Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique used to reduce the dimension of a data set.

The data set typically comes in the form of an $n \times p$ matrix

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}.$$

The purpose of PCA is to reduce the size of the data matrix while retaining as much information as possible.

Common Uses of PCA

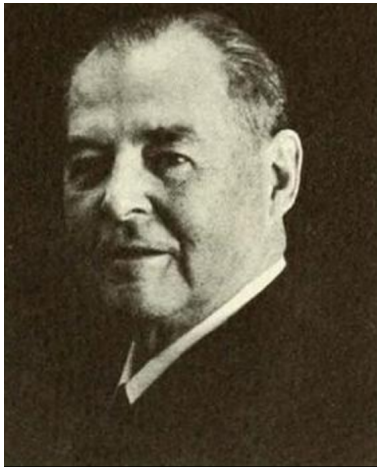
- ✓ Exploration data analysis
- ✓ Data visualization
- ✓ Data compression
- ✓ "Axis rotation" in order to simplify the correlation structure
- ✓ Reduction of the number of variables to ease model construction

Examples of areas where dimension reduction include genetics and text mining, where typically $p \gg n$.

Origin

Harold Hotelling introduced PCA in the following article:

Hotelling, H. (1933).
Analysis of a complex of statistical variables into principal components.
Journal of Educational Psychology, vol. 24, pp. 417–441, 498–520.



Harold Hotelling (b: September 29, 1895; d: December 26, 1973) was an American mathematical statistician and an influential economic theorist. He was on faculty at Stanford (1927–31), Columbia (1931–46) and Chapel Hill (1946–73). A street in Chapel Hill bears his name.

Motivating Example (1–7)

Consider the offensive performance of players in the American Baseball League in 2018.

The data are available here:

http://www.baseball-reference.com/leagues/AL/2018-standard-batting.shtml

```
> dat <- read.csv("2018mlbbatters.csv",row.names="Rk")
> dim(dat)
[1] 604  28
```

Motivating Example (2–7)

List of variables

```
> colnames(dat)
 [1] "Name"        "Age"         "Tm"          "G"           "PA"          "AB"
 [7] "R"           "H"           "X2B"         "X3B"         "HR"          "RBI"
[13] "SB"          "CS"          "BB"          "SO"          "BA"          "OBP"
[19] "SLG"         "OPS"         "OPS."        "TB"          "GDP"         "HBP"
[25] "SH"          "SF"          "IBB"         "Pos_Summary"
```

Examples

G (Games), PA (Plate Appearances), AB (At Bats), R (Runs), H (Hits), RBI (Runs Batted In), SB (Stolen Bases), CS (Caught Stealing), etc.

There are 24 quantitative variables in all.

Motivating Example (3–7)

For example, for Kevin Pillar, former Toronto Blue Jays outfielder, one has


```
> dat[which(substr(dat[,1],1,12)=="Kevin_Pillar"),]
```

```
430 Kevin_Pillar\\pillake01  29 TOR 142 542 512 65 129  40   2 15  59 14  3
18 98 0.252 0.282  SLG   OPS OPS.  TB GDP HBP SH SF IBB Pos_Summary
430 0.426 0.708   93 218   8   6  0  6   0          *8
```

As some players played very little, there are missing data. So let's concentrate on the observations for which complete records are available.

```
dat <- na.omit(dat)
> dim(dat)
[1] 442  28
```

Motivating Example (4–7)

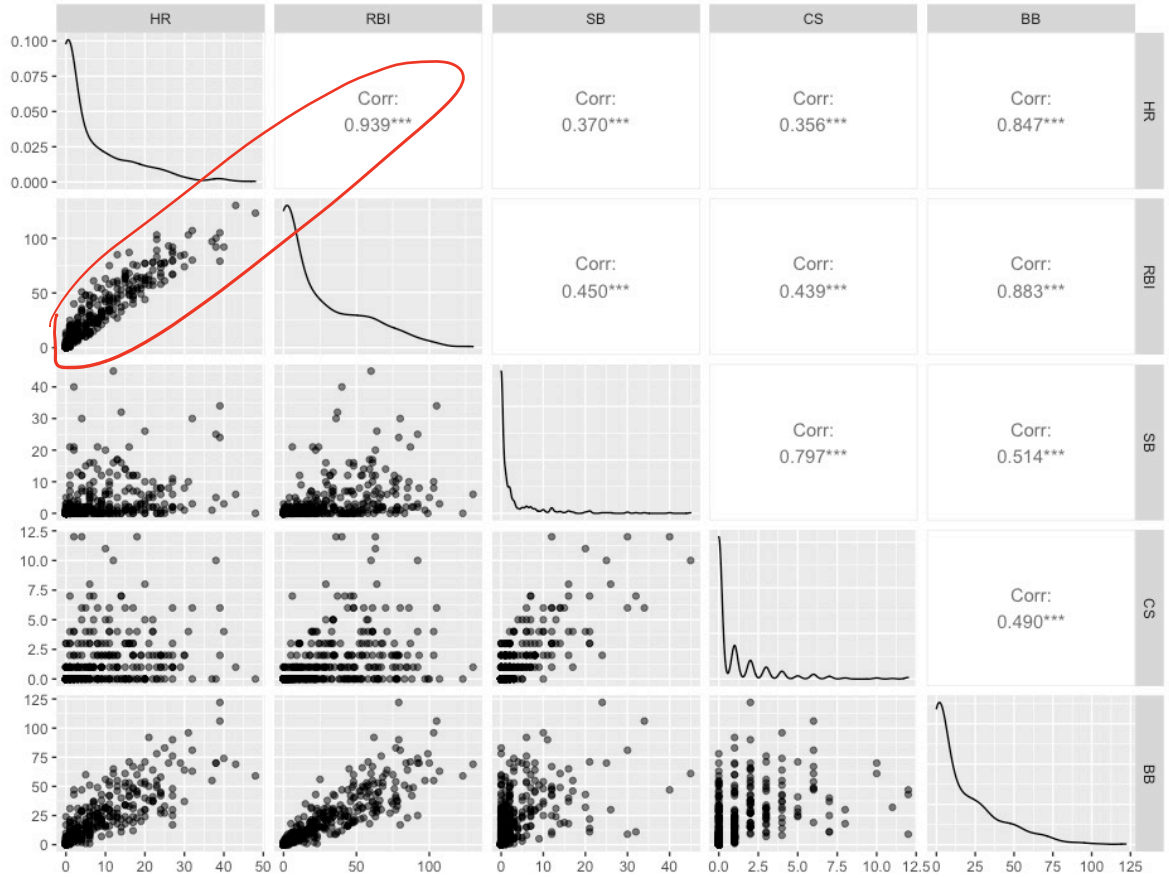
For the 24 quantitative variables available, one could look at bivariate scatterplots, e.g., as follows:

```
library(GGally)
ggpairs(dat[,5:9], aes( alpha = 0.4))
```

However, this makes for

$$\binom{24}{2} = 276 \text{ graphs to look at!}$$

Motivating Example (5–7)



Motivating Example (6–7)

Each player's statistics are highly correlated.

This can be checked by looking at the Pearson correlation matrix.

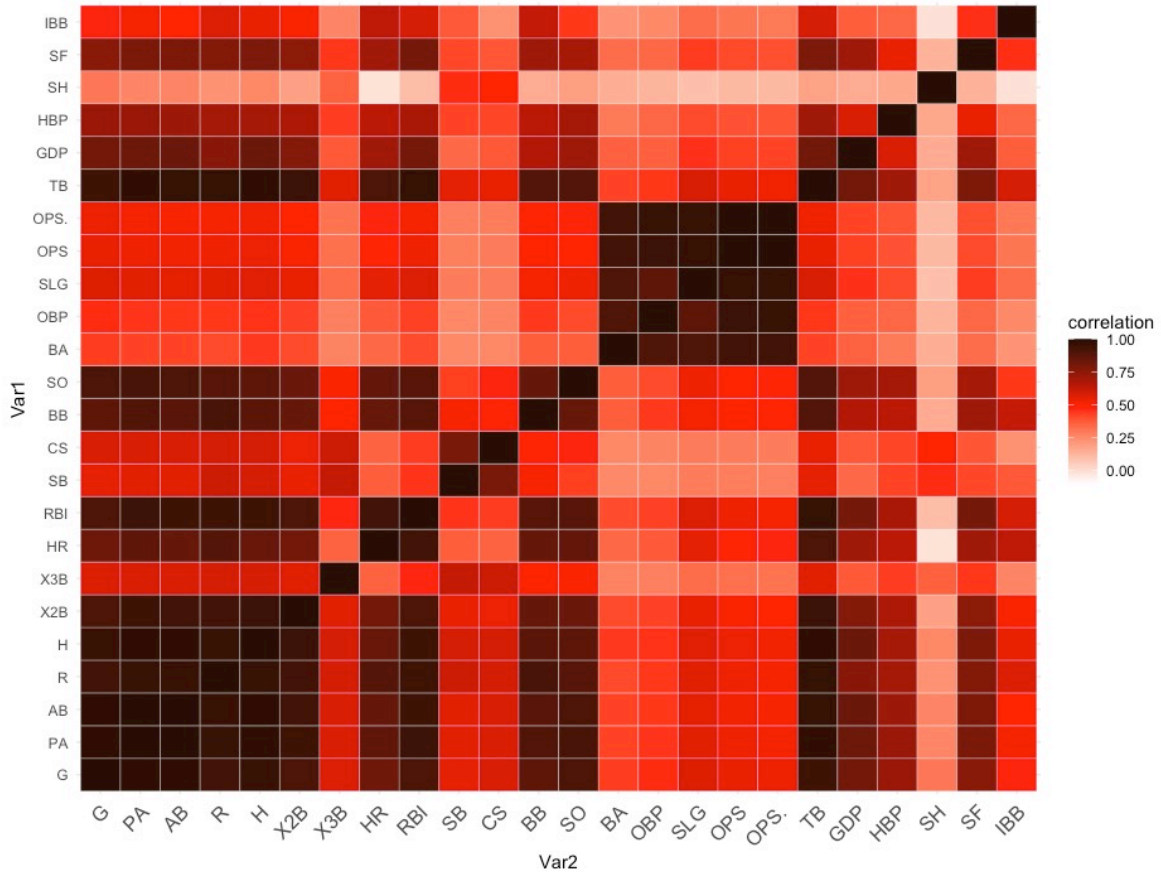
```
> cormat <- cor(dat[,-c(1:3,28)], method="pearson")
> library(reshape2)
> cormat.long <- melt(cormat)
```

