Principal Components Analysis

STAT 315 6/08-7/08

Two types of analysis of multivariate data:

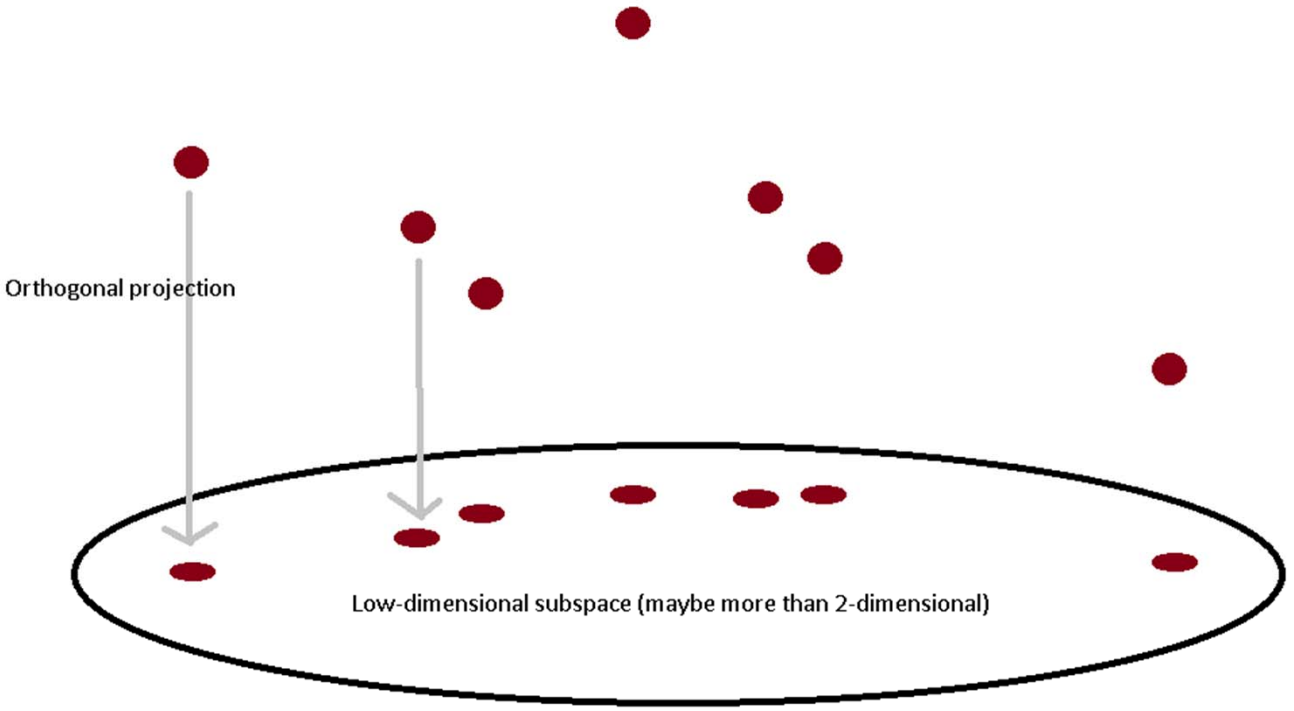
- One variable is “y” and the others are “x” ’s. You try to **predict** the y using the x’s.

  aka **Supervised learning**

- No particular variable is the “y”. You are trying to understand the data. Maybe even just **visualise** it (cf. beginning of the course) maybe find **cluster**s of related observations.

  aka **Unsupervised learning**

Principal components analysis (PCA) is a very useful unsupervised technique for visualising your data. It is a great starting point for almost any statistical work.



Do Xenon demo here with rgl

The most basic form of PCA is just projection into a lower-dimensional subspace. Usually we want to choose the subspace so that we lose as little information as possible. Recall the formula for projection of a vector *v* in the direction of a vector *w*

$$\frac{< v, w >}{< w, w >} w$$

Suppose we want to project all our data onto a single *w.* We impose the condition $\|w\| = 1$. (why?)

UC investigating NZ's gross national happiness

16 April 2013 A University of Canterbury statistics researcher is taking a statistical approach to gauge the level of New Zealand's gross national happiness. (read article)

More News

Assume data are centered (mean of each x_i is zero.) The problem we want to solve is:

Find $w = (w_1, w_2, \ldots, w_d)$ so that $\|w\| = 1$ and

$$var\left(\sum x_{ij} w_j\right)$$

is as large as possible. [Calculate on board here]. We need to maximise

$$w' COV(X) w \qquad\qquad \|w\| = 1$$

If *w* is an **eigenvector** of *COV(X)* with eigenvalue λ then

$$w'COV(X)w = w'\lambda w = \lambda w'w = \lambda$$

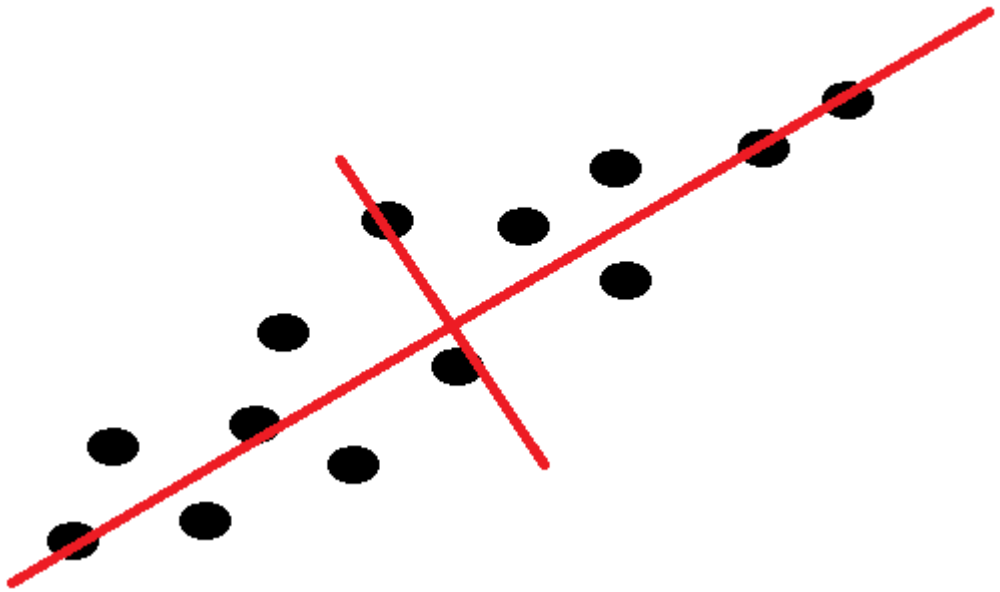
It turns out that the biggest possible value is the same as the biggest eigenvalue of $COV(X)$.

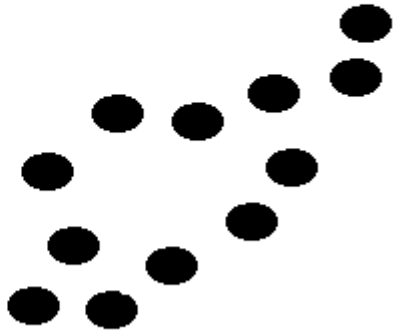
Worked example: data *(x,y) = (1,0), (0,1), (1,1).*

What is the first principal component? What are the *loadings*?

