Lecture 9

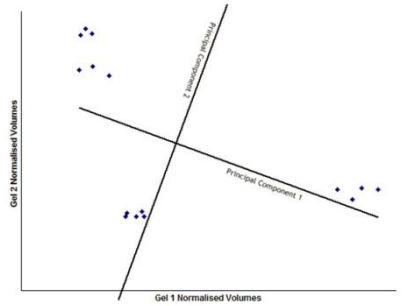
Multivariate statistics: Principal Component Analysis

多变量统计：主成分分析

Principal component analysis (PCA)

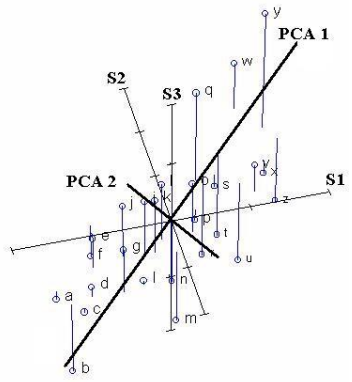
PCA是一种将大量变量简化成一组新的轴或者主成分的技术

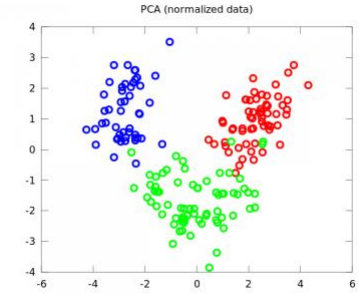
- PCA is a technique to reduce a large number of variables to a reduced set of new axes, or *principal components*



当测量的变量高度相关时，PCA非常有用

- PCA is useful when measured variables are highly correlated





Applications

应用于数据多维化的各个领域

- PCA is applied in various fields where data are multidimensional
  - Psychology (behaviour, personality, clinical assessment)
  - Image analysis (recognition, compression)
  - Finance (risk management)
  - Usually in association with machine learning

PCA: intuitively 直观

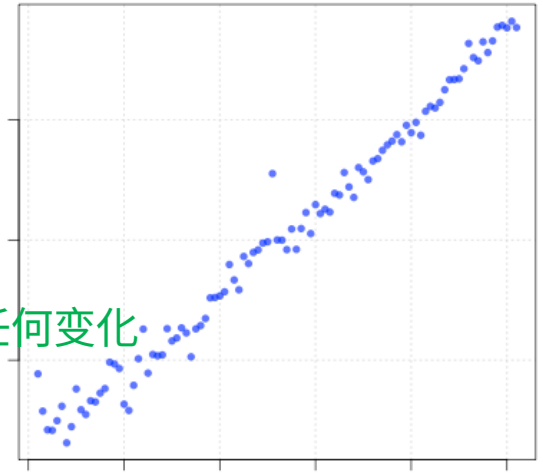
- You have two measurements from 100 trousers
  - **leg length *x*, ankle width *y***;
  - plotting shows a straight line

- How to describe the differences across trousers?
  - we can use the two measurements ("do you have a length 105 and ankle width 20?")
  - or we define the diagonal line as a new axis

  将对角线定义为一个新轴

- The new axis is a 'principal component' 新轴是一个“主成分”
  - new axis or variable may be called 'trousers size' ("do you have a size 10?") 新轴或变量可称为“裤子尺码”
    - notice that 'size' was not an original measurement!
  - new size axis reduced dimensionality (from two measurements to one)

  新轴降低了维度(从两次测量到一次测量)



有第二个轴但在此案例中几乎不能解释任何变化

- ps. there is a second axis but it explains little variation in this case
  - orthogonal to first
  - (possibly random error in measurements)

与第一个轴正交
(可能是测量中的随机误差)

PCA: intuitively

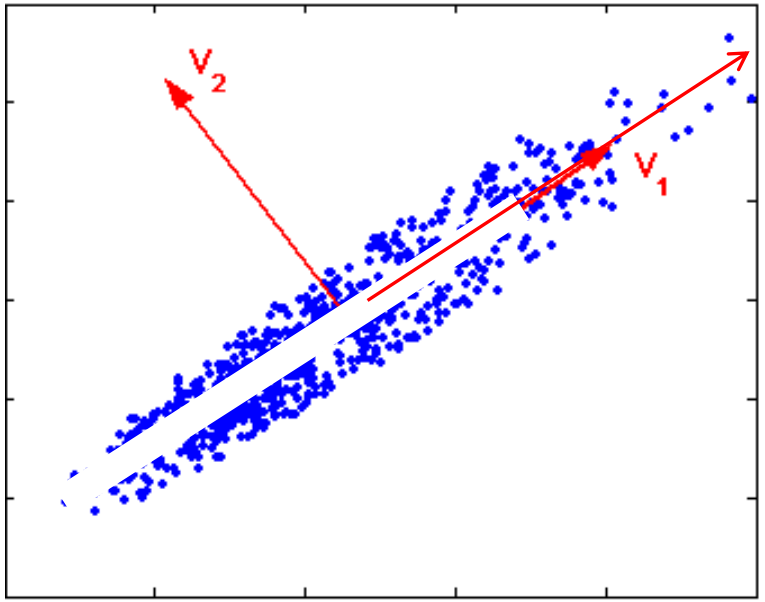
- Now assume that in addition to different ‘sizes’, there are two types of trousers: bootleg and skinny

