SM2_Portfolio_1

Henry Bourne

2023-01-23

Principal Component analysis

Task 1

First let's look at the structure of the data:

```
USA <- USArrests[,-3]
summary(USA)
```

```
##      Murder          Assault           Rape      
##  Min.   : 0.800   Min.   : 45.0   Min.   : 7.30  
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:15.07  
##  Median : 7.250   Median :159.0   Median :20.10  
##  Mean   : 7.788   Mean   :170.8   Mean   :21.23  
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:26.18  
##  Max.   :17.400   Max.   :337.0   Max.   :46.00  
```

We have four columns and would like to perform PCA on the data minus the UrbanPop feature. First we will carry out PCA using the covariance matrix, S, manually:

```
S.USA <- cov(USA)
ev.USA <- eigen(S.USA)
ev.USA
```

```
## eigen() decomposition
## $values
## [1] 6996.480738   48.658639    6.725962
## 
## $vectors
##             [,1]        [,2]        [,3]
## [1,] -0.04180743  0.02555358  0.99879886
## [2,] -0.99630506 -0.07612980 -0.03975532
## [3,] -0.07502247  0.99677042 -0.02864195
```

Notice that the eigenvectors in the matrix are already sorted by the size of their eigenvalues and so have decreasing sample variance. Thus the matrix of our principal components is exactly the matrix of eigenvalues above.

Now we will carry out PCA using the correlation matrix, R, and the help of the *prcomp* command.

```
pc_USArrests <- prcomp(~Murder + Assault + Rape, USArrests, scale. = TRUE, retx=TRUE); pc_USArrests # W
```

```
## Standard deviations (1, .., p=3):
## [1] 1.5357670 0.6767949 0.4282154
## 
## Rotation (n x k) = (3 x 3):
```

```
##                PC1        PC2        PC3
## Murder  -0.5826006  0.5339532 -0.6127565
## Assault -0.6079818  0.2140236  0.7645600
## Rape    -0.5393836 -0.8179779 -0.1999436
```

```
pc_USArrests$x
```

```
##                         PC1         PC2          PC3
## Alabama        -1.198027832  0.83381177 -0.162178476
## Alaska         -2.308747325 -1.52396221  0.038335742
## Arizona        -1.503330652 -0.49830384  0.878223112
## Arkansas       -0.175989446  0.32473260  0.071111741
## California     -2.045235843 -1.27257704  0.381539326
## Colorado       -1.263413283 -1.42640632 -0.083693139
## Connecticut     1.627064626  0.17860374  0.290256038
## Delaware        0.074812801  0.41561083  0.998446677
## Florida        -2.830731325  0.42331809  0.208151641
## Georgia        -1.842343065  0.88277323 -1.080609032
## Hawaii          1.302403647 -0.53528688 -0.772523404
## Idaho           1.469223571 -0.15225639  0.414302742
## Illinois       -1.079579292  0.27941102  0.291234322
## Indiana         0.513394603 -0.20015992 -0.442228830
## Iowa            2.156636865 -0.11239443 -0.054668600
## Kansas          0.832079409 -0.08014103 -0.191017094
## Kentucky        0.478831591  0.50650575 -0.730308649
## Louisiana      -1.644731071  1.04957018 -0.373768062
## Maine           2.174592271  0.25034567  0.281818786
## Maryland       -1.790861674  0.18886129  0.551385569
## Massachusetts   0.895952687 -0.04050899  0.382293578
## Michigan       -1.989964308 -0.46614937 -0.129836314
## Minnesota       1.765715718 -0.32440018 -0.055072348
## Mississippi    -1.517623726  1.60645578 -0.271636024
## Missouri       -0.616205380 -0.44134841 -0.252834582
## Montana         0.967991307  0.04418005 -0.211907462
## Nebraska        1.240695366 -0.19093752 -0.039096727
## Nevada         -2.609154480 -1.41350501 -0.404109160
## New Hampshire   2.266374551  0.03511068  0.006998662
## New Jersey      0.277745503  0.13462220 -0.001387439
## New Mexico     -1.942431655 -0.21292600  0.307911561
## New York       -1.330622591  0.19467099  0.193797219
## North Carolina -1.614416593  1.51406566  0.901426577
## North Dakota    2.654501432  0.03705123  0.126761976
## Ohio            0.425915739 -0.20485611 -0.400616435
## Oklahoma        0.374013508 -0.08879455  0.012150589
## Oregon          0.007484513 -1.08883723  0.126183190
## Pennsylvania    1.036129757  0.20425023 -0.249615405
## Rhode Island    1.308026751  0.59975203  0.923110514
## South Carolina -1.747107574  0.97782271  0.035740543
## South Dakota    1.637375231  0.02980135 -0.036558286
## Tennessee      -1.176095386  0.21275256 -0.724219698
## Texas          -1.123432710  0.30710594 -0.504726135
## Utah            0.887958106 -0.83848257  0.144173194
## Vermont         2.220758807 -0.12420650 -0.125927856
## Virginia        0.043078167  0.09584025 -0.224223254
## Washington      0.408525962 -0.96439783  0.190536108
```

```
## West Virginia   1.621260055  0.55554584 -0.275017445
## Wisconsin       2.153811966 -0.02739623 -0.127792014
## Wyoming         0.527690701  0.34566289  0.169682461
```

The **sdev** component tells us the standard deviations of the principal components which corresponds to the squareroot of the eigenvalues of the covariance matrix, $(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3})^T = \text{diag}(\Lambda)$. The **rotation** component contains the matrix of the principal components, ie. the eigenvectors of the correlation matrix, R, $\mathbf{A} = [v_1 \quad v_2 \quad v_3]$ where $\mathbf{A}$ is the matrix of principal components (the eigenvectors of R) and v_i is the i-th eigenvector. Finally **x** contains the data transformed by the principal component matrix (ie. our now uncorrelated data), in the notes: $\mathbf{Y} = \mathbf{XA}$.

The correlation matrix can be interpreted as the sample covariance matrix of the scaled data. So wether we should use the covariance matrix or correlation matrix depends on the variances of the predictor variables. Recalling the variances:

```r
diag(S.USA)
```