(In practice, we don't need to go through a calculation as the software does it for us.)

Important: usually the variables are scaled to have variance 1 before doing PCA.





Scaling corresponds to finding eigenvalues and eigenvectors of the **correlation** matrix rather than the **covariance** matrix. This works better if your data are on very different scales, but it does mean that the result of PCA is no longer a projection of your original data set; it is a re-scaling followed by a projection.

Since PCA is a purely exploratory technique, it is up to you to decide which version to use.

Another important note:

The principal components are only defined up to +/-1

Higher Principal Components

Usually want to project our data onto a space of more than one dimension (can plot up to 6 dimensions using colour, but only 2 if we want to be “unbiased”) The higher principal components are just the other eigenvectors of the correlation (or covariance) matrix. These correspond to the directions in which variance is maximised conditional on being orthogonal to the previous ones.

Example

crabs data in R. A simple example of a **morphological** data set.

```
> head(crabs)
  sp sex index   FL  RW   CL   CW  BD
1  B   M     1  8.1 6.7 16.1 19.0 7.0
2  B   M     2  8.8 7.7 18.1 20.8 7.4
3  B   M     3  9.2 7.8 19.0 22.4 7.7
4  B   M     4  9.6 7.9 20.1 23.1 8.2
5  B   M     5  9.8 8.0 20.3 23.0 8.2
6  B   M     6 10.8 9.0 23.0 26.5 9.8
```

Body depth

Species (B or O)

Frontal lobe size

Rear Width

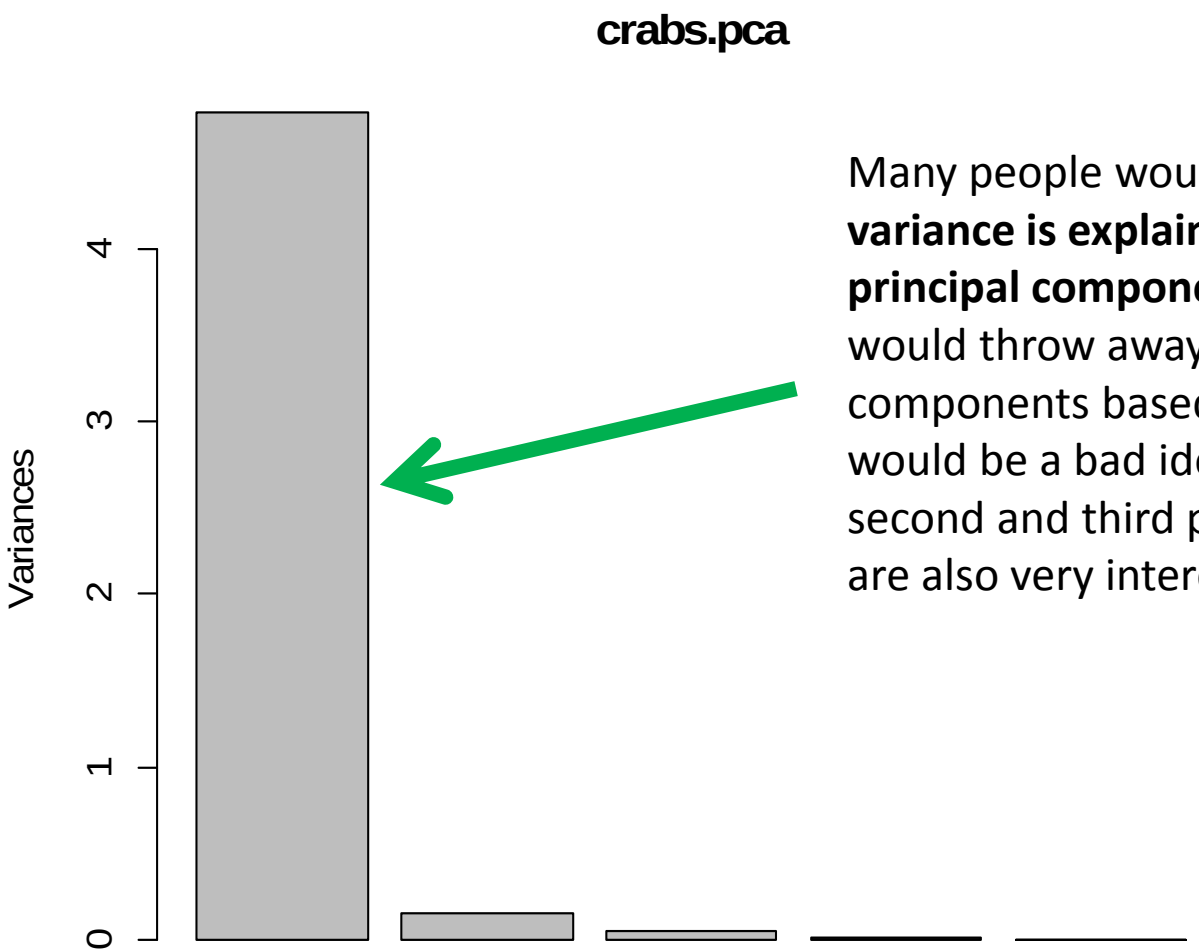
Carapace length

Carapace width

```
> crabs.mm <- crabs[,4:8]
> crabs.pca <- prcomp(crabs.mm, center=T, scale=T)

> str(crabs.pca)
List of 5
 $ sdev    : num [1:5] 2.1883 0.3895 0.2159 0.1055 0.0414
 $ rotation: num [1:5, 1:5] 0.452 0.428 0.453 0.451 0.451 ...
 $ center  : Named num [1:5] 15.6 12.7 32.1 36.4 14
 $ scale   : Named num [1:5] 3.5 2.57 7.12 7.87 3.42
 $ x       : num [1:200, 1:5] -4.92 -4.38 -4.12 -3.87 -3.82 ...
```