除了大小的不同，还将裤子分为两种，会看到两条平行线，阔腿裤是高的那条线

- You would see two parallel lines
  - ‘bootleg’ as the higher line

可以定义两个新的坐标轴，第一个是裤子尺寸，第二个是裤子形状

- You can define two new axes
  - the first is ‘trousers size’
  - the second axis reflects ‘trousers shape’ (ankle width for a given leg length), separating bootleg from skinny trousers



在这种情况下没有降为，但两个新的坐标轴更好地描述了裤子的变化(PC1 PC2)

- In this case, there is no reduction of dimensionality, but the two new axes ‘size’ (‘PC1’) and ‘shape’ (‘PC2’) provide a better description of variation in trousers
  - instead of generic centimetre dimensions, specific trouser size and shape measures

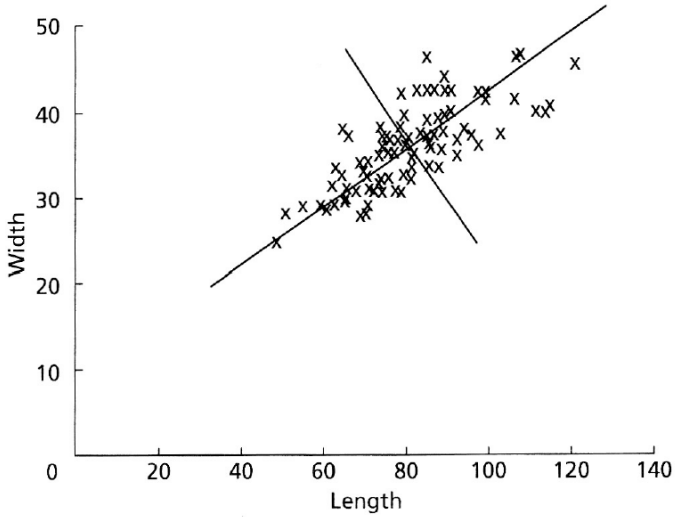
PCA: intuitively

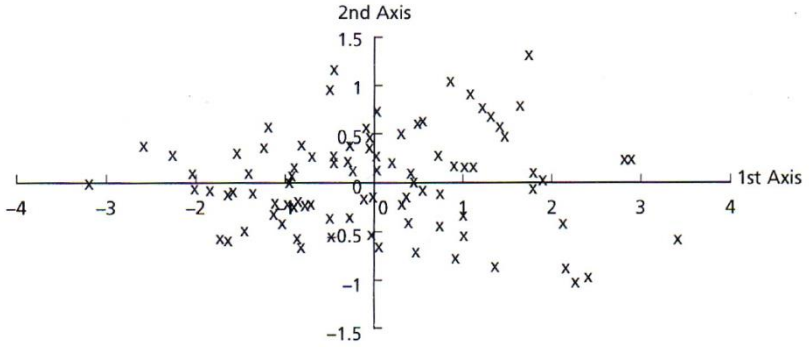
每个PC将占样本原始变化(方差)的一部分

- Each PC will account for a fraction of original variation (variance) in the sample

PC是相互独立的，按重要性(即解释的方差分数)排序的

- PCs are:
  - mutually independent (orthogonal)
  - ordered by importance (i.e. fraction of variance explained): PC1, PC2 etc.

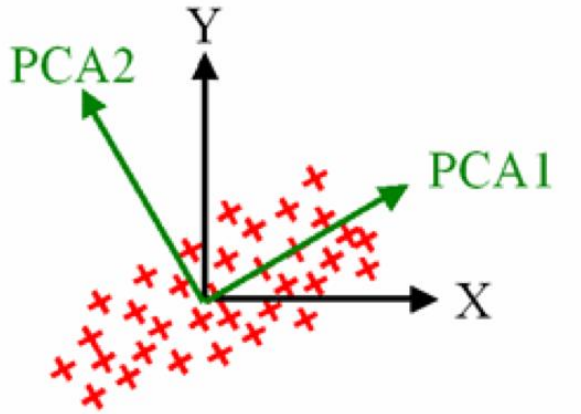




PCA: geometry 几何

- PCs are obtained through rotation of original axes
  - *x* and *y* are orthogonal (perpendicular), therefore PCs remain orthogonal
- Axes rotation is done so that first PC accounts for most variation, and so on
  - PC1, PC2, PC3 are axes of decreasing importance

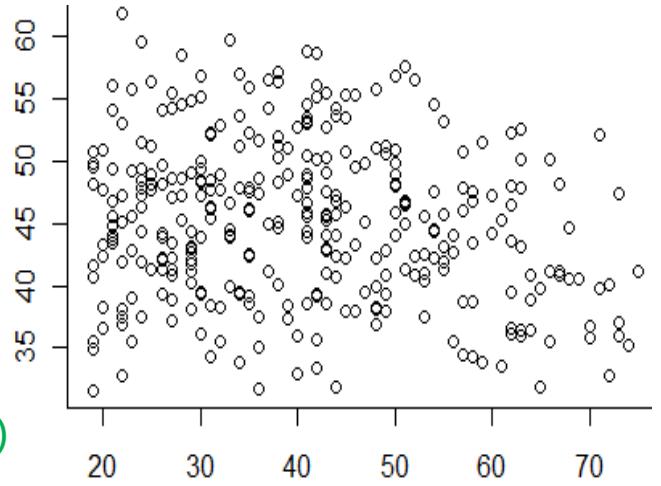
通过旋转原坐标轴得到PC，所以PC仍是正交的



进行轴旋转时，PC1是班花最大的，依此类推，123是重要性递减的坐标轴

- Note: PCA is only useful when original variables are correlated
  - otherwise no rotation will be good

