Multicolinearity

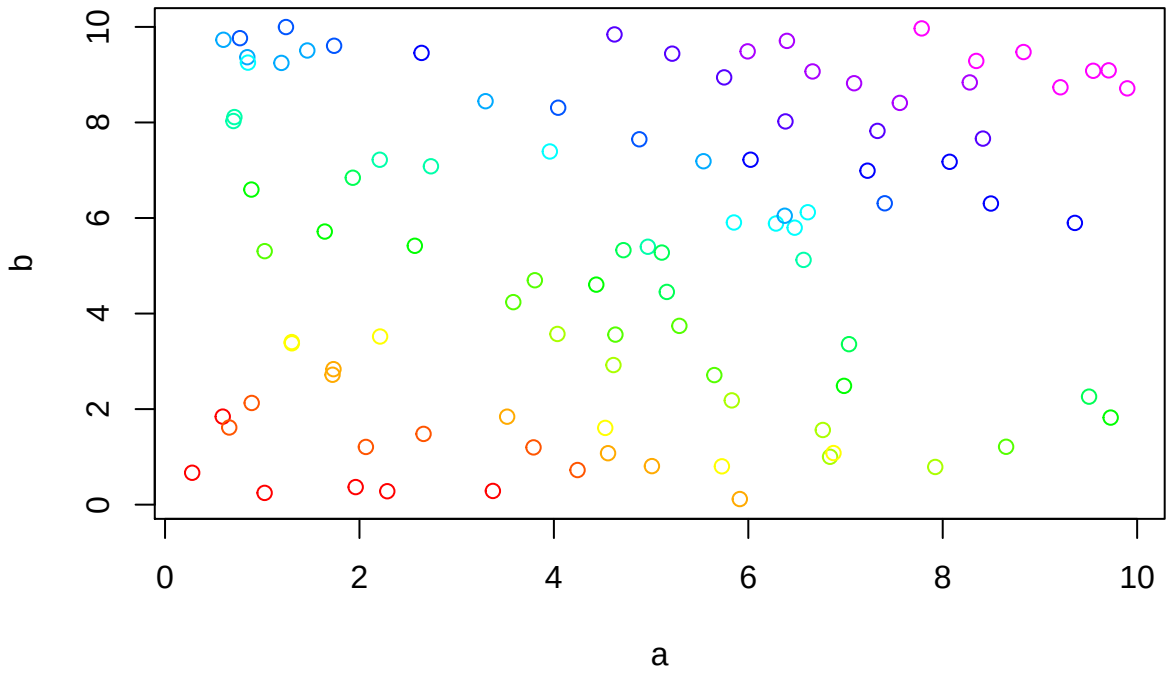
Bob Horton

March 18, 2015

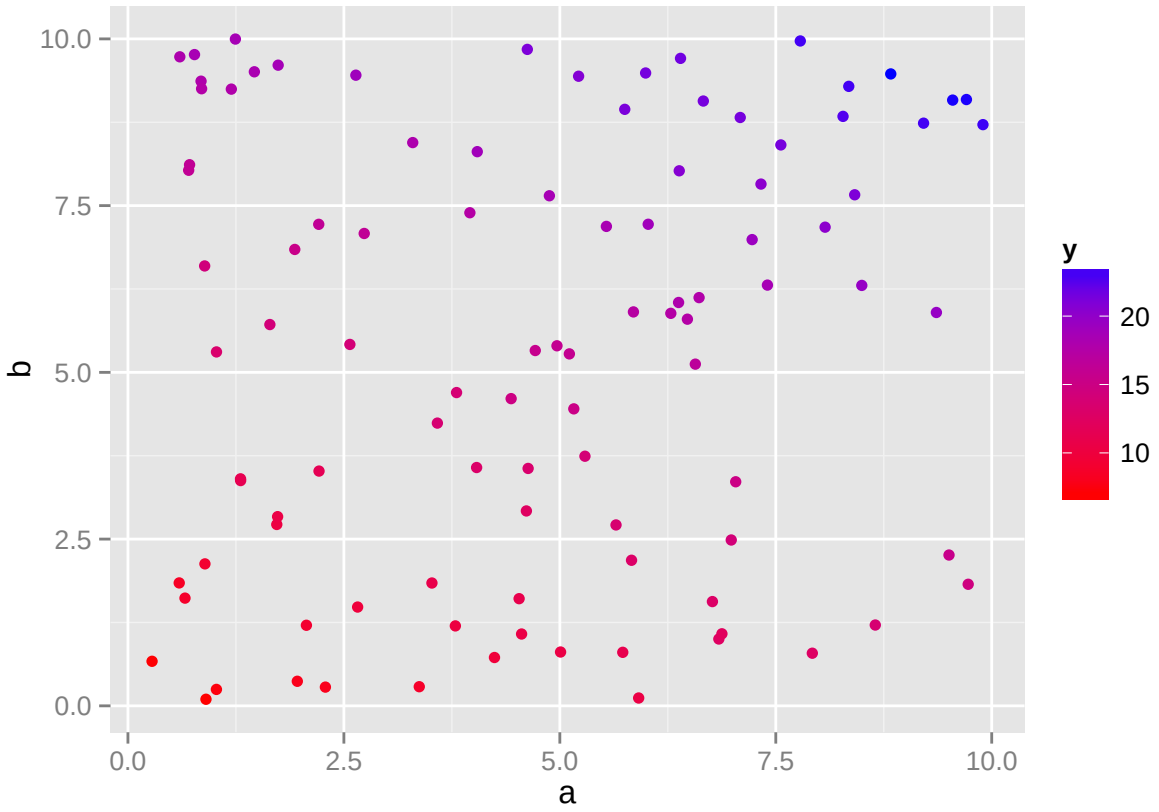
```
N <- 1e2    # 1e5

a <- runif(N, min=0, max=10)
b <- runif(N, min=0, max=10)

s <- 1.2

y <- 6 + 0.7 * a + 1.2 * b + rnorm(N, sd=0.2)

df1 <- data.frame(a, b, y)
```

To display three dimensions, we can use color. Here we break the y values into a series of ranges, and assign a color to each range. Colors are made by the `rainbow` function, which makes a series of hues spanning the spectrum.

