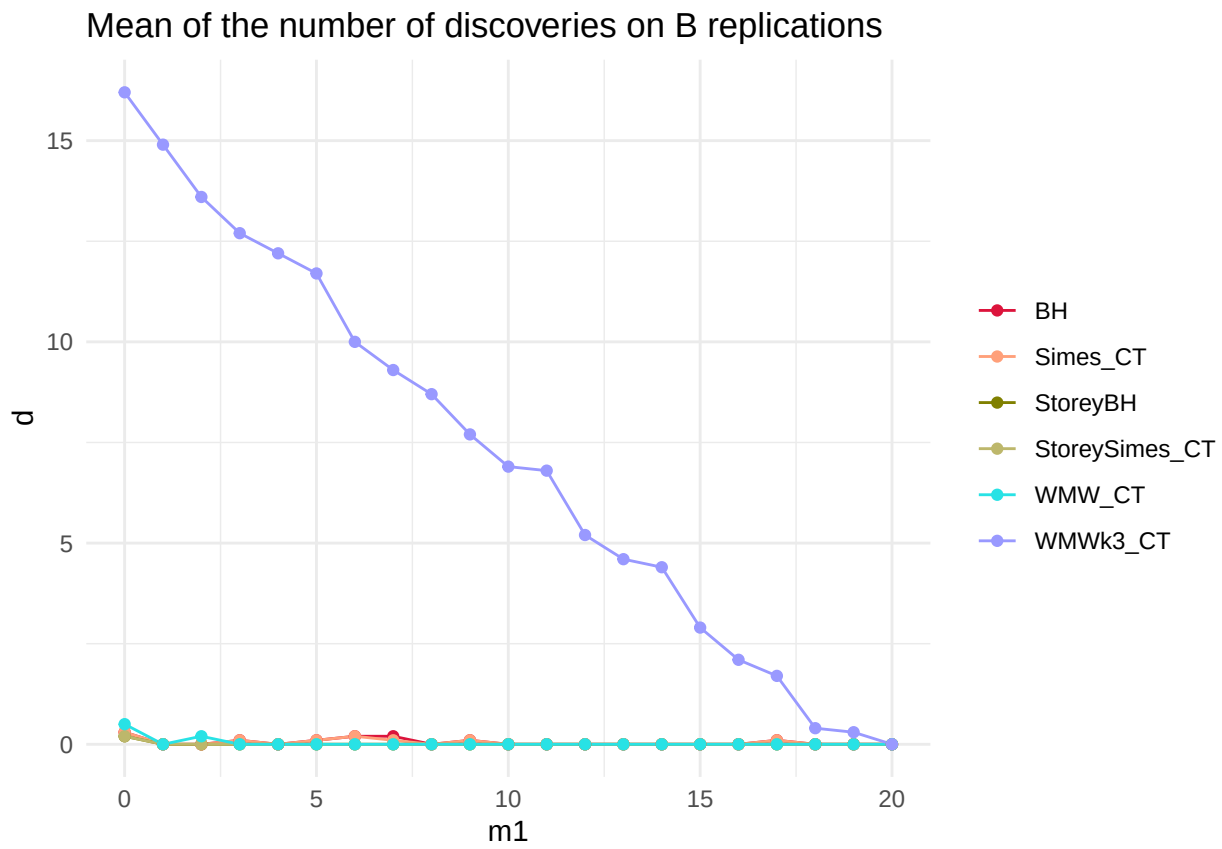
```
store_res$mean.powerGlobalNull
```

```
##          BH StoBH Simes StoSimes WMW WMWk3
## m1 = 0  0.3   0.2   0.3      0.2 0.2   1.0
## m1 = 1  0.0   0.0   0.0      0.0 0.0   1.0
```

```
## m1 = 2  0.0   0.0   0.0      0.0 0.1   1.0
## m1 = 3  0.1   0.0   0.1      0.0 0.0   1.0
## m1 = 4  0.0   0.0   0.0      0.0 0.0   1.0
## m1 = 5  0.1   0.0   0.1      0.0 0.0   1.0
## m1 = 6  0.2   0.0   0.2      0.0 0.0   1.0
## m1 = 7  0.1   0.0   0.1      0.0 0.0   1.0
## m1 = 8  0.0   0.0   0.0      0.0 0.0   1.0
## m1 = 9  0.1   0.0   0.1      0.0 0.0   1.0
## m1 = 10 0.0   0.0   0.0      0.0 0.0   1.0
## m1 = 11 0.0   0.0   0.0      0.0 0.0   1.0
## m1 = 12 0.0   0.0   0.0      0.0 0.0   1.0
## m1 = 13 0.0   0.0   0.0      0.0 0.0   1.0
## m1 = 14 0.0   0.0   0.0      0.0 0.0   1.0
## m1 = 15 0.0   0.0   0.0      0.0 0.0   0.9
## m1 = 16 0.0   0.0   0.0      0.0 0.0   1.0
## m1 = 17 0.1   0.0   0.1      0.0 0.0   1.0
## m1 = 18 0.0   0.0   0.0      0.0 0.0   0.3
## m1 = 19 0.0   0.0   0.0      0.0 0.0   0.3
## m1 = 20 0.0   0.0   0.0      0.0 0.0   0.0
```

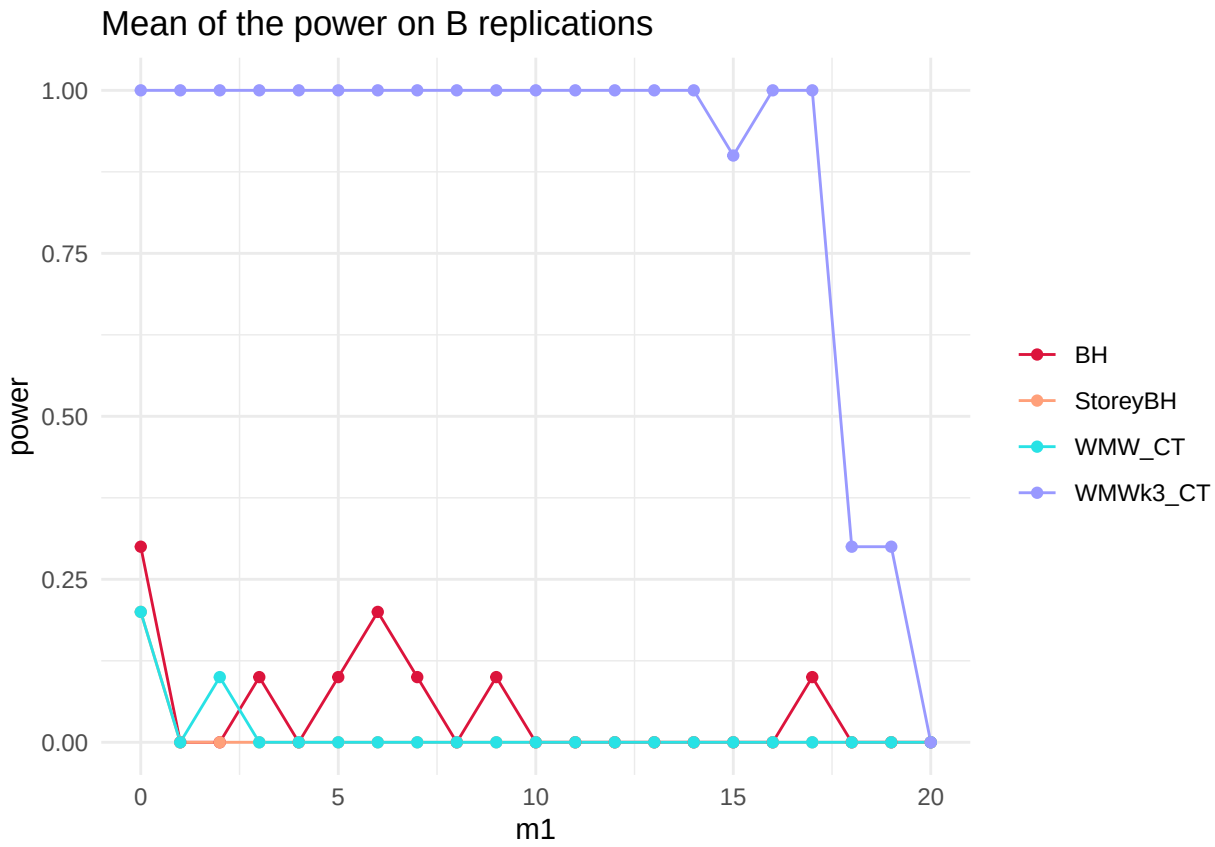
```r
# Plot discoveries
df <- data.frame(
  x = m1s,
  BH = store_res$mean.discov[, 1],
  StoreyBH = store_res$mean.discov[, 2],
  Simes_CT = store_res$mean.discov[, 3],
  StoreySimes_CT = store_res$mean.discov[, 4],