PCA只有在原始变量相互关联时才有用，不然旋转原坐标轴也没用(如右图)



数学上

PCA: mathematics

- A principal component is a linear combination of existing variables

主成分是现有变量的线性组合

- PC1 is:

$$Y_{1,1} = b_{1,1}X_1 + b_{2,1}X_2 \ldots + b_{n,1}X_n$$

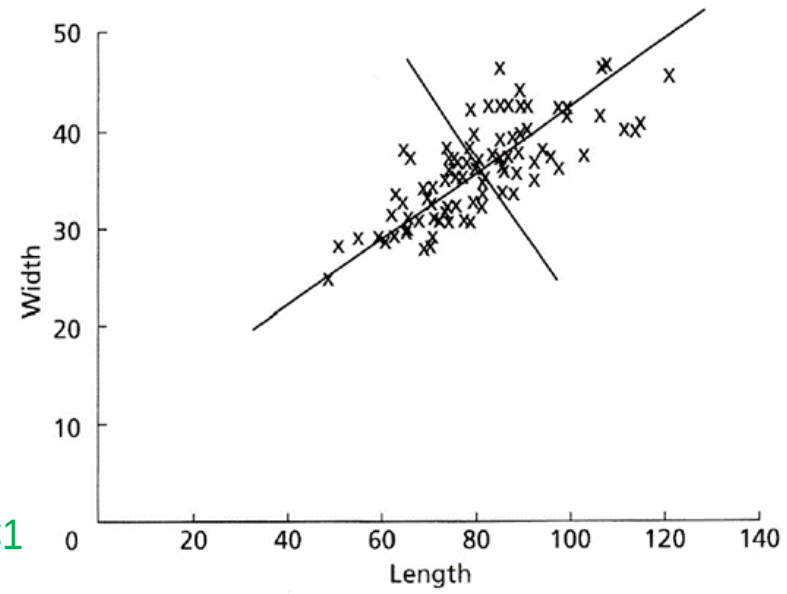
where

Y1是新坐标轴上的“测量值”

- Y_1 = *score* of case 1 on PC1, or its ‘measurement’ on the new axis
- $b_{i,1}$ = *loading* of variable X_i on PC1 (and correlation between PC1 score and X_1 if values are standardised as *z*-scores)
- X_i = original measurements of case 1

如果数值标准化为z分数，则表示PC1和X1之间的相关性

Xi是案例一的原始测量值



And PC2 will be

$$Y_{1,2} = b_{1,2}X_1 + b_{2,2}X_2 \ldots + b_{n,2}X_n$$

Break

Example in R: life history variation

- file *lifehistory*
  - primate data (log values, base 10) on 4 variables: *weight*, *brain size*, *lifespan* and *age at first reproduction*

分析需要从方差-协方差矩阵，或是计算所有变量的协方差开始

- Analysis starts with a *variance-covariance matrix*, or calculation of covariances between all variables
  - note: to calculate covariance and correlation matrix and run PCA, data file must exclude both NAs and non-numerical variables

要计算协方差和相关矩阵并运行PCA，数据文件必须排除NA和非数值变量

- First create a NEW file *lifehistory2* including the four numerical variables of interest:

  > lifehistory2 <- subset(lifehistory, select= c(lifespan, weight, brain, firstrep))

- or just exclude *species* and *group* columns (notice "-" before c)

  > lifehistory2 <- subset(lifehistory, select= -c(group, species))

- then eliminate all NAs in *lifehistory2*　　消除所有NA

  > lifehistory2 <- lifehistory2[complete.cases(lifehistory2),]

如果在其他情况下，指向选择一个变量的完整案例

- note: if in another context you wanted to select complete cases of one variable only:

  > newfile <- oldfile[complete.cases(oldfile$variable)]

方差-协方差矩阵

Variance-covariance matrix

- PCA calculates all covariances between *lifehistory2* variables

```
> cov(lifehistory2)
```

	lifespan	weight	brain	firstrep
lifespan	0.03727196	0.09491868	0.08595568	0.03403697
weight	0.09491868	0.42676653	0.34460016	0.13945175
brain	0.08595568	0.34460016	0.29760857	0.12096556
firstrep	0.03403697	0.13945175	0.12096556	0.06359508

X和X的协方差是X的方差

- Covariance of *x* and *x* is the variance of *x*; 对角线分别是四个变量的方差
- Diagonal of our square 4 x 4 matrix shows variances of *weight*, *brain*, *lifespan* and *first reproduction*
  - hence ‘variance-covariance matrix’

所以叫“ 方差-协方差矩阵 ”

相关矩阵

Correlation matrix

协方差反映的是变化的幅度，因此最好使用相关矩阵

- However, covariances reflect magnitude of variation; it is thus preferable to use a ***correlation matrix***
  - **remember**: correlation is covariance after variables are standardised (mean=0 and sd=1)

相关性就是变量标准化之后的协方差
平均值=0 标准差=1

```
> cor(lifehistory2)
```

	lifespan	weight	brain	firstrep
lifespan	1.0000000	0.7526021	0.8161327	0.6991141
weight	0.7526021	1.0000000	0.9669357	0.8464807
brain	0.8161327	0.9669357	1.0000000	0.8792802
firstrep	0.6991141	0.8464807	0.8792802	1.0000000