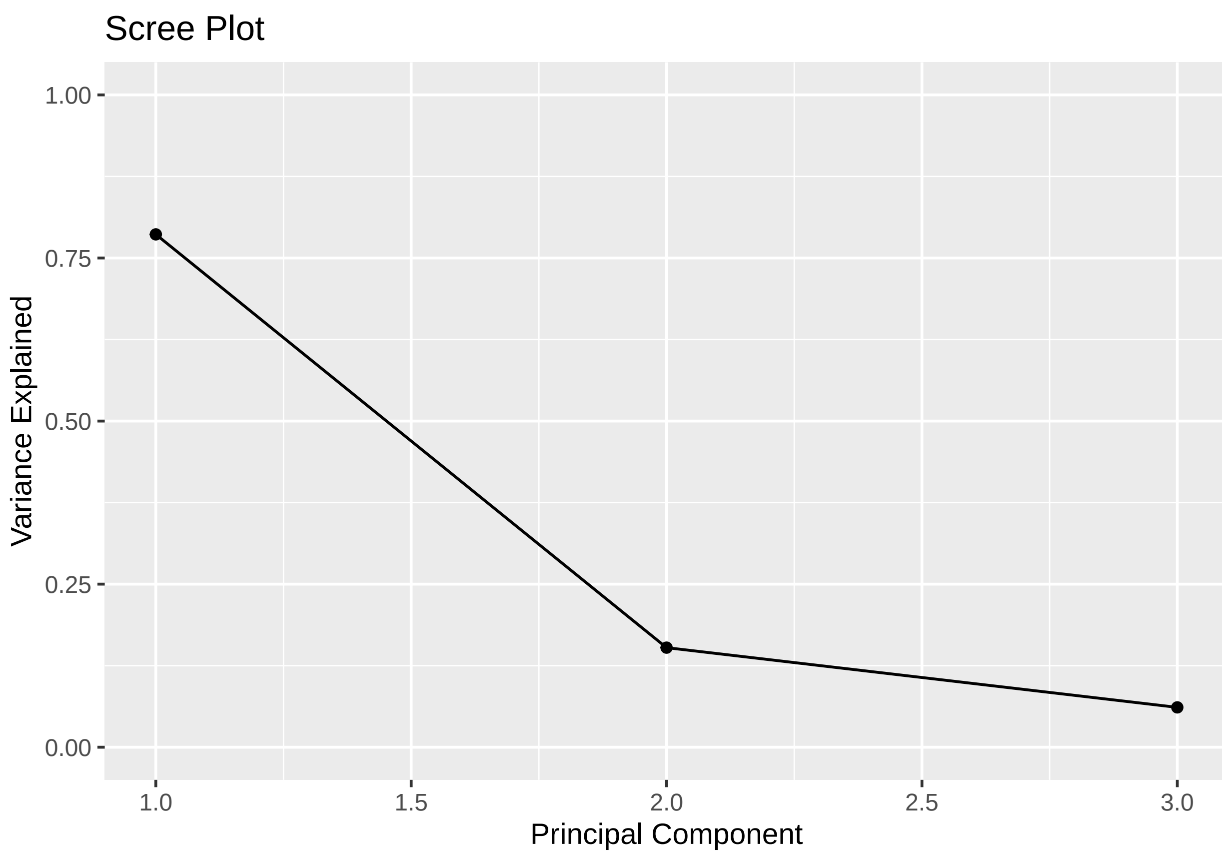
```
##     Murder    Assault       Rape
##   18.97047 6945.16571   87.72916
```

We see that the variances of the predictor variables greatly differ by orders of magnitude, therefore it is sensible to work with the scaled data, ie. the correlation matrix R.

Now we will plot a **scree plot** which is a plot of the variance explained by each principal component, ie. the percentage of the variance accounted for by the principal component.

```r
var_prcnt = pc_USArrests$sdev^2 / sum(pc_USArrests$sdev^2)

qplot(c(1:3), var_prcnt) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot") +
  ylim(0, 1)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```



From the scree plot we can see the first principal component accounts for a very large percentage of the variance with the 3rd principal component accounting for less than 12% of the variance.

We will now compute the Kaiser's criterion:

```
max(which(pc_USArrests$sdev > sum(pc_USArrests$sdev)/length(pc_USArrests$sdev)))
```

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

and now the number of PC's to keep according to Horn's parallel analysis:

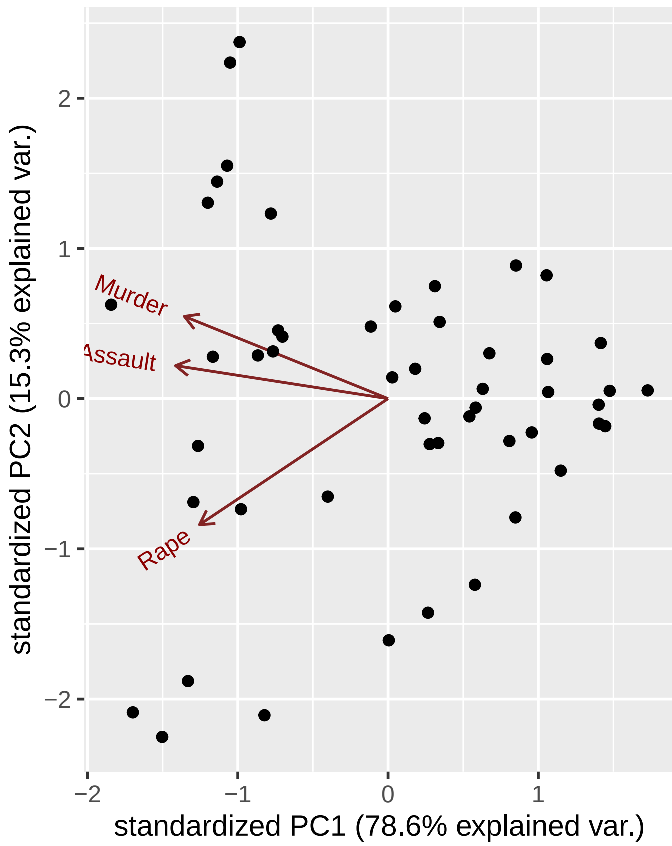
```
M <-1000
n <- nrow(USA)
p <- ncol(USA)
lambdas <- matrix(NA, M, p)
for(i in 1:M){
  M <- matrix( rnorm(p*n,mean=0,sd=1), n, p)  # genrate matrix with N(0,1) entries
  R <- cor(M) # find correlation matix
  S <- diag(pc_USArrests$sdev) %*% R %*% diag(pc_USArrests$sdev) # scale
  lambdas[i,] <- eigen(S)$values # find the eigenvalues
}
mean_lambda <- colMeans(lambdas) # find the mean of the eigenvalues

# find the largest index of eigenvalues larger than the mean eigenvalue of our M standard normal matric
max(which(pc_USArrests$sdev > mean_lambda))
```

```
## [1] 3
```

Kaiser's criterion tells us to only keep the first principal component, ie. let $q = 1$, whereas Horn's parallel analysis suggests not throwing away any of the principal components, ie. $q = 3$. Horn's parallel analysis takes into account the sampling error that arises from the fact that we don't have infinite observations (only n) unlike Kaiser's criterion, therefore we will select $q = 3$. Looking at the scree plot we can back up this decision as throwing away the second two principal components would amount to loosing close to 25% of the varaince.

Now we will produce a biplot:

```
ggbiplot(pc_USArrests)
```



The black dots are the datapoints transformed by our PC's, here we plot the first component versus the second. The red arrows tell us which predictor variables contribute to which principal components. Here we see that higher values in murder and assault contribute to a larger value of PC2 for example. From the plot we see that all the features have negative values for the first component which means the first component is an average of all three features and suggests that all three features are correlated with each other. For the second component we see that larger values in murder and assault lead to a higher value and rape to a lower value, ie. it contrasts rape against the other features.

Task 2

We will be working with the iris dataset:

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
```

```
##        Species
##   setosa    :50
##   versicolor:50
##   virginica :50
##
##
##
```

Ideally we would like to use the covariance matrix as this will preserve variance, however, if the predictor variables are scaled differently this could lead to one predictor variable accounting for a large percentage of the variance. In this case we would prefer to standardize the data and use the correlation matrix. Let us asses the variances of the predictor variables:

```
diag(cov(iris[,-5]))
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.6856935    0.1899794    3.1162779    0.5810063
```

