Comparison between different local tests: Simes, Simes with Storey and Wilcoxon-Mann-Whitney using the natural outliers distribution

2023-07-23

The aim is to compare on real datasets the performance of three closed testing procedures, which respectively use Simes local test with and without Storey estimator for the proportion of true null hypotheses and Wilcoxon-Mann-Whitney local test. We will consider outlier population to be the set of observations tagged as "outlier" in the dataset of interest.

R functions and libraries

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

compact_results = function(res){
  resT=as.data.frame(t(res))

  results = list()
  for(j in 1:length(n1s)){
    discoveries = as.data.frame(
      cbind("d_BH"=unlist(res[[j]][rownames(res[[j]])=="d_BH",]),
            "d_StoBH"=unlist(res[[j]][rownames(res[[j]])=="d_StoBH",]),
            "d_Sim"=unlist(res[[j]][rownames(res[[j]])=="d_Sim",]),
            "d_StoSimes"=unlist(res[[j]][rownames(res[[j]])=="d_StoSimes",]),
            "d_WMW"=unlist(res[[j]][rownames(res[[j]])=="d_WMW",])
            )
    )
    mean.discoveries = apply(discoveries, MARGIN = 2, FUN = mean)

    power.GlobalNull = as.data.frame(discoveries>0)
    mean.powerGlobalNull = apply(power.GlobalNull, MARGIN = 2, FUN = mean)

    out_identification = as.data.frame(
      cbind("out.identif_WMW"=
              unlist(res[[j]][rownames(res[[j]])=="outlier.identified_WMW",]),
            "out.identif_StoSimes"=
              unlist(res[[j]][rownames(res[[j]])=="outlier.identified_StoSimes",]),
            "out.identif_Simes"=
              unlist(res[[j]][rownames(res[[j]])=="outlier.identified_Simes",])
            )
    )
```

```
    mean.out_identification = apply(out_identification, MARGIN = 2, FUN = mean)
    mean.out_identification_pos = apply(out_identification>0, MARGIN = 2, FUN = mean)

    results[[j]] = list("discoveries" = discoveries,
                        "mean.discoveries" = mean.discoveries,
                        "power.GlobalNull" = power.GlobalNull,
                        "mean.powerGlobalNull" = mean.powerGlobalNull,
                        "out_identification" = out_identification,
                        "mean.out_identification" = mean.out_identification,
                        "mean.out_identification>0" = mean.out_identification_pos,
                        "pi.not" = res[[j]][rownames(res[[j]])=="pi.not",],
                        "uniques" = res[[j]][rownames(res[[j]])=="uniques",],
                        "n1" = res[[j]][rownames(res[[j]])=="n1",1],
                        "alpha" = res[[j]][rownames(res[[j]])=="alpha",1])
  }
  return(results)
}


TrainingIsoForest = function(l, dataset){

  tr_ind = sample(in_ind, size = l)
  tr = dataset[tr_ind,]
  isofo.model = isotree::isolation.forest(tr, ndim=ncol(dataset), ntrees=10, nthreads=1,
                            scoring_metric = "depth", output_score = TRUE)$model
  in_index2 = setdiff(in_ind, tr_ind)

  return(list("model"=isofo.model, "inlier_remaining" = in_index2))

}



CompareMethodNaturalOutliers = function(B, n1, n, out_ind, inlier_remaining, isofo.model, dataset){

  n0 = n-n1
  foreach(b = 1:B, .combine=cbind) %dopar% {
    if(n1==0){
      N = n0 + m
      in_index3 = sample(inlier_remaining, size = N)
      cal_ind = in_index3[1:m]
      te_ind = in_index3[(m+1):N]
      cal = dataset[cal_ind,]
      te = dataset[te_ind,]
      S_cal = predict.isolation_forest(isofo.model, cal, type = "score")
      S_te = predict.isolation_forest(isofo.model, te, type = "score")

      d_WMW = nout::d_MannWhitney(S_Y = S_te, S_X = S_cal, alpha=alpha)
      d_Sim = nout::d_Simes(S_X = S_cal, S_Y = S_te, alpha = alpha)
      StoSimes = nout::d_StoreySimes(S_X = S_cal, S_Y = S_te, alpha = alpha)
      d_StoSimes = StoSimes$d
      pi.not = StoSimes$pi.not
      d_BH = nout::d_benjhoch(S_X = S_cal, S_Y = S_te, alpha = alpha)
```