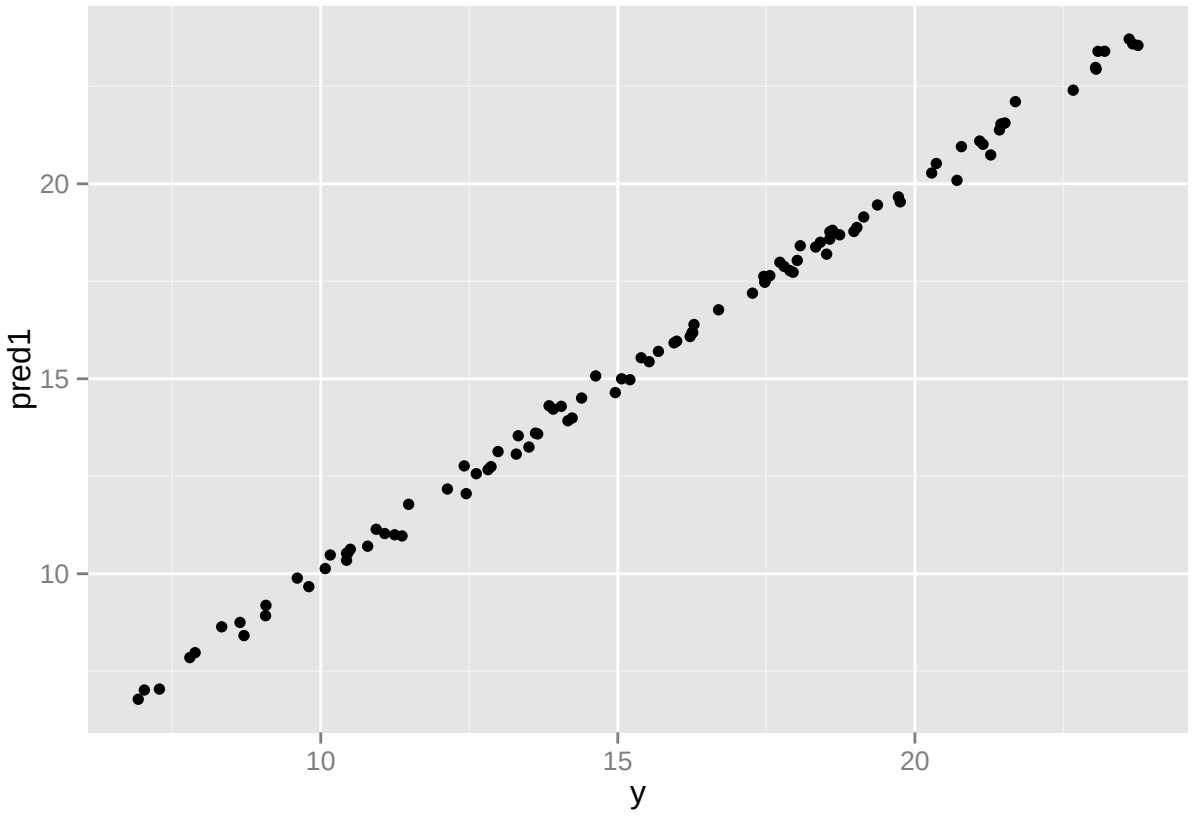
```
df1$y_bucket <- cut(y, breaks=quantile(y, probs=0:16/16))
rbow <- rainbow(16, end=5/6)

with(df1, plot(x=a, y=b, col=rbow[y_bucket]))
```




```
library(ggplot2)
ggplot(df1, aes(x=a, y=b, col=y)) +
    geom_point() +
    scale_colour_gradient(low="red", high="blue")
```




```
fit1 <- lm( y ~ a + b, data=df1)

df1$pred1 <- fit1$fitted

ggplot(df1, aes(x=y, y=pred1)) + geom_point()
```




```
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ a + b, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46413 -0.11276 -0.00421  0.13112  0.61943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.025442   0.050667  118.92   <2e-16 ***
## a           0.707009   0.007601   93.02   <2e-16 ***
## b           1.190614   0.006488  183.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2074 on 97 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.9979
## F-statistic: 2.39e+04 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
###
# (I want the fit to be good, but the coefficients to be non-significant.
# Try more multicollinear columns:
```

