

Updated 9-12-04

Dimensionality

This document includes discussion of two different ways of understanding (mathematically and empirically) the dimensionality of multiple observed variables:

1. Geometric interpretation
 2. Eigenvalue-Eigenvector decomposition of covariance (or correlation) matrices
-

"Reducing the dimension of data" – What does this really mean?

Think of a line (which is a **one** dimensional object) in 2-dimensions, 3-dimensions, >3 dimensions.

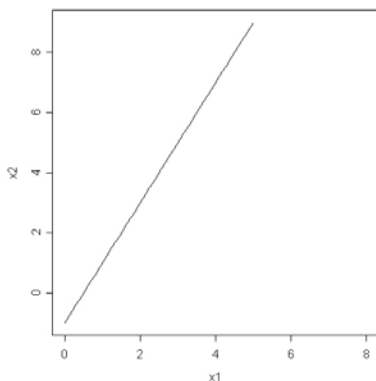
In 2-dimensions, the line can be represented by ordered pairs (x1, x2).

In 3-dimensions, the line can be represented by ordered triplets (x1, x2, x3)

In 4-dimensions, the line can be represented by ordered quadrats (x1, x2, x3, x4) etc.

But actually, no matter how many dimensions there are, a line can be represented as a function of just **one** "underlying" variable.

Here this line below can be represented as $X_2 = 2 X_1 - 1$.



But, instead it also can be represented by thinking of X1 and X2 both as functions of the same one variable f, that is.....

$$X_1(f) = 2 + f$$

$$X_2(f) = 3 + 2f$$

where f goes from negative infinity to positive infinity

When $f = 0$, we have the point $(2,3)$

When $f = -1$, we have the point $(1,1)$

Intuition:

1. X_1 and X_2 change because f changes
 2. X_1 and X_2 are related because they are both related to f
-

The equations above as a function of f represent just one way to parameterize the linear relation between X_1 and X_2 .

NOTE, THIS IS NOT UNIQUE.

In other words, there is more than one way to parameterize the exact same line.

For example,

$$X_1 = 6 - 13f$$

$$X_2 = 11 - 26f$$

gives the same relation between X_1 and X_2 .

When $f = 4/13$, we have the point $(2,3)$

When $f = 5/13$, we have the point $(1,1)$

ALSO, another way to parameterize this same line is

$$X_1 = f$$

$$X_2 = 2f - 1$$

This parameterization might be considered the most simple way, that is, we fix one of the observed variables to be equal to the underlying variable.

The fact that the parameterization is not unique is related to the idea of "factor rotation" which will be talked about more later.

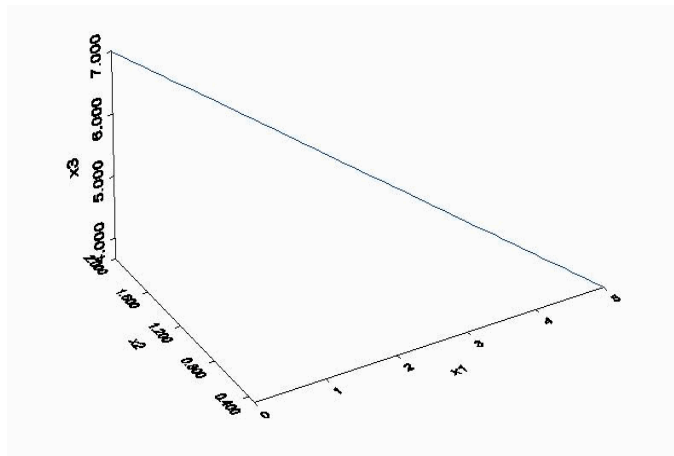
What about if we have 3-dimensions, i.e. 3 observed X variables. A line is defined by two restrictions of the 3-dimensional space.

