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

2023-07-24

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){
      n0 = n
      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 $B = 10^4$.

Mammography dataset

The dataset is available at http://odds.cs.stonybrook.edu/mammography-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 mammography/mammography.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()
```

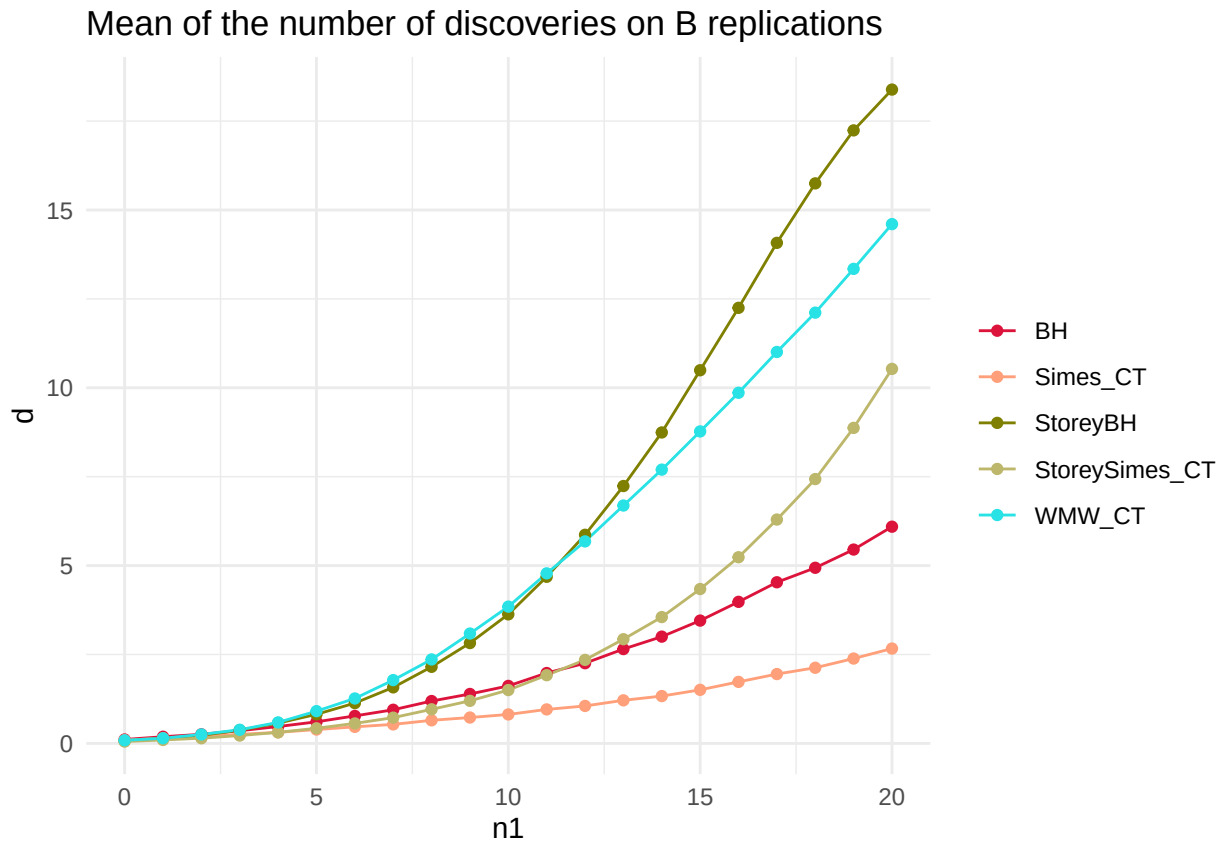
```
## 10031.29 sec elapsed
```