在相关矩阵中，所有变量的标准差均为1，因此方差也是1，总方差为n，即变量数

- **Important**: in the correlation matrix, standard deviation of all variables is 1 (therefore variance or squared sd is 1 too), and total variance will be *n,* i.e. variable number
  - In the example, total variance=4, because there are 4 variables

在示例中，总方差=4，因为有四个变量

特征向量和特征值

Eigenvectors and eigenvalues

- Now we need to understand a mathematical property of square matrices (our correlation matrix):

如果A是一个正方形矩阵，那么有n对数字λ和向量v，使得：

- If *A* is a square matrix (*n* rows by *n* columns), then there are *n* pairs of numbers λ and vectors *v* such that:

$$Av=\lambda v$$

向量v(一列n行)是一个特征向量，λ是特征值

  - vector *v* (a single column with *n* rows) is an *eigenvector*
  - number λ (lambda) is an *eigenvalue*

- What does that mean?
  - if you multiply square matrix *A* by *v*, you get the same vector *v* times λ ('eigen' is German for 'same')

示例

Example

- Square matrix *A* has 2 eigenvectors; *v* is one of them

A × v = λ × v

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

(Note

- To find the two *v*, λ pairs, solve equation
  - $Av=\lambda v$
  - $Av - \lambda Iv = 0$
  - $(A - \lambda I)v = 0$

(A = n x n matrix, I = identity matrix)

- determinant of $(A - \lambda I)$ must be 0
- i.e. you get a quadratic equation and two λs
- for each λ, a *v* can be calculated)

ps. you don't need to know this)

PCA

计算平方相关(方差-协方差)矩阵n × n的特征向量和特征值

- So what is PCA?
- It is the calculation of the eigenvectors and eigenvalues of the square correlation (variance-covariance) matrix *n* x *n*
  - *n* is the number of variables n是变量个数

- For *n* variables, we obtain *n* PCs (=eigenvectors) and n eigenvalues

对于n个变量，我们可以得到n个PC(=特征向量)和n个特征值
每个特征向量都有那个PC自己的b

- Each eigenvector has the loadings (the b coefficients) on that PC
  - $Y_1 = b_1X_1 + b_2X_2 \ldots + b_nX_n$

每个特征值表示该PC能解释多少方差

- And each eigenvalue is how much variance is explained by that PC
…and so on until the last (nth) PC

PCA in R

要运行PCA，我们可以用函数prcomp或者princomp

- To run PCA, functions *prcomp* or *princomp* can be used
- We will be using *prcomp*

我们希望PCA基于相关矩阵而不是协方差矩阵，所以要加上参数scale=T

```
> pca1 <- prcomp(lifehistory2, scale=T)
> pca1
Standard deviations:
[1] 1.8673661 0.5681131 0.4068723 0.1569917

Rotation:
             PC1        PC2        PC3        PC4
lifespan  0.4663392  0.8530973  0.1961560 -0.1275759
weight    0.5132214 -0.2365816 -0.5756064 -0.5910248
brain     0.5264700 -0.1008899 -0.2995261  0.7892621
firstrep  0.4918952 -0.4539553  0.7351764 -0.1071420
```

- We want PCA based on a correlation matrix (rather than covariance matrix); therefore **don't forget the argument**:

- scale=T
  - scales variance to 1 and mean to 0

将方差缩放为1，平均值缩放为0

Output

四个特征向量或者PC

- Rotation: the 4 eigenvectors or PCs

  PC1栏是PC1的各个B系数

- Column PC1: *b* loadings of PC1

> pca1$rotation会只显示下面那个表格

```
> pca1 <- prcomp(lifehistory2, scale.=T, retx=T)
> pca1-
Standard deviations:
[1] 1.8673661 0.5681131 0.4068723 0.1569917
```

Rotation:

	PC1	PC2	PC3	PC4
lifespan	0.4663392	0.8530973	0.1961560	-0.1275759
weight	0.5132214	-0.2365816	-0.5756064	-0.5910248
brain	0.5264700	-0.1008899	-0.2995261	0.7892621
firstrep	0.4918952	-0.4539553	0.7351764	-0.1071420

- PC1 is :
  - Y_1 = 0.466(lifespan) + 0.513(weight)+ 0.526(brain)+ 0.492(firstrep)

所有的b均为正值

  - all *b* loadings are positive

  = PC1 seems to be a 'size' axis

    - (like ¨trousers size¨)

如果PC1是大小，则应与weight变量密切相关

- If PC1 is 'size', it should strongly correlate with variable *weight*

```
> cor(pcscores$PC1, lifehistory$weight)
[1] 0.9583722
```

- On PC2, *lifespan* is very positive while *firstrep* is very negative
  - (we'll interpret PCs later)

cor是计算相关性系数的，取值范围是-1到+1
当其为正值，表示x和y正相关，值越大正相关性越强
当其为负值，表示x和y负相关，值越小负相关性越强
当其趋于0，表示x和y基本不相关，=0表示不相关

先subset选出三个需要的列，然后再消除NA
因为如果先消除NA的话，有的不需要的列有NA，就会使本可以用的那三列数据的那一行也被消除
PPT改了hh

Exercise:

- Run a PCA using the *hdr* dataset, using only three variables: *lifespan*, *schoolingyears*, and *GNI*. and then write down the equation describing PC1 (i.e. Y1 = b1X1 + b2X2+…etc)

Break

PC的保留标准

PC retention criteria

- Since PCs decrease in order of importance, do we need all of them to explain patterns?