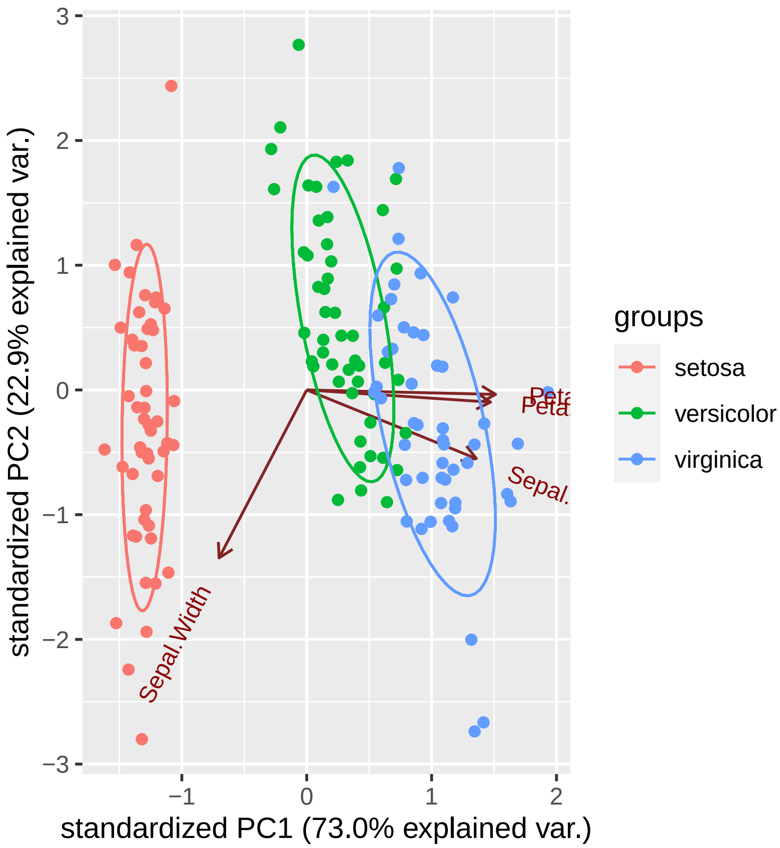
We see that the variances do differ and even though they all use the same measurement on the same scale the variances are very different, for example Petal.length has a variance of 13 times Sepal.Width so we will use the correlation matrix:

```
pc_iris <- prcomp(~Sepal.Length + Sepal.Width + Petal.Length + Petal.Width , iris, scale. = TRUE, retx='
```

```
## Standard deviations (1, .., p=4):
## [1] 1.7083611 0.9560494 0.3830886 0.1439265
##
## Rotation (n x k) = (4 x 4):
##                     PC1         PC2        PC3        PC4
## Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
## Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
## Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
## Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Let's now plot the two dimensional reduction of observations and color the data-points according to their species:

```
ggbiplot(pc_iris, ellipse=TRUE, groups = iris$Species)
```



From the plot we see that using our first two principal components has resulted in a two-dimensional reduction of the data-set where the points are clustered according to group, ie. all the points belonging to the setosa species appear together in the feature space and the same can be said for versicolor and virginica, although the between group scatterness between these two is much lower (ie. they appear much closer together in feature space).

Task 3

For this task we will be working with the communities and crime dataset, let's start by loading in and summarising the dataset:

```
library(mogavs)
data(crimeData)
summary(crimeData)
```

```
##       x.V6               x.V7              x.V8              x.V9
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.01000   1st Qu.:0.3500   1st Qu.:0.0200   1st Qu.:0.6300
##  Median :0.02000   Median :0.4400   Median :0.0600   Median :0.8500
##  Mean   :0.05759   Mean   :0.4634   Mean   :0.1796   Mean   :0.7537
##  3rd Qu.:0.05000   3rd Qu.:0.5400   3rd Qu.:0.2300   3rd Qu.:0.9400
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V10             x.V11             x.V12             x.V13
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0400   1st Qu.:0.010   1st Qu.:0.3400   1st Qu.:0.4100
##  Median :0.0700   Median :0.040   Median :0.4000   Median :0.4800
##  Mean   :0.1537   Mean   :0.144   Mean   :0.4242   Mean   :0.4939
##  3rd Qu.:0.1700   3rd Qu.:0.160   3rd Qu.:0.4700   3rd Qu.:0.5400
```

```
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##      x.V14            x.V15            x.V16             x.V17
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.2500   1st Qu.:0.3000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.2900   Median :0.4200   Median :0.03000   Median :1.0000
##  Mean   :0.3363   Mean   :0.4232   Mean   :0.06407   Mean   :0.6963
##  3rd Qu.:0.3600   3rd Qu.:0.5300   3rd Qu.:0.07000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##      x.V18            x.V19            x.V20            x.V21
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2000   1st Qu.:0.4400   1st Qu.:0.1600   1st Qu.:0.3700
##  Median :0.3200   Median :0.5600   Median :0.2300   Median :0.4800
##  Mean   :0.3611   Mean   :0.5582   Mean   :0.2916   Mean   :0.4957
##  3rd Qu.:0.4900   3rd Qu.:0.6900   3rd Qu.:0.3700   3rd Qu.:0.6200
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V22            x.V23            x.V24            x.V25
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3500   1st Qu.:0.1425   1st Qu.:0.3600   1st Qu.:0.2300
##  Median :0.4750   Median :0.2600   Median :0.4700   Median :0.3300
##  Mean   :0.4711   Mean   :0.3178   Mean   :0.4792   Mean   :0.3757
##  3rd Qu.:0.5800   3rd Qu.:0.4400   3rd Qu.:0.5800   3rd Qu.:0.4800
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V26            x.V27           x.V28            x.V29
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2200   1st Qu.:0.240   1st Qu.:0.1725   1st Qu.:0.1100
##  Median :0.3000   Median :0.320   Median :0.2500   Median :0.1700
##  Mean   :0.3503   Mean   :0.368   Mean   :0.2911   Mean   :0.2035
##  3rd Qu.:0.4300   3rd Qu.:0.440   3rd Qu.:0.3800   3rd Qu.:0.2500
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##      x.V30            x.V31              x.V32            x.V33
##  Min.   :0.0000   Min.   :  0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.1900   1st Qu.:  0.1700   1st Qu.:0.2600   1st Qu.:0.01000
##  Median :0.2800   Median :  0.2500   Median :0.3450   Median :0.02000
##  Mean   :0.3224   Mean   :  0.3804   Mean   :0.3863   Mean   :0.05551
##  3rd Qu.:0.4000   3rd Qu.:  0.3600   3rd Qu.:0.4800   3rd Qu.:0.05000
##  Max.   :1.0000   Max.   :191.0542   Max.   :1.0000   Max.   :1.00000
##      x.V34           x.V35            x.V36            x.V37
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.110   1st Qu.:0.1600   1st Qu.:0.2300   1st Qu.:0.2100
##  Median :0.250   Median :0.2700   Median :0.3600   Median :0.3100
##  Mean   :0.303   Mean   :0.3158   Mean   :0.3833   Mean   :0.3617
##  3rd Qu.:0.450   3rd Qu.:0.4200   3rd Qu.:0.5100   3rd Qu.:0.4600
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V38            x.V39            x.V40            x.V41
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2200   1st Qu.:0.3800   1st Qu.:0.2500   1st Qu.:0.3200
##  Median :0.3200   Median :0.5100   Median :0.3700   Median :0.4100
##  Mean   :0.3635   Mean   :0.5011   Mean   :0.3964   Mean   :0.4406
##  3rd Qu.:0.4800   3rd Qu.:0.6275   3rd Qu.:0.5200   3rd Qu.:0.5300
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V42            x.V43            x.V44            x.V45
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2400   1st Qu.:0.3100   1st Qu.:0.3300   1st Qu.:0.3100
##  Median :0.3700   Median :0.4000   Median :0.4700   Median :0.4000
```