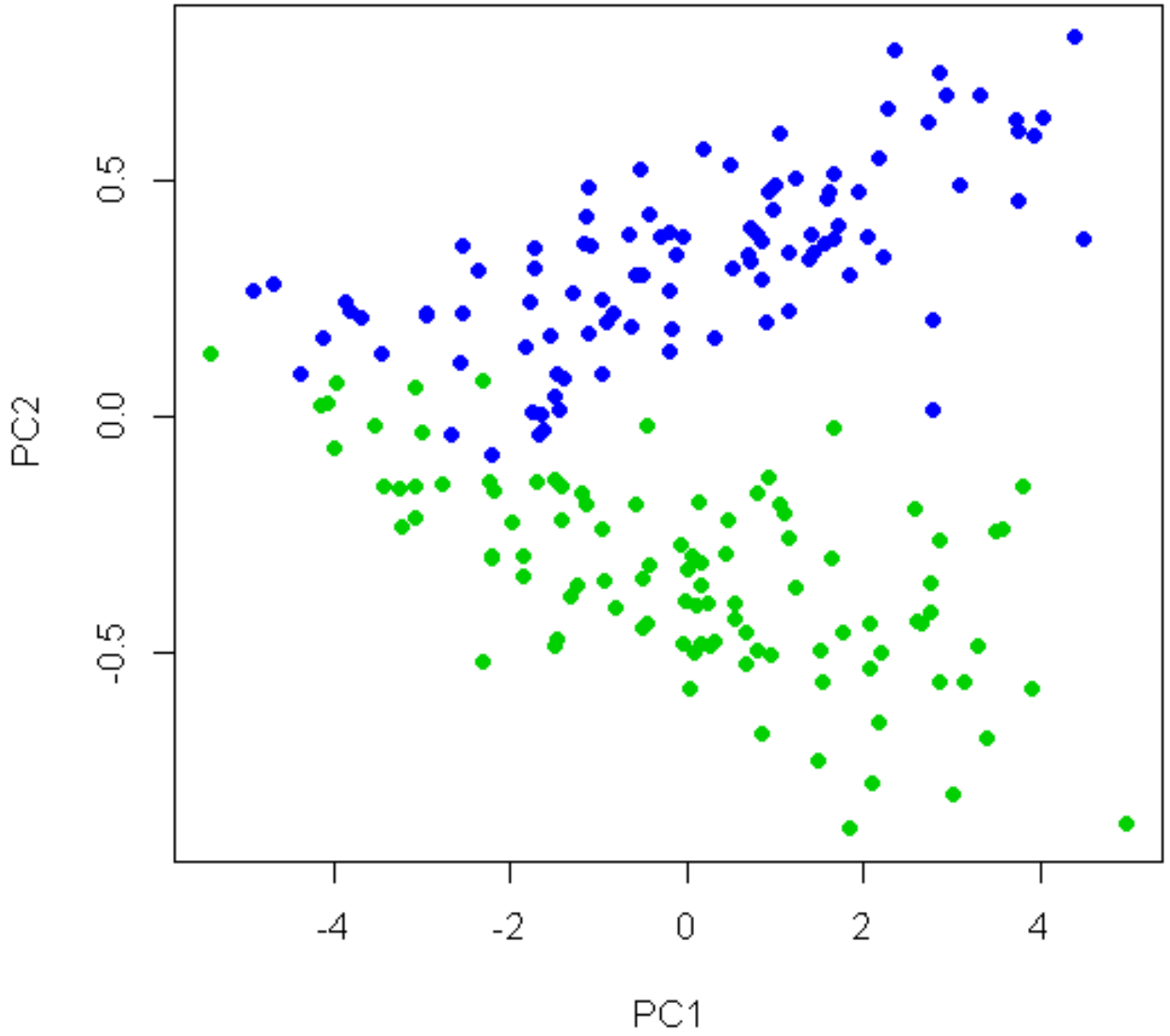
Here, sdev is the square roots of the eigenvalues. The eigenvalues are often called the "variance explained". You can get a plot of them with `plot(crabs.pca)`. This is called a **scree plot.** It is often used to choose how many of the principal components are "important".

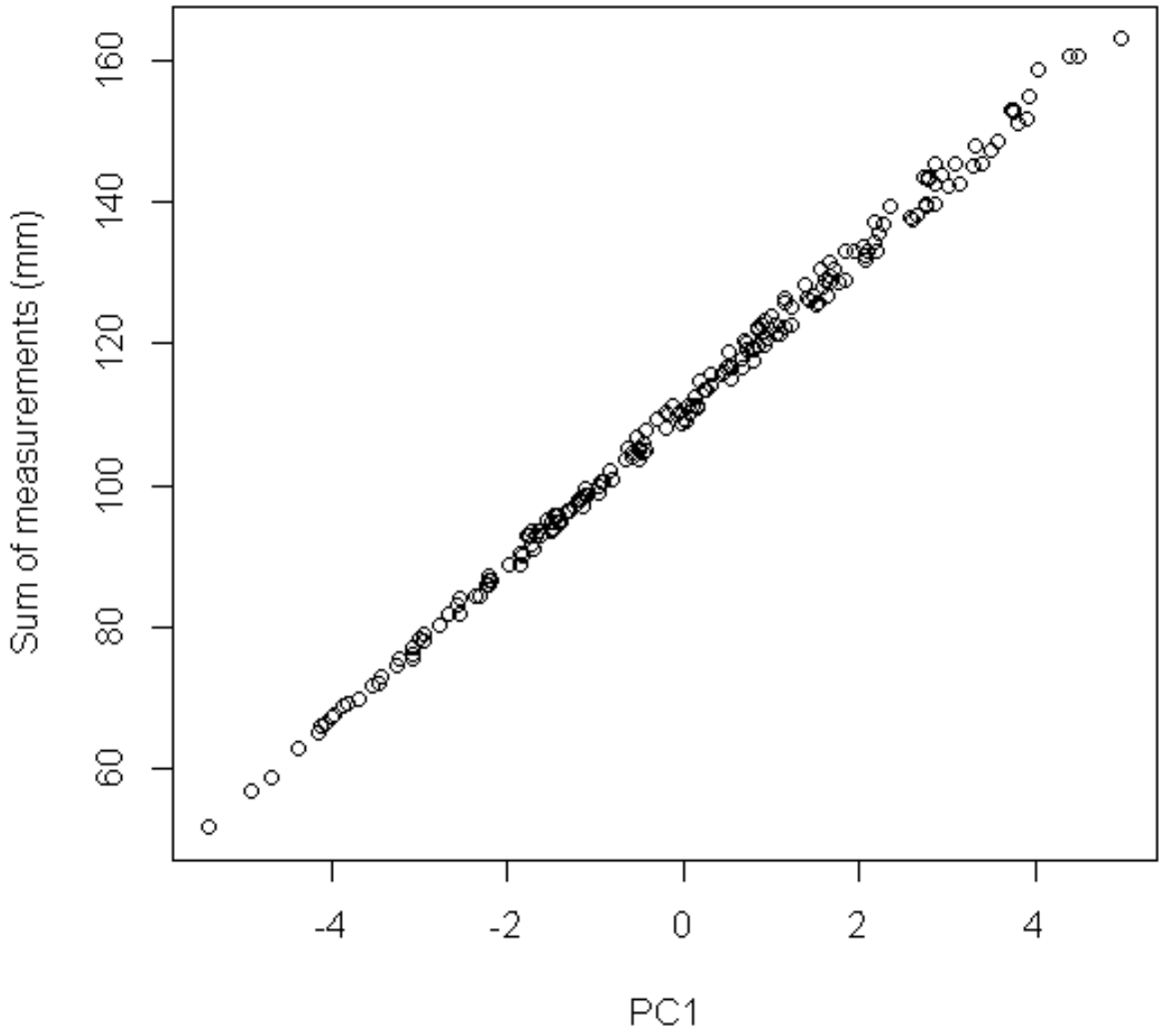


Many people would say **"96% of the variance is explained by the first principal component."** Some people would throw away the other components based on this, which would be a bad idea in this case, as the second and third principal components are also very interesting.

```
> cumsum(crabs.pca$sdev^2)/sum(crabs.pca$sdev^2)
[1] 0.9577670 0.9881040 0.9974306 0.9996577 1.0000000
```

```
> plot(crabs.pca$x[,1], crabs.pca$x[,2], col=as.numeric(crabs$sex)+2, pch=19,
xlab="PC1", ylab="PC2")
```





The matrix `crabs.pca$x` contains the **loadings** for the principal components. It is a 200 x 5 matrix.