```
stopCluster(cluster)

kest
```

```
## [1] 6.062147
```

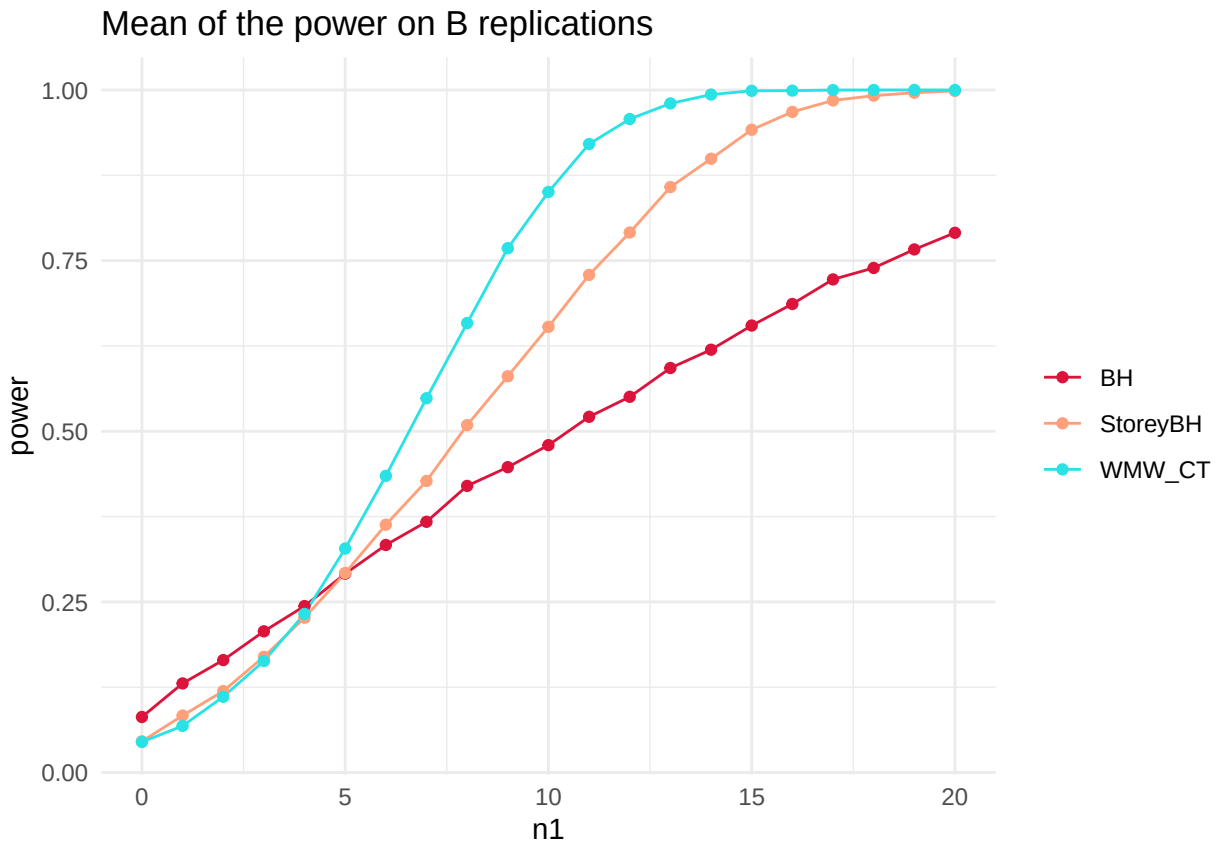
```
results = compact_results(res)

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]
```

```
}

# 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())
```




```
# Plot power
dfpower <- data.frame(
  x = n1s,
  BH = pow_BH,
  StoreyBH = pow_StoBH,
```

```
  WMW_CT = pow_WMW
)
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())
```