```
##  Mean   :0.3912   Mean   :0.4413   Mean   :0.4612   Mean   :0.4345
##  3rd Qu.:0.5100   3rd Qu.:0.5400   3rd Qu.:0.5900   3rd Qu.:0.5000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V46            x.V47            x.V48            x.V49
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3600   1st Qu.:0.3600   1st Qu.:0.4000   1st Qu.:0.4900
##  Median :0.5000   Median :0.5000   Median :0.4700   Median :0.6300
##  Mean   :0.4876   Mean   :0.4943   Mean   :0.4877   Mean   :0.6109
##  3rd Qu.:0.6200   3rd Qu.:0.6300   3rd Qu.:0.5600   3rd Qu.:0.7600
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V50            x.V51           x.V52            x.V53
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.4900   1st Qu.:0.530   1st Qu.:0.4800   1st Qu.:0.3900
##  Median :0.6400   Median :0.700   Median :0.6100   Median :0.5100
##  Mean   :0.6207   Mean   :0.664   Mean   :0.5829   Mean   :0.5014
##  3rd Qu.:0.7800   3rd Qu.:0.840   3rd Qu.:0.7200   3rd Qu.:0.6200
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##      x.V54            x.V55             x.V56          x.V57
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00   Min.   :0.00000
##  1st Qu.:0.4200   1st Qu.:0.00000   1st Qu.:0.09   1st Qu.:0.00000
##  Median :0.5400   Median :0.01000   Median :0.17   Median :0.01000
##  Mean   :0.5267   Mean   :0.03629   Mean   :0.25   Mean   :0.03006
##  3rd Qu.:0.6500   3rd Qu.:0.02000   3rd Qu.:0.32   3rd Qu.:0.02000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00   Max.   :1.00000
##      x.V58            x.V59            x.V60            x.V61
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1600   1st Qu.:0.2000   1st Qu.:0.2500   1st Qu.:0.2800
##  Median :0.2900   Median :0.3400   Median :0.3900   Median :0.4300
##  Mean   :0.3202   Mean   :0.3606   Mean   :0.3991   Mean   :0.4279
##  3rd Qu.:0.4300   3rd Qu.:0.4800   3rd Qu.:0.5300   3rd Qu.:0.5600
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V62            x.V63            x.V64            x.V65
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0300   1st Qu.:0.0300   1st Qu.:0.0300   1st Qu.:0.0300
##  Median :0.0900   Median :0.0800   Median :0.0900   Median :0.0900
##  Mean   :0.1814   Mean   :0.1821   Mean   :0.1848   Mean   :0.1829
##  3rd Qu.:0.2300   3rd Qu.:0.2300   3rd Qu.:0.2300   3rd Qu.:0.2300
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V66            x.V67            x.V68            x.V69
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.7300   1st Qu.:0.0300   1st Qu.:0.1500   1st Qu.:0.1400
##  Median :0.8700   Median :0.0600   Median :0.2000   Median :0.1900
##  Mean   :0.7859   Mean   :0.1506   Mean   :0.2676   Mean   :0.2519
##  3rd Qu.:0.9400   3rd Qu.:0.1600   3rd Qu.:0.3100   3rd Qu.:0.2900
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V70            x.V71            x.V72            x.V73
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3400   1st Qu.:0.3900   1st Qu.:0.2700   1st Qu.:0.4400
##  Median :0.4400   Median :0.4800   Median :0.3600   Median :0.5600
##  Mean   :0.4621   Mean   :0.4944   Mean   :0.4041   Mean   :0.5626
##  3rd Qu.:0.5500   3rd Qu.:0.5800   3rd Qu.:0.4900   3rd Qu.:0.7000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V74            x.V75            x.V76            x.V77
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
```

```
##  1st Qu.:0.0600   1st Qu.:0.4000   1st Qu.:0.0000   1st Qu.:0.01000
##  Median :0.1100   Median :0.5100   Median :0.5000   Median :0.03000
##  Mean   :0.1863   Mean   :0.4952   Mean   :0.3147   Mean   :0.07682
##  3rd Qu.:0.2200   3rd Qu.:0.6000   3rd Qu.:0.5000   3rd Qu.:0.07000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##      x.V78            x.V79            x.V80            x.V81
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.6300   1st Qu.:0.4300   1st Qu.:0.0600   1st Qu.:0.2900
##  Median :0.7700   Median :0.5400   Median :0.1300   Median :0.4200
##  Mean   :0.7195   Mean   :0.5487   Mean   :0.2045   Mean   :0.4333
##  3rd Qu.:0.8600   3rd Qu.:0.6700   3rd Qu.:0.2700   3rd Qu.:0.5600
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V82            x.V83            x.V84            x.V85
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3500   1st Qu.:0.0600   1st Qu.:0.1000   1st Qu.:0.0900
##  Median :0.5200   Median :0.1850   Median :0.1900   Median :0.1800
##  Mean   :0.4942   Mean   :0.2645   Mean   :0.2431   Mean   :0.2647
##  3rd Qu.:0.6700   3rd Qu.:0.4200   3rd Qu.:0.3300   3rd Qu.:0.4000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V86            x.V87            x.V88            x.V89
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0900   1st Qu.:0.0900   1st Qu.:0.1700   1st Qu.:0.2000
##  Median :0.1700   Median :0.1800   Median :0.3100   Median :0.3300
##  Mean   :0.2635   Mean   :0.2689   Mean   :0.3464   Mean   :0.3725
##  3rd Qu.:0.3900   3rd Qu.:0.3800   3rd Qu.:0.4900   3rd Qu.:0.5200
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V90           x.V91            x.V92            x.V93
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.220   1st Qu.:0.2100   1st Qu.:0.3700   1st Qu.:0.3200
##  Median :0.370   Median :0.3400   Median :0.4800   Median :0.4500
##  Mean   :0.423   Mean   :0.3841   Mean   :0.4901   Mean   :0.4498
##  3rd Qu.:0.590   3rd Qu.:0.5300   3rd Qu.:0.5900   3rd Qu.:0.5800
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V94            x.V95             x.V96             x.V97
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.2500   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0600
##  Median :0.3700   Median :0.00000   Median :0.00000   Median :0.1300
##  Mean   :0.4038   Mean   :0.02944   Mean   :0.02278   Mean   :0.2156
##  3rd Qu.:0.5100   3rd Qu.:0.01000   3rd Qu.:0.00000   3rd Qu.:0.2800
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      x.V98            x.V99            x.V100           x.V101
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.4700   1st Qu.:0.4200   1st Qu.:0.5200   1st Qu.:0.5600
##  Median :0.6300   Median :0.5400   Median :0.6700   Median :0.7000
##  Mean   :0.6089   Mean   :0.5351   Mean   :0.6264   Mean   :0.6515
##  3rd Qu.:0.7775   3rd Qu.:0.6600   3rd Qu.:0.7700   3rd Qu.:0.7900
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      x.V102             x.V103            x.V104           x.V105
##  Min.   :-0.53937   Min.   :-0.5378   Min.   :0.0000   Min.   :-0.6192
##  1st Qu.:-0.02460   1st Qu.: 0.1400   1st Qu.:0.8852   1st Qu.: 0.1600
##  Median : 0.03000   Median : 0.2963   Median :0.9600   Median : 0.3200
##  Mean   : 0.03604   Mean   : 0.3776   Mean   :0.9551   Mean   : 0.3664
##  3rd Qu.: 0.10392   3rd Qu.: 0.6238   3rd Qu.:1.0181   3rd Qu.: 0.5688
##  Max.   : 1.00000   Max.   : 1.4801   Max.   :1.5258   Max.   : 1.3772
```