不需要保留所有PC
因为我们希望其中的几个因素可能足以解释样本中的大部分差异
这就是降维的意义所在

Not necessarily;

- if you start with 20 variables, you end up with 20 PCs
- But we hope that a few of them may be enough to explain most variation in sample
  - this is the point of reducing dimensionality!

根据特征值，有多种保留PC的方法

- There are various criteria for retention of PCs based on eigenvalues (proportion of variance explained):
  - $\lambda > 1$
  - scree plots
  - individual/cumulative variance thresholds
  - ‘interpretability’

Rule 1: λ >1

第一条标准是只保留 λ >1的PC

- A first criterion is only to keep PCs with λ >1
  - rationale: variances of each original variable are scaled to one; it makes no sense to use a PC explaining less variance than an original variable

理由：每个原始变量的方差都按照比例变成了1，这使得解释方差小于原始变量的PC没有意义

- To show eigenvalues (=variance, or sd^2): 显示特征值

```
> pca1$sdev^2
[1] 3.48705601 0.32275255 0.16554505 0.02464639
```

- Confirming that total variance is n=4: 确认总方差为4

```
> sum(pca1$sdev^2)
[1] 4
```

- Based on this criterion, we should retain only PC1 (λ = 3.487 = total variance explained by PC1)

根据这一标准，应只保留PC1(λ=3.487=PC1解释的总方差)

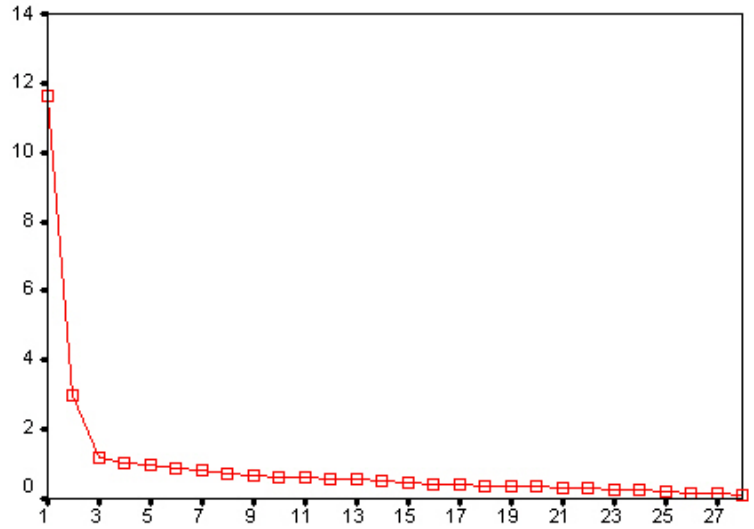
Rule 2: Scree plots

可以绘制出每个PC的特征值，观察其变化规律

- We can also plot eigenvalues λ from each PC and look at the pattern

scree是指堆积在悬崖底部的碎屑

- (ps. ‘scree’ is the debris accumulated at the base of a cliff)



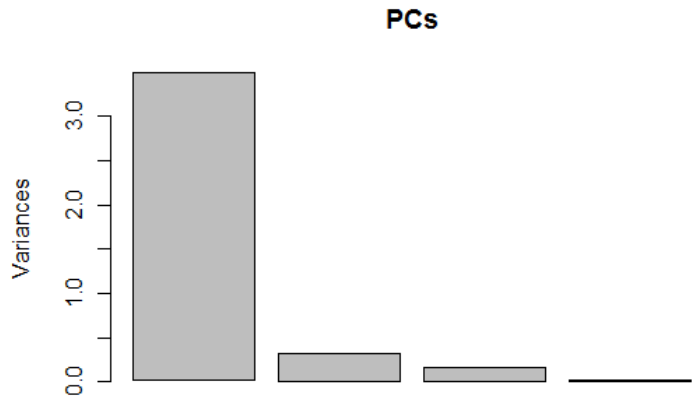
途中凹陷表明有截断点，在截断之前的PC应该爆裂，之后的丢弃

- A dip in plot suggests cut-off point:
  - PCs before dip should be kept
  - the others (‘scree’), discarded

Scree plots

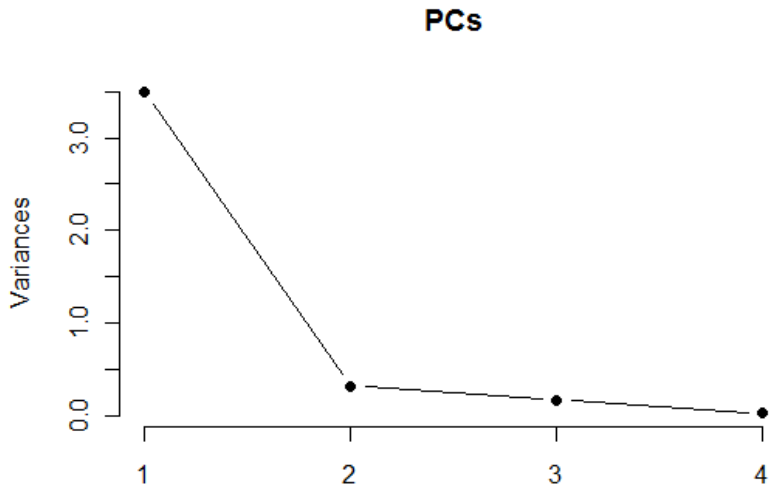
- Scree plot as bars: 柱状图

```
>screeplot(pca1, main="PCs", pch=16)
```



- Scree plot as line: 线图
  - add *type=“line*

```
>screeplot(pca1, main="PCs", type="line", pch=16)
```