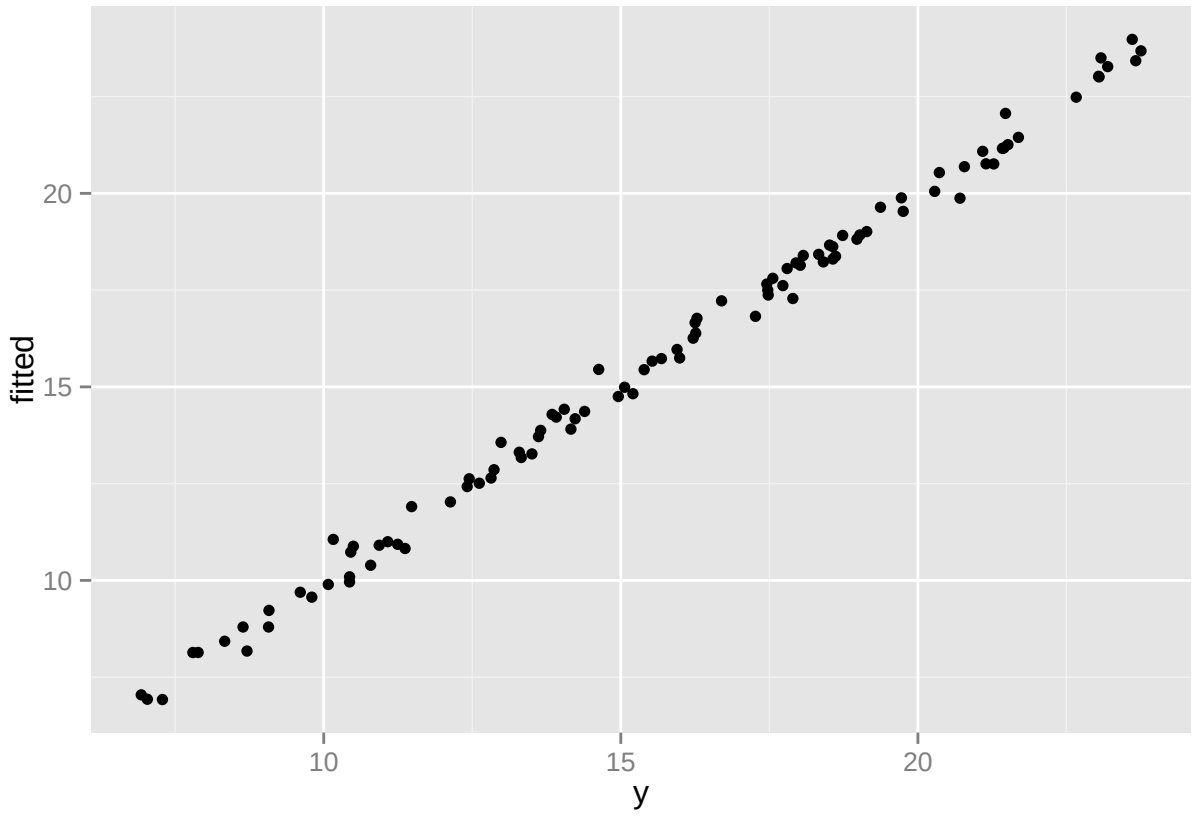
Matrix version

```
num_a_cols <- 20
num_b_cols <- 20

X_signal <- matrix( c(rep(a, num_a_cols), rep(b, num_b_cols)), ncol=(num_a_cols + num_b_cols) )
X_noise <- matrix( rnorm( (num_a_cols + num_b_cols) * N), ncol=(num_a_cols + num_b_cols) )
X <- X_signal + X_noise
df2 <- cbind(data.frame(X), y)

fitA <- lm(y ~ ., data=df2)
summary(fitA)
```

```
##
## Call:
## lm(formula = y ~ ., data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89915 -0.17752  0.01199  0.23160  0.83308
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.133031   0.118062  51.948  < 2e-16 ***
## X1           0.055086   0.044137   1.248  0.21694
## X2          -0.011065   0.045692  -0.242  0.80950
## X3          -0.001581   0.051742  -0.031  0.97572
## X4           0.079500   0.051053   1.557  0.12477
## X5           0.011193   0.045600   0.245  0.80695
## X6           0.022033   0.048339   0.456  0.65022
## X7           0.114104   0.047098   2.423  0.01849 *
## X8           0.024753   0.053120   0.466  0.64295
## X9           0.064318   0.048954   1.314  0.19398
## X10          0.097650   0.052834   1.848  0.06958 .
## X11         -0.025880   0.049787  -0.520  0.60514
## X12          0.012657   0.048736   0.260  0.79600
## X13          0.080667   0.045254   1.783  0.07981 .
## X14          0.020522   0.051121   0.401  0.68954
## X15          0.030191   0.053860   0.561  0.57723
## X16          0.008788   0.048030   0.183  0.85544
## X17          0.034722   0.048889   0.710  0.48037
## X18          0.016917   0.050376   