```
##      x.V106             x.V107            x.V108             x.V109
##  Min.   :-0.63745   Min.   :-0.8905   Min.   :-0.858872   Min.   :-0.5377
##  1st Qu.:-0.05187   1st Qu.: 0.0800   1st Qu.: 0.005023   1st Qu.: 0.1400
##  Median : 0.03000   Median : 0.2022   Median : 0.183860   Median : 0.2964
##  Mean   : 0.02585   Mean   : 0.2395   Mean   : 0.173225   Mean   : 0.3775
##  3rd Qu.: 0.10100   3rd Qu.: 0.4048   3rd Qu.: 0.328242   3rd Qu.: 0.6238
##  Max.   : 1.09156   Max.   : 2.2631   Max.   : 3.074837   Max.   : 1.4799
##      x.V110            x.V111            x.V112             x.V113
##  Min.   :-0.7876   Min.   :-0.2461   Min.   :-0.73744   Min.   :-0.43847
##  1st Qu.: 0.5704   1st Qu.: 0.6571   1st Qu.:-0.04077   1st Qu.:-0.03827
##  Median : 0.7444   Median : 0.8415   Median : 0.07946   Median : 0.04984
##  Mean   : 0.7225   Mean   : 0.8149   Mean   : 0.10467   Mean   : 0.08496
##  3rd Qu.: 0.8935   3rd Qu.: 0.9831   3rd Qu.: 0.22227   3rd Qu.: 0.16339
##  Max.   : 1.5377   Max.   : 1.6100   Max.   : 1.18392   Max.   : 1.16911
##      x.V114             x.V115             x.V116              x.V117
##  Min.   :-0.99924   Min.   :-0.70307   Min.   :-0.488758   Min.   :-1.4252
##  1st Qu.:-0.13445   1st Qu.:-0.02781   1st Qu.:-0.069858   1st Qu.: 0.2955
##  Median : 0.01267   Median : 0.10000   Median : 0.014724   Median : 0.4752
##  Mean   : 0.04463   Mean   : 0.12893   Mean   : 0.003826   Mean   : 0.4718
##  3rd Qu.: 0.21179   3rd Qu.: 0.25901   3rd Qu.: 0.072606   3rd Qu.: 0.6400
##  Max.   : 1.71153   Max.   : 1.13316   Max.   : 1.000000   Max.   : 6.5607
##      x.V118             x.V119            x.V120           x.V121
##  Min.   :-2.04213   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 0.01546   1st Qu.:0.02000   1st Qu.:0.1000   1st Qu.:0.0200
##  Median : 0.20044   Median :0.04000   Median :0.1700   Median :0.0700
##  Mean   : 0.20550   Mean   :0.06523   Mean   :0.2329   Mean   :0.1617
##  3rd Qu.: 0.42598   3rd Qu.:0.07000   3rd Qu.:0.2800   3rd Qu.:0.1900
##  Max.   : 1.53559   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##      x.V122              x.V123             x.V124             x.V125
##  Min.   :-0.708426   Min.   :-0.44286   Min.   :-0.9850   Min.   :-1.9221
##  1st Qu.: 0.005435   1st Qu.:-0.01398   1st Qu.: 0.4115   1st Qu.: 0.0000
##  Median : 0.093085   Median : 0.04000   Median : 0.6622   Median : 0.3896
##  Mean   : 0.110848   Mean   : 0.04666   Mean   : 0.6364   Mean   : 0.3753
##  3rd Qu.: 0.224945   3rd Qu.: 0.11472   3rd Qu.: 0.8490   3rd Qu.: 0.7615
##  Max.   : 1.000000   Max.   : 1.00000   Max.   : 1.9686   Max.   : 2.4168
##      x.V126            x.V127            y
##  Min.   :0.00000   Min.   :-0.6075   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.: 0.1200   1st Qu.:0.070
##  Median :0.00000   Median : 0.2800   Median :0.150
##  Mean   :0.09405   Mean   : 0.3560   Mean   :0.238
##  3rd Qu.:0.00000   3rd Qu.: 0.5883   3rd Qu.:0.330
##  Max.   :1.00000   Max.   : 1.7307   Max.   :1.000
```

We see that this dataset contains a large number of variables, each of these variables tell us something different about a community such as its state, population, percentage of people under the poverty level and wether or not a gang unit is deployed in that community among a total of 127 variables (not including the target variable). Among these the first 5 variables are non predictive and so won't be used in constructing a model, leaving 122 predictor variables (note that the dataframe from the *mogavs* package does not include the first 5 attributes). The target variable is y which is the total number of violent crimes per 100k population.

We would like to fit a regression model of the form $Z_i \sim f(\alpha + \beta^T x_i^0)$ using PCR to estimate the model parameters. The first thing to do is decide wether PCA should be applied to the covariance matrix or the correlation matrix. One thing to note is that in the original dataset all values were standardized between 0 and 1 and the dataset in *mogavs* is the same bar the fact that they impute missing values so we may find some values outside of this range. Knowing this it would make sense to use the covariance matrix for PCA as it will preserve variance and as all the attributes are already scaled.

Let's now carry out our PCA and print out a summary of our results:

```
pc_crimeData <- prcomp(~ . -y , crimeData, scale. = TRUE, retx=TRUE); summary(pc_crimeData)
```

```
## Importance of components:
##                           PC1    PC2     PC3    PC4     PC5     PC6    PC7
## Standard deviation     5.1263 4.2852 3.14023 2.8484 2.52280 2.44983 2.1274
## Proportion of Variance 0.2154 0.1505 0.08083 0.0665 0.05217 0.04919 0.0371
## Cumulative Proportion  0.2154 0.3659 0.44674 0.5132 0.56542 0.61461 0.6517
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.92926 1.79021 1.56346 1.46739 1.39980 1.30804 1.28740
## Proportion of Variance 0.03051 0.02627 0.02004 0.01765 0.01606 0.01402 0.01359
## Cumulative Proportion  0.68222 0.70849 0.72852 0.74617 0.76223 0.77626 0.78984
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     1.26287 1.19862 1.14023 1.12564 1.05175 0.99584 0.97054
## Proportion of Variance 0.01307 0.01178 0.01066 0.01039 0.00907 0.00813 0.00772
## Cumulative Proportion  0.80291 0.81469 0.82535 0.83573 0.84480 0.85293 0.86065
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.96633 0.92222 0.89262 0.85634 0.84920 0.82465 0.77804
## Proportion of Variance 0.00765 0.00697 0.00653 0.00601 0.00591 0.00557 0.00496
## Cumulative Proportion  0.86830 0.87527 0.88181 0.88782 0.89373 0.89930 0.90426
##                           PC29    PC30    PC31    PC32   PC33    PC34    PC35
## Standard deviation     0.76580 0.73427 0.72819 0.70541 0.6899 0.68415 0.67038
## Proportion of Variance 0.00481 0.00442 0.00435 0.00408 0.0039 0.00384 0.00368
## Cumulative Proportion  0.90907 0.91349 0.91784 0.92191 0.9258 0.92965 0.93334
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.66397 0.62569 0.60702 0.58972 0.57603 0.56650 0.55393
## Proportion of Variance 0.00361 0.00321 0.00302 0.00285 0.00272 0.00263 0.00252
## Cumulative Proportion  0.93695 0.94016 0.94318 0.94603 0.94875 0.95138 0.95389
##                           PC43    PC44   PC45    PC46    PC47   PC48    PC49
## Standard deviation     0.54304 0.52869 0.5180 0.49787 0.49653 0.4687 0.45403
## Proportion of Variance 0.00242 0.00229 0.0022 0.00203 0.00202 0.0018 0.00169
## Cumulative Proportion  0.95631 0.95860 0.9608 0.96283 0.96486 0.9667 0.96835
##                           PC50    PC51    PC52    PC53   PC54    PC55   PC56
## Standard deviation     0.45164 0.44451 0.43885 0.42861 0.4130 0.41005 0.3987
## Proportion of Variance 0.00167 0.00162 0.00158 0.00151 0.0014 0.00138 0.0013
## Cumulative Proportion  0.97002 0.97164 0.97322 0.97472 0.9761 0.97750 0.9788
##                           PC57    PC58    PC59    PC60   PC61    PC62    PC63
## Standard deviation     0.38490 0.37296 0.37131 0.36040 0.3487 0.32963 0.32066
## Proportion of Variance 0.00121 0.00114 0.00113 0.00106 0.0010 0.00089 0.00084
## Cumulative Proportion  0.98001 0.98115 0.98228 0.98335 0.9843 0.98524 0.98608
##                           PC64    PC65    PC66   PC67    PC68    PC69    PC70
## Standard deviation     0.31633 0.30719 0.29810 0.2913 0.28788 0.27641 0.26830
## Proportion of Variance 0.00082 0.00077 0.00073 0.0007 0.00068 0.00063 0.00059
## Cumulative Proportion  0.98690 0.98767 0.98840 0.9891 0.98978 0.99040 0.99099
##                           PC71    PC72    PC73    PC74    PC75    PC76    PC77
## Standard deviation     0.25982 0.25450 0.25103 0.23801 0.23588 0.22980 0.22319
## Proportion of Variance 0.00055 0.00053 0.00052 0.00046 0.00046 0.00043 0.00041
## Cumulative Proportion  0.99155 0.99208 0.99259 0.99306 0.99351 0.99395 0.99435
##                           PC78    PC79    PC80    PC81    PC82    PC83    PC84
## Standard deviation     0.21832 0.21411 0.21056 0.20020 0.19450 0.18859 0.18256
## Proportion of Variance 0.00039 0.00038 0.00036 0.00033 0.00031 0.00029 0.00027
## Cumulative Proportion  0.99475 0.99512 0.99548 0.99581 0.99612 0.99641 0.99669
```