- ps. all graphical parameters apply (colour, x and y labels etc.)
- Conclusion: only PC1 should be retained (PCs 2, 3 and 4 are ‘scree’)

结论：应该只保留PC1，其他的舍弃，是碎石

解释的方差分数

Rule 3: fraction of variance explained

另一条规则是只能保留能解释特定部分方差的PC

- Another rule is to only retain PCs explaining over a given fraction of variance
  - *individually*: keep PCs that explain >10% or 15% of total variance
    - but this may clash with the first rule ($\lambda>1$)
  - *as a set*: keep PCs that *cumulatively* explain >80% or 90%
    - more useful as a supplementary rule

个别：保留能解释10%或15%以上总变量的PC
但这可能和第一条规则冲突

作为一组：保留累积解释率大于80%或90%的PC

作为补充规则更有用

- To check for individual and cumulative percentages of variance:

```
> summary(pca1)
Importance of components:
                          PC1     PC2     PC3     PC4
Standard deviation     1.8674 0.56811 0.40687 0.15699
Proportion of Variance 0.8718 0.08069 0.04139 0.00616
Cumulative Proportion  0.8718 0.95245 0.99384 1.00000
```

Exercise:

- Based on the table on this slide, manually calculate the proportion of variance explained by each PC
- (i.e. show that PC1 explains 87.2% of variance etc.)

```
(sd**2)/n = proportion of variance
```

- PC1 explains 87% of all variance and should be kept
- OR you may want to keep PC1 and PC2 (that cumulatively explain 95% of variance)
  - but PC2 explains less than 10% individually (and is in the scree); rules must be used together

规则必须一起使用

可解释性

Rule 4: ‘Interpretability’

- As we’ve seen, there is no test for identifying acceptable PCs

我们应该一起使用这些规则，他们往往意味着不同的选择

- We should use the rules together; they often imply different choices
  - e.g a PC may explain <10% of total variance (‘don’t keep it’), but may be required to explain >80% of total variance cumulatively (‘keep it’)

这意味着：保留PC的一个原因是其潜在的“ 意义 ”

- This means that a reason for keeping a PC is its potential ‘meaning’
  - i.e. if I think that variation in sample is basically size and dimorphism, it makes sense to keep and interpret PC1 as ‘size’ (if all variables have positive loading on it), and PC2 as ‘shape’ or ‘economic development’ etc.

Interpretability

```
> pca1 <- prcomp(lifehistory2, scale.=T, retx=T)
> pca1
Standard deviations:
[1] 1.8673661 0.5681131 0.4068723 0.1569917

Rotation:
              PC1        PC2        PC3        PC4
lifespan 0.4663392  0.8530973  0.1961560 -0.1275759
weight   0.5132214 -0.2365816 -0.5756064 -0.5910248
brain    0.5264700 -0.1008899 -0.2995261  0.7892621
firstrep 0.4918952 -0.4539553  0.7351764 -0.1071420
```

在有疑问的情况下，最好不要让PC无法解释

- Back to our PCs and loadings:
- PC1: 所有的loading均为正的
  - all variables have positive loadings
  - i.e. PC1 represents some property that is large when everything else is large too
  - this suggest a ‘size’ or ‘scale’ component

即PC1代表某种属性，当其他所有属性都很大时，这种属性也很大表明他有“大小”或者“规模”的成分

- PC2:
  - strong positive loading on *lifespan*, strong negative loading on *firstrep*
  - weak loadings (near 0) from weight and brain
    - i.e. a PC based mostly on timing variables, not size variables
  - PC2 score is higher for species where lifespan is long but first reproduction is early
- In doubt, it is wiser not to keep PCs with no straightforward interpretation

分组案例

Grouping cases

我们还可以研究PC的分组模式

- We can also look at grouping patterns formed by PCs

- We can extract PC scores (Y_1, Y_2,…) for each case 可以为每组情况提取PC分数
  - remember: each case has *n* variables, *n* new PCs and *n* new measurements ('PC scores')

- Extracting PC scores into a matrix: command *$x*

  > matrixpc <- pca1$x 将PC分数提取到矩阵中

- We saved matrix with scores; now we convert matrix into data frame

  > pcscores <- data.frame(matrixpc) 转换成数据框

现在使用函数data.frame将pc分数添加到原始数据集中，这个操作穿件了新的数据集

- Now we add *pcscores* (with the new variables or PCs) to our original *lifehistory* dataset (which still has *group* and *species* columns) using the function *data.frame*, which creates new datasets

(ps. Now we are using file *lifehistory*, not *lifehistory2*)

> lifehistorypc <- data.frame(lifehistory, pcscores)