```
  d_StoBH = nout::d_StoreyBH(S_X = S_cal, S_Y = S_te, alpha = alpha)
  uniques = length(unique(c(S_cal, S_te)))
  return(list("d_BH" = d_BH,
              "d_StoBH" = d_StoBH,
              "d_Sim" = d_Sim,
              "d_StoSimes" = d_StoSimes,
              "d_WMW" = d_WMW,
              "outlier.identified_WMW" = 0,
              "outlier.identified_Simes" = 0,
              "outlier.identified_StoSimes" = 0,
              "uniques" = uniques,
              "n1" = n1,
              "pi.not" = pi.not,
              "alpha" = alpha))
}

else{
  N = n0 + m
  in_index3 = sample(inlier_remaining, size = N)
  cal_ind = in_index3[1:m]
  if(n0!=0)
    tein_ind = in_index3[(m+1):N]
  else
    tein_ind = NULL
  teout_ind = sample(out_ind, size = n1)
  cal = dataset[cal_ind,]
  te = dataset[c(tein_ind, teout_ind),]
  S_cal = predict.isolation_forest(isofo.model, cal, type = "score")
  S_te = predict.isolation_forest(isofo.model, te, type = "score")

  d_WMW = nout::d_MannWhitney(S_Y = S_te, S_X = S_cal, alpha=alpha)
  d_Sim = nout::d_Simes(S_X = S_cal, S_Y = S_te, alpha = alpha)
  StoSimes = nout::d_StoreySimes(S_X = S_cal, S_Y = S_te, alpha = alpha)
  d_StoSimes = StoSimes$d
  pi.not = StoSimes$pi.not
  d_BH = nout::d_benjhoch(S_X = S_cal, S_Y = S_te, alpha = alpha)
  d_StoBH = nout::d_StoreyBH(S_X = S_cal, S_Y = S_te, alpha = alpha)
  uniques = length(unique(c(S_cal, S_te)))

  # outlier identification with WMW
  conf.pval = sapply(1:n, function(j) (1+sum(S_cal >= S_te[j]))/(m+1))
  confvalid.pval = conf.pval<alpha
  confvalid.index = which(conf.pval<alpha)

  if(d_WMW>0){
    outlierTF = sapply(confvalid.index, function(h)
        nout::dselection_MannWhitney(S_Y = S_te, S_X = S_cal, S = h, alpha=alpha))
    outlier.identified_WMW = confvalid.index[as.logical(outlierTF)]
  }
  else  outlier.identified_WMW = NULL

  # outlier identification with Simes
  if(d_Sim>0){
```

```
        outlierTF = sapply(confvalid.index, function(h)
            nout::dselection_Simes(S_Y = S_te, S_X = S_cal, S = h, alpha=alpha))
        outlier.identified_Simes = confvalid.index[as.logical(outlierTF)]
      }
      else  outlier.identified_Simes = NULL

      # outlier identification with StoreySimes
      if(d_StoSimes>0){
        outlierTF = sapply(confvalid.index, function(h)
            nout::dselection_StoreySimes(S_Y = S_te, S_X = S_cal, S = h, alpha=alpha))
        outlier.identified_StoSimes = confvalid.index[as.logical(outlierTF)]
      }
      else  outlier.identified_StoSimes = NULL

      return(list("d_BH" = d_BH,
                  "d_StoBH" = d_StoBH,
                  "d_Sim" = d_Sim,
                  "d_StoSimes" = d_StoSimes,
                  "d_WMW" = d_WMW,
                  "outlier.identified_WMW" = length(outlier.identified_WMW),
                  "outlier.identified_Simes" = length(outlier.identified_Simes),
                  "outlier.identified_StoSimes" = length(outlier.identified_StoSimes),
                  "uniques" = uniques,
                  "n1" = n1,
                  "pi.not" = pi.not,
                  "alpha" = alpha))
    }
  }
}


estimatek = function(B, inlier_remaining, out_ind, isofo.model, dataset){
  ress = foreach(b = 1:B, .combine=c) %dopar% {
   inlier_ind = sample(inlier_remaining, size = 1)
   outlier_ind = sample(out_ind, size = 1)
   inlier = dataset[inlier_ind,]
   outlier = dataset[outlier_ind,]
   S_inlier = predict.isolation_forest(isofo.model, inlier, type = "score")
   S_outlier = predict.isolation_forest(isofo.model, outlier, type = "score")

   greater.logi = S_inlier<S_outlier

   return(greater.logi)
  }

  greater.prob = mean(ress)
  k=greater.prob/(1-greater.prob)
  return(k)
}
```

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 $B = 10^4$.

Digits dataset

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