```
##                            PC85    PC86    PC87    PC88   PC89   PC90    PC91
## Standard deviation     0.17283 0.16885 0.16837 0.16166 0.1563 0.1556 0.14596
## Proportion of Variance 0.00024 0.00023 0.00023 0.00021 0.0002 0.0002 0.00017
## Cumulative Proportion  0.99693 0.99717 0.99740 0.99761 0.9978 0.9980 0.99819
##                            PC92    PC93    PC94    PC95    PC96    PC97    PC98
## Standard deviation     0.14342 0.13589 0.13405 0.13088 0.12553 0.11681 0.11606
## Proportion of Variance 0.00017 0.00015 0.00015 0.00014 0.00013 0.00011 0.00011
## Cumulative Proportion  0.99835 0.99851 0.99865 0.99879 0.99892 0.99903 0.99915
##                            PC99   PC100   PC101   PC102   PC103   PC104   PC105
## Standard deviation     0.11383 0.10552 0.10317 0.09484 0.09310 0.08754 0.08136
## Proportion of Variance 0.00011 0.00009 0.00009 0.00007 0.00007 0.00006 0.00005
## Cumulative Proportion  0.99925 0.99934 0.99943 0.99950 0.99957 0.99964 0.99969
##                           PC106   PC107   PC108   PC109   PC110   PC111   PC112
## Standard deviation     0.07377 0.07086 0.06453 0.06077 0.05462 0.05390 0.04860
## Proportion of Variance 0.00004 0.00004 0.00003 0.00003 0.00002 0.00002 0.00002
## Cumulative Proportion  0.99974 0.99978 0.99981 0.99984 0.99987 0.99989 0.99991
##                           PC113   PC114   PC115   PC116   PC117   PC118   PC119
## Standard deviation     0.04829 0.04163 0.03993 0.03839 0.03532 0.02948 0.02670
## Proportion of Variance 0.00002 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99993 0.99994 0.99996 0.99997 0.99998 0.99999 0.99999
##                           PC120   PC121     PC122
## Standard deviation     0.02461 0.02171 0.0009488
## Proportion of Variance 0.00000 0.00000 0.0000000
## Cumulative Proportion  1.00000 1.00000 1.0000000
```

Let's first fit a linear model (carry out PCR) using all the principal components and print out a summary:

```
Regr_data <- data.frame(y = crimeData$y, pc_crimeData$x) # Create data frame from target variable and o
lmodel.all <- lm(y ~ ., data = Regr_data)
summary(lmodel.all)
```

```
##
## Call:
## lm(formula = y ~ ., data = Regr_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53580 -0.07223 -0.01285  0.05017  0.74159
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.380e-01  2.960e-03  80.405  < 2e-16 ***
## PC1         -2.824e-02  5.775e-04 -48.899  < 2e-16 ***
## PC2          1.689e-02  6.909e-04  24.453  < 2e-16 ***
## PC3         -1.483e-02  9.428e-04 -15.728  < 2e-16 ***
## PC4          1.046e-02  1.039e-03  10.066  < 2e-16 ***
## PC5          1.147e-02  1.173e-03   9.773  < 2e-16 ***
## PC6          6.699e-03  1.208e-03   5.543 3.39e-08 ***
## PC7          8.480e-03  1.392e-03   6.093 1.34e-09 ***
## PC8          2.970e-02  1.535e-03  19.355  < 2e-16 ***
## PC9          2.270e-03  1.654e-03   1.373 0.170058
## PC10        -7.410e-03  1.894e-03  -3.913 9.43e-05 ***
## PC11        -5.313e-03  2.018e-03  -2.634 0.008519 **
## PC12         7.854e-03  2.115e-03   3.713 0.000210 ***
## PC13         4.048e-03  2.263e-03   1.789 0.073829 .
```

