# Data Reduction I : PCA and Factor Analysis

Carlos Pinto[1]

[1]Neuropsychiatric Genetics Group
University of Dublin

Data Analysis Seminars 11 November 2009

# Overview

1. Distortions in Data
2. Variance and Covariance
3. PCA: A Toy Example
4. Running PCA in R
5. Factor Analysis
6. Running Factor Analysis in R
7. Conclusions

# Sources of Confusion in Data

1. Noise
2. Rotation
3. Redundancy

in *linear* systems . . .

# Sources of Confusion in Data

1. Noise
2. Rotation
3. Redundancy

in *linear* systems . . .

# Sources of Confusion in Data

1. Noise
2. Rotation
3. Redundancy

in *linear* systems . . .

# Sources of Confusion in Data

1. Noise
2. Rotation
3. Redundancy

in *linear* systems . . .

# Noise

1. Noise arises from inaccuracies in our measurements
2. Noise is measured relative to the signal

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

The $SNR$ must be large if we are to extract useful information from our experiment.

# Noise

1. Noise arises from inaccuracies in our measurements
2. Noise is measured relative to the signal

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

The $SNR$ must be large if we are to extract useful information from our experiment.

# Noise

1. Noise arises from inaccuracies in our measurements
2. Noise is measured relative to the signal

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

The $SNR$ must be large if we are to extract useful information from our experiment.

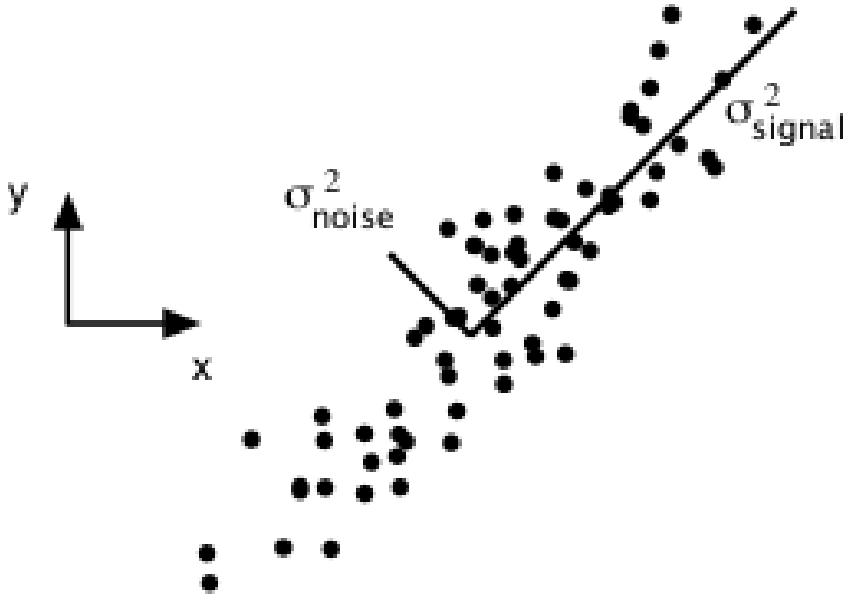
# Noise

1. Noise arises from inaccuracies in our measurements
2. Noise is measured relative to the signal

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

The $SNR$ must be large if we are to extract useful information from our experiment.

# Variance in Signal and Noise



# Rotation

1. We don't always know the true underlying variables in the system we are studying
2. We may actually be measuring some (hopefully linear!) combination of the true variables.

# Rotation

1. We don't always know the true underlying variables in the system we are studying
2. We may actually be measuring some (hopefully linear!) combination of the true variables.

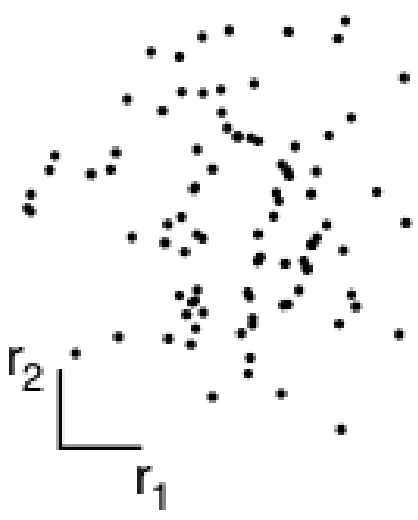
# Redundancy

1. We don't always know whether the variables we are measuring are independent.
2. We may actually be measuring one or more (hopefully linear!) combinations of the *same* set of variables.