This matrix can then be visualized with the command

```
> ggplot(data = cormat.long, aes(Var2, Var1, fill = value))
```

and the geometry `geom_tile`.

Motivating Example (7–7)



What can be Done?

How can one visualize efficiently high-dimensional data?

Can one summarize the information in the 24 variables efficiently?

Can one identify the best players?

PCA is designed to try and answer some of these questions.

In what follows, we will

- ✓ study the mathematics of PCA;
- ✓ learn how to do it in R.

Definition (1–2)

Consider a p-dimensional random vector

$$\mathbf{X} = (X_1, \ldots, X_p)^\top$$

with covariance matrix $\text{var}(\mathbf{X}) = \mathbf{\Sigma}$.

The first principal component is the linear combination

$$Y_1 = \mathbf{w}_1^\top \mathbf{X} = \sum_{j=1}^{p} w_{1j} X_j, \tag{1}$$

$(w_{11}, w_{12} \cdots w_{1p})$

whose variance is largest.

Idea: Capture the greatest amount of variability possible by combining the p variables $X_1, \ldots, X_p$ into a single one.

Definition (2–2)

Upon reflection, there is no solution to the previous problem because if

$$Y_1 = \mathbf{w}_1^\top \mathbf{X} = \sum_{j=1}^{p} w_{1j} X_j$$

were one, one could get a better one by taking

$$kY_1 = k\mathbf{w}_1^\top \mathbf{X} = (k\mathbf{w}_1)^\top \mathbf{X}$$

for any $|k| > 1$. Indeed,

$$\text{var}(kY_1) = k^2 \text{var}(Y_1) > \text{var}(Y_1).$$

Morale: One must add a constraint, viz.

$$\sum_{j=1}^{p} w_{1j}^2 = \mathbf{w}_1^\top \mathbf{w}_1 = 1.$$

Computation of the First Principal Component

It was already seen in the review that

$$\text{var}(Y_1) = \mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1.$$

Therefore, the problem consists in maximizing

$$F(\mathbf{w}_1) = \mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1 \tag{2}$$

under the constraint $\mathbf{w}_1^\top \mathbf{w}_1 = 1$.

Equivalently, one must maximize

$$F(\mathbf{w}_1, \lambda) = \mathbf{w}_1^\top \boldsymbol{\Sigma} \mathbf{w}_1 - \lambda(\mathbf{w}_1^\top \mathbf{w}_1 - 1), \tag{3}$$

where λ is a Lagrange multiplier.

Solution (1–2)

Differentiate F with respect to $w_{11}, \ldots, w_{1p}$ and λ, which makes it possible to take the constraint $\|\mathbf{w}_1\| = 1$ into account, viz.

$$\frac{\partial}{\partial \lambda} F(\mathbf{w}_1, \lambda) = 1 - \mathbf{w}_1^\top \mathbf{w}_1 = 0.$$

Let

$$\frac{\partial}{\partial \mathbf{w}_1} F(\mathbf{w}_1, \lambda) = \left(\frac{\partial}{\partial w_{11}} F(\mathbf{w}_1, \lambda), \ldots, \frac{\partial}{\partial w_{1p}} F(\mathbf{w}_1, \lambda) \right)^\top.$$

Using the results reviewed earlier, one finds

$$\frac{\partial}{\partial \mathbf{w}_1} F(\mathbf{w}_1, \lambda) = 2\boldsymbol{\Sigma}\mathbf{w}_1 - 2\lambda \mathbf{w}_1 = 0. \tag{4}$$

Solution (2–2)

Equation (4), viz.

$$\frac{\partial}{\partial \mathbf{w}_1} F(\mathbf{w}_1, \lambda) = 2\boldsymbol{\Sigma}\mathbf{w}_1 - 2\lambda\mathbf{w}_1 = 0,$$

holds true if and only if

$$\boldsymbol{\Sigma}\mathbf{w}_1 = \lambda\mathbf{w}_1. \tag{5}$$

From the definition of eigenvalue and eigenvector, one deduces that

- ✓ $\mathbf{w}_1$ is a (normed) eigenvector of $\boldsymbol{\Sigma}$;
- ✓ λ is the corresponding eigenvalue.

Maximizing the Variance of Y_1

Given that Y_1 is of the form

$$Y_1 = \mathbf{w}_1^\top \mathbf{X}$$

and that $\mathbf{\Sigma}\mathbf{w}_1 = \lambda \mathbf{w}_1$, one has

$$\text{var}(Y_1) = \mathbf{w}_1^\top \mathbf{\Sigma}\mathbf{w}_1 = \lambda \mathbf{w}_1^\top \mathbf{w}_1 = \lambda.$$

Therefore, the way to maximize this quantity is to take

- ✓ $\lambda = \lambda_1$ is the largest eigenvalue of $\mathbf{\Sigma}$;
- ✓ $\mathbf{w}_1$ is the corresponding normed eigenvector.

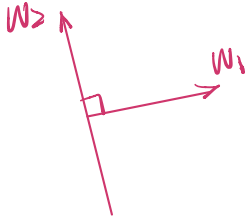
Second Principal Component

The search for the second principal component has two objectives:

- ✓ keep the maximum of variation present in $\mathbf{X}$;
- ✓ simplify the dependence structure to facilitate the interpretation.

Concretely, given Y_1, the second principal component is defined as

$$Y_2 = \mathbf{w}_2^\top \mathbf{X},$$



where

(i) $\text{var}(Y_2) = \mathbf{w}_2^\top \boldsymbol{\Sigma} \mathbf{w}_2$ is maximized given $\mathbf{w}_2^\top \mathbf{w}_2 = 1$;

(ii) $\text{cov}(Y_1, Y_2) = 0$.

Geometric Interpretation (James et al., 2013)

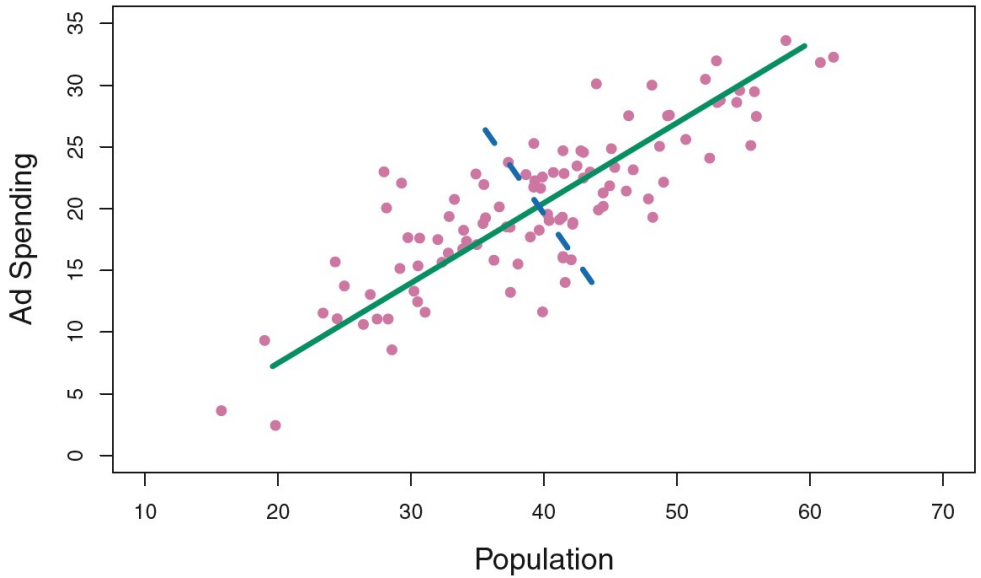


FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Source: G. James, D. Witten, T. Hastie & R.J. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer, New York.

Computation of the Second Principal Component

One has

$$\begin{aligned}\text{cov}(Y_1, Y_2) &= \text{cov}(\mathbf{w}_1^\top \mathbf{X}, \mathbf{w}_2^\top \mathbf{X}) \\ &= \mathbf{w}_1^\top \mathbf{\Sigma} \mathbf{w}_2 = \mathbf{w}_2^\top \mathbf{\Sigma} \mathbf{w}_1 = \lambda_1 \mathbf{w}_2^\top \mathbf{w}_1 = 0\end{aligned}$$

if and only if $\mathbf{w}_2^\top \mathbf{w}_1 = 0$.