```
set.seed(321)

# Initializing parameters
B = 10^4
m = 199
l = 199
n = 20
alpha = n/(l+1)
n1s = seq(from=0, to=n, by=1)

data = readMat("~/nout/trials/RealData/Datasets/Dataset digits/pendigits.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)

cluster <- makeCluster(parallel::detectCores())
registerDoSNOW(cluster)
clusterEvalQ(cluster, {list(library(isotree), library(nout))})
```

```
## [[1]]
## [[1]][[1]]
## [1] "isotree"   "snow"      "stats"     "graphics"  "grDevices" "utils"
## [7] "datasets"  "methods"   "base"
##
## [[1]][[2]]
##  [1] "nout"      "isotree"   "snow"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"
##
##
## [[2]]
## [[2]][[1]]
## [1] "isotree"   "snow"      "stats"     "graphics"  "grDevices" "utils"
## [7] "datasets"  "methods"   "base"
##
## [[2]][[2]]
##  [1] "nout"      "isotree"   "snow"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"
##
##
## [[3]]
## [[3]][[1]]
## [1] "isotree"   "snow"      "stats"     "graphics"  "grDevices" "utils"
## [7] "datasets"  "methods"   "base"
##
## [[3]][[2]]
##  [1] "nout"      "isotree"   "snow"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"
##
##
## [[4]]
## [[4]][[1]]
```

```
## [1] "isotree"   "snow"      "stats"     "graphics"  "grDevices" "utils"
## [7] "datasets"  "methods"   "base"
##
## [[4]][[2]]
##  [1] "nout"      "isotree"   "snow"      "stats"     "graphics"  "grDevices"
##  [7] "utils"     "datasets"  "methods"   "base"
```

```
clusterExport(cluster, list("n", "m", "l", "in_ind", "out_ind", "dataset", "alpha"))

tic()
modeltrain = TrainingIsoForest(l=l, dataset=dataset)
kest = estimatek(B=B, inlier_remaining=modeltrain$inlier_remaining,
          out_ind=out_ind, isofo.model=modeltrain$model, dataset=dataset)
res = lapply(1:length(n1s),
             function(j) CompareMethodNaturalOutliers(B=B, n1=n1s[j], n=n, dataset=dataset,
                                isofo.model=modeltrain$model,
                                out_ind=out_ind,
                                inlier_remaining=modeltrain$inlier_remaining))
toc()
```

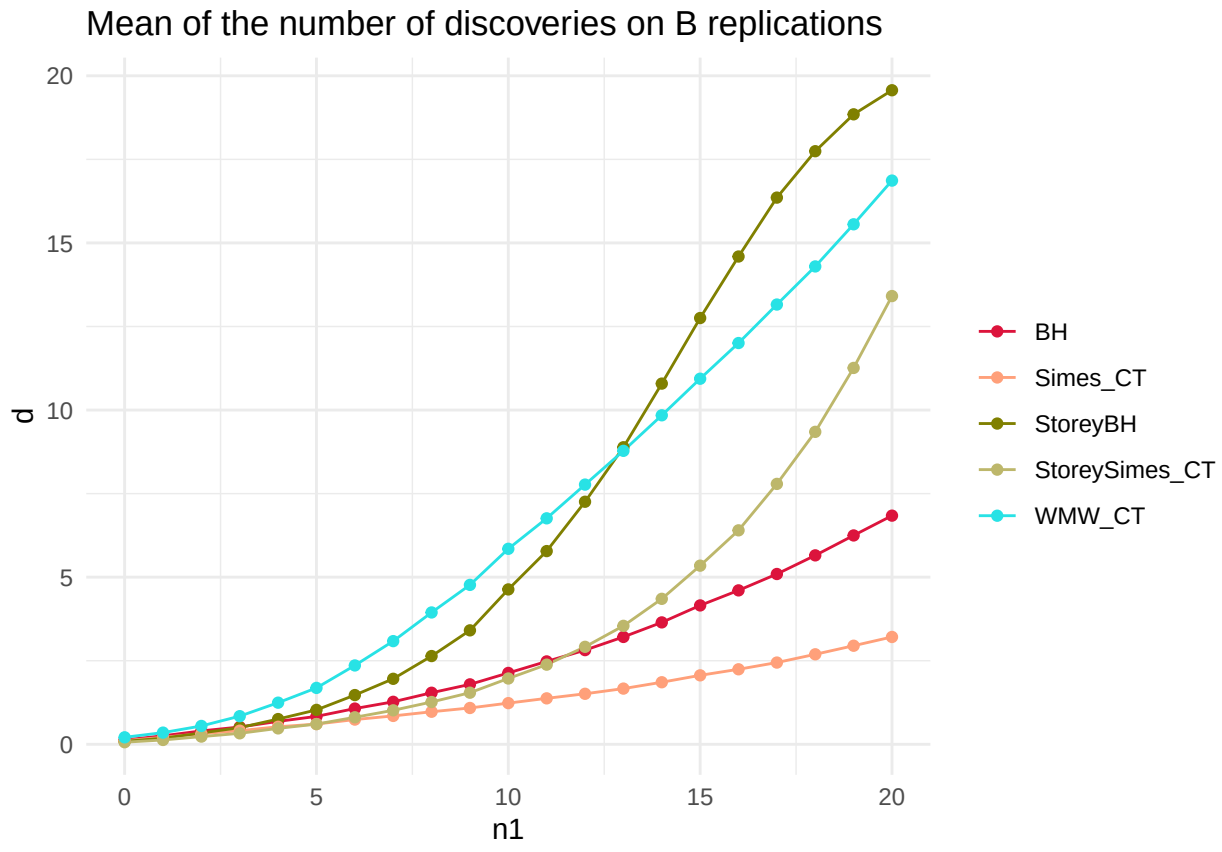
```
## 8926 sec elapsed
```