  WMW_CT = store_res$mean.discov[, 5],
  WMWk3_CT = store_res$mean.discov[, 6]
)
df_long <- tidyr::pivot_longer(df, cols = -x, names_to = "group", values_to = "y")

ggplot(df_long, aes(x = x, y = y, color = group)) +
  geom_line() +
  geom_point()+
  scale_color_manual(values = c("#DC143C", "#FFA07A", "#808000", "#BDB76B", 5, "#9999FF")) +
  labs(x = "m1", y = "d", title = "Mean of the number of discoveries on B replications") +
  theme_minimal() +
  theme(legend.title = element_blank())
```




```r
# Plot power
dfpower <- data.frame(
  x = m1s,
  BH = store_res$mean.powerGlobalNull[, 1],
  StoreyBH = store_res$mean.powerGlobalNull[, 2],
  WMW_CT = store_res$mean.powerGlobalNull[, 5],
  WMWk3_CT = store_res$mean.powerGlobalNull[, 6]
)
df_long_power <- tidyr::pivot_longer(dfpower, cols = -x, names_to = "group", values_to = "y")

# Plot the lines with different colors and legends
ggplot(df_long_power, aes(x = x, y = y, color = group)) +
  geom_line() +
  geom_point()+
  scale_color_manual(values = c("#DC143C","#FFA07A",5, "#9999FF")) +
  labs(x = "m1", y = "power", title = "Mean of the power on B replications") +
  theme_minimal() +
  theme(legend.title = element_blank())
```




```
resShuttle = res
save(resShuttle, file="~/nout/trials/RealData/PowerStudy/New!/alpha0.2/ShuttleOnly0.2/resShuttlek3")
```