Grouping cases

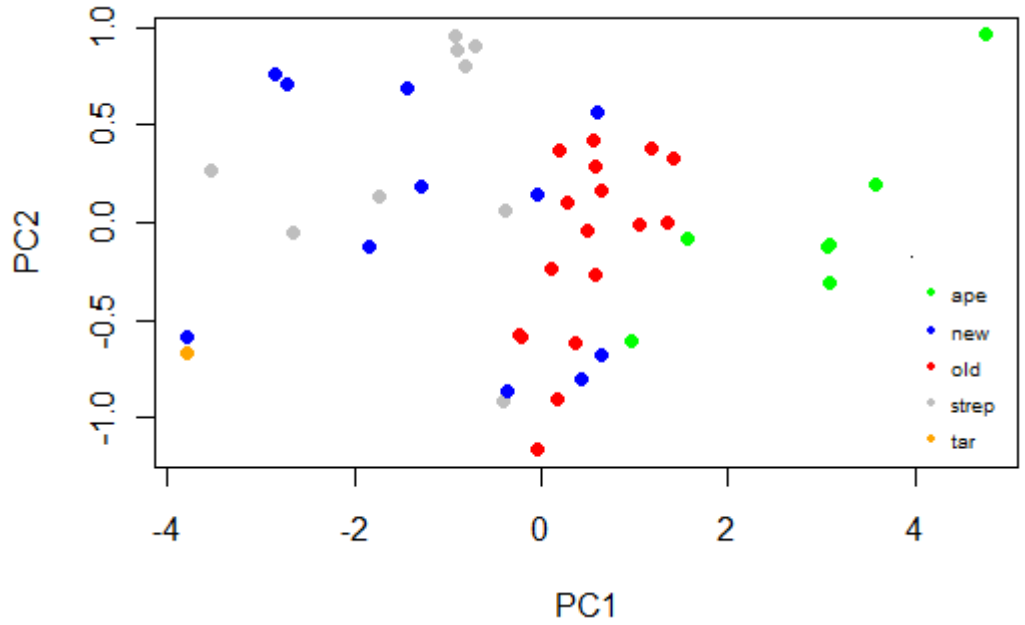
- Now we can plot PCs against each other and label groups

现在可以绘制PC之间的对比图，并标注组别

- PC1 x PC2 by each of 5 primate groups

pch是点的样式

```
> plot(PC2~PC1, pch=16,
col=c("green", "blue", "red",
"grey", "orange")[group],
  data=lifehistorypc)
```



- Remember that *R* reads factor levels *alphabetically*: apes, New World monkeys, Old World monkeys, strepshirrhines, tarsier

R按照字母顺序读取因子级别

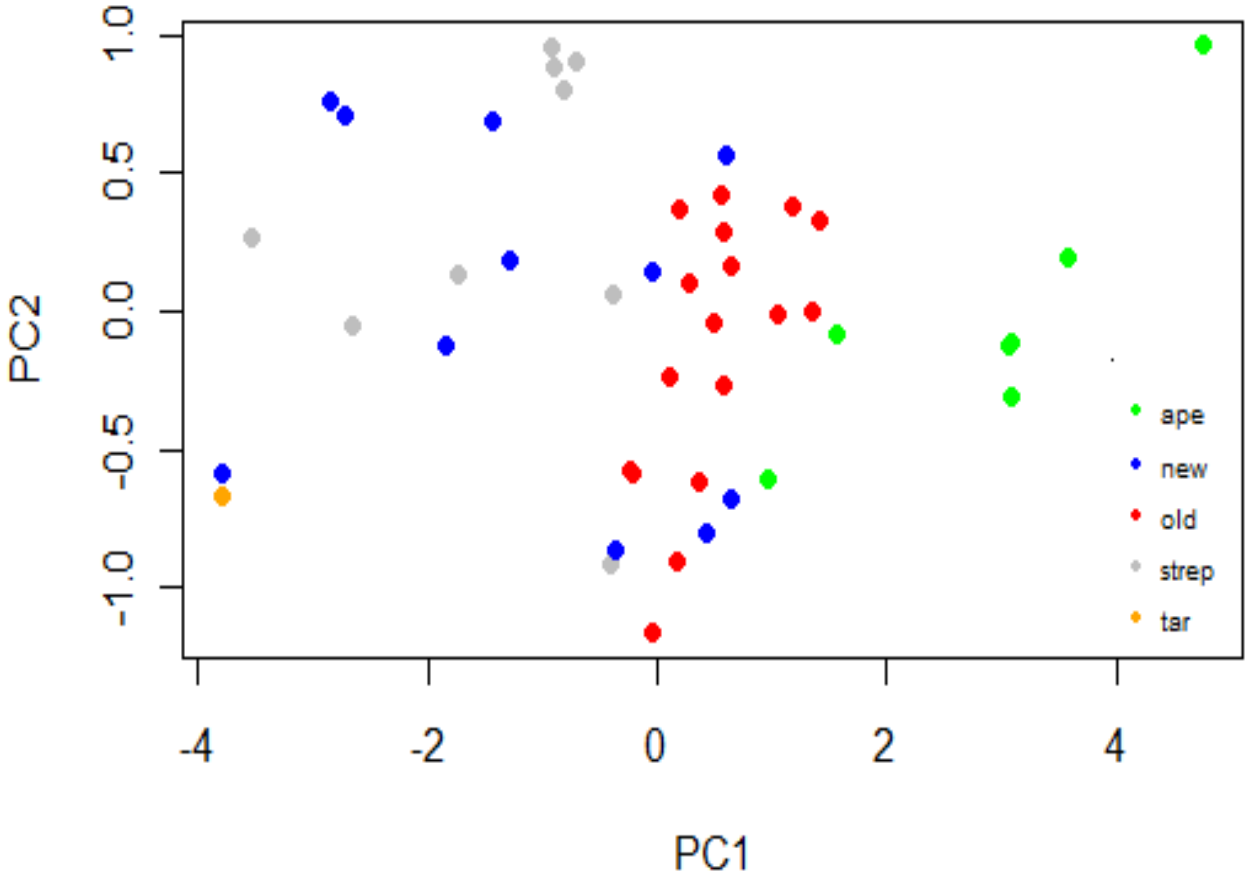
Grouping cases

PC1从狨猴和长尾猴增加到新世界猴和旧世界猴，最后是猿猴

- PC1 increases from tarsier and strepsirrhines to New and Old World monkeys and finally apes
  - it makes sense to interpret it as ‘body size’