```
stopCluster(cluster)

kest
```

```
## [1] 8.569378
```

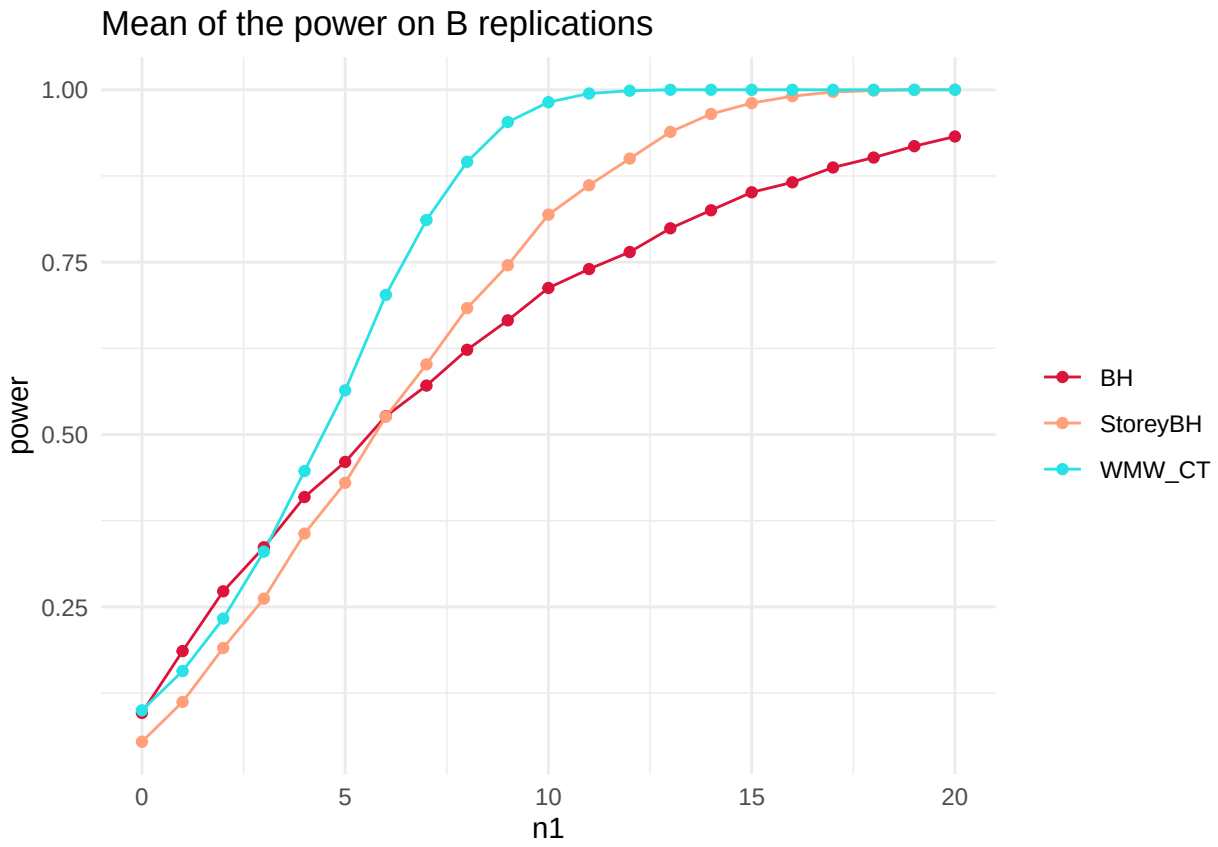
```
results = compact_results(res)