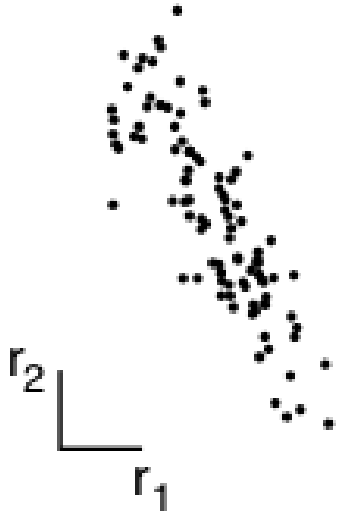
# Redundancy

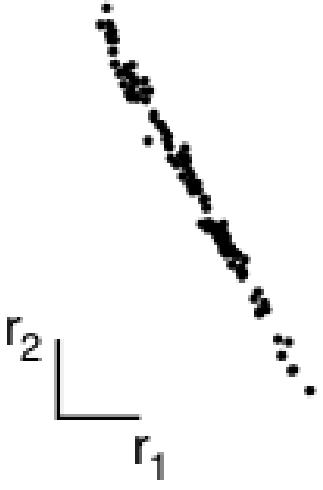
1. We don't always know whether the variables we are measuring are independent.
2. We may actually be measuring one or more (hopefully linear!) combinations of the *same* set of variables.