Therefore, one looks for the vector $\mathbf{w}_2$ which maximizes

$$F(\mathbf{w}_2, \lambda, \kappa) = \mathbf{w}_2^\top \mathbf{\Sigma} \mathbf{w}_2 - \lambda(\mathbf{w}_2^\top \mathbf{w}_2 - 1) - \kappa(\mathbf{w}_2^\top \mathbf{w}_1 - 0), \qquad (6)$$

where λ and κ are two Lagrange multipliers.

Derivatives must be taken with respect to $\mathbf{w}_2$, as well as λ and κ.

Purpose of the Lagrange Multipliers

Upon differentiation with respect to λ, one finds

$$\frac{\partial}{\partial \lambda} F(\mathbf{w}_2, \lambda, \kappa) = 1 - \mathbf{w}_2^\top \mathbf{w}_2 = 0,$$

which ensures that $\mathbf{w}_2$ has norm 1.

Upon differentiation with respect to κ, one finds

$$\frac{\partial}{\partial \kappa} F(\mathbf{w}_2, \lambda, \kappa) = -\mathbf{w}_2^\top \mathbf{w}_1 = -\mathbf{w}_1^\top \mathbf{w}_2 = 0,$$

which ensures that $\mathbf{w}_1$ and $\mathbf{w}_2$ are linearly independent and that

$$\text{cov}(Y_1, Y_2) = 0.$$

Other Derivatives

A simple calculation shows that

$$\frac{\partial}{\partial \mathbf{w}_2} F(\mathbf{w}_2, \lambda, \kappa) = 2\boldsymbol{\Sigma}\mathbf{w}_2 - 2\lambda\mathbf{w}_2 - \kappa\mathbf{w}_1 = 0. \tag{7}$$

Upon pre-multiplying Eq. (7) by $\mathbf{w}_1^\top$, one finds

$$2\mathbf{w}_1^\top \boldsymbol{\Sigma}\mathbf{w}_2 - 2\lambda\mathbf{w}_1^\top \mathbf{w}_2 - \kappa\mathbf{w}_1^\top \mathbf{w}_1 = 0.$$

However,

$$\mathbf{w}_1^\top \boldsymbol{\Sigma} = \lambda_1 \mathbf{w}_1^\top, \quad \mathbf{w}_1^\top \mathbf{w}_1 = 1,$$

and $\mathbf{w}_1^\top \mathbf{w}_2 = 0$, whence

$$2\lambda_1 \mathbf{w}_1^\top \mathbf{w}_2 - 2\lambda\mathbf{w}_1^\top \mathbf{w}_2 - \kappa\mathbf{w}_1^\top \mathbf{w}_1 = 0 \quad \Leftrightarrow \quad \kappa = 0.$$

Solution

Given that $\kappa = 0$, Eq. (7), viz.

$$\frac{\partial}{\partial \mathbf{w}_2} F(\mathbf{w}_2, \lambda, \kappa) = 2\boldsymbol{\Sigma}\mathbf{w}_2 - 2\lambda\mathbf{w}_2 - \kappa\mathbf{w}_1 = 0.$$

becomes

$$\boldsymbol{\Sigma}\mathbf{w}_2 = \lambda\mathbf{w}_2,$$

which shows that λ is another eigenvalue of $\boldsymbol{\Sigma}$.

Given that

$$\text{var}(Y_2) = \mathbf{w}_2^\top \boldsymbol{\Sigma}\mathbf{w}_2 = \mathbf{w}_2^\top \lambda\mathbf{w}_2 = \lambda,$$

the variance of Y_2 is maximized when $\lambda = \lambda_2$ is the second largest eigenvalue of $\boldsymbol{\Sigma}$ and $\mathbf{w}_2$ is the corresponding normed eigenvector. (We implicitly assume here that $\lambda_2 < \lambda_1$.)

Geometric interpretation (James et al., 2013)

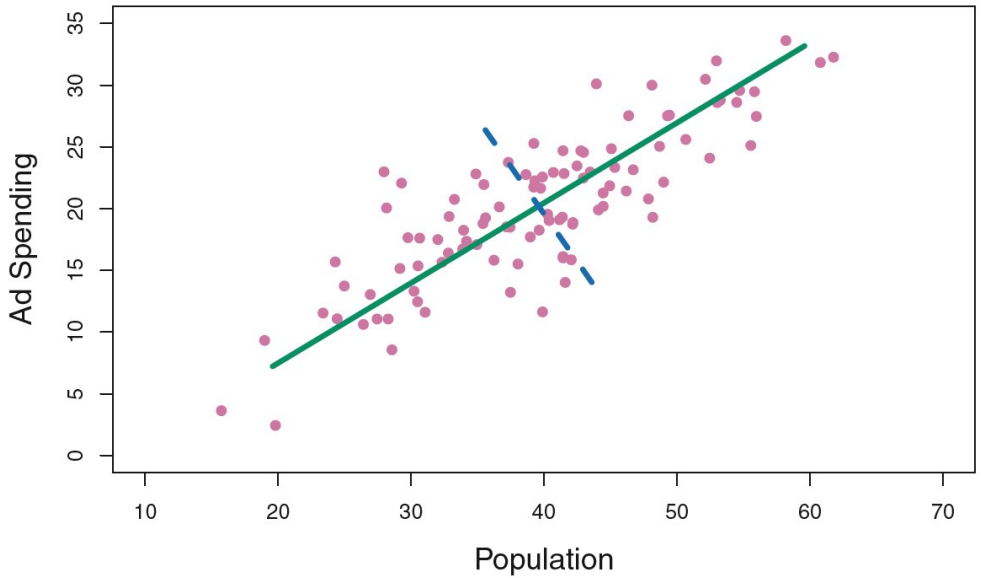


FIGURE 6.14. *The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Source: G. James, D. Witten, T. Hastie & R.J. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer, New York.

Extension

Proceeding in the same way, one finds that

$$Y_k = k\text{th principal component} = \mathbf{w}_k^\top \mathbf{X},$$

where $\mathbf{w}_k$ is the normed eigenvector associated with the kth largest eigenvalue λ_k of $\mathbf{\Sigma}$.

In total, there are then p orthogonal principal components

$$Y_1, \ldots, Y_p$$

which are mutually orthogonal and such that

$$\text{var}(Y_1) = \lambda_1, \ldots, \text{var}(Y_p) = \lambda_p,$$

where $\lambda_1 > \cdots > \lambda_p$ are the p eigenvalues of $\mathbf{\Sigma}$ (assumed to be distinct).

Principal Component Analysis (Matrix Form)

Let $\mathbf{X}$ be $p \times 1$ random vector with $\text{var}(\mathbf{X}) = \mathbf{\Sigma}$. For simplicity, assume that the eigenvalues of $\mathbf{\Sigma}$ are such that $\lambda_1 > \cdots > \lambda_p > 0$.

The corresponding $p \times 1$ vector $\mathbf{Y}$ of principal components is given by

$$\mathbf{Y} = \mathbf{W}^\top \mathbf{X},$$

where

$$\mathbf{W} = (\mathbf{w}_1 | \cdots | \mathbf{w}_p) = \begin{pmatrix} w_{11} & \cdots & w_{p1} \\ \vdots & \cdots & \vdots \\ w_{1p} & \cdots & w_{pp} \end{pmatrix}$$

is a $p \times p$ matrix such that, for each $j \in \{1, \ldots, p\}$,

$$\mathbf{w}_j = (w_{j1}, \ldots, w_{jp})^\top$$

is a normed eigenvector corresponding to λ_j.

Properties of $\mathbf{W}$

- ✓ Each of the columns of $\mathbf{W}$ is unique up to a sign;
- ✓ $\mathbf{W}^\top \mathbf{W} = \mathbf{W}\mathbf{W}^\top = \mathbf{I}_p$ and hence $\mathbf{W}^\top = \mathbf{W}^{-1}$;
- ✓ $\mathbf{\Sigma W} = \mathbf{W\Lambda}$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$.

Given that

$$\text{var}(\mathbf{Y}) = \mathbf{W}^\top \mathbf{\Sigma W} = \mathbf{W}^\top (\mathbf{\Sigma W}) = \mathbf{W}^\top (\mathbf{W\Lambda}) = (\mathbf{W}^\top \mathbf{W})\mathbf{\Lambda} = \mathbf{\Lambda},$$

it follows that for all $i, j \in \{1, \ldots, p\}$,

$$i \neq j \quad \Rightarrow \quad \text{cov}(Y_i, Y_j) = 0,$$

$$i \leq j \quad \Rightarrow \quad \text{var}(Y_i) = \lambda_i \geq \text{var}(Y_j) = \lambda_j.$$

Interpretation

Given that $\mathbf{W}$ is orthonormal,

$$\mathbf{Y} = \mathbf{W}^\top \mathbf{X}$$

represents a rotation of the (random) vector $\mathbf{X}$.

