Testing `choose_pc` Function

Robin

10/14/2019

Function Choose PC

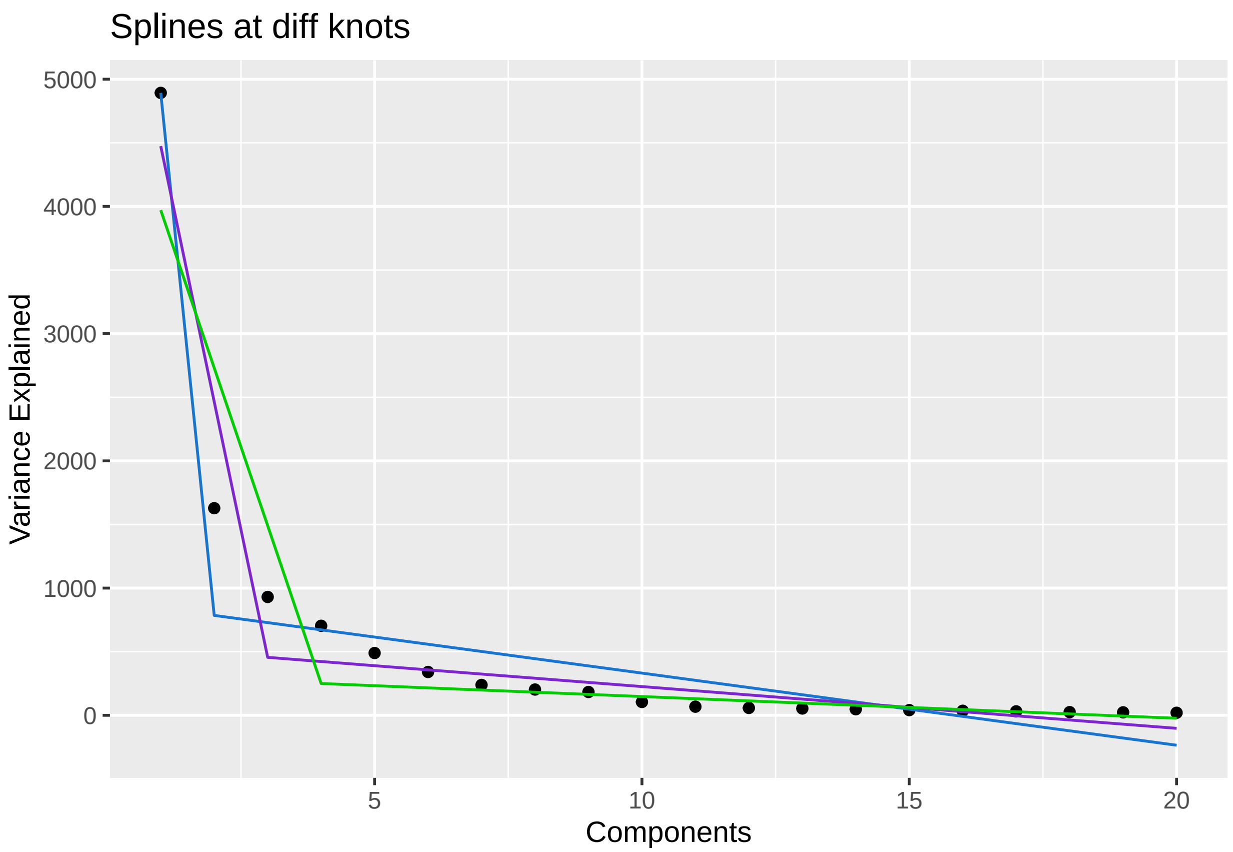
```
#' Choose PC Function
#'
#' An automatic method of chooseing the number of principal components using the elbow method.
#' From the top 20 components, choose the best linear single knot spline. The linear spline with knot a
#'  is the number of components we should choose.
#'
#' @param d the vector of diagonals from an SVD
#' @param total_component default is 20. The most number of components to include when fitting linear s
#' @param return_all default FALSE.  If TRUE, then data.frame of mean squared error for the linear spli
#' @return either return the number optimal number of principal components or additionally, the data.fr
#' with knot for each principal component is outputted

choose_pc <- function(d, total_component = 20, return_all = F){

  n_diag <- min(length(d), total_component) # capping at the number of components we look at to be the
  upper <- n_diag - 1
  x <- 1:n_diag
  d.square <- d[1:n_diag]^2

  fit_list <- sapply(2:upper, function(xx) mean((d.square - lm(d.square ~ bs(x, degree = 1, knots = xx)
  res_dat <- data.frame(n_component = 2:upper, mse = fit_list)

  if(return_all == T){
    return(list("n_component" = res_dat[which.min(res_dat$mse), "n_component"], "MSE" = res_dat))
  }else{
    return(c("n_component" = res_dat[which.min(res_dat$mse), "n_component"]));
  }

}
```

```
mouse <- fread('../contrastive/experiments/datasets/Data_Cortex_Nuclear.csv')
mouse <- NAtoZero(mouse)

targ <- mouse[Behavior == "S/C" & Treatment == "Saline" & Genotype %in% c("Control", "Ts65Dn"), .SD, .S
background <- mouse[Behavior == "S/C" & Treatment == "Saline" & Genotype == "Ts65Dn", .SD, .SDcols = -c
mouse_labels <- targ[, .SD, .SDcols = c("Genotype")]

some_d <- svd(scale(background[, .SD, .SDcols = -c("Genotype", "Treatment", "Behavior")]))$d
some_d2 <- some_d[1:20]^2
x <- 1:20

choose_pc(some_d, total_component = 20, return_all = T)
```

```
## $n_component
## [1] 2
```

```
##
## $MSE
##    n_component       mse
## 1            2  62962.23
## 2            3  64405.17
## 3            4 134158.24
## 4            5 202551.60
## 5            6 265625.32
## 6            7 323274.25
## 7            8 375878.96
## 8            9 422599.20
## 9           10 463312.77
## 10          11 500417.87
## 11          12 534679.95
## 12          13 566183.64
## 13          14 595018.79
## 14          15 621411.40
## 15          16 645659.87
## 16          17 668021.20
## 17          18 688797.67
## 18          19 708658.79
```

```
n_knots <- choose_pc(some_d, total_component = 20)
predicted2 <- lm(some_d2 ~ bs(x, degree = 1, knots = n_knots))$fitted.value
predicted3 <- lm(some_d2 ~ bs(x, degree = 1, knots = 3))$fitted.value
predicted4 <- lm(some_d2 ~ bs(x, degree = 1, knots = 4))$fitted.value
plot_dat <- data.table( x, "d.squared" = some_d2, predicted2, predicted3, predicted4)

ggplot(data = plot_dat)+
  geom_point(aes(x = x, y = d.squared))+
  geom_line(aes(x = x, y = predicted2), color = "dodgerblue3")+
  geom_line(aes(x = x, y = predicted3), color = "purple3")+
  geom_line(aes(x = x, y = predicted4), color = "green3")+
  labs(title = "Splines at diff knots", x = "Components", y = "Variance Explained")
```


Splines at diff knots
Variance Explained
5000
4000
3000
2000
1000
0
5
10
15
20
Components