PC1 is `crabs.pca$x[ ,1]` etc.

The first principal component, PC1, corresponds to the overall size of the crab. This is very common in morphological data.

PC2 seems to correspond to the sex of the crab.

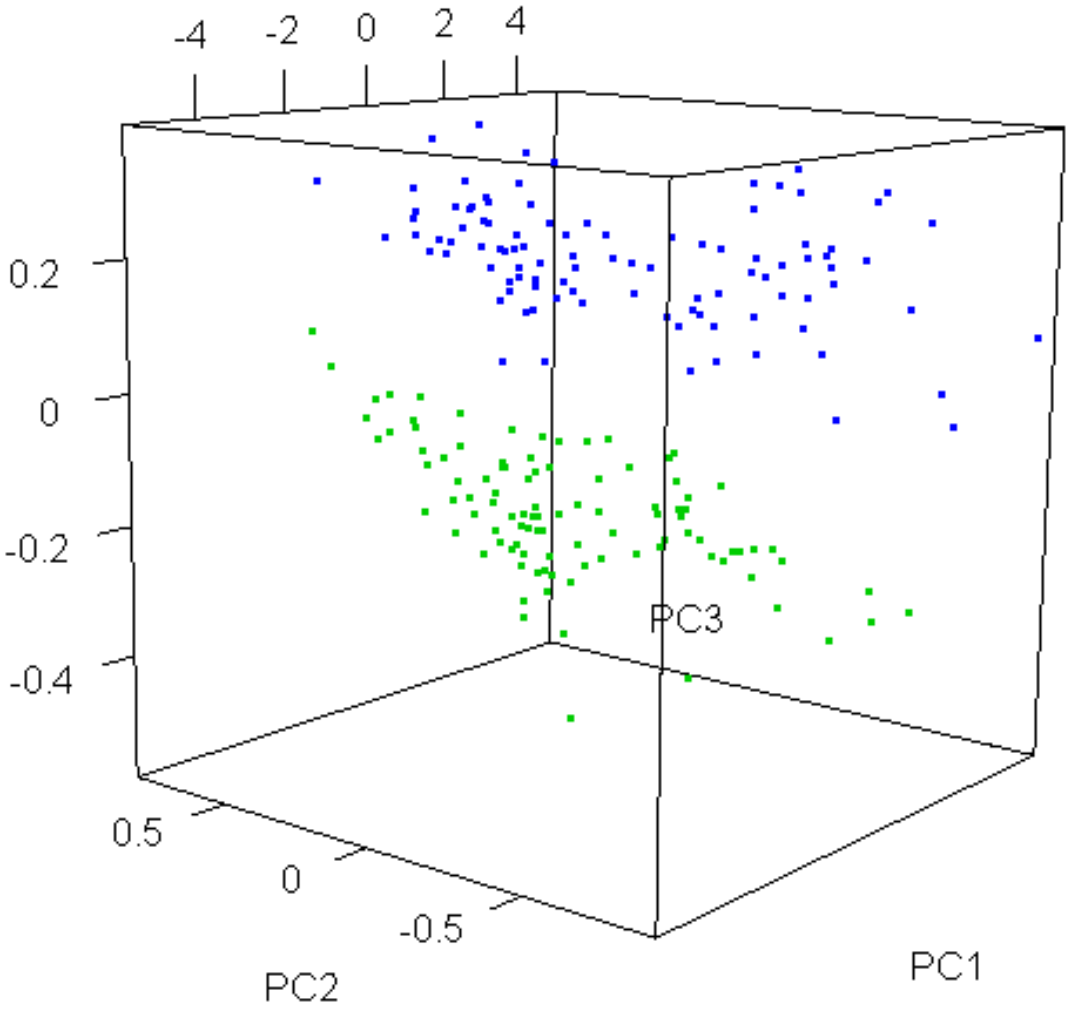
The matrix `crabs.pca$rotation` shows how the principal components are made up from the x's.

```
> crabs.pca$rotation
          PC1        PC2         PC3          PC4         PC5
FL 0.4520437  0.1375813  0.53076841  0.696923372  0.09649156
RW 0.4280774 -0.8981307 -0.01197915 -0.083703203 -0.05441759
CL 0.4531910  0.2682381 -0.30968155 -0.001444633 -0.79168267
CW 0.4511127  0.1805959 -0.65256956  0.089187816  0.57452672
BD 0.4511336  0.2643219  0.44316103 -0.706636423  0.17574331
```

e. g. PC1 = 0.45*FL + 0.43*RW + 0.45*CL + 0.45*CW + 0.45*BD

Rear width (RW) is a strong factor in determining whether a crab is male or female (useful observation for a scientist!)

```
> library(rgl)
> plot3d(crabs.pca$x[,1:3], col=as.numeric(crabs$sp) + 2)
```



PC3 corresponds to the crab's species!

Potentially very useful (imagine that we didn't know that there were two species of crab here. PCA might have helped us to discover this!)

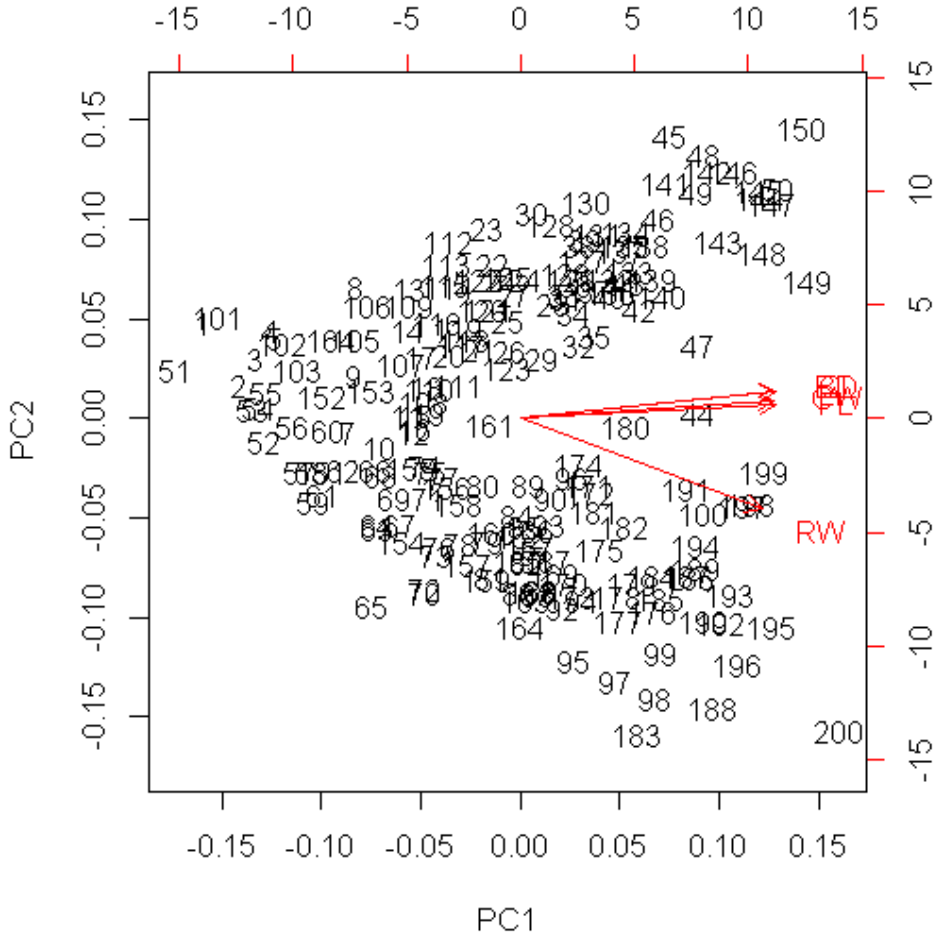
Caution!