d_BH = vector()
d_StoBH = vector()
d_Sim = vector()
d_StoSimes = vector()
d_WMW = vector()

pow_BH = vector()
pow_StoBH = vector()
pow_Sim = vector()
pow_StoSimes = vector()
pow_WMW = vector()

for(j in 1:length(n1s)){
  d_BH[j] = results[[j]]$mean.discoveries[1]
  d_StoBH[j] = results[[j]]$mean.discoveries[2]
  d_Sim[j] = results[[j]]$mean.discoveries[3]
  d_StoSimes[j] = results[[j]]$mean.discoveries[4]
  d_WMW[j] = results[[j]]$mean.discoveries[5]

  pow_BH[j] = results[[j]]$mean.powerGlobalNull[1]
  pow_StoBH[j] = results[[j]]$mean.powerGlobalNull[2]
  pow_Sim[j] = results[[j]]$mean.powerGlobalNull[3]
  pow_StoSimes[j] = results[[j]]$mean.powerGlobalNull[4]
  pow_WMW[j] = results[[j]]$mean.powerGlobalNull[5]

}
```

```r
# Plot discoveries
df <- data.frame(
  x = n1s,
  BH = d_BH,
  StoreyBH = d_StoBH,
  Simes_CT = d_Sim,
  StoreySimes_CT = d_StoSimes,
  WMW_CT = d_WMW
)
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)) +
  labs(x = "n1", 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 = n1s,
  BH = pow_BH,
  StoreyBH = pow_StoBH,
  WMW_CT = pow_WMW
)
```

```r
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)) +
  labs(x = "n1", y = "power", title = "Mean of the power on B replications") +
  theme_minimal() +
  theme(legend.title = element_blank())
```




```r
outlier.identification = list()
for(i in 1:length(n1s)){
  outlier.identification[[i]] = matrix(nrow = 3, ncol = 4)
  rownames(outlier.identification[[i]]) = c("WMW", "Simes", "StoSimes")
  colnames(outlier.identification[[i]]) = c("mean.out.identif", "%successful.identification",
                                            "mean.d", "mean.d>0(power)")
  outlier.identification[[i]][,1] = apply(
    results[[i]][["out_identification"]], MARGIN = 2, FUN = mean)
  outlier.identification[[i]][,2] = apply(
    results[[i]][["out_identification"]]>0, MARGIN = 2, FUN = mean)
  outlier.identification[[i]][,3] = results[[i]]$mean.discoveries[c(3,4,5)]
  outlier.identification[[i]][,4] = results[[i]]$mean.powerGlobalNull[c(3,4,5)]
}

for(i in 1:length(n1s)){
```

```
  cat("\n")
  cat(paste("n1=", n1s[i]))
  print(outlier.identification[[i]])
}
```

