FurtherOptimizationSci

xiucheng quek

22 June 2016

Read in tables, there are 3 tables. ranked.tsv contains the cumulative score of the top parameters. combination.index.txt contains the table that maps the combination the the combination\_id in ranked.tsv. seq.index.txt containst the index for the sequence pair

rankTable <- read.table('ranked.tsv', sep='\t', header=TRUE)  
comb <- read.table('combination.index.txt', sep='\t', header=TRUE)  
seq\_pair <- read.table('seq.index.txt', sep='\t', header=TRUE)

## Get the top 10 sci score

mergedTable <- merge(rankTable, comb, by.x='combination', by.y='index')  
mergedTable <- mergedTable[order(mergedTable$score),]  
mergedTable$meanScore <- mergedTable$score / 8976  
write.table(head(mergedTable, n = 100) , file='top100.tsv', sep ='\t', quote = F, col.names =T, row.names = F)  
write.table(mergedTable, file='all\_ranked\_with\_comb.tsv' , sep='\t', quote=F, col.names=T, row.names=F)

## Factor Analysis to Confirm findgs

Based on the rank score above, it appears that varying T and S has no impact on the SCI score. t of 0.4 should be used instesad

To confirm our findings, I did a factor analysis using anova

# First read in table containing sci score  
sci\_scores <- read.table('merged\_sci\_diff.txt', sep='\t', header=TRUE)  
# factorize   
as.factor(sci\_scores$sequence) -> sci\_scores$sequence  
as.factor(sci\_scores$combination) -> sci\_scores$combination  
  
  
## merged data  
merged\_scores <- merge(sci\_scores, comb, by.x='combination', by.y ='index')

Due to the large number of combination of factors, I did one way anova for first without testing for factors independence. They will be corrected using tukey honestly significant difference test for multiple comparisions for each factors

# One way anova

Appears only that t values score are the main factor for difference in score.