It is important not to get carried away when interpreting the principal components, especially those which carry a low proportion of the variance.

Often, these can just be noise.

```
> biplot(crabs.pca)
```

Biplot: a plot of the first two principal components with arrows showing how the variables are made up from the principal components.



For example, the RW arrow points in the direction

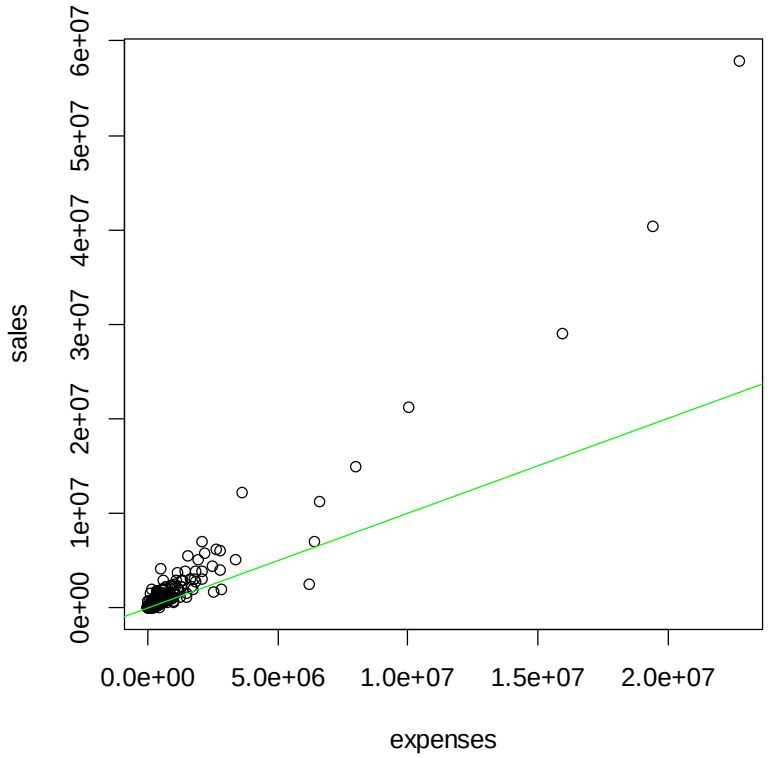
(0.43, -0.90)

Example:

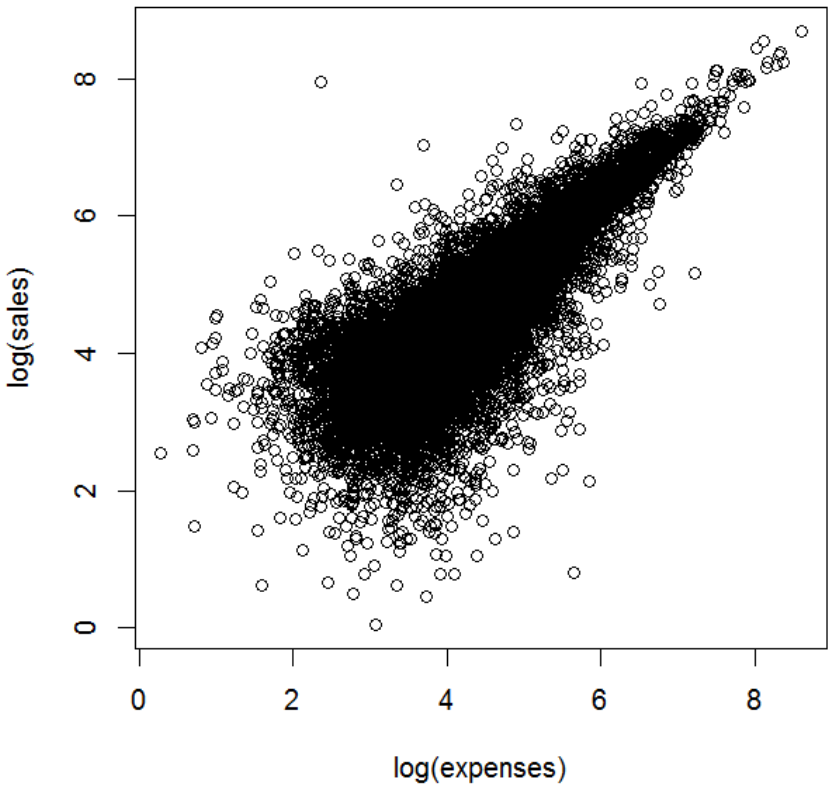
Using PCA to visualise a six-dimensional time series.

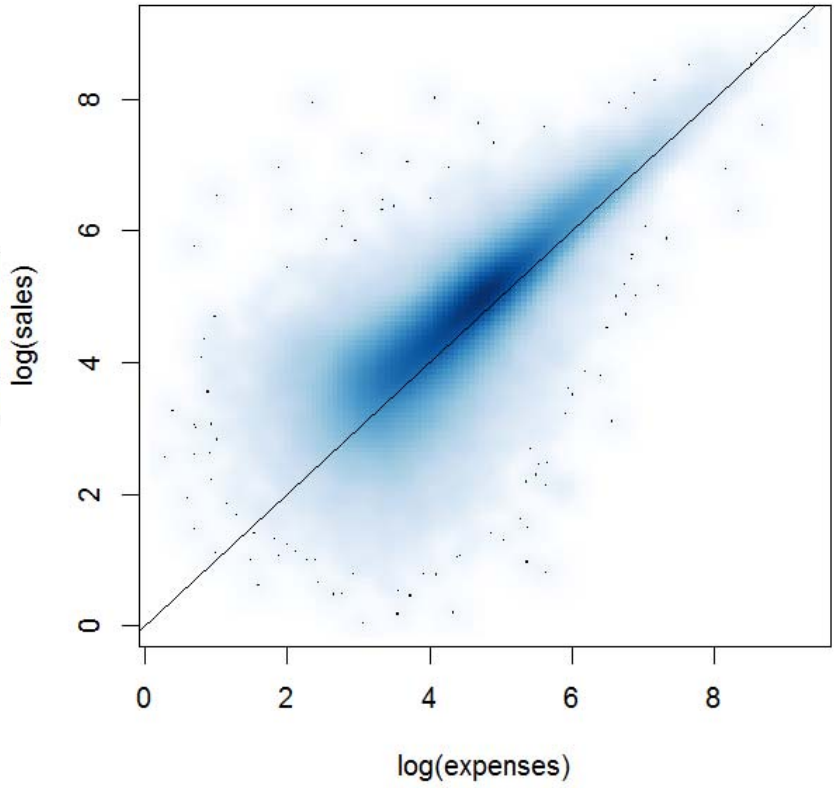
Data:

Name	expenses	sales
Small Coffee Co	4323	15357
United Poultry	23388	41494
Xmas Holdings	30450	0
OzCorp	5426038	1782330
…		



`smoothScatter()`

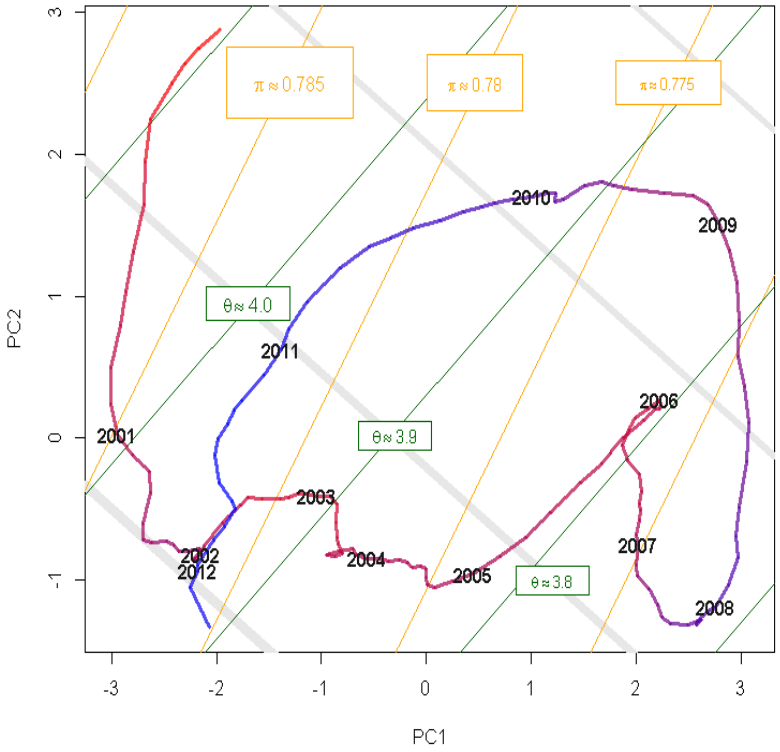


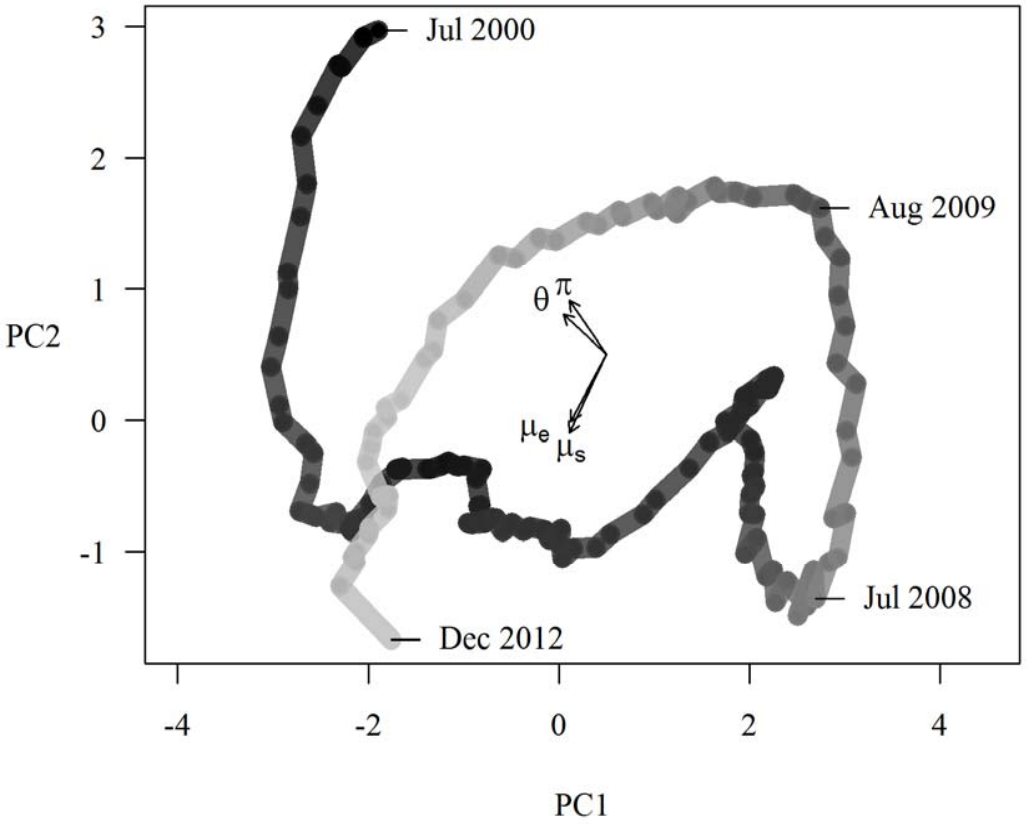


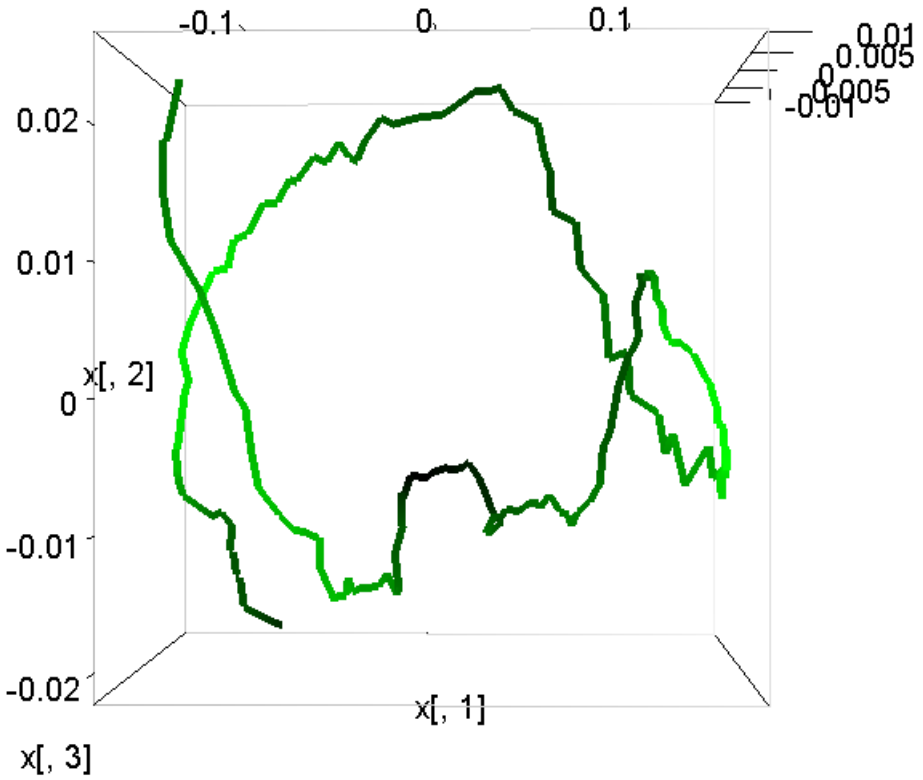
Aim:

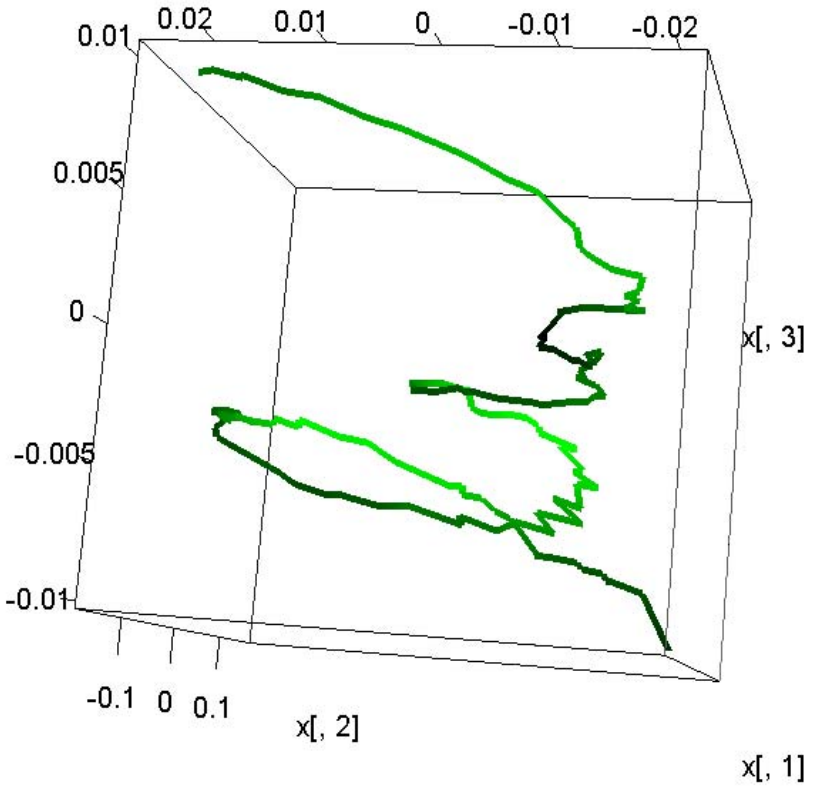
- Reduce the dimension by fitting a model with a few parameters.
- Use the time series of the model parameters to visualise how the joint distribution changes from month to month

- Fit to each month from Jan 2000 to Jul 2013.
- Adjust for inflation.
- Take moving average (because of seasonality).
- Use PCA to reduce from six to three dimensions.
- Plot the first 2 principal components (using colour for the third dimension).





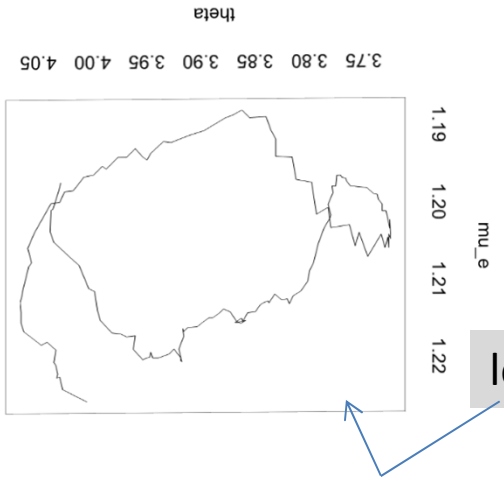




PCA on covariance matrix visualised using `rgl`

$$PC1 \approx -\theta$$
$$PC2 \approx -\mu_e - \mu_s$$
$$PC3 \approx \pi$$
$$PC4 \approx \mu_s - \mu_e$$