For example $X_1 = 6 - 3 X_2$ and $X_3 = 3 + 2 X_2$ defines a line in three dimensions. Basically two intersecting planes. Another way to write this is

$$X_1 = f$$

$$X_2 = 2 - 1/3 f$$

$$X_3 = 7 - 2/3 f$$



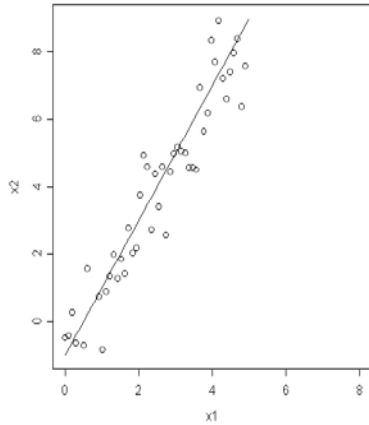
When $f = 0$, we get the point $(0, 2, 7)$

When $f = 3$, we get the point $(3, 1, 5)$

What about if we have data???????

$$X_1 = 2 + f + e_1, \quad \text{Var}(e_1) = \psi_1$$

$$X_2 = 3 + 2f + e_2, \quad \text{Var}(e_2) = \psi_2$$



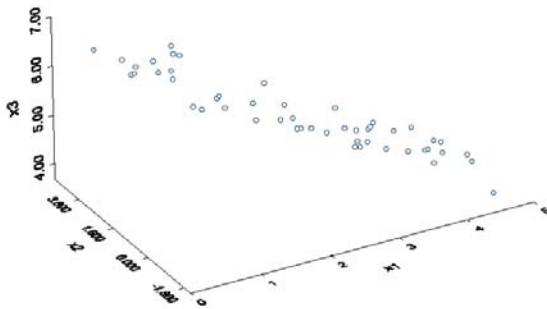
There is one dimension (or one factor) underlying this data. Most of the variability can be explained by a projection of the data onto a line.

What about if we have data in 3-dimensions???????

$$X1 = f + e1$$

$$X2 = 2 - 1/3 f + e2$$

$$X3 = 7 - 2/3 f + e3$$



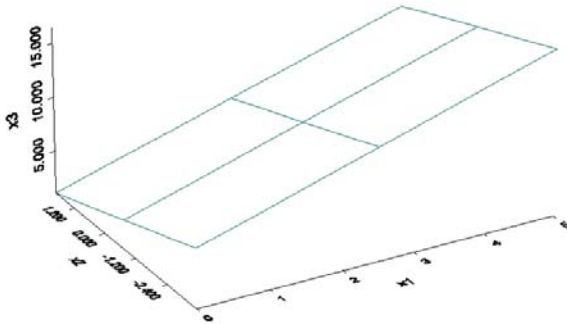
So, there is one dimension underlying the data. In other words the data follow a line.

What if there are two dimensions underlying the data. -----> the data will follow a plane.

$$X1 = f1$$

$$X_2 = f_2$$

$$X_3 = 3 + 2f_1 - f_2$$

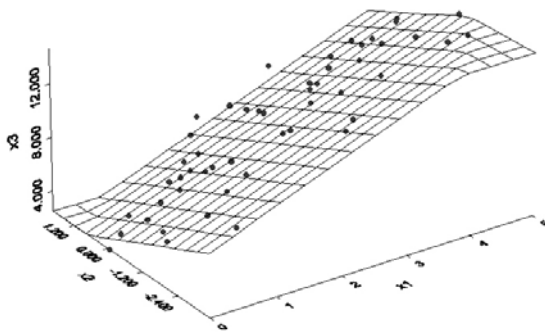


And here's what it would look like if we had data with two underlying dimensions. (ignore the bends in the plane at the edges this is an artifact of the graphing software I was using).

$$X_1 = f_1 + e_1$$

$$X_2 = f_2 + e_2$$

$$X_3 = 3 + 2f_1 - f_2 + e_3$$



What if there are > 2 underlying dimensions?

Geometric interpretation is not clear anymore, but form of the equations is the same.

In general, we are going to need something other than data visualization and geometry in order to determine number of factors underlying data.....Thus, we are going to investigate the Covariance structure via eigenvalues and eigenvectors next.