# Degrees of Redundancy



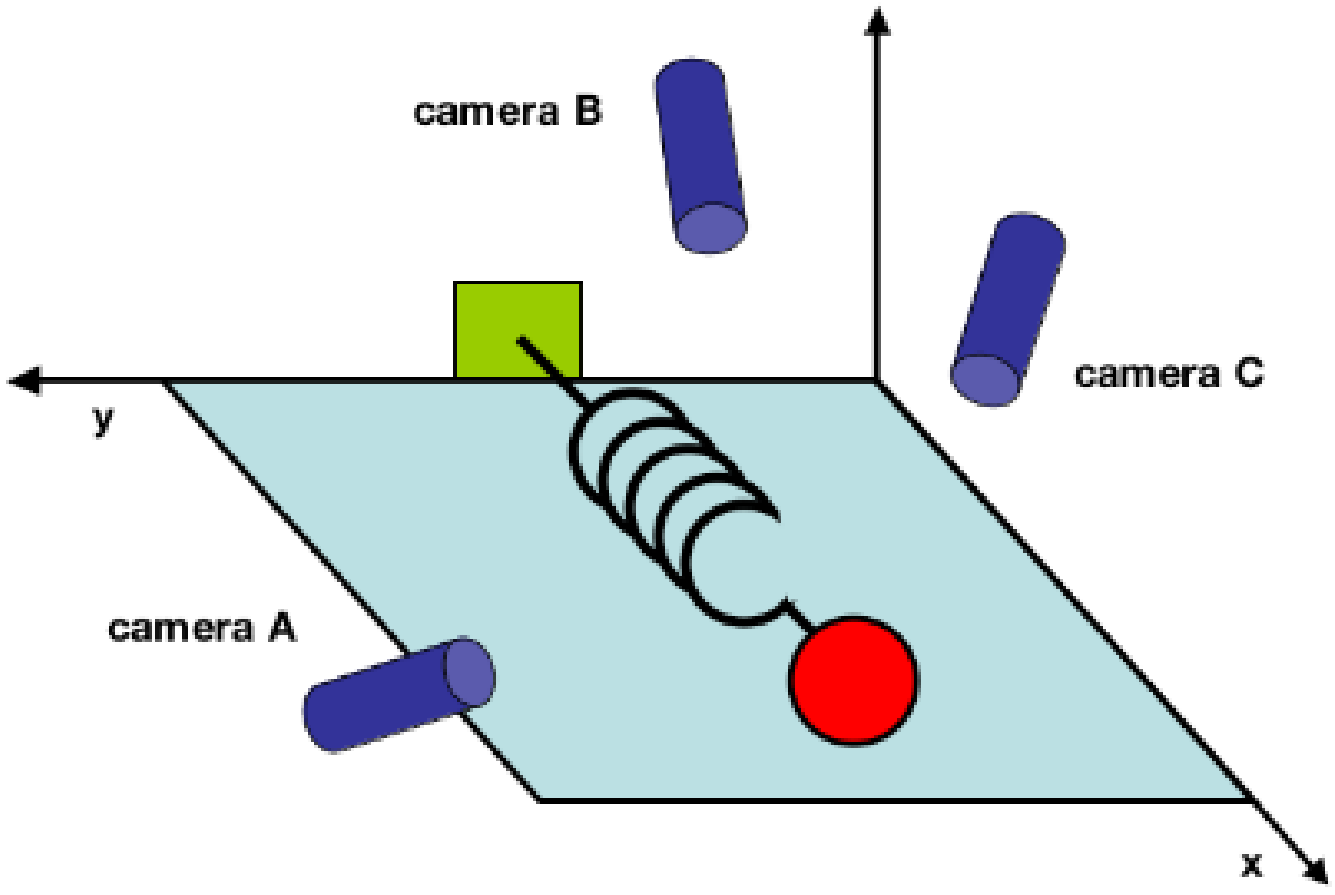

low redundancy




high redundancy

# An Illustrative Experiment



# Goals of PCA

1. To isolate noise
2. To separate out the redundant degrees of freedom
3. To eliminate the effects of rotation

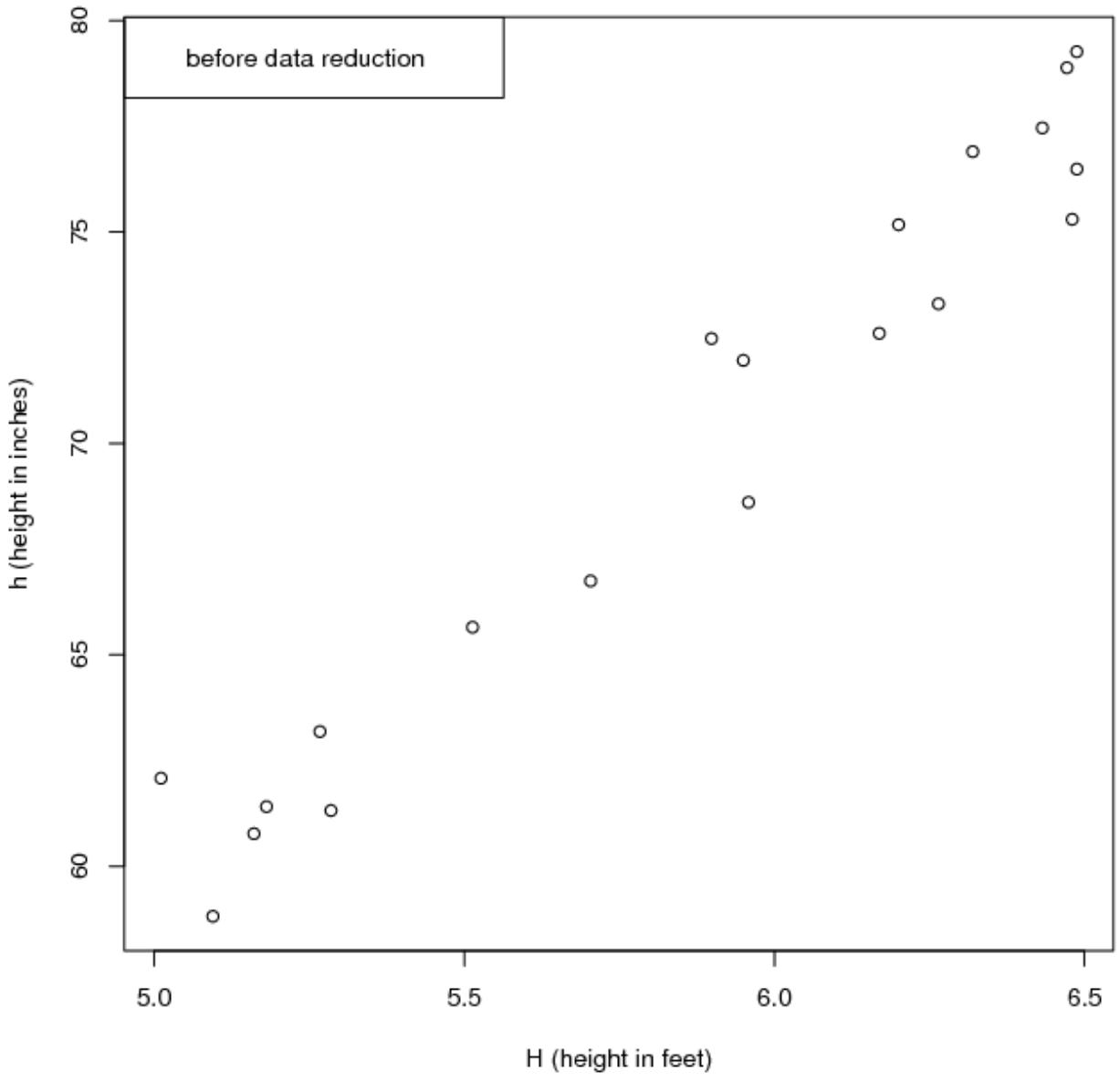
# Goals of PCA

1. To isolate noise
2. To separate out the redundant degrees of freedom
3. To eliminate the effects of rotation

