BERTINORO 3 (JVW)

Correlation and PCA

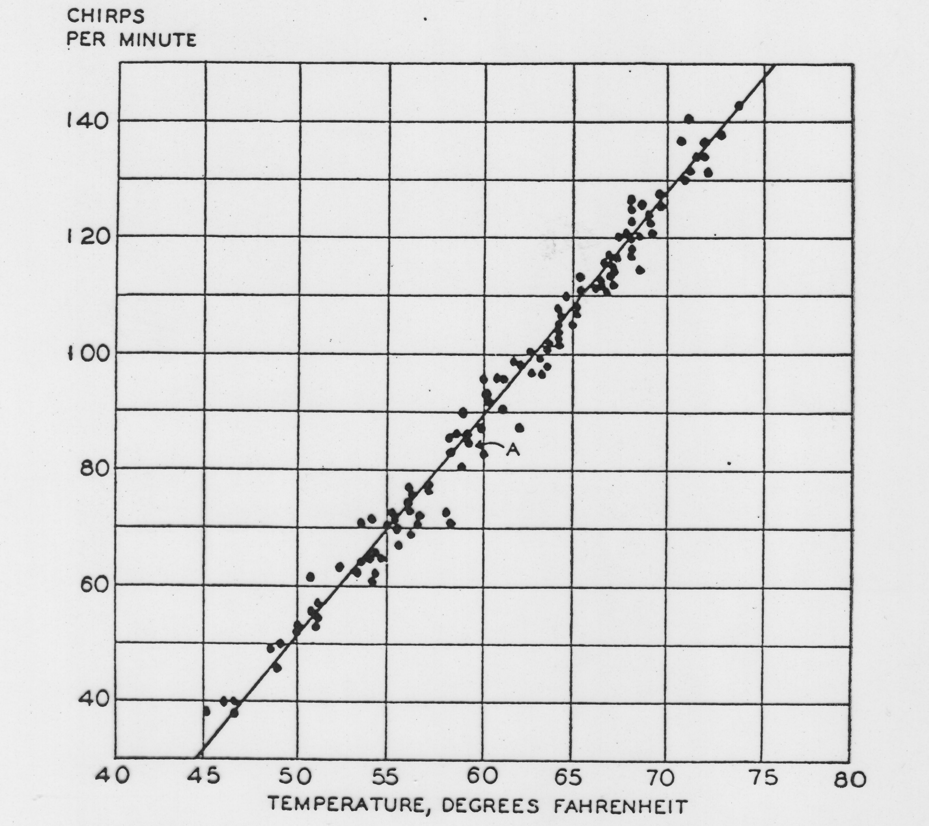
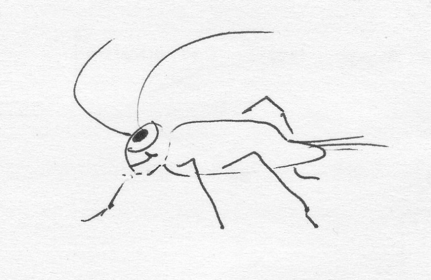


Chart 19.1. Temperature and Chirps per Minute of 115 Crickets. Data provided by Mr. Bert. E. Holmes.



Correlation - why do we try it?

When we make a set of measurements, it is instinct to try to correlate the observations with other results. We might wish

(1) to check that other observers' measurements are reasonable,

(2) to check that our measurements are reasonable,

(3) to test a hypothesis, perhaps one for which the observations were explicitly made,

(4) in the absence of any hypothesis, any knowledge, or anything better to do with the data, to find if they are correlated with other results in the hope of discovering some New and Universal Truth.