Thus if $\mathbf{x} = (x_1, \ldots, x_p)^\top$ is a realized value of $\mathbf{X}$, then

$$\mathbf{y} = (y_1, \ldots, y_p)^\top = \mathbf{W}^\top \mathbf{x}^\top$$

gives the coordinates of $\mathbf{x}$ in the new system of axes. By definition,

$$y_j = \mathbf{w}_j^\top \mathbf{x} = w_{j1} x_1 + \cdots + w_{jp} x_p$$

is called the score of $\mathbf{x}$ on principal axis $j \in \{1, \ldots, p\}$.

Estimation of $\mathbf{\Sigma}$ (1–2)

In practice, the matrix $\mathbf{\Sigma}$ is unknown.

As seen in the review section of the course, $\mathbf{\Sigma}$ can be estimated from a random sample

$$\mathbf{X}_1, \ldots, \mathbf{X}_n$$

by

$$\mathbf{S}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right) \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)^\top,$$

where

$$\bar{\mathbf{X}} = (\mathbf{X}_1 + \cdots + \mathbf{X}_n)/n.$$

$\mathbf{S}^2$ is in fact a consistent estimator of $\mathbf{\Sigma}$.

Estimation of $\mathbf{\Sigma}$ (2–2)

Typically, $\mathbf{S}^2$ is invertible and its eigenvalues are distinct, viz.

$$\ell_1 > \cdots > \ell_p > 0.$$

These eigenvalues are the estimates of $\lambda_1, \ldots, \lambda_p$, respectively.

Under this assumption, one can write $\mathbf{S}^2$ in the form

$$\mathbf{S}^2 = \hat{\mathbf{W}}\mathbf{L}\hat{\mathbf{W}}^\top,$$

where

$$\widehat{\mathbf{W}} = (\hat{w}_{ij}) \quad \text{and} \quad \mathbf{L} = \text{diag}(\ell_1, \ldots, \ell_p).$$

If $\ell_1 > \cdots > \ell_p$, the principal components are then unique, up to a sign.

Sensitivity to the Scale of $X_1, \ldots, X_p$

Given that principal component analysis (PCA) looks for combinations of variables that maximize variance, a variable X_i that has a large variance is likely to carry a large weight.

Example: Measuring a variable X in meters rather than kilometers multiplies its variance by 10^6, given that

$$X \mapsto 1000\,X = 10^3\,X \quad \Rightarrow \quad \text{var}(10^3 X) = 10^6 \text{var}(X).$$

This conversion would ensure that it has an important weight in all principal components.

To avoid such problems, it is strongly recommended to carry out PCA on the standardized variables, unless the variances are naturally similar.

Recommendation: Standardize the Variables

For each $j \in \{1, \ldots, p\}$, define

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad S_j^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2, \quad X_{ij}^* = (X_{ij} - \bar{X}_j)/\sqrt{S_j^2}.$$

Next, let $\mathbf{X}^*$ be the $n \times p$ matrix with entries X_{ij}^*.

The sample covariance matrix computed from the standardized data, viz.

$$(\mathbf{S}^*)^2 = \mathbf{X}^*\mathbf{X}^{*\top}/n$$

is the same as the sample correlation matrix $\hat{\mathbf{R}}$ computed from $\mathbf{X}$.

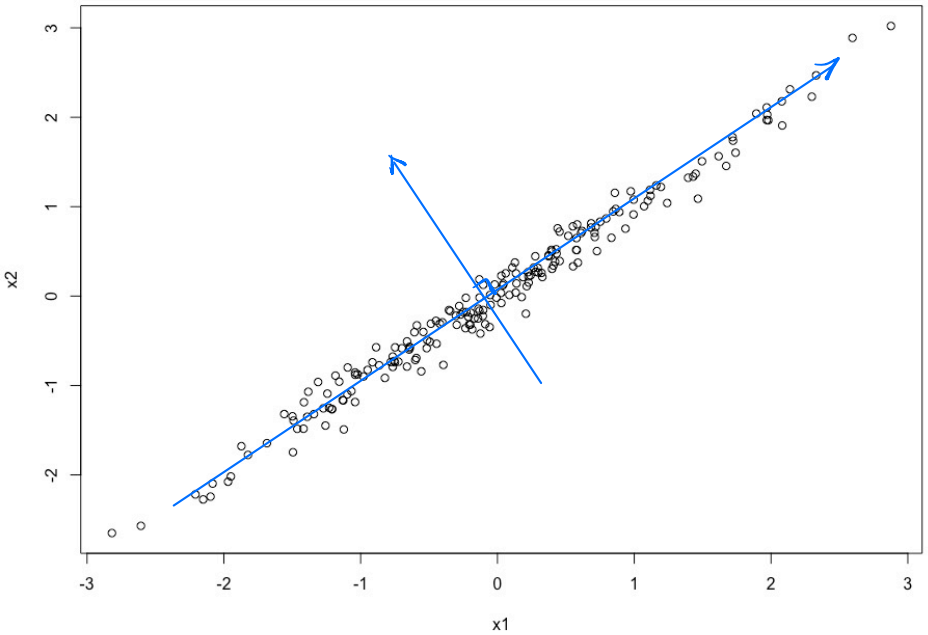
Again, it is strongly recommended to perform PCA on the correlation matrix, $\hat{\mathbf{R}}$, rather than on the raw covariance matrix.

Simulated Example (1–5)

Generate a random sample of size 200 from the bivariate Normal distribution with correlation 0.99:

```
> library(MASS)
> mu <- c(0,0)
> rho <- 0.99
> sigma <- matrix(c(1,rho,rho,1),2)
> N <- 200
> bvn <- mvrnorm(N, mu = mu, Sigma = sigma)
> plot(bvn, xlab="x1", ylab="x2")
```

Simulated Example (2–5)



Because the correlation is very large (0.99), the data are nearly all aligned and a regression model would actually provide a close fit.

Simulated Example (3–5)

Perform the PCA:

```
> acp.v1 <- prcomp(bvn, center=TRUE, scale=TRUE)
> acp.v1
Standard deviations (1, .., p=2):
[1] 1.4108394 0.0976328

Rotation (n x k) = (2 x 2):
            PC1        PC2
[1,] -0.7071068  0.7071068
[2,] -0.7071068 -0.7071068

> var.expl <- acp.v1$sdev^2
> var.expl
[1] 1.990467837 0.009532163
```

$(1.41)^2 + (0.097)^2 = \lambda_1 + \lambda_2 = tr(R) \Rightarrow$

$Var(R_1) \quad Var(R_2)$

Simulated Example (4–5)

Eigenvalue decomposition of $\hat{\mathbf{R}}$: Note $\lambda_1 = 1+\rho = 1.99$, $\lambda_2 = 1-\rho = 0.01$

$\hat{\lambda}_1$ $\hat{\lambda}_2$

- ✓ Eigenvalues: $\ell_1 = 1.9905$, $\ell_2 = 0.0095$; their sum is $2 = \text{tr}(\hat{\mathbf{R}})$.
- ✓ Corresponding eigenvectors (up to a sign): sign can be changed

$$\hat{\mathbf{w}}_1^\top = (0.707, 0.707) \quad \text{and} \quad \hat{\mathbf{w}}_2^\top = (-0.707, 0.707)$$

- ✓ First principal component:

$$\mathbf{Y}_1 = \frac{1}{\sqrt{2}} \mathbf{X}_1^* + \frac{1}{\sqrt{2}} \mathbf{X}_2^*.$$

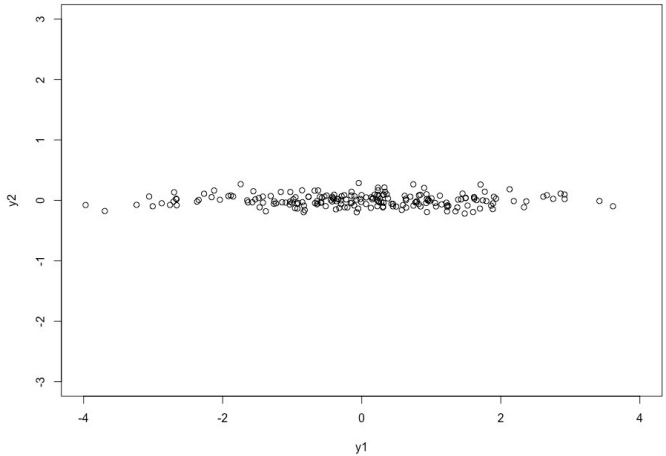
- ✓ Second principal component:

$$\mathbf{Y}_2 = \frac{-1}{\sqrt{2}} \mathbf{X}_1^* + \frac{1}{\sqrt{2}} \mathbf{X}_2^*.$$

Simulated Example (5–5)

Plot the transformed data:

```
> plot(acp.v1$x, xlim=c(-4,4), ylim=c(-3,3))
```



This is just a rotation of the data, so that the variability is maximized on the first axis.