0.336  0.73820
## X19          0.021827   0.052303   0.417  0.67796
## X20          0.030398   0.050787   0.599  0.55177
## X21          0.026390   0.051089   0.517  0.60740
## X22          0.045101   0.058766   0.767  0.44586
## X23          0.101677   0.046013   2.210  0.03102 *
## X24          0.148191   0.046768   3.169  0.00243 **
## X25          0.137351   0.049669   2.765  0.00758 **
## X26          0.003463   0.050277   0.069  0.94532
## X27          0.102189   0.045662   2.238  0.02901 *
## X28          0.032087   0.054745   0.586  0.56004
## X29          0.055258   0.048217   1.146  0.25642
## X30          0.032695   0.047939   0.682  0.49791
## X31          0.077780   0.056593   1.374  0.17452
```

```
## X32          0.101028   0.048240   2.094  0.04054 *
## X33          0.069847   0.046223   1.511  0.13611
## X34          0.020837   0.059237   0.352  0.72628
## X35          0.065873   0.051535   1.278  0.20617
## X36         -0.016279   0.058713  -0.277  0.78255
## X37          0.018915   0.051486   0.367  0.71465
## X38          0.049415   0.057593   0.858  0.39436
## X39          0.037509   0.052277   0.718  0.47590
## X40          0.085809   0.049364   1.738  0.08738 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3966 on 59 degrees of freedom
## Multiple R-squared:  0.9955, Adjusted R-squared:  0.9924
## F-statistic: 325.9 on 40 and 59 DF,  p-value: < 2.2e-16
```

```
df2$fitted <- fitA$fitted
# plot(y ~ fitted, data=df2)
ggplot(df2, aes(x=y, y=fitted)) + geom_point()
```




```
df2$fitted <- fitA$fitted
```