EIGENVALUE EIGENVECTOR DECOMPOSITION

Symmetric matrices can be factored in a way that gives more insight about their structure

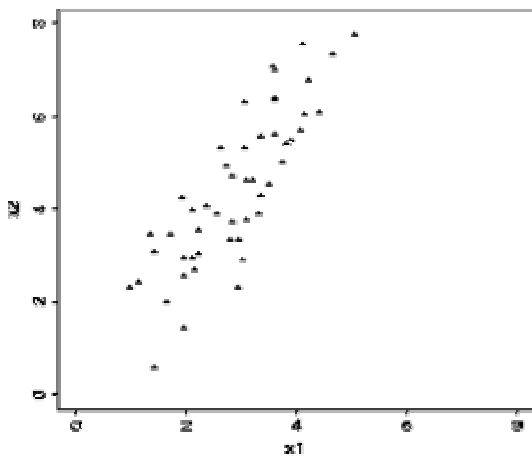
Take A to be a symmetric $p \times p$ matrix. Then

$$A = U D U'$$

where the columns of U contain the eigenvectors of A and the diagonal matrix D contains the associated eigenvalues of A .

- Eigenvectors give the **direction of variability**. The first eigenvector gives the direction of maximum variability in the data, while the second eigenvector looks in the orthogonal directions from the first and gives the direction (out of all the orthogonal directions) which has the most variability, etc.
 - Eigenvalues quantify **how much variability** is in the direction of the corresponding eigenvector.
-

EXAMPLE Given the following 50 data points represented by (X_1, X_2)

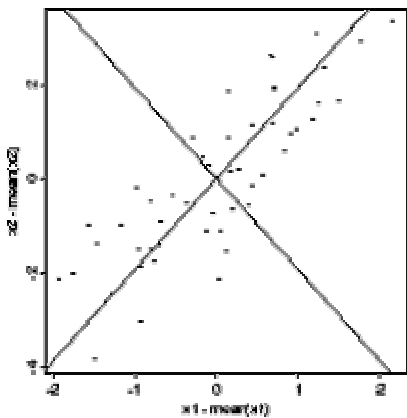


$$\widehat{Var} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = S = \begin{pmatrix} s_{X_1}^2 & \widehat{Cov}(X_1, X_2) \\ \widehat{Cov}(X_1, X_2) & s_{X_2}^2 \end{pmatrix} = \begin{pmatrix} 0.9336 & 1.3672 \\ 1.3672 & 2.8274 \end{pmatrix}$$

And the eigenvalue eigenvector decomposition of S is given by...

U		D		U'	
0.4640	0.8858	3.54359	0	0.4640	0.8858
0.8858	-0.4640	0	0.21741	0.8858	-0.4640

So the 1st eigenvector is the direction (.4640, .8858) (i.e. the line with positive slope going through Figure 2). And the eigenvalue associated with this eigenvector is 3.54359. This number is the variance of the data if it were projected onto the first eigenvalue. Compare this variance to the variance of X_2 . And the second eigenvector is the direction (.8858, -.4640) (i.e. the line with negative slope going through Figure 2). Its associated eigenvalue is .21741, that is, if we projected the data onto this eigenvector the variance would be .21741.



Notice that the total variability in the data is

$$S_{x_1}^2 + S_{x_2}^2 = .9336 + 2.8274 = 3.761$$

AND also can be calculated as the sum of the eigenvalues

$$3.54359 + 0.21741 = 3.761$$

So, we can say that the first eigenvector (dimension) explains $3.54359 / 3.761 = 94.2\%$ of the variability in the data.

Since the data is only two dimensional (i.e. only (X_1, X_2)) there are only two eigenvector/eigenvalue pairs.

If there are more than 3 dimensions graphical interpretation becomes very difficult but eigenvector/eigenvalue idea is extendable.

PRINCIPAL COMPONENT ANALYSIS is just eigenvector/eigenvalue decomposition.

The principle components are linear combinations of the observed variables where the weights of the sum are taken to be the eigenvectors.

So in the 2-variable example above, the 1st principal component is calculated by taking

$$.4640 * X_1 + .8858 * X_2$$

and the second principal component is calculated by

$$.8858 * X_1 - .4640 * X_2$$

Then the variance of the first principal component is 3.54359 (1st eigenvalue) and the variance of the second principal component is 0.21741 (2nd eigenvalue)