```
## PC14          5.263e-03  2.300e-03   2.289 0.022200 *
## PC15          6.344e-03  2.344e-03   2.706 0.006865 **
## PC16          6.188e-03  2.470e-03   2.505 0.012321 *
## PC17         -7.919e-03  2.596e-03  -3.050 0.002322 **
## PC18          4.489e-03  2.630e-03   1.707 0.088016 .
## PC19          5.542e-03  2.815e-03   1.969 0.049102 *
## PC20          8.705e-03  2.973e-03   2.928 0.003450 **
## PC21          1.290e-02  3.050e-03   4.227 2.48e-05 ***
## PC22         -2.973e-03  3.064e-03  -0.970 0.332027
## PC23          2.298e-03  3.210e-03   0.716 0.474090
## PC24         -5.209e-03  3.317e-03  -1.570 0.116471
## PC25         -4.470e-03  3.457e-03  -1.293 0.196176
## PC26          9.165e-03  3.486e-03   2.629 0.008635 **
## PC27         -1.379e-02  3.590e-03  -3.842 0.000126 ***
## PC28         -1.758e-02  3.805e-03  -4.621 4.08e-06 ***
## PC29         -4.117e-03  3.866e-03  -1.065 0.286982
## PC30          6.348e-04  4.032e-03   0.157 0.874918
## PC31         -1.347e-02  4.066e-03  -3.314 0.000939 ***
## PC32         -4.061e-03  4.197e-03  -0.968 0.333316
## PC33         -1.522e-02  4.291e-03  -3.546 0.000400 ***
## PC34          3.100e-03  4.327e-03   0.716 0.473816
## PC35         -1.231e-02  4.416e-03  -2.787 0.005377 **
## PC36          1.203e-02  4.459e-03   2.699 0.007016 **
## PC37         -4.692e-03  4.732e-03  -0.992 0.321556
## PC38          9.079e-03  4.877e-03   1.862 0.062808 .
## PC39         -1.714e-02  5.020e-03  -3.414 0.000655 ***
## PC40          5.089e-03  5.139e-03   0.990 0.322231
## PC41          1.344e-02  5.226e-03   2.572 0.010180 *
## PC42          1.112e-02  5.344e-03   2.081 0.037585 *
## PC43         -1.348e-02  5.452e-03  -2.472 0.013534 *
## PC44          2.579e-02  5.600e-03   4.605 4.40e-06 ***
## PC45          2.223e-03  5.715e-03   0.389 0.697364
## PC46         -1.482e-02  5.946e-03  -2.493 0.012767 *
## PC47         -1.077e-02  5.962e-03  -1.806 0.071123 .
## PC48          7.530e-03  6.316e-03   1.192 0.233337
## PC49         -1.258e-02  6.520e-03  -1.930 0.053758 .
## PC50         -5.006e-03  6.555e-03  -0.764 0.445139
## PC51          7.567e-03  6.660e-03   1.136 0.256063
## PC52         -1.773e-03  6.746e-03  -0.263 0.792723
## PC53          9.487e-03  6.907e-03   1.374 0.169753
## PC54          1.004e-02  7.169e-03   1.400 0.161694
## PC55         -4.562e-03  7.220e-03  -0.632 0.527561
## PC56         -1.955e-02  7.426e-03  -2.632 0.008551 **
## PC57          6.022e-03  7.692e-03   0.783 0.433741
## PC58          3.771e-03  7.938e-03   0.475 0.634808
## PC59          1.252e-02  7.973e-03   1.571 0.116417
## PC60         -1.736e-02  8.214e-03  -2.113 0.034750 *
## PC61         -1.016e-02  8.491e-03  -1.196 0.231810
## PC62          2.453e-02  8.981e-03   2.731 0.006367 **
## PC63         -3.102e-02  9.232e-03  -3.360 0.000794 ***
## PC64          2.753e-02  9.359e-03   2.941 0.003308 **
## PC65          1.188e-02  9.637e-03   1.233 0.217691
## PC66          2.647e-02  9.931e-03   2.665 0.007756 **
## PC67          8.890e-05  1.016e-02   0.009 0.993022
```

```
## PC68          1.083e-02  1.028e-02   1.053 0.292650
## PC69         -2.372e-02  1.071e-02  -2.215 0.026877 *
## PC70         -2.273e-03  1.103e-02  -0.206 0.836848
## PC71         -1.116e-02  1.139e-02  -0.979 0.327532
## PC72         -1.188e-02  1.163e-02  -1.021 0.307182
## PC73         -7.115e-05  1.179e-02  -0.006 0.995187
## PC74          3.207e-02  1.244e-02   2.579 0.009996 **
## PC75         -1.077e-02  1.255e-02  -0.858 0.390975
## PC76          2.725e-02  1.288e-02   2.116 0.034515 *
## PC77         -1.002e-02  1.326e-02  -0.755 0.450263
## PC78          3.872e-02  1.356e-02   2.855 0.004350 **
## PC79          4.661e-03  1.383e-02   0.337 0.736077
## PC80          1.701e-02  1.406e-02   1.210 0.226598
## PC81         -4.873e-03  1.479e-02  -0.330 0.741771
## PC82          3.826e-02  1.522e-02   2.514 0.012034 *
## PC83          2.435e-02  1.570e-02   1.551 0.121102
## PC84         -2.334e-02  1.622e-02  -1.439 0.150313
## PC85          2.667e-03  1.713e-02   0.156 0.876301
## PC86          3.643e-02  1.753e-02   2.078 0.037851 *
## PC87          3.902e-02  1.758e-02   2.219 0.026601 *
## PC88          1.098e-02  1.831e-02   0.600 0.548873
## PC89          9.069e-02  1.894e-02   4.788 1.82e-06 ***
## PC90          9.816e-03  1.902e-02   0.516 0.605940
## PC91         -3.004e-03  2.028e-02  -0.148 0.882274
## PC92          2.258e-02  2.064e-02   1.094 0.274180
## PC93          1.265e-02  2.179e-02   0.581 0.561580
## PC94         -3.359e-02  2.209e-02  -1.521 0.128442
## PC95         -5.407e-02  2.262e-02  -2.390 0.016934 *
## PC96          1.111e-02  2.358e-02   0.471 0.637640
## PC97         -2.393e-02  2.534e-02  -0.944 0.345234
## PC98          8.628e-03  2.551e-02   0.338 0.735218
## PC99         -1.765e-02  2.601e-02  -0.679 0.497512
## PC100         2.800e-02  2.806e-02   0.998 0.318424
## PC101         4.068e-02  2.869e-02   1.418 0.156473
## PC102        -2.381e-02  3.122e-02  -0.763 0.445652
## PC103        -7.881e-03  3.180e-02  -0.248 0.804281
## PC104        -1.229e-01  3.382e-02  -3.635 0.000285 ***
## PC105        -2.535e-02  3.639e-02  -0.697 0.486131
## PC106        -4.821e-03  4.013e-02  -0.120 0.904394
## PC107         7.911e-03  4.178e-02   0.189 0.849838
## PC108         4.691e-03  4.588e-02   0.102 0.918574
## PC109         8.734e-02  4.871e-02   1.793 0.073135 .
## PC110         7.491e-02  5.420e-02   1.382 0.167131
## PC111         5.931e-02  5.493e-02   1.080 0.280417
## PC112        -1.108e-01  6.092e-02  -1.819 0.069127 .
## PC113        -9.026e-02  6.131e-02  -1.472 0.141163
## PC114        -2.550e-02  7.112e-02  -0.359 0.719951
## PC115        -1.710e-01  7.414e-02  -2.307 0.021184 *
## PC116         2.912e-02  7.712e-02   0.378 0.705779
## PC117         2.628e-02  8.382e-02   0.313 0.753953
## PC118        -7.354e-02  1.004e-01  -0.732 0.464013
## PC119         1.062e-01  1.109e-01   0.958 0.338237
## PC120        -1.459e-01  1.203e-01  -1.213 0.225353
## PC121         2.200e-01  1.364e-01   1.613 0.106926
```

```
## PC122       -1.510e+00  3.120e+00  -0.484 0.628602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1322 on 1871 degrees of freedom
## Multiple R-squared:  0.6979, Adjusted R-squared:  0.6782
## F-statistic: 35.43 on 122 and 1871 DF,  p-value: < 2.2e-16
```

We can then obtain our values for α, β from our a, γ:

```
gamma <- as.matrix(lmodel.all$coefficients[2:123])
A <- as.matrix(pc_crimeData$rotation)
beta <- A %*% gamma
beta
```