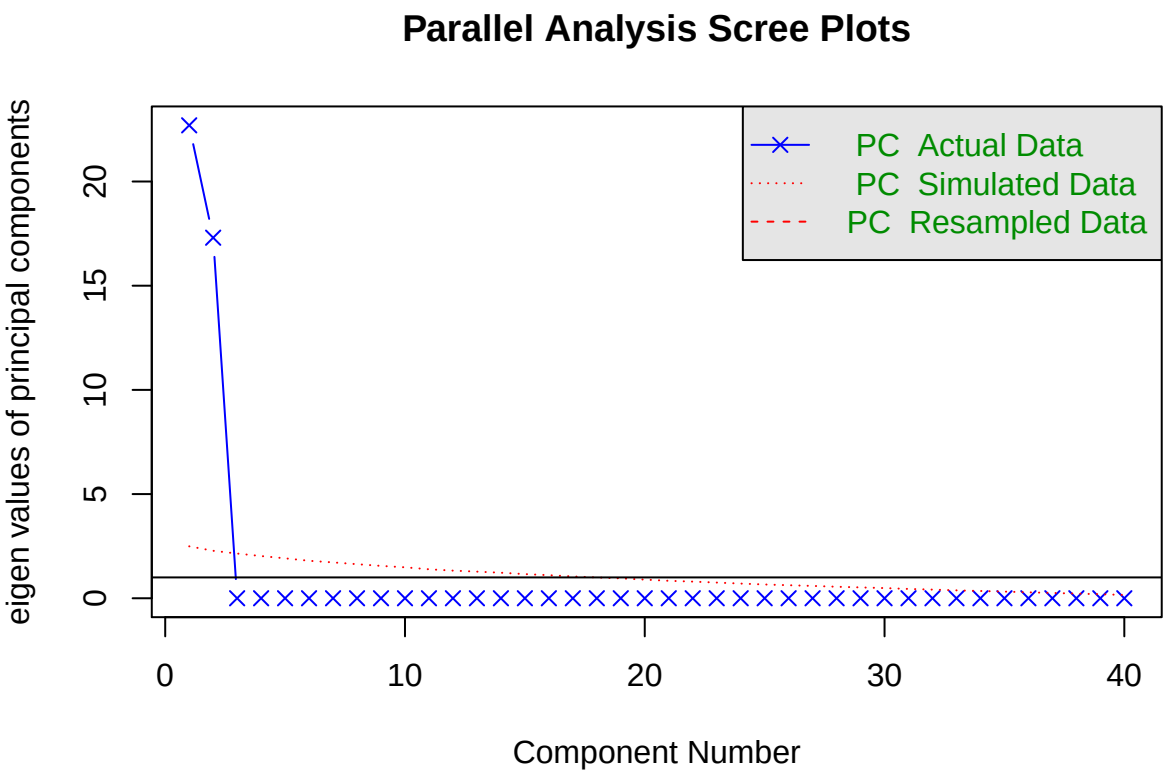
Number of components

```r
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

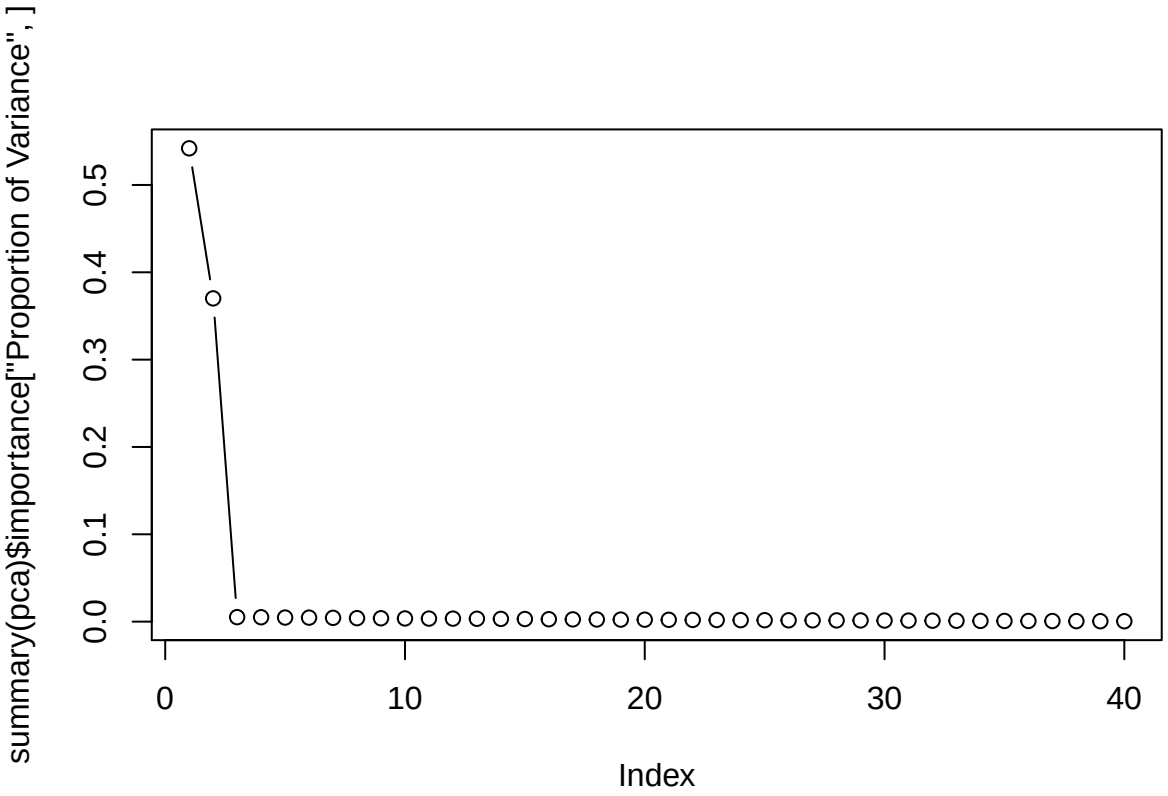
```r
fa.parallel(X_signal, fa="pc")
```