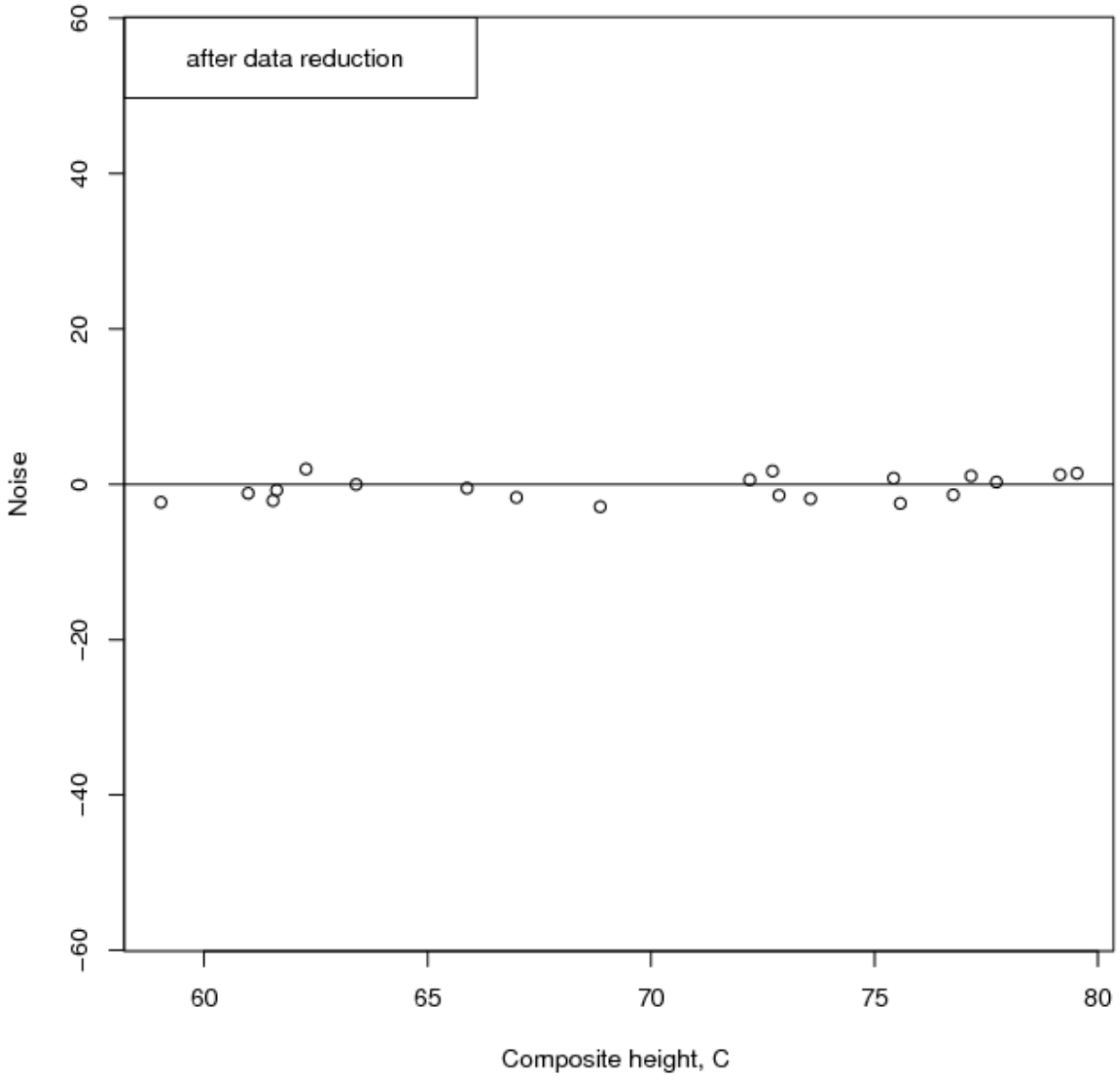
# Goals of PCA

1. To isolate noise
2. To separate out the redundant degrees of freedom
3. To eliminate the effects of rotation

# Example: Measuring Two Correlated Variables



# Reduced Data



# Variance

Variable : $X$ Measurements: $\{x_1 \ldots x_n\}$

Variable : $Y$ Measurements: $\{y_1 \ldots y_n\}$

Variance of $X$: $\sigma^2_{XX} = \frac{\sum_i (x_i - \bar{x})(x_i - \bar{x})}{n-1}$