Remark

The PCA representation preserves the distance between points because

$$\widehat{\mathbf{W}}^\top = \widehat{\mathbf{W}}^{-1}.$$

Indeed, for all $i, j \in \{1, \ldots, n\}$,

$$\begin{aligned}
\|\mathbf{Y}_i - \mathbf{Y}_j\|^2 &= (\mathbf{Y}_i - \mathbf{Y}_j)^\top (\mathbf{Y}_i - \mathbf{Y}_j) \\
&= \{\widehat{\mathbf{W}}^\top (\mathbf{X}_i - \mathbf{X}_j)\}^\top \widehat{\mathbf{W}}^\top (\mathbf{X}_i - \mathbf{X}_j) \\
&= (\mathbf{X}_i - \mathbf{X}_j)^\top \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top (\mathbf{X}_i - \mathbf{X}_j) \\
&= (\mathbf{X}_i - \mathbf{X}_j)^\top (\mathbf{X}_i - \mathbf{X}_j) \\
&= \|\mathbf{X}_i - \mathbf{X}_j\|^2.
\end{aligned}$$

Exercise from Last Week

Given $\rho \in (-1, 1)$, consider the 2×2 correlation matrix given by

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

The eigenvalues of $\mathbf{\Sigma}$ are then

$$\lambda_1 = 1 + \rho, \quad \lambda = 1 - \rho$$

and that the corresponding eigenvectors are (up to a sign)

$$\hat{\mathbf{w}}_1^\top = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) \quad \text{and} \quad \hat{\mathbf{w}}_2^\top = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right).$$

Measure of the Quality of PCA

Pillai's trace is a global measure of variation in a random vector $\mathbf{X}$, viz.

$$\text{tr}(\mathbf{\Sigma}) = \text{tr}(\mathbf{\Lambda}) = \lambda_1 + \cdots + \lambda_p. \tag{1}$$

The proportion of variation explained by principal component $\mathbf{Y}_i$ is

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}.$$

Similarly, the first m principal components together explain

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p} \times 100\%$$

of the variability in the vector $\mathbf{X}$.

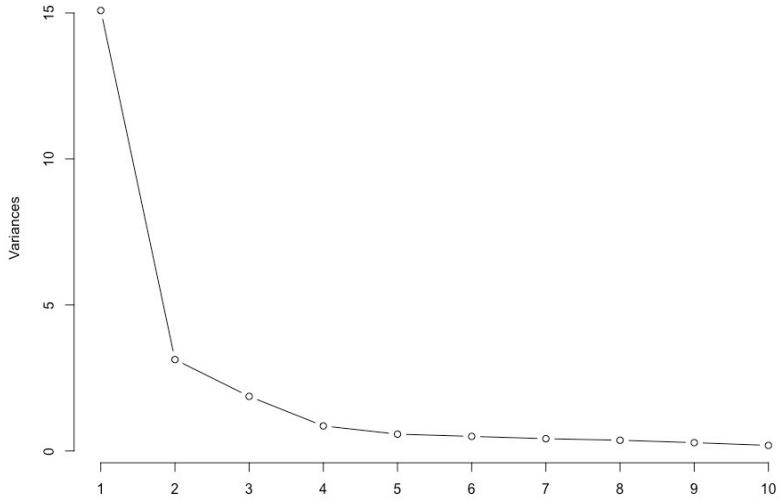
K. C. Sreedharan Pillai

K C Sreedharan Pillai (1920–85) is an Indian statistician who worked for the United Nations (1954–62) and at Purdue University (1962–85). He contributed to multivariate analysis and was a keen golfer, too.

Baseball Example (1–12)

A graph of the eigenvalues in decreasing order is called a scree plot.

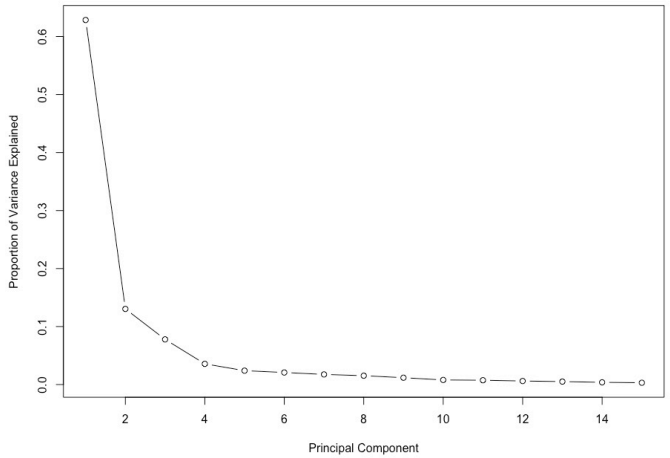
```
> acp.v1 <- prcomp(dat[,-c(1:3,28)], center=TRUE, scale=TRUE)
> screeplot(acp.v1, type = "lines", main="")
```



Baseball Example (2–16)

Equivalently, one can plot a graph of the proportion of variability explained by each eigenvalue.

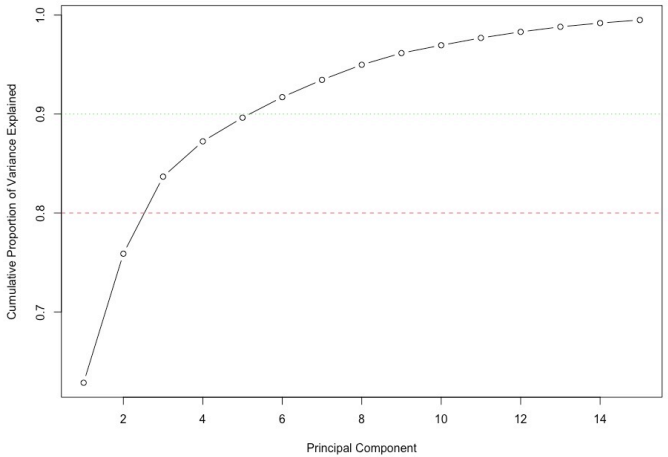
```
> var.expl <- acp.v1$sdev^2
> plot((var.expl/sum(var.expl))[1:15], xlab="Principal Component",
     ylab= "Proportion of Variance Explained", type="b")
```



Baseball Example (3–16)

Yet another option is to plot the cumulative proportion of variability explained by the eigenvalues.

```
> plot(cumsum(var.expl/sum(var.expl))[1:15], xlab="Principal Component",
     ylab= "Cumulative Proportion of Variance Explained", type="b")
> abline(h=0.8,col=2,lty=2)
> abline(h=0.9,col=3,lty=3)
```



Baseball Example (4–16)

One can deduce from these graphs that

- ✓ the first principal component explains 62.8% of the variability;
- ✓ the second principal component explains 13.0% of the variability;
- ✓ the third principal component explains 7.8% of the variability;

Therefore, 83.67% of the entire variability in the data set is explained by the first three principal components.

Baseball Example (5–16)

To look at a plot of the first three principal components (PC), say, one can proceed as follows.

We first bind the PC scores to the data.

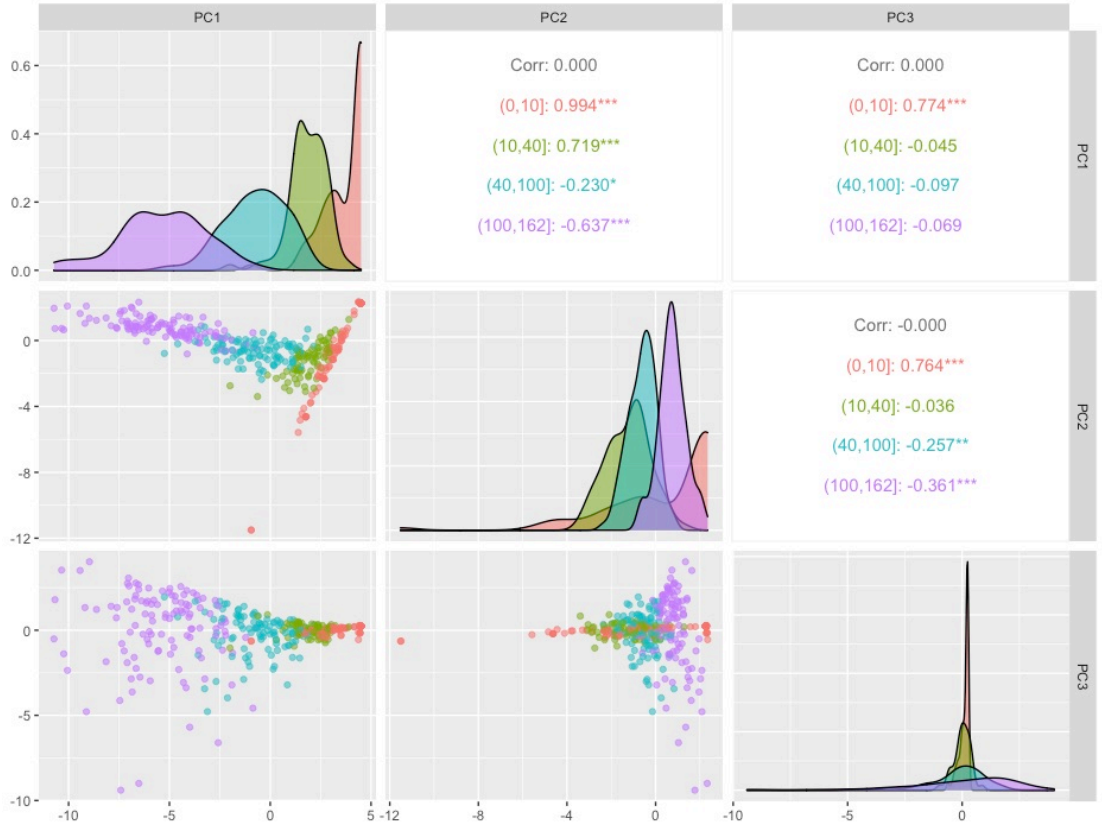
```
> dat2 <- cbind(dat, acp.v1$x)
```

The transformed data are then in columns 29, 30, 31, ...