```
## Loading required package: parallel
## Loading required package: MASS

## Warning in cor.smooth(R): Matrix was not positive definite, smoothing was
## done

## In smc, the correlation matrix was not invertible, smc's returned as 1s

## Warning in cor.smooth(R): Matrix was not positive definite, smoothing was
## done

## In smc, the correlation matrix was not invertible, smc's returned as 1s

## Warning in cor.smooth(r): Matrix was not positive definite, smoothing was
## done

## The determinant of the smoothed correlation was zero.
## This means the objective function is not defined.
## Chi square is based upon observed residuals.
## The determinant of the smoothed correlation was zero.
## This means the objective function is not defined for the null model either.
## The Chi square is thus based upon observed correlations.

## Warning in cor.smooth(r): Matrix was not positive definite, smoothing was
## done

## In factor.stats, the correlation matrix is singular, an approximation is used

## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs =
## np.obs, : In factor.stats, the correlation matrix is singular, and we
## could not calculate the beta weights for factor score estimates

## In factor.scores, the correlation matrix is singular, an approximation is used

## Warning in cor.smooth(r): Matrix was not positive definite, smoothing was
## done

## I was unable to calculate the factor score weights, factor loadings used instead
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  2
```

```
pca <- prcomp(X)
plot(summary(pca)$importance["Proportion of Variance",], type="b")
```



```r
cbind(pca$x[, c("PC1", "PC2")], y=df2$y)
```

```
##                 PC1          PC2         y
##   [1,] -19.96742796   1.04843693 21.691757
##   [2,] -10.19540518  -6.88976050 18.973691
##   [3,]   7.29622592   0.83327020 12.984279
##   [4,] -24.72256054   5.75393360 23.081459
##   [5,]   3.21415195  13.82685027 15.063384
##   [6,] -18.03136450  -8.15104280 21.146380
##   [7,]   7.32568515  13.69153049 14.160038
##   [8,] -12.41162532 -23.44429917 18.019976
##   [9,]  11.47281139  -7.14891265 11.478973
##  [10,]  -9.53277988 -23.09968345 17.895613
##  [11,]   2.92436125 -13.37120359 14.231130
##  [12,]  17.94713822   5.08214391 10.454685
##  [13,]  26.92417397  -8.47392238  6.928486
##  [14,]  -7.81316978   8.88204792 18.568697
##  [15,]   6.48376544  28.04233225 14.627389
##  [16,]  19.35825671  -2.13823249  9.799341
##  [17,]  10.75158747  23.41530964 13.650704
##  [18,] -12.41012619   8.26466957 19.370464
##  [19,]  12.87804527  15.19573251 12.448344
##  [20,] -12.02743186   1.71327088 20.708240
##  [21,]   4.55385384   2.74204993 15.203687
##  [22,]  14.12285746 -12.54225776 11.368787
##  [23,]  -4.27134303   6.87506951 17.480896
##  [24,]  -6.57902955   1.42946955 18.408608
##  [25,]  10.40026235   9.65434269 12.866052
##  [26,] -25.58023163  13.94027522 23.607091
##  [27,] -16.69769924  -3.20389252 20.783062
##  [28,] -21.93060579  13.78353874 23.049267
##  [29,] -23.50817508  10.80506197 23.754437
##  [30,]  -3.57423332   3.80921606 17.266508
##  [31,]  27.92038036  -7.90056511  7.284992
##  [32,]  27.28276816 -10.65399961  7.030943
##  [33,] -17.95580331  -2.09387077 21.515531
##  [34,]   4.60043147  26.14717022 15.527558
##  [35,]  -9.33531969  19.96701565 19.722105
##  [36,] -22.54805772  14.73729636 23.666293
##  [37,]   6.47413848   5.54877240 13.913319
##  [38,]  10.86727649   4.85992952 12.414253
##  [39,]  -9.91109928   2.81630302 19.023148
##  [40,]  13.90838970  -8.00582315 10.160379
##  [41,]  14.89679312  19.43534778 12.617953
##  [42,]  -0.01427354   0.28613885 15.990348
##  [43,]  15.42339452   4.88306944 10.933293
##  [44,]   2.45672382  -7.31328165 14.047873
##  [45,]   5.21766968  -1.24495237 13.843417
##  [46,]  -4.12227994 -13.45959663 16.260898
##  [47,] -21.73358057  17.34913632 23.194189
##  [48,]  -6.55824732   7.30086967 17.727646
##  [49,]  20.55227952 -13.95255184  8.332974
##  [50,] -11.94408510 -19.89953575 18.329899
```