Variance of $Y$: $\sigma^2_{YY} = \frac{\sum_i (y_i - \bar{y})(y_i - \bar{y})}{n-1}$

# Covariance

Covariance of $X$ and $Y$:

$$\sigma_{XY}^2 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \sigma_{YX}^2$$

The covariance measures the degree of the (linear!) relationship between $X$ and $Y$

Small values of the covariance indicate that the variables are independent (uncorrelated)

# Covariance Matrix

We can write the four variances associated with our two variables in the form of a matrix:

$$\begin{pmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{pmatrix}$$

The diagonal elements are the variances and the off–diagonal elements are the covariances

The covariance matrix is *symmetric*, since $C_{XY} = C_{YX}$

# Three Variables

For $3$ variables, we get a $3 \times 3$ matrix:

$$\begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix}$$

As before, the diagonal elements are the variances and the off–diagonal elements are the covariances, and the matrix is symmetric

# Interpretation of the Covariance Matrix

1. The diagonal terms are the variances of our different variables.
2. Large diagonal values correspond to strong signals
3. The off-diagonal terms are the covariances *between* different variables. They reflect distortions in the data (noise, rotation, redundancy).
4. Large off-diagonal values correspond to distortions in our data

# Interpretation of the Covariance Matrix

1. The diagonal terms are the variances of our different variables.
2. Large diagonal values correspond to strong signals
3. The off-diagonal terms are the covariances *between* different variables. They reflect distortions in the data (noise, rotation, redundancy).
4. Large off-diagonal values correspond to distortions in our data

# Minimizing Distortion

1. If we redefine our variables (as linear combinations of each other) the covariance matrix will change.
2. We want to change the covariance matrix so that the off–diagonal elements are close to zero.
3. That is, we want to *diagonalize* the covariance matrix.

# Minimizing Distortion

1. If we redefine our variables (as linear combinations of each other) the covariance matrix will change.
2. We want to change the covariance matrix so that the off–diagonal elements are close to zero.
3. That is, we want to *diagonalize* the covariance matrix.

# Minimizing Distortion

1. If we redefine our variables (as linear combinations of each other) the covariance matrix will change.
2. We want to change the covariance matrix so that the off–diagonal elements are close to zero.
3. That is, we want to *diagonalize* the covariance matrix.

# Example 1: Toy Example

$$
\begin{aligned}
X &= (1,2) & \overline{X} &= 3/2 \\
Y &= (3,6) & \overline{Y} &= 9/2
\end{aligned}
$$

Subtract the means, since we are only interested in the variances

$$
\begin{aligned}
X' &= \left(-\frac{1}{2}, \frac{1}{2}\right) \\
Y' &= \left(-\frac{3}{2}, \frac{3}{2}\right)
\end{aligned}
$$

# Example 1: Toy Example

$$X = (1,2) \qquad \overline{X} = 3/2$$
$$Y = (3,6) \qquad \overline{Y} = 9/2$$

Subtract the means, since we are only interested in the variances

$$X' = \left(-\frac{1}{2}, \frac{1}{2}\right)$$
$$Y' = \left(-\frac{3}{2}, \frac{3}{2}\right)$$

# Variances

$$
\begin{aligned}
C'_{11} &= \left(-\tfrac{1}{2} \times -\tfrac{1}{2}\right) + \left(\tfrac{1}{2} \times \tfrac{1}{2}\right) &= \frac{1}{2} \\
C'_{12} &= \left(-\tfrac{1}{2} \times -\tfrac{3}{2}\right) + \left(\tfrac{1}{2} \times \tfrac{3}{2}\right) &= \frac{3}{2} \\
C'_{21} &= \left(-\tfrac{3}{2} \times -\tfrac{1}{2}\right) + \left(\tfrac{3}{2} \times \tfrac{1}{2}\right) &= \frac{3}{2} \\
C'_{22} &= \left(-\tfrac{3}{2} \times -\tfrac{3}{2}\right) + \left(\tfrac{3}{2} \times \tfrac{3}{2}\right) &= \frac{9}{2}
\end{aligned}
$$

# Covariance Matrix

The covariance matrix is:

$$C' = \begin{pmatrix} \frac{1}{2} & \frac{3}{2} \\ \\ \frac{3}{2} & \frac{9}{2} \end{pmatrix}$$

This is not diagonal – the non–zero off-diagonal terms indicate the presence of distortion, in this case because the two variables are correlated

# Rotation of Axes I

We need to redefine our variables to diagonalize the covariance matrix.