taov<-aov(sci.diff ~ as.factor(t) + factor(k) + factor(S) + factor(T), data = merged\_scores)  
TukeyHSD(taov)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = sci.diff ~ as.factor(t) + factor(k) + factor(S) + factor(T), data = merged\_scores)  
##   
## $`as.factor(t)`  
## diff lwr upr p adj  
## 0.5-0.4 -0.002698883 -0.003504169 -0.001893597 0  
## 0.6-0.4 -0.012728001 -0.013533287 -0.011922714 0  
## 0.8-0.4 -0.050624180 -0.051429466 -0.049818893 0  
## 0.6-0.5 -0.010029118 -0.010834404 -0.009223831 0  
## 0.8-0.5 -0.047925297 -0.048730583 -0.047120010 0  
## 0.8-0.6 -0.037896179 -0.038701465 -0.037090893 0  
##   
## $`factor(k)`  
## diff lwr upr p adj  
## 0.2-0.1 1.841778e-04 -0.001383814 0.001752170 0.9999978  
## 0.3-0.1 1.840073e-04 -0.001383985 0.001751999 0.9999978  
## 0.4-0.1 1.840073e-04 -0.001383985 0.001751999 0.9999978  
## 0.5-0.1 1.840073e-04 -0.001383985 0.001751999 0.9999978  
## 0.6-0.1 1.840073e-04 -0.001383985 0.001751999 0.9999978  
## 0.7-0.1 1.840073e-04 -0.001383985 0.001751999 0.9999978  
## 0.8-0.1 1.840073e-04 -0.001383985 0.001751999 0.9999978  
## 0.9-0.1 1.840073e-04 -0.001383985 0.001751999 0.9999978  
## 1-0.1 1.840073e-04 -0.001383985 0.001751999 0.9999978  
## 0.3-0.2 -1.704685e-07 -0.001568163 0.001567822 1.0000000  
## 0.4-0.2 -1.704685e-07 -0.001568163 0.001567822 1.0000000  
## 0.5-0.2 -1.704685e-07 -0.001568163 0.001567822 1.0000000  
## 0.6-0.2 -1.704685e-07 -0.001568163 0.001567822 1.0000000  
## 0.7-0.2 -1.704685e-07 -0.001568163 0.001567822 1.0000000  
## 0.8-0.2 -1.704685e-07 -0.001568163 0.001567822 1.0000000  
## 0.9-0.2 -1.704685e-07 -0.001568163 0.001567822 1.0000000  
## 1-0.2 -1.704685e-07 -0.001568163 0.001567822 1.0000000  
## 0.4-0.3 -7.743806e-15 -0.001567992 0.001567992 1.0000000  
## 0.5-0.3 -1.804112e-14 -0.001567992 0.001567992 1.0000000  
## 0.6-0.3 -1.207368e-14 -0.001567992 0.001567992 1.0000000  
## 0.7-0.3 -1.348921e-14 -0.001567992 0.001567992 1.0000000  
## 0.8-0.3 -1.695866e-14 -0.001567992 0.001567992 1.0000000  
## 0.9-0.3 -7.716050e-15 -0.001567992 0.001567992 1.0000000  
## 1-0.3 -4.857226e-15 -0.001567992 0.001567992 1.0000000  
## 0.5-0.4 -1.029732e-14 -0.001567992 0.001567992 1.0000000  
## 0.6-0.4 -4.329870e-15 -0.001567992 0.001567992 1.0000000  
## 0.7-0.4 -5.745404e-15 -0.001567992 0.001567992 1.0000000  
## 0.8-0.4 -9.214851e-15 -0.001567992 0.001567992 1.0000000  
## 0.9-0.4 2.775558e-17 -0.001567992 0.001567992 1.0000000  
## 1-0.4 2.886580e-15 -0.001567992 0.001567992 1.0000000  
## 0.6-0.5 5.967449e-15 -0.001567992 0.001567992 1.0000000  
## 0.7-0.5 4.551914e-15 -0.001567992 0.001567992 1.0000000  
## 0.8-0.5 1.082467e-15 -0.001567992 0.001567992 1.0000000  
## 0.9-0.5 1.032507e-14 -0.001567992 0.001567992 1.0000000  
## 1-0.5 1.318390e-14 -0.001567992 0.001567992 1.0000000  
## 0.7-0.6 -1.415534e-15 -0.001567992 0.001567992 1.0000000  
## 0.8-0.6 -4.884981e-15 -0.001567992 0.001567992 1.0000000  
## 0.9-0.6 4.357625e-15 -0.001567992 0.001567992 1.0000000  
## 1-0.6 7.216450e-15 -0.001567992 0.001567992 1.0000000  
## 0.8-0.7 -3.469447e-15 -0.001567992 0.001567992 1.0000000  
## 0.9-0.7 5.773160e-15 -0.001567992 0.001567992 1.0000000  
## 1-0.7 8.631984e-15 -0.001567992 0.001567992 1.0000000  
## 0.9-0.8 9.242607e-15 -0.001567992 0.001567992 1.0000000  
## 1-0.8 1.210143e-14 -0.001567992 0.001567992 1.0000000  
## 1-0.9 2.858824e-15 -0.001567992 0.001567992 1.0000000  
##   
## $`factor(S)`  
## diff lwr upr p adj  
## 5-1 -4.318788e-06 -0.0008096051 0.0008009675 0.9999991  
## 10-1 -9.240951e-06 -0.0008145272 0.0007960453 0.9999908  
## 50-1 -5.270059e-05 -0.0008579869 0.0007525857 0.9983111  
## 10-5 -4.922163e-06 -0.0008102084 0.0008003641 0.9999986  
## 50-5 -4.838181e-05 -0.0008536681 0.0007569045 0.9986910  
## 50-10 -4.345964e-05 -0.0008487459 0.0007618266 0.9990496  
##   
## $`factor(T)`  
## diff lwr upr p adj  
## 5-1 5.515105e-05 -0.0005810781 0.0006913802 0.9775029  
## 10-1 5.515105e-05 -0.0005810781 0.0006913802 0.9775029  
## 10-5 2.969847e-15 -0.0006362292 0.0006362292 1.0000000

## Lets bootstrap sequences and do anova test within each sequence pair

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.