```
##  [51,] -11.50417913 -20.33333699 17.799538
##  [52,]   1.07802955 -18.12173407 14.390451
##  [53,] -20.90192830   7.49266721 22.664634
##  [54,]  12.44297297 -12.41125177 11.244343
##  [55,]   4.65442531 -15.41755443 13.291061
##  [56,] -11.19768682 -11.69742511 18.070618
##  [57,] -12.25973173  10.23594789 20.282507
##  [58,] -17.55738111   1.40326364 21.447916
##  [59,]  20.14656817  13.48022917 10.434321
##  [60,]  -3.32293527   7.83419664 16.695685
##  [61,] -10.33179612  -1.93057120 18.566197
##  [62,]  -5.15149752   6.19979865 17.473149
##  [63,]  19.72833042   3.67196703 10.076280
##  [64,]   0.21518575  -1.24500986 15.683162
##  [65,] -10.71857315 -22.68284265 17.558228
##  [66,] -16.98003080  -0.65234948 21.276181
##  [67,]  19.93769554   6.55897186 10.434667
##  [68,]  20.45834688   6.40679069  9.604647
##  [69,]  -6.72557101   6.77706288 17.949527
##  [70,]   9.06264474   9.24939691 13.503940
##  [71,] -14.16524191 -18.82623815 18.733456
##  [72,]  17.22441432  -0.89796738 10.789611
##  [73,]  21.67317686  -4.17772974  9.071024
##  [74,]  -6.32167142 -20.87669813 16.250610
##  [75,] -22.43609798   9.90983847 23.041106
##  [76,]  12.63318414  14.02719050 12.815979
##  [77,]  -1.14423353 -13.86402657 15.392426
##  [78,]  24.82106047  -2.75575827  7.885522
##  [79,]  -1.09384239   2.68592835 15.946902
##  [80,] -20.33557317   5.63686036 21.474662
##  [81,]   2.37534159   0.07924486 14.958001
##  [82,]  -6.36163544 -20.68177939 16.217599
##  [83,] -11.90510387 -20.95613117 18.613468
##  [84,] -15.41450829  11.54127645 21.090693
##  [85,]  24.08292296  -4.30510749  7.797186
##  [86,] -15.30494203   7.72650572 20.361050
##  [87,]  14.23782215  -8.66078123 10.498248
##  [88,] -12.27103878 -24.67728633 18.514131
##  [89,] -14.31552399 -17.45788345 19.138551
##  [90,]  -8.57557979  14.40821199 19.752880
##  [91,]  18.24185641 -10.57348123  9.078128
##  [92,]  -5.76855223 -11.08386551 16.281602
##  [93,]  -7.84904625  -7.06029396 17.457046
##  [94,]  16.95012659  12.35896795 11.076870
##  [95,]  14.81676928  15.64030463 12.132039
##  [96,]  23.14287743 -11.18859721  8.708341
##  [97,] -17.04130463   6.66645217 21.422746
##  [98,]   8.67098103   2.14276306 13.324292
##  [99,]  21.21185377   4.54611814  8.641729
## [100,]   7.27115391  -3.27647134 13.614490
```