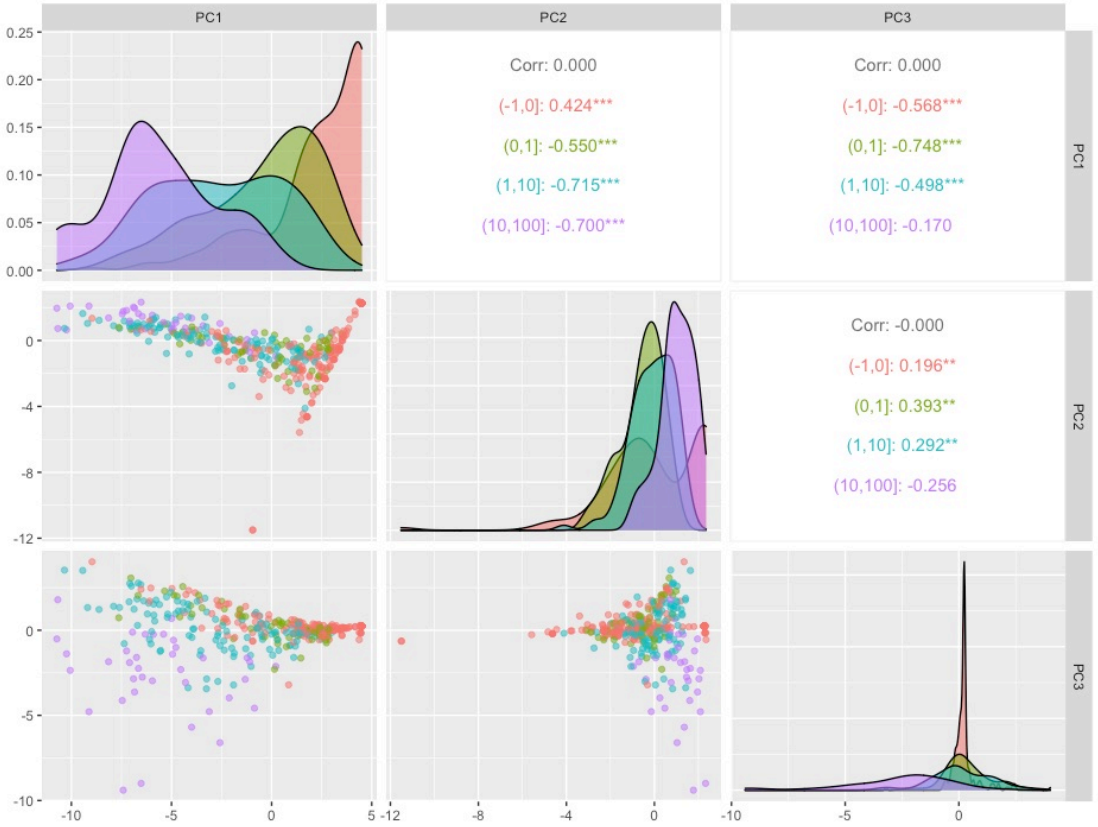
```
> ggpairs(dat2[,29:31], aes(alpha = 0.4,
                            col=cut(dat2$G, breaks=c(0,10,40,100,162))))

> ggpairs(dat2[,29:31], aes(alpha = 0.4,
                            col=cut(dat2$SB, breaks=c(-1,0,1,10,100))))
```

Baseball Example (6–16)



Baseball Example (7–16)



Baseball Example (8–16)

It can be seen on the previous slides that the principal components are indeed uncorrelated.

Their interpretation is the key to a better understanding of the nature of the data at hand.

To help with interpretation, you can

- ✓ compute the loadings;
- ✓ visualize the initial variables in the PC space.

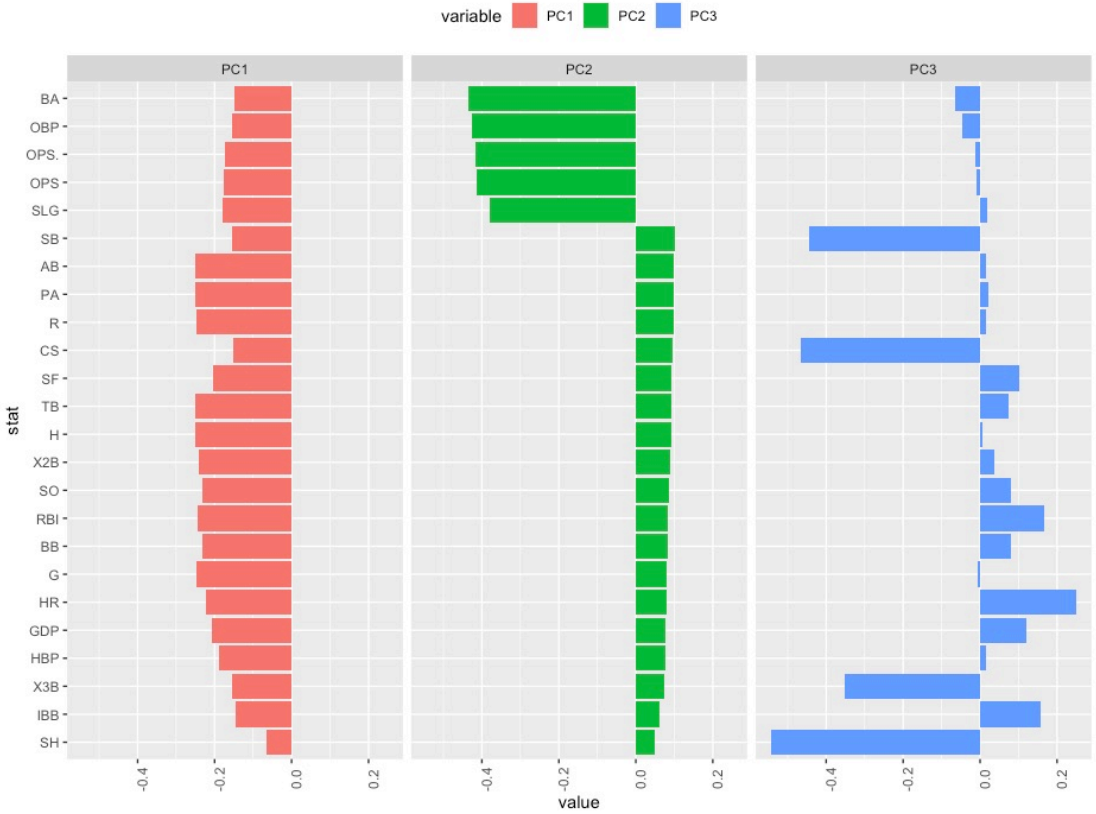
Baseball Example (9–16)

Computation and plot of the loadings:

```
> contrib <- data.frame(stat=rownames(acp.v1$rotation),
                        acp.v1$rotation[,1:3])
> contrib$stat <- factor(contrib$stat, levels =
                        contrib$stat[order(contrib$PC2^2)])
> contrib.long <- melt(contrib)

> ggplot(contrib.long, aes(x=stat,fill=variable, y=value))+
  geom_bar(stat="identity",position=PositionDodge)+
  facet_grid(~variable)+
  theme(legend.position="top",axis.text.x = element_text(angle = 90))+
  coord_flip()
```

Baseball Example (10–16)



Baseball Example (11–16)

PC1 allocates positive weights to all the variables and the differences in weights are not that great.

A precise interpretation is difficult; it seems to say only that "the more you play, the larger your statistics."

PC2 and PC3 are far more interesting:

- ✓ PC2 contrasts BA (Batting Average: Hits/At Bats), OBP (On-Base Percentage), OPS (On-Base Percentage and Slugging), and SLG (Slugging Average: Total Bases/At Bats) with the others variables.
- ✓ PC3 contrasts SB (stolen bases), CS (Caught Stealing), X3B (Triple Hits) and SH (Sacrifice Hits) with the others variables.