可以解释为“体型”

- PC2 does not show primate group patterning
  - this does not necessarily mean that interpreting PC2 as ‘adult lifespan’ or ‘reproductive span’ is wrong



PC2解释不了什么，但不一定意味着将其解释为成人寿命或者生育期是错误的

确定案例

Identifying cases

- We may want to look at specific cases on the plot

- Done with function *identify*
  - and right-clicking point of interest

> identify(PC2~PC1)

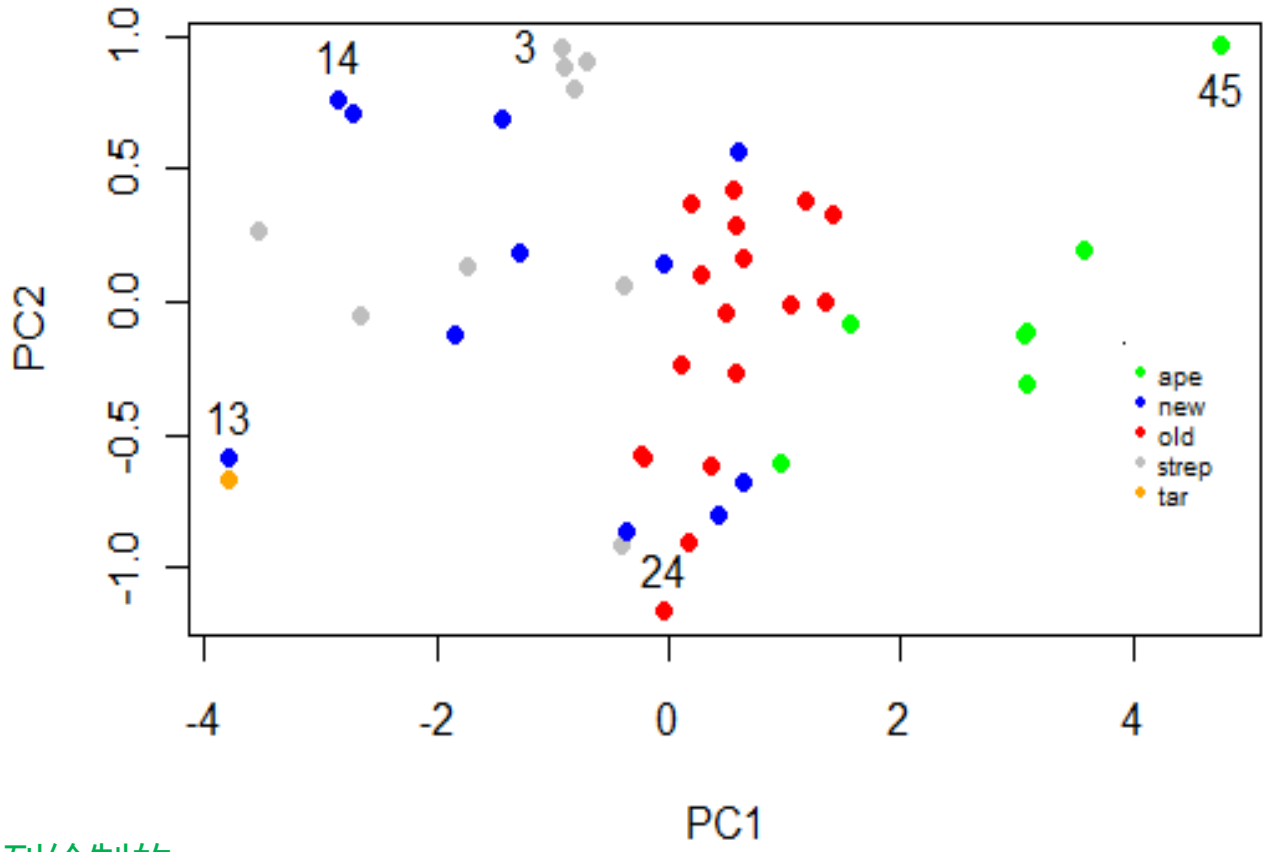
用这个函数之后，可以点击感兴趣的点

- Using *lifehistory* file, we can identify the species:
  - 3 V. variegata
  - 13 C pygmaea (pygmy marmoset)
  - 14 C. jacchus
  - 24 C. aethiops
  - 45 H. sapiens

  - (note: this plot is based on line number, not column *row.names*)

该图是根据行号绘制的，而不是根据行名那一列绘制的

- Try plotting PC3~PC1, PC3~PC2



Notes

估计PC值和各变量之间的相关性可能很有意义，这可以显示出这些变量在PC上有显著的loading

- It may be relevant to estimate correlations between PC scores and each variable; this may show whether variable has a ‘significant’ loading on a PC
  - some authors argue that PCs with less than 3 significant variable loadings should not be kept

有些作者认为不应保留显著变量loading少于三个的PC

- PCA is most useful when dataset includes many variables

当数据集包括很多变量时，PCA最为有效
切记使用相关矩阵而非协方差矩阵，除非所有变量都具有相同的标度和方差(非常罕见)

- Remember to use a correlation matrix rather than a covariance matrix, unless all variables have the same scale and variance (which is very rare!)

Exercise:

- Based on your PCA using the *hdr* dataset, how many PCs should you keep?
- What would you call the kept PC(s)?