Choose new variables which are a linear combination of the old ones:

$$X'' = a_1 X' + a_2 Y'$$

$$Y'' = b_1 X' + b_2 Y'$$

We have to determine the four constants $a_1$, $a_2$, $b_1$, $b_2$ such that the covariance matrix is diagonal.

# Rotation of Axes I

We need to redefine our variables to diagonalize the covariance matrix.

Choose new variables which are a linear combination of the old ones:

$$X'' = a_1 X' + a_2 Y'$$

$$Y'' = b_1 X' + b_2 Y'$$

We have to determine the four constants $a_1$, $a_2$, $b_1$, $b_2$ such that the covariance matrix is diagonal.

# Rotation of Axes I

We need to redefine our variables to diagonalize the covariance matrix.

Choose new variables which are a linear combination of the old ones:

$$X'' = a_1 X' + a_2 Y'$$

$$Y'' = b_1 X' + b_2 Y'$$

We have to determine the four constants $a_1$, $a_2$, $b_1$, $b_2$ such that the covariance matrix is diagonal.

# Rotation of Axes II

In this case, the constants are easy to find:

$$X'' = X' + 3Y'$$

$$Y'' = -3X' + Y'$$

# Recalculating the Data Points

Our data points relative to the new axes are now:

$$X_1'' = -\tfrac{1}{2} + \left(3 \times -\tfrac{3}{2}\right) = -5$$

$$X_2'' = \tfrac{1}{2} + \left(3 \times \tfrac{3}{2}\right) = 5$$

$$Y_1'' = \left(-3 \times -\tfrac{1}{2}\right) + -\tfrac{3}{2} = 0$$

$$Y_2'' = \left(-3 \times \tfrac{1}{2}\right) + \tfrac{3}{2} = 0$$

# Recalculating the Variances

The new variances are now:

$$C''_{11} = (-5 \times -5) + (5 \times 5) = 25$$

$$C''_{12} = (-5 \times 0) + (5 \times 0) = 0$$

$$C''_{21} = (0 \times -5) + (0 \times 5) = 0$$

$$C''_{22} = (0 \times 0) + (0 \times 0) = 0$$

# Diagonalised Covariance Matrix

The covariance matrix is now:

$$C'' = \begin{pmatrix} 25 & 0 \\ 0 & 0 \end{pmatrix}$$

1. All off–diagonal entries are now zero, so data distortions have been removed
2. The variance of $Y''$ (second entry on the diagonal) is now zero. This variable has now been eliminated, so the data has been dimensionally reduced (from 2 dimensions to 1)

# Eigenvalues

1. The numbers on the diagonal of $C''$ are called the eigenvalues of the covariance matrix.
2. Large eigenvalues correspond to large variances – these are the important variables to consider.
3. Small eigenvalues correspond to small variances – these variables can sometimes be neglected.

# Eigenvectors

The directions of the new rotated axes are called the eigenvectors of the covariance matrix

**Etymological Note:**

*vector*: from Latin, *carrier*

*eigen*: from German, *proper to, belonging to*

# Example 2: A More Realistic Example

3 variables, 5 observations

| Finance, $X$ | Marketing, $Y$ | Policy, $Z$ |
|---|---|---|
| 3 | 6 | 5 |
| 7 | 3 | 3 |
| 10 | 9 | 8 |
| 3 | 9 | 7 |
| 10 | 6 | 5 |

# Running PCA in R

```
# read in the  data

data <- read.table("factor1.dat", header=T, sep = "\t")

# run the PCA

pca_out <- princomp(data, cor = FALSE)

# print results

cat("\n", "Call was:" , "\n")
print (pca_out$call)
print (pca_out$loadings)
cat("\n", "Standard deviations:", "\n")
print (pca_out$sdev)
cat("\n", "Centre:" , " \n")
print (pca_out$center)
cat("\n",  "Number of observations:" , "\n")
print (pca_out$n.obs)

rm(list = ls())
```