} Much the same as the correlation matrix version

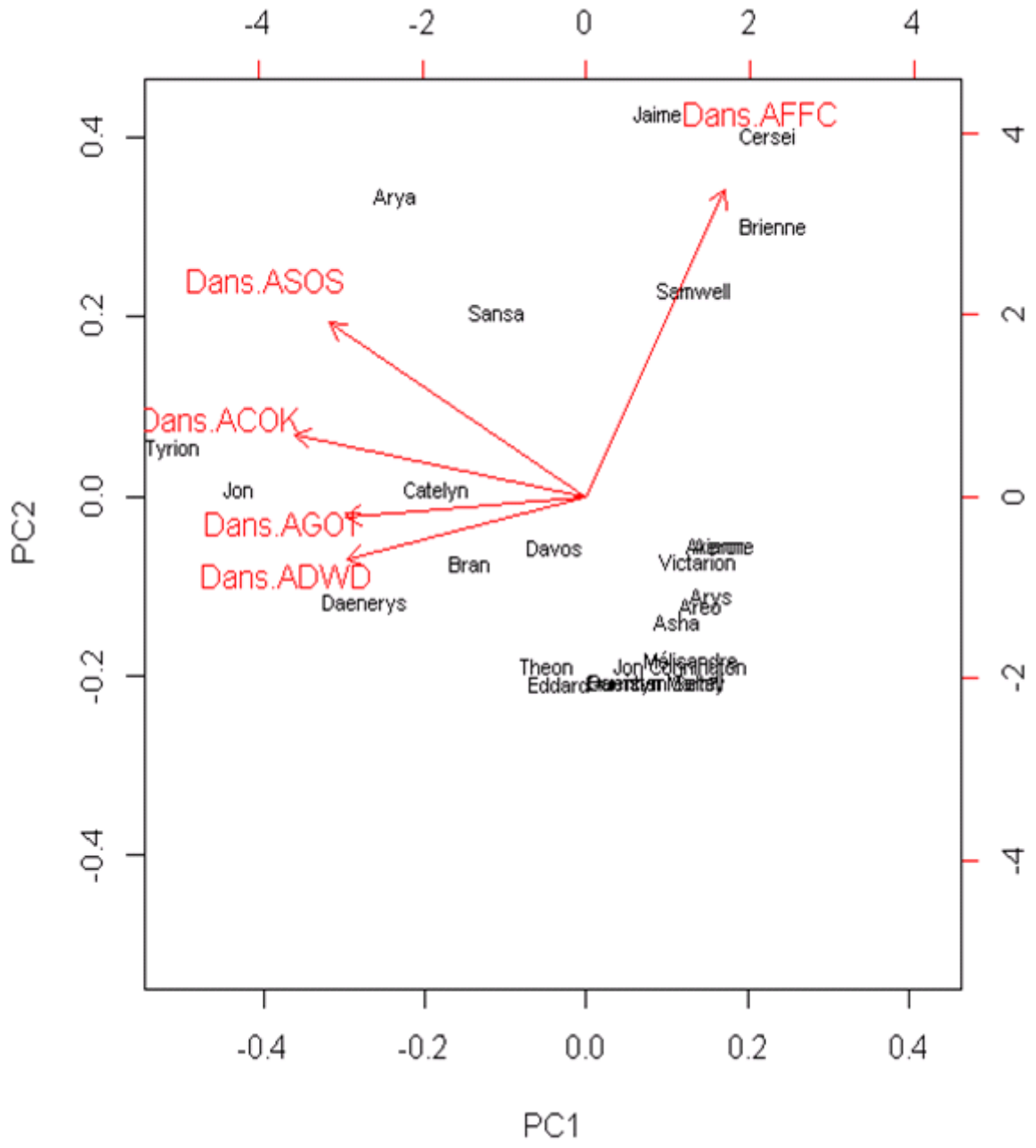


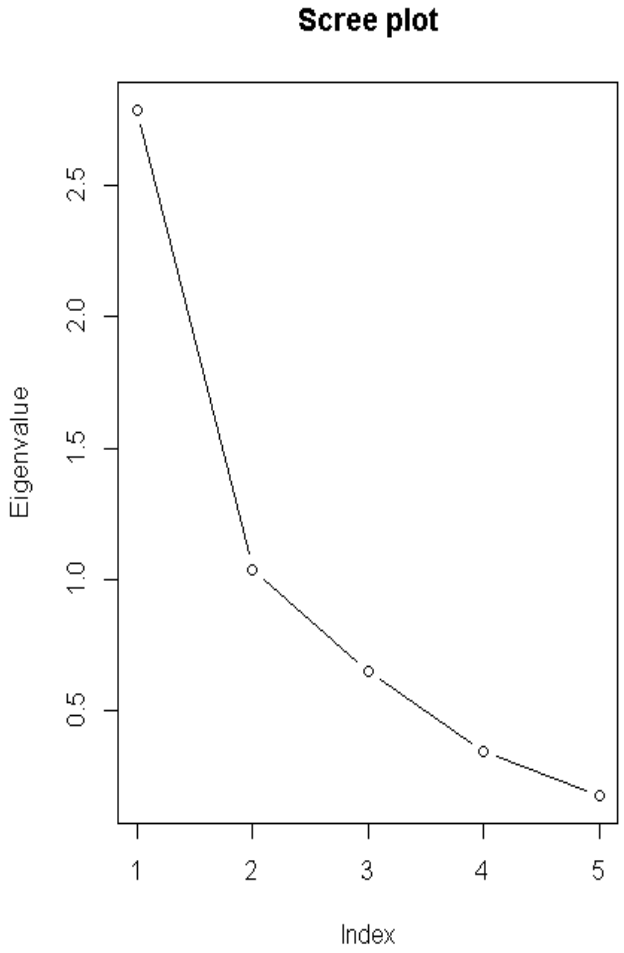
looks much the same shape as this plot of theta vs mu_e

Example: using PCA to visualise Game of Thrones.

http://www.lagardedenuit.com/wiki/index.php?title=Personnages_PoV

Personnage	Dans AGOT	Dans ACOK	Dans ASOS	Dans AFFC	Dans ADWD	Chapitres au total
Eddard	15	0	0	0	0	15
Catelyn	11	7	7	0	0	25
Sansa	6	8	7	3[4]	0	24
Arya	5	10	13	3[5]	2[6]	33
Bran	7	7	4	0	3	21
Jon	9	8	12	0	13	42
Daenerys	10	5	6	0	10	31
Tyrion	9	15	11	0	12	47
Theon	0	6	0	0	7[7]	13
Davos	0	3	6	0	4	13
Samwell	0	0	5	5	0	10
Jaime	0	0	9	7	1	17
Cersei	0	0	0	10	2	12
Brienne	0	0	0	8	0	8
Areo	0	0	0	1[8]	1[9]	2
Arys	0	0	0	1[10]	0	1
Arianne	0	0	0	2[11]	0	2
Asha	0	0	0	1[12]	3[13]	4
Aeron	0	0	0	2[14]	0	2
Victarion	0	0	0	2[15]	2[16]	4
Quentyn Martell	0	0	0	0	4[17]	4
Jon Connington	0	0	0	0	2[18]	2
Mélisandre	0	0	0	0	1	1
Barristan Selmy	0	0	0	0	4[19]	4


-4
-2
0
2
4
0.4
0.2
0.0
-0.2
-0.4
PC2
PC1
Jaime
Dans.AFFC
Cersei
Arya
Brienne
Dans.ASOS
Sansa
Samwell
Dans.ACOK
Tyrion
Jon
Catelyn
Dans.AGOT
Dans.ADWD
Bran
Davos
Victarion
Daenerys
Asha
Theon


Scree plot
Eigenvalue
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
Index

	PC1	PC2	PC3	PC4	PC5
Dans.AGOT	-0.4508665	-0.0566973	-0.71665896	-0.5258787	-0.05793421
Dans.ACOK	-0.5464952	0.1710553	0.03468803	0.3197532	0.75408071
Dans.ASOS	-0.4797854	0.4762526	0.02677051	0.3921822	-0.62327028
Dans.AFFC	0.2598717	0.8431337	0.07148840	-0.4305571	0.17635831
Dans.ADWD	-0.4475881	-0.1727268	0.69236509	-0.5310585	-0.09185754

	PC1	PC2	PC3	PC4	PC5
Eddard	-0.26	-1.04	-2.47	-1.16	-0.23
Catelyn	-1.52	0.04	-1.75	0.42	0.07
Sansa	-0.89	1.03	-0.90	0.61	0.49
Arya	-1.92	1.67	-0.34	1.12	-0.03
Bran	-1.16	-0.37	-0.62	0.19	0.47
Jon	-3.51	0.05	0.94	-0.64	-0.74
Daenerys	-2.23	-0.59	0.19	-1.09	-0.37
Tyrion	-4.18	0.27	0.81	-0.07	0.66
Theon	-0.39	-0.93	1.15	0.00	0.84
Davos	-0.31	-0.28	0.62	0.72	-0.45
Samwell	1.11	1.16	-0.01	0.21	-0.44
Jaime	0.74	2.13	0.25	0.12	-0.89
Cersei	1.87	2.01	0.45	-1.26	0.52
Brienne	1.92	1.51	0.04	-0.68	0.45
Areo	1.16	-0.60	0.04	0.23	-0.01
Arys	1.28	-0.56	-0.14	0.37	0.01
Arianne	1.37	-0.26	-0.12	0.22	0.08
Asha	0.93	-0.69	0.40	-0.04	-0.06
Aeron	1.37	-0.26	-0.12	0.22	0.08
Victarion	1.14	-0.35	0.25	-0.06	0.03
Quentyn Martell	0.72	-1.04	0.56	-0.03	-0.14
Jon Connington	0.95	-0.95	0.20	0.25	-0.10
Mélisandre	1.07	-0.90	0.02	0.39	-0.07
Barristan Selmy	0.72	-1.04	0.56	-0.03	-0.14

Many interesting features in these data. Notice how *A Feast for Crows* (the least popular book) is quite different from the others (see biplot.)

How would you interpret the different Principal Components?

When PCA goes bad?

NZDep2006 combines the following census data (calculated as proportions for each small area):

Variable description (in order of decreasing weight)

- People aged 18-64 receiving a means tested benefit
- People living in equivalised* households with income below a threshold
- People not living in own home
- People aged <65 living in a single parent family
- People aged 18-64 unemployed
- People aged 18-64 without any qualifications
- People living in equivalised* households below a bedroom occupancy threshold
- People with no access to a telephone
- People with no access to a car