```
##
## n1= 0         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                        0                           0 0.1065          0.0963
## Simes                      0                           0 0.0620          0.0545
## StoSimes                   0                           0 0.2091          0.1001
##
## n1= 1         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   0.5406                      0.1546 0.2086          0.1860
## Simes                 0.0000                      0.0000 0.1294          0.1122
## StoSimes              0.1900                      0.1711 0.3491          0.1570
##
## n1= 2         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   0.8796                      0.2313 0.3166          0.2727
## Simes                 0.0000                      0.0000 0.2300          0.1905
## StoSimes              0.2885                      0.2523 0.5483          0.2331
##
## n1= 3         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   1.3419                      0.3292 0.4068          0.3365
## Simes                 0.0000                      0.0000 0.3283          0.2620
## StoSimes              0.3758                      0.3165 0.8415          0.3302
##
## n1= 4         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                    1.962                      0.4461 0.5240          0.4094
## Simes                  0.000                      0.0000 0.4762          0.3564
## StoSimes               0.479                      0.3824 1.2450          0.4471
##
## n1= 5         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                    2.658                      0.5629 0.6051          0.4603
## Simes                  0.000                      0.0000 0.6045          0.4300
## StoSimes               0.550                      0.4295 1.6869          0.5642
##
## n1= 6         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   3.5232                      0.7010 0.7397          0.5266
## Simes                 0.0000                      0.0000 0.8076          0.5260
## StoSimes              0.6575                      0.4898 2.3590          0.7025
##
## n1= 7         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   4.3676                      0.8102 0.8487          0.5710
## Simes                 0.0000                      0.0000 1.0162          0.6018
## StoSimes              0.7498                      0.5300 3.0867          0.8111
##
## n1= 8         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   5.1696                      0.8954 0.9739          0.6229
## Simes                 0.0000                      0.0000 1.2652          0.6833
## StoSimes              0.8390                      0.5742 3.9439          0.8956
##
## n1= 9         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   5.8725                      0.9530 1.0866          0.6656
## Simes                 0.0000                      0.0000 1.5433          0.7455
```

```
## StoSimes          0.9277                     0.6144 4.7689          0.9532
##
## n1= 10        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               6.5420                     0.9816 1.2317          0.7126
## Simes             0.0000                     0.0000 1.9689          0.8189
## StoSimes          1.0286                     0.6553 5.8484          0.9820
##
## n1= 11        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               7.0426                     0.9945 1.3745          0.7400
## Simes             0.0000                     0.0000 2.3832          0.8615
## StoSimes          1.1162                     0.6800 6.7603          0.9946
##
## n1= 12        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               7.5314                     0.9984 1.5109          0.7647
## Simes             0.0000                     0.0000 2.9157          0.9002
## StoSimes          1.1903                     0.6989 7.7694          0.9986
##
## n1= 13        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               8.0290                     1.0000 1.6662          0.7991
## Simes             0.0000                     0.0000 3.5440          0.9389
## StoSimes          1.2858                     0.7328 8.7801          1.0000
##
## n1= 14        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               8.4931                     1.0000 1.8566          0.8253
## Simes             0.0000                     0.0000 4.3519          0.9649
## StoSimes          1.3889                     0.7601 9.8429          1.0000
##
## n1= 15        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW               9.0256                     1.0000  2.0633          0.8513
## Simes             0.0000                     0.0000  5.3442          0.9807
## StoSimes          1.4957                     0.7816 10.9374          1.0000
##
## n1= 16        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW               9.4928                     0.9999  2.2443          0.8658
## Simes             0.0000                     0.0000  6.4011          0.9907
## StoSimes          1.5899                     0.8001 12.0050          1.0000
##
## n1= 17        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW               9.9547                     1.0000  2.4454          0.8873
## Simes             0.0000                     0.0000  7.7904          0.9969
## StoSimes          1.6760                     0.8195 13.1570          1.0000
##
## n1= 18        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW              10.4097                       1.00  2.6922          0.9017
## Simes             0.0000                       0.00  9.3479          0.9988
## StoSimes          1.7711                       0.83 14.2971          1.0000
##
## n1= 19        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW              10.9167                     1.0000  2.9486          0.9182
## Simes             0.0000                     0.0000 11.2605          0.9999
## StoSimes          1.8907                     0.8552 15.5574          1.0000
##
## n1= 20        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW              11.3955                     1.0000  3.2111          0.9322
```

```
## Simes              0.0000                    0.0000 13.4100          1.0000
## StoSimes           1.9889                    0.8656 16.8653          1.0000
```

```
resDigits0.1 = list("raw.res"=res,
                    "k.est" = kest,
                    "compact.results" = results,
                    "outlier.identification" = outlier.identification)
save(resDigits0.1, file="~/nout/trials/RealData/PowerStudy/FinalSimu/Digits/resDigits0.1")
```