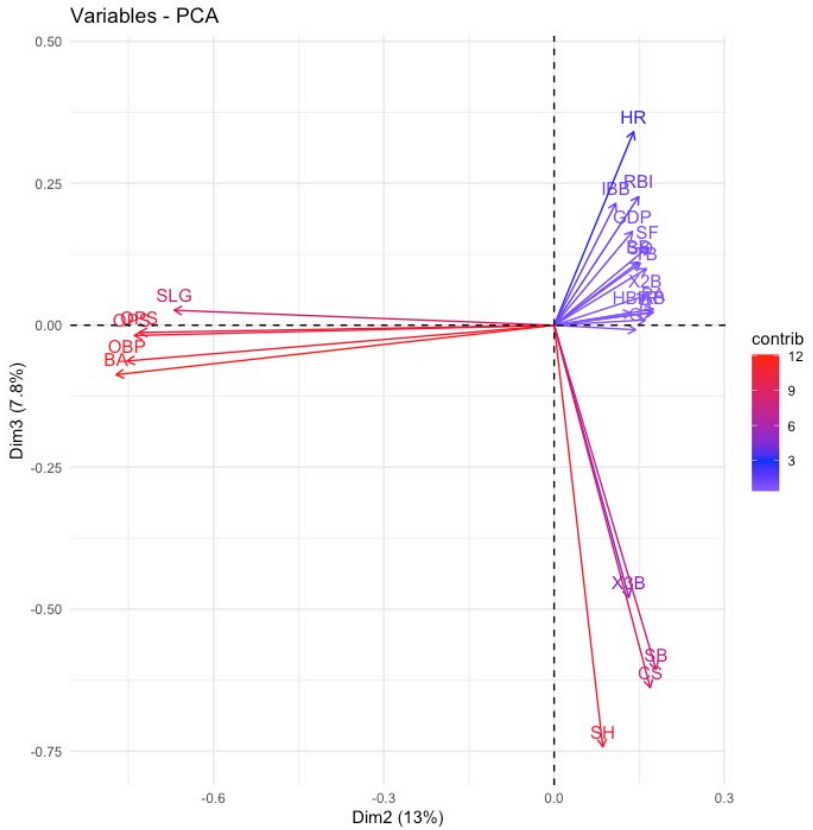
Baseball Example (12–16)

This can be visualized as follows:

```
> library(factoextra)
> fviz_pca_var(acp.v1,col.var="contrib")+
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=3) +
  theme_minimal()

> fviz_pca_var(acp.v1,col.var="contrib", axes=c(2,3))+
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=3) +
  xlim(c(-0.8,0.25))+ylim(c(-0.75,0.45))+
  theme_minimal()
```

Baseball Example (13–16)



Baseball Example (14–16)

Here is a list of the exceptional players based on P2:

```
dat2[dat2$PC1< -8,c(1,3,11,12,17)]
```

	Name	Tm	HR	RBI	BA
49	Andrew_Benintendi*\\beninan01	BOS	16	87	0.290
54	Mookie_Betts\\bettsmo01	BOS	32	80	0.346
70	Alex_Bregman\\bregmal01	HOU	31	103	0.286
135	Khris_Davis\\daviskh01	OAK	48	123	0.247
237	Mitch_Haniger\\hanigmi01	SEA	26	93	0.285
310	Francisco_Lindor#\\lindofr01	CLE	38	92	0.277
338	J.D._Martinez\\martijd02	BOS	43	130	0.330
363	Whit_Merrifield\\merriwh01	KCR	12	60	0.304
446	Jose_Ramirez#\\ramirjo01	CLE	39	105	0.270
528	Giancarlo_Stanton\\stantmi03	NYY	38	100	0.266
556	Mike_Trout\\troutmi01	LAA	39	79	0.312

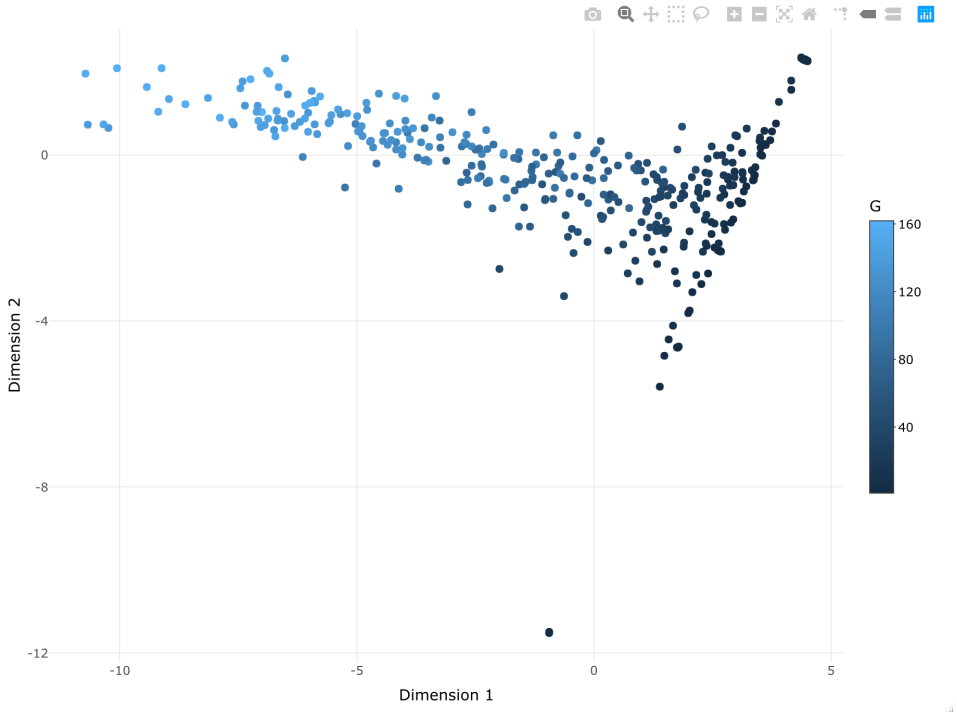
By the way, the Boston Red Sox won the World Series in 2018!

Baseball Example (15–16)

A nice plot with `plotly`:

```
> library(plotly)
> g_ind <- ggplot() +
  geom_point(data = dat2,
             aes(PC1, PC2, text=Name, col=G)) +
  xlab('Dimension 1') +
  ylab('Dimension 2')  +
  theme_minimal()
> ggplotly(g_ind)
```

Baseball Example (16–16)



Determining the Number of Factors to Retain

Selecting the correct number of factors to retain in exploratory data analysis through PCA is of vital importance to researchers.

Often, 2–3 components may suffice when the interest lies strictly in data visualization. However, this may not be sufficient for subsequent analyses based on dimension-reduced data derived from PCA.

Failure to select the "correct" number of components may have dire consequences on subsequent analyses, e.g., on the measurement and interpretation of psychological constructs in psychoanalysis.

In this segment, various rules will be reviewed for determining the number of principal components to retain.

The 80% Rule

Dozens of different methods have been developed for selecting the number of factors.

All the methods employed are heuristics: none can be shown to be universally valid and each can make sense in some circumstances.

Possibly the simplest and most common criterion is the "80% rule," which simply suggests to

> keep as many principal components as necessary to explain 80% of the overall variability in the data set.

The choice of 80% as a cut-off point in not unreasonable but *ad hoc*.

Kaiser's Rule

Suppose that PCA is carried out with the correlation matrix, viz.

$$\hat{\mathbf{R}} = \frac{1}{n} \mathbf{X}^* \mathbf{X}^{*\top}.$$

American psychologist Henry F. Kaiser (1927–92) proposed to

retain all principal components Y_k
with corresponding eigenvalue $\ell_k \geq 1$.

In general, the trace of the correlation matrix $\hat{\mathbf{R}}$ equals

$$1 + \cdots + 1 = p = \ell_1 + \cdots + \ell_p,$$

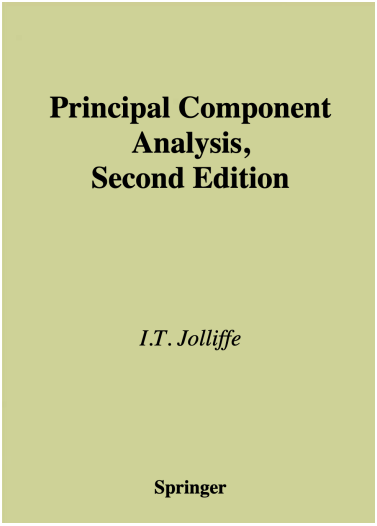
and hence the average of all eigenvalues is $p/p = 1$. Therefore, Kaiser's rule is to keep all eigenvalues that are above average.

Jolliffe's Rule

In his 1986 Springer book, the British statistician Ian T. Jolliffe proposes a slightly more general rule, viz.

keep Y_j if and only if $\ell_j \geq \bar{\ell}$,

where $\bar{\ell} = (\ell_1 + \cdots + \ell_p)/p$.



Jolliffe's rule is slightly more conservative, in the sense that it tends to keep fewer principal components.

When the PCA is carried out with the correlation matrix, this rule states

this rule keeps $Y_j \Leftrightarrow \ell_j \geq 0.7$.

Cattell's Rule

This rule is more pragmatic: it suggests to use the "scree plot" as a basis for decision. Specifically,

- ✓ plott the pairs (j, ℓ_j), i.e., the eigenvalues in decreasing order;
- ✓ retain only those components above the point of inflection.

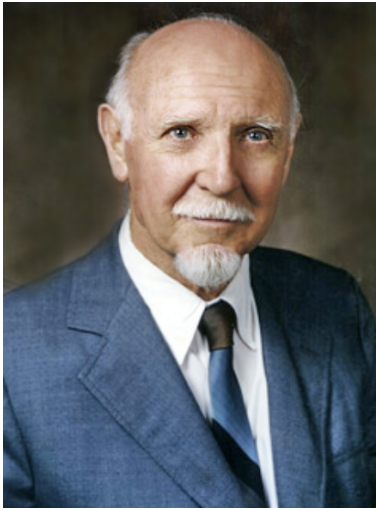
The justification for acting this way is that while ℓ_j decreases (sharply), the corresponding principal component (or eigenvector) explains some of the variability in the data.

When a plateau is reached, it is no longer productive to add principal components.

Raymond Cattell (1905–98)

Cattell's rule is advocated in a 1966 book by the American-British psychologist Raymond Cattell (1905–98).

He was an early proponent of using factor analytic methods instead of what he called "subjective verbal theorizing" to explore empirically the basic dimensions of personality, motivation, and cognitive abilities.



One application of PCA was Cattell's discovery of 16 separate primary trait factors within the normal personality. He called these factors "source traits." It is the basis for the standard "16PF Questionnaire (16PF)."