```
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 0.0900          0.0814
## Simes                      0                          0 0.0520          0.0457
## StoSimes                   0                          0 0.0866          0.0448
##
## n1= 1         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   0.2441                     0.0671 0.1460          0.1306
## Simes                 0.0000                     0.0000 0.0966          0.0835
## StoSimes              0.1273                     0.1153 0.1453          0.0683
##
## n1= 2         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   0.4392                     0.1101 0.1923          0.1648
## Simes                 0.0000                     0.0000 0.1453          0.1194
## StoSimes              0.1600                     0.1399 0.2519          0.1110
##
## n1= 3         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   0.6976                     0.1627 0.2551          0.2069
## Simes                 0.0000                     0.0000 0.2209          0.1696
## StoSimes              0.2063                     0.1725 0.3836          0.1634
##
## n1= 4         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   1.0528                     0.2317 0.3110          0.2439
## Simes                 0.0000                     0.0000 0.3105          0.2269
## StoSimes              0.2386                     0.1923 0.5908          0.2324
##
## n1= 5         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   1.5796                     0.3276 0.3891          0.2913
## Simes                 0.0000                     0.0000 0.4248          0.2923
## StoSimes              0.2974                     0.2300 0.9071          0.3282
##
## n1= 6         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   2.2191                     0.4341 0.4653          0.3334
## Simes                 0.0000                     0.0000 0.5622          0.3631
## StoSimes              0.3341                     0.2466 1.2632          0.4347
##
## n1= 7         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   2.9725                     0.5480 0.5345          0.3674
## Simes                 0.0000                     0.0000 0.7241          0.4272
## StoSimes              0.3628                     0.2648 1.7797          0.5484
##
## n1= 8         mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW                   3.7822                     0.6584 0.6491          0.4200
## Simes                 0.0000                     0.0000 0.9583          0.5090
## StoSimes              0.4293                     0.3002 2.3609          0.6585
##
## n1= 9         mean.out.identif %successful.identification mean.d mean.d>0(power)
```

```
## WMW               4.6178                     0.7680 0.7276          0.4474
## Simes             0.0000                     0.0000 1.1965          0.5806
## StoSimes          0.4627                     0.3119 3.0889          0.7682
##
## n1= 10        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               5.4499                     0.8504 0.8145          0.4797
## Simes             0.0000                     0.0000 1.5003          0.6530
## StoSimes          0.4908                     0.3249 3.8460          0.8506
##
## n1= 11        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               6.1948                     0.9202 0.9553          0.5212
## Simes             0.0000                     0.0000 1.9176          0.7291
## StoSimes          0.5465                     0.3440 4.7816          0.9209
##
## n1= 12        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               6.8441                     0.9575 1.0529          0.5505
## Simes             0.0000                     0.0000 2.3482          0.7912
## StoSimes          0.5697                     0.3481 5.6806          0.9575
##
## n1= 13        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               7.4198                     0.9802 1.2115          0.5927
## Simes             0.0000                     0.0000 2.9296          0.8578
## StoSimes          0.6410                     0.3737 6.6890          0.9803
##
## n1= 14        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               7.8814                     0.9933 1.3304          0.6195
## Simes             0.0000                     0.0000 3.5527          0.8993
## StoSimes          0.6418                     0.3745 7.6990          0.9933
##
## n1= 15        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               8.3870                     0.9987 1.5062          0.6549
## Simes             0.0000                     0.0000 4.3406          0.9418
## StoSimes          0.6912                     0.3912 8.7730          0.9988
##
## n1= 16        mean.out.identif %successful.identification mean.d mean.d>0(power)
## WMW               8.8493                     0.9991 1.7296          0.6865
## Simes             0.0000                     0.0000 5.2352          0.9679
## StoSimes          0.7438                     0.3948 9.8600          0.9991
##
## n1= 17        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW               9.3266                     0.9999  1.9498          0.7225
## Simes             0.0000                     0.0000  6.2948          0.9847
## StoSimes          0.8058                     0.4173 11.0076          0.9999
##
## n1= 18        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW               9.6857                     1.0000  2.1272          0.7393
## Simes             0.0000                     0.0000  7.4332          0.9917
## StoSimes          0.8195                     0.4158 12.1130          1.0000
##
## n1= 19        mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW              10.1539                     1.0000  2.3863          0.7664
## Simes             0.0000                     0.0000  8.8726          0.9960
## StoSimes          0.9086                     0.4381 13.3442          1.0000
##
```

```
## n1= 20         mean.out.identif %successful.identification  mean.d mean.d>0(power)
## WMW               10.6486                     1.0000  2.6670          0.7908
## Simes              0.0000                     0.0000 10.5311          0.9987
## StoSimes           0.9524                     0.4424 14.6032          1.0000
```

```
resMammo0.1 = list("raw.res"=res,
                   "k.est" = kest,
                   "compact.results" = results,
                   "outlier.identification" = outlier.identification)
save(resMammo0.1, file="~/nout/trials/RealData/PowerStudy/FinalSimu/Cover/resMammo0.1")
```