# PCA Output from R

```
> source("make_pca")

 Call was:
princomp(x = data, cor = FALSE)

Loadings:
        Comp.1 Comp.2 Comp.3
Physics  0.998
History          0.791 -0.611
English          0.612  0.790

               Comp.1 Comp.2 Comp.3
SS loadings     1.000  1.000  1.000
Proportion Var  0.333  0.333  0.333
Cumulative Var  0.333  0.667  1.000

 Standard deviations:
   Comp.1    Comp.2    Comp.3
3.1421463 2.8298304 0.1974246

 Centre:
Physics History English
    6.6     6.6     5.6

 Number of observations:
[1] 5
```

# Assumptions and Limitations of PCA

1. Large variances represent signal. Small signals represent noise.
2. The new axes are *linear combinations* of the original axes
3. The principal components are orthogonal (perpendicular) to each other
4. The distributions of our measurements are fully described by their mean and variance

# Advantages of PCA

1. The procedure is hypothesis free.
2. The procedure is well defined
3. The solution is (essentially) unique

# Factor Analysis

There are a multiplicity of procedures which go under the name of factor analysis.

Factor analysis is *hypothesis driven* and tries to find *a priori* structure in the data.

We'll use the data from Example 2 in the next few slides...

# Factor Model

**Hypothesis:** A significant part of the variance of our 3 variables can be expressed as linear combinations of two underlying factors, $F_1$ and $F_2$:

$$X = a_1F_1 + a_2F_2 + e_X$$

$$Y = b_1F_1 + b_2F_2 + e_Y$$

$$Z = c_1F_1 + c_2F_2 + e_Z$$

$e_X$, $e_Y$ and $e_Z$ allow for errors in the measuements of our 3 variables.

# Factor Model

**Hypothesis:** A significant part of the variance of our 3 variables can be expressed as linear combinations of two underlying factors, $F_1$ and $F_2$:

$$
\begin{aligned}
X &= a_1 F_1 + a_2 F_2 + e_X \\
Y &= b_1 F_1 + b_2 F_2 + e_Y \\
Z &= c_1 F_1 + c_2 F_2 + e_Z
\end{aligned}
$$

$e_X$, $e_Y$ and $e_Z$ allow for errors in the measuements of our 3 variables.

# Factor Model

**Hypothesis:** A significant part of the variance of our 3 variables can be expressed as linear combinations of two underlying factors, $F_1$ and $F_2$:

$$
\begin{aligned}
X &= a_1 F_1 + a_2 F_2 + e_X \\
Y &= b_1 F_1 + b_2 F_2 + e_Y \\
Z &= c_1 F_1 + c_2 F_2 + e_Z
\end{aligned}
$$

$e_X$, $e_Y$ and $e_Z$ allow for errors in the measuements of our 3 variables.

# Assumptions

1. The factors $F_i$ are independent with mean 0 and variance 1

2. The error terms are independent with mean 0 and variance $\sigma_i^2$

With these assumptions, we can calculate the variances in terms of our hypothesised loadings.

# Assumptions

1. The factors $F_i$ are independent with mean 0 and variance 1

2. The error terms are independent with mean 0 and variance $\sigma_i^2$

With these assumptions, we can calculate the variances in terms of our hypothesised loadings.

# Factor Model Variances

$$C_{XX} = a_1^2 + a_2^2 + \sigma_{e_X}^2$$

$$C_{YY} = b_1^2 + b_2^2 + \sigma_{e_Y}^2$$

$$C_{ZZ} = c_1^2 + c_2^2 + \sigma_{e_Z}^2$$

---

$$C_{XY} = a_1 b_1 + a_2 b_2$$

$$C_{XZ} = a_1 c_1 + a_2 c_2$$

$$C_{YZ} = b_1 c_1 + b_2 c_2$$

# Observed Variances for Example 2

$$C_{XX} = 9.84$$

$$C_{XX} = 5.04$$

$$C_{XX} = 3.04$$

---

$$C_{XY} = -0.36$$

$$C_{XZ} = 0.44$$

$$C_{YZ} = 3.84$$

Note: Observations are not mean-adjusted

# Fitting the Model

To fit the model we would like to choose the loadings $a_i$, $b_i$ and $c_i$ so that the predicted variances match the observed variances.

**Problem:** The solution is not unique

That is, there are *different* choices for the loadings that yield the *same* values for the variances; in fact, an infinite number.....

# Fitting the Model

To fit the model we would like to choose the loadings $a_i$, $b_i$ and $c_i$ so that the predicted variances match the observed variances.