```
##                 [,1]
## x.V6    1.158459e-03
## x.V7   -1.361135e-02
## x.V8    4.538361e-02
## x.V9   -7.049684e-03
## x.V10  -2.308081e-03
## x.V11   1.909909e-02
## x.V12   2.332894e-02
## x.V13  -5.082177e-02
## x.V14  -2.634251e-02
## x.V15   4.441465e-02
## x.V16  -1.391306e-02
## x.V17   2.451993e-02
## x.V18  -2.539889e-02
## x.V19  -5.788102e-03
## x.V20   1.330137e-02
## x.V21  -3.069719e-02
## x.V22   1.141787e-03
## x.V23  -2.309006e-03
## x.V24  -9.657565e-03
## x.V25   6.832491e-02
## x.V26   1.132749e-02
## x.V27  -7.277088e-02
## x.V28  -6.069770e-03
## x.V29  -8.368736e-03
## x.V30   4.862996e-03
## x.V31  -6.854617e-03
## x.V32   8.908148e-03
## x.V33   1.194985e-02
## x.V34  -2.305494e-02
## x.V35  -2.510318e-02
## x.V36   1.133176e-02
## x.V37   2.425019e-03
## x.V38   8.277392e-03
## x.V39   4.280229e-02
## x.V40  -1.248538e-02
## x.V41  -5.203969e-03
## x.V42   1.900786e-02
## x.V43   2.879510e-02
## x.V44   1.181022e-01
```

```
## x.V45   5.023253e-02
## x.V46   8.924443e-02
## x.V47  -2.055924e-01
## x.V48  -2.387217e-02
## x.V49  -1.805277e-02
## x.V50  -6.916394e-02
## x.V51   3.868461e-03
## x.V52  -2.083202e-03
## x.V53   1.515117e-02
## x.V54  -3.667834e-02
## x.V55  -6.307376e-03
## x.V56   2.246601e-02
## x.V57  -1.644652e-02
## x.V58   2.109581e-02
## x.V59  -2.244550e-02
## x.V60  -8.421989e-03
## x.V61   1.283928e-02
## x.V62  -1.988285e-02
## x.V63   1.180013e-02
## x.V64   5.019217e-02
## x.V65  -4.273388e-02
## x.V66  -1.601103e-02
## x.V67  -4.444079e-02
## x.V68  -2.758476e-03
## x.V69  -1.709937e-02
## x.V70   1.293402e-01
## x.V71  -1.540354e-02
## x.V72  -5.931697e-02
## x.V73  -1.339625e-01
## x.V74   4.501633e-02
## x.V75   3.197326e-02
## x.V76   1.249994e-02
## x.V77   1.919219e-02
## x.V78  -1.087915e-02
## x.V79   1.003147e-01
## x.V80   1.261731e-02
## x.V81  -1.354420e-02
## x.V82  -4.995937e-03
## x.V83  -4.340051e-03
## x.V84   1.609763e-04
## x.V85  -7.506953e-02
## x.V86   3.010632e-02
## x.V87   3.264191e-02
## x.V88  -5.903674e-02
## x.V89   2.062651e-02
## x.V90  -5.571917e-03
## x.V91   4.690232e-02
## x.V92   2.657209e-03
## x.V93  -9.194261e-03
## x.V94  -1.535766e-02
## x.V95   1.135041e-02
## x.V96   1.643288e-02
## x.V97   2.548713e-02
## x.V98   8.313771e-03
```

```
## x.V99  -9.888430e-03
## x.V100 -8.435677e-04
## x.V101  5.405756e-04
## x.V102 -1.379793e-01
## x.V103 -1.053692e+00
## x.V104 -8.016928e-02
## x.V105 -3.247578e-03
## x.V106 -6.656172e-03
## x.V107 -6.407702e-03
## x.V108  1.279992e-02
## x.V109  1.081014e+00
## x.V110 -1.020243e-02
## x.V111 -2.399313e-03
## x.V112 -8.275559e-03
## x.V113 -1.245445e-02
## x.V114 -2.289675e-03
## x.V115  1.026542e-02
## x.V116  9.385233e-04
## x.V117 -2.249574e-03
## x.V118 -4.328249e-03
## x.V119  5.114529e-04
## x.V120 -2.539817e-03
## x.V121 -1.471767e-02
## x.V122  1.344588e-02
## x.V123  6.582734e-02
## x.V124 -9.821415e-03
## x.V125  6.403888e-05
## x.V126  7.002727e-03
## x.V127 -2.750836e-02
```

```
a <- lmodel.all$coefficients[1]
alpha <- a - t(beta) %*% colMeans(crimeData[,-ncol(crimeData)])
alpha
```

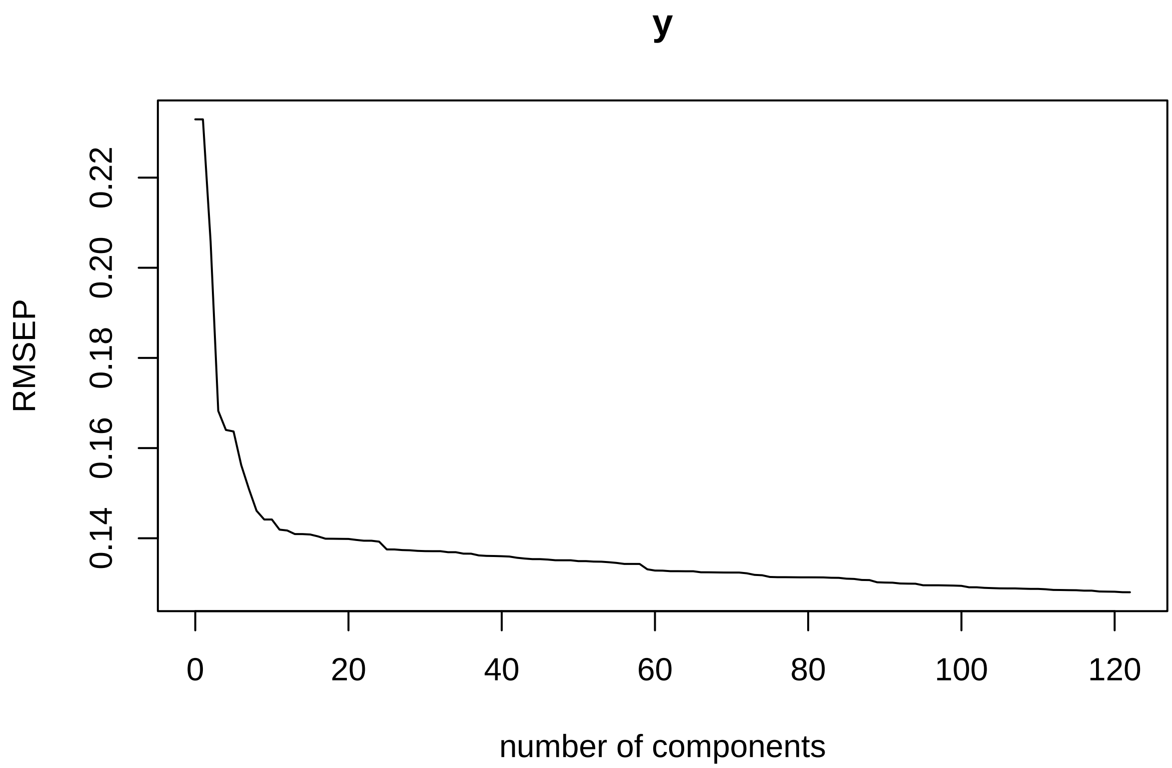
```
##           [,1]
## [1,] 0.3812796
```

Now we would like to see how the performance of our PCR differs as we change the number of principal components. To analyse this we will fit multiple PCR's and measure their performance in terms of the Root Mean Squared Error of Prediction (RMSEP) for different numbers of components:

```
library(pls)
```

```
##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##     loadings
```

```
model <- pcr(y~ ., ncol(crimeData)-1, data=crimeData)
validationplot(model)
```



We see a very quick initial decrease in RMSEP when we increase the number of components, followed by a more gradual decrease. This indicates that somewhere around 10 components the performance increase we get from including more principal components has diminished very significantly. So ideally we should set q=10 as this is a good trade off between performance and keeping dimensionality low.