Horn's Rule

If it is reasonable to assume that the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are multivariate Normal, then one could proceed as follows.

1. Simulate n observations from $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$.
2. Compute the correlation matrix and its eigenvalues $m_{11}, \ldots, m_{1p}$.
3. Repeat Steps 1–2 to obtain K sets of eigenvalues, viz.

$$m_{k1}, \ldots, m_{kp} \text{ for all } k \in \{1, \ldots, K\}.$$

4. Compute, for each $j \in \{1, \ldots, p\}$, $\bar{m}_j = (m_{1j} + \cdots + m_{Kj})/K$.
5. Retain Y_k if and only if $\ell_k \geq \bar{m}_k$.

Given that the eigenvalues of $\mathbf{I}_p$ are $\mu_1 = \cdots = \mu_p = 1$, one can expect $\bar{m}_i > 1$ for roughly $p/2$ value, leading to retain eigenvalues $\ell_j > 1$.

John L. Horn (1928–2006)

Horn's rule is advocated in a 1965 article by the cognitive American psychologist John Leonard Horn (1928–2006).

He identified broad intellectual abilities to supplement fluid reasoning ability (g_f) and crystallized ability (g_c) postulated by his supervisor Raymond Cattell at the University of Illinois.

The Cattell–Horn–Carroll (CHC) theory is the basis for many modern IQ tests. Horn's parallel analysis, a method for determining the number of factors to keep in an exploratory factor analysis, is also named after him.

Another Example (1–7)

Consider crime rates per 100,000 people in seven categories for each of the fifty states of the USA in 1977:

Murder, Rape, Robbery, Assault, Burglary, Larceny, Auto Theft

These data are stored in the file `crime.txt` on `myCourses`.

```
library(calibrate)
md<-read.table("crime.txt",header=TRUE)
attach(md)
```

We could look at the correlation matrix to begin with:

```
cor(md[,2:8])
```

Another Example (2–7)

	MURDER	RAPE	ROBBE	ASSAU	BURGLA	LARCEN	AUTO
MURDER	1.000	0.601	0.484	0.649	0.386	0.102	0.069
RAPE	0.601	1.000	0.592	0.740	0.712	0.614	0.349
ROBBE	0.484	0.592	1.000	0.557	0.637	0.447	0.591
ASSAU	0.649	0.740	0.557	1.000	0.623	0.404	0.276
BURGLA	0.386	0.712	0.637	0.623	1.000	0.792	0.558
LARCEN	0.102	0.614	0.447	0.404	0.792	1.000	0.444
AUTO	0.069	0.349	0.591	0.276	0.558	0.444	1.000

Note that some correlations are fairly high.

```
objPCA<-prcomp(md[,2:8],scale.=TRUE)
objPCA
```

Another Example (3–7)

```
summary(objPCA)

Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     2.0285 1.1130 0.8519 0.5625 0.50791 0.47121 0.35222
Proportion of Variance 0.5878 0.1770 0.1037 0.0452 0.03685 0.03172 0.01772
Cumulative Proportion  0.5878 0.7648 0.8685 0.9137 0.95056 0.98228 1.00000
```

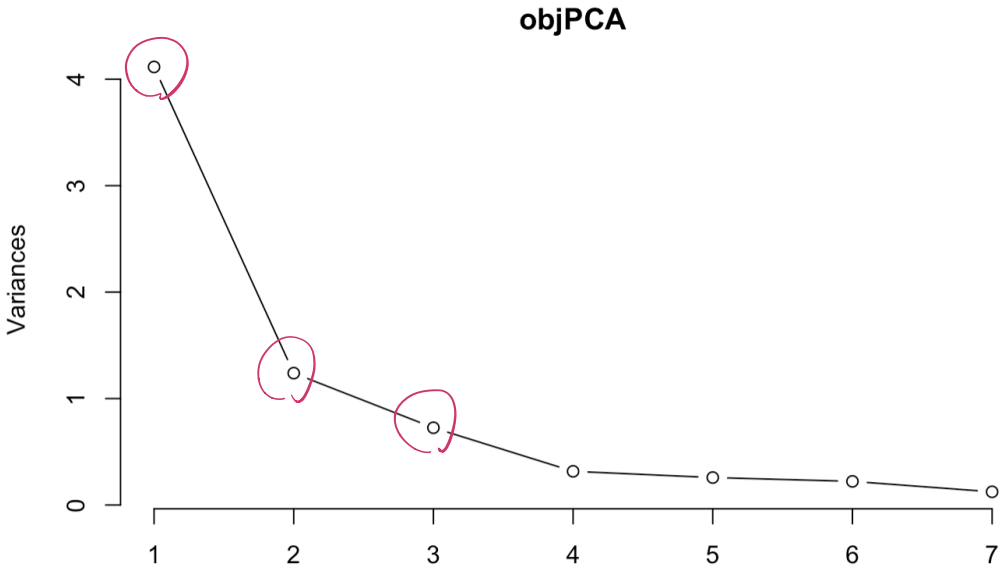
$\sum s^2 = 7$

The first two principal components already account for 76.5% of the variability.

The first three principal components already account for 86.9% of the variability.

Another Example (4–7)

```
screeplot(objPCA,type="lines")
```



Another Example (5–7)

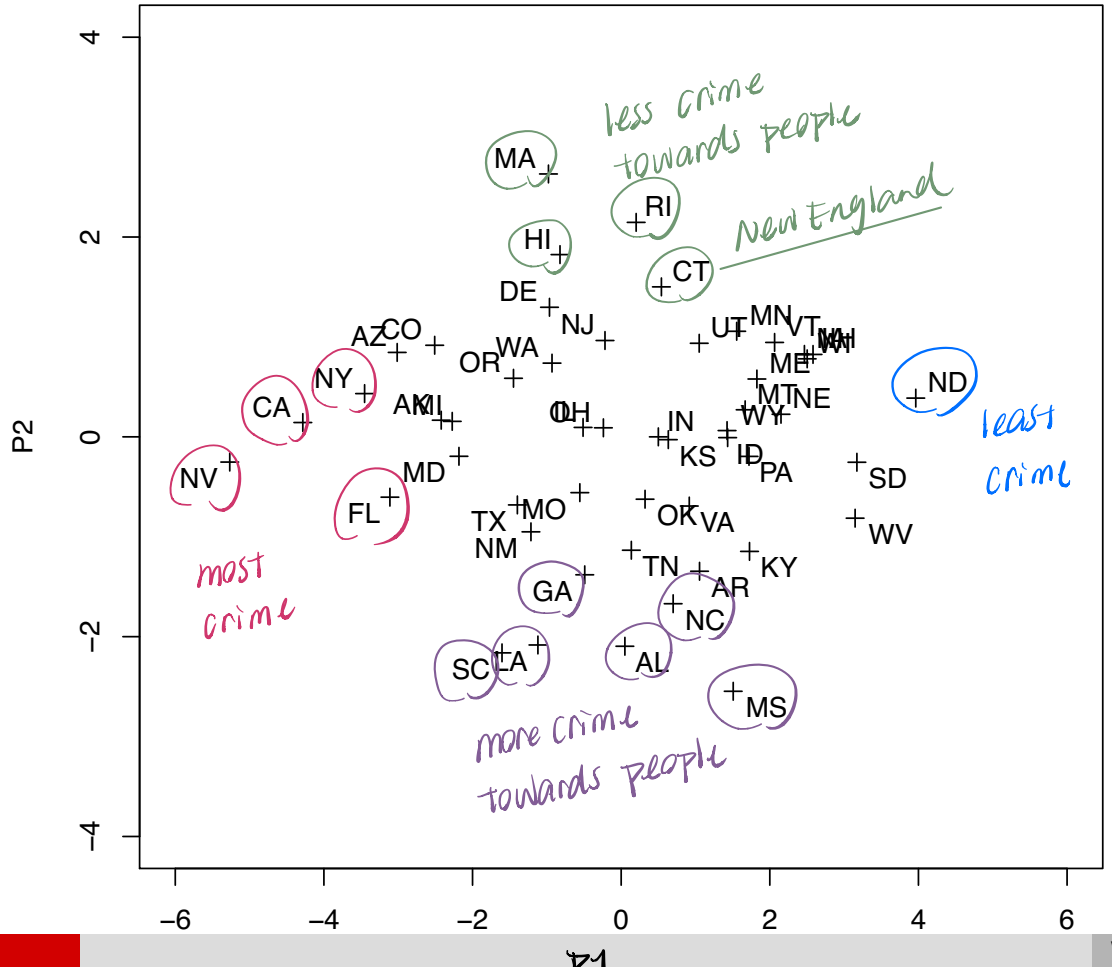
$\sum v^2 = 1$

	PC1	PC2	PC3
MURDER	-0.300	-0.629	0.178
RAPE	-0.432	-0.169	-0.244
ROBBE	-0.397	0.042	0.496
ASSAU	-0.397	-0.344	-0.069
BURGLA	-0.440	0.203	-0.209
LARCEN	-0.357	0.402	-0.539
AUTO	-0.295	0.502	0.568

A state with a high score on the first principal component has high crime rates in all categories.

A state with a high score on the second principal component has a high proportion of crimes against property compared to crimes against people.

Another Example (6–7)



Another Example (7–7)

```
biplot(objPCA)
```