**Problem:** The solution is not unique

That is, there are *different* choices for the loadings that yield the *same* values for the variances; in fact, an infinite number.....

# So ... what does the computer do ?

In essence, it finds an initial set of loadings and then rotates these to find the “best” solution according to a user specified criterion

- **Varimax:** Tends to maximise the difference in loading between *factors*.
- **Quartimax:** Tends to maximise the difference in loading between *variables*.

# So ... what does the computer do ?

In essence, it finds an initial set of loadings and then rotates these to find the “best” solution according to a user specified criterion

- **Varimax:** Tends to maximise the difference in loading between *factors*.
- **Quartimax:** Tends to maximise the difference in loading between *variables*.

# Example 3

| Physics | History | English | Maths | Politics |
|---|---|---|---|---|
| 8 | 4 | 3 | 8 | 5 |
| 7 | 3 | 3 | 9 | 3 |
| 10 | 4 | 3 | 9 | 4 |
| 3 | 8 | 9 | 4 | 7 |
| 4 | 8 | 6 | 4 | 7 |
| 5 | 9 | 8 | 5 | 9 |

# Running Factor Analysis in R

```
# read in the  data

data <- read.table("factor2.dat", header=T, sep = "\t")

# run the factor analysis with no rotation

none_out <- factanal(data,  factors=2, rotation="none")

# run the factor analysis with varimax rotation

varimax_out <- factanal(data,  factors=2, rotation="varimax")

# print results

print(none_out)
cat("\n\n")
print(varimax_out)


rm(list = ls())
```

# PCA Analysis of Example 3

```
 Call was:
princomp(x = data, cor = FALSE)

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Physics   0.448  0.728 -0.145 -0.463  0.186
History  -0.472  0.360  0.164 -0.243 -0.749
English  -0.489        -0.842 -0.154  0.170
Maths     0.440  0.129 -0.462  0.570 -0.501
Politics -0.380  0.569  0.174  0.615  0.352

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
SS loadings       1.0    1.0    1.0    1.0    1.0
Proportion Var    0.2    0.2    0.2    0.2    0.2
Cumulative Var    0.2    0.4    0.6    0.8    1.0

 Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
4.9208397 1.3061940 0.7743450 0.4235666 0.2239976

 Centre:
 Physics  History  English    Maths Politics
6.166667 6.000000 5.333333 6.500000 5.833333

 Number of observations:
[1] 6
```

# Factor Analysis of Example 3

```
Call:
factanal(x = data, factors = 2, rotation = "none")

Uniquenesses:
 Physics  History  English    Maths Politics
   0.005    0.005    0.104    0.036    0.050

Loadings:
         Factor1 Factor2
Physics  -0.943   0.325
History   0.958   0.277
English   0.944
Maths    -0.981
Politics  0.889   0.401

               Factor1 Factor2
SS loadings      4.452   0.348
Proportion Var   0.890   0.070
Cumulative Var   0.890   0.960

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.38 on 1 degree of freedom.
The p-value is 0.538
```

```
Call:
factanal(x = data, factors = 2, rotation = "varimax")

Uniquenesses:
 Physics  History  English    Maths Politics
   0.005    0.005    0.104    0.036    0.050

Loadings:
         Factor1 Factor2
Physics  -0.397   0.915
History   0.851  -0.520
English   0.689  -0.649
Maths    -0.675   0.713
Politics  0.896  -0.385

               Factor1 Factor2
SS loadings      2.614   2.186
Proportion Var   0.523   0.437
Cumulative Var   0.523   0.960

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.38 on 1 degree of freedom.
The p-value is 0.538
```

# Some Opinions on Factor Analysis...

*FA is not worth the time necessary to understand it and carry it out.*
-Hills, 1977

*Factor analysis should not be used in most practical situations.*
-Chatfield and Collins, 1980, pg. 89.

*At the present time, factor analysis still maintains the flavor of an art, and no single strategy should yet be "chiseled into stone".*
-Johnson and Wichern, 2002, pg. 517.

# Take Home

1. Measurements are subject to many distortions including noise, rotation and redundancy
2. PCA is a robust, well defined method for reducing data and investigating underlying structure.
3. There are many approaches to factor analysis; all of them are subject to a certain degree of arbitrariness and should be used with caution

# Further Reading

“A Tutorial on Principal Components Analysis”
Jonathon Shlens
http://www.snl.salk.edu/ shlens/notes.html

“Factor Analysis”
Peter Tryfos
www.yorku.ca/ptryfos/f1400.pdf