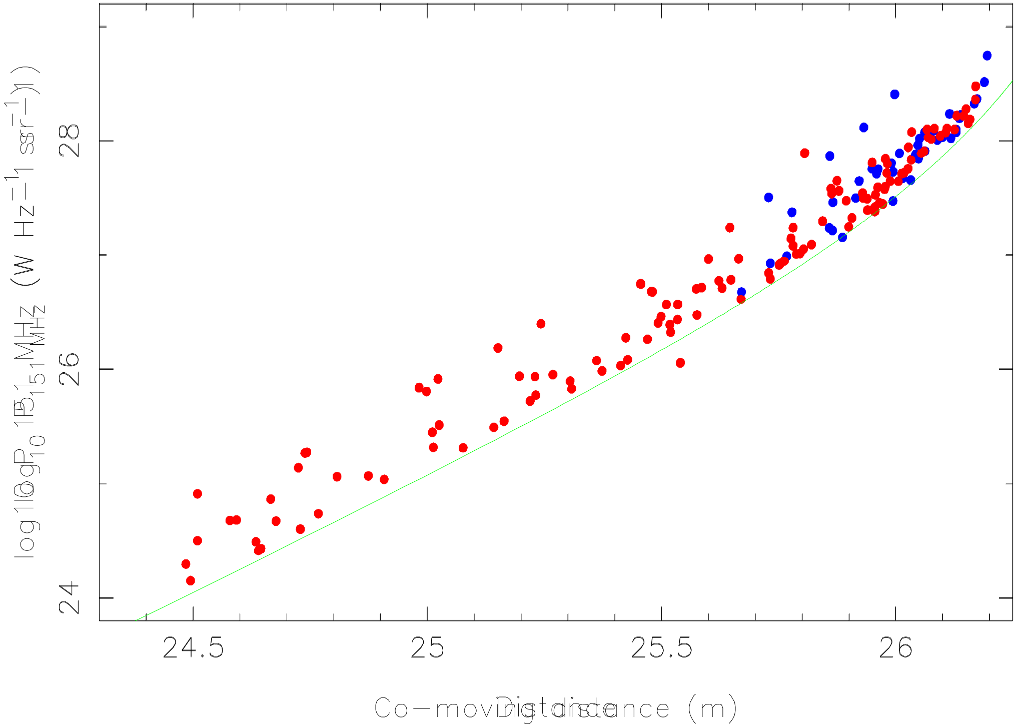
We are gonna do it - and we are going to fall into some deadly traps. We already have.

The fishing trip

Suppose that we have plotted something against something, on a Fishing Expedition.

Does the eye see much correlation? If not, formal testing for correlation is probably a waste of time.

Could the apparent correlation be due to selection effects? Consider for instance the beautiful correlation obtained by Sandage (1972): 3CR radio luminosities vs distance.



Radio luminosities of 3CR radio sources versus distance modulus.

Still on the fishing trip…

The plot **proves** luminosity evolution for radio sources? Are the more distant objects (at earlier epochs) clearly not the more powerful?

No! The sample is flux- (or apparent intensity) limited; the solid line shows the flux-density limit of the 3CR catalogue. The lower right-hand region can never be populated.

But the upper left? Provided that **the luminosity function** (the true space density in objects per Mpc^3) **slopes downward with increasing luminosity**, the objects are bound to crowd towards the line. **The only conclusion** from the diagram!

Astronomers produce many plots of this type, and say things like terms like
'The lower right-hand region of the diagram is unpopulated because of the detection limit, but there is no reason why objects in the upper left-hand region should have escaped detection....'

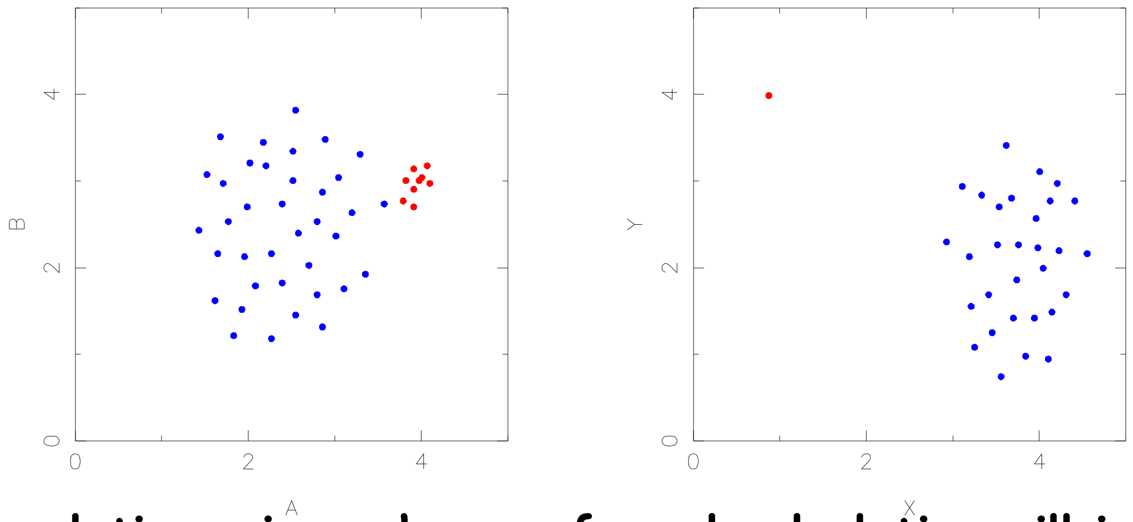
Nonsense – probabilities rule! There are only low-luminosity sources to be seen at low redshifts because there's not enough volume to pick up the high-fliers.

This applies to any proposed correlation for variables with steep probability functions dependent upon one of the variables plotted.

Still fishing….

If we are happy, we can try formal calculation of the significance of the correlation. But, if there is a correlation, does the regression line (the fit) make sense?

If we are still happy - is the formal result is realistic?
Wall's rule of thumb: if 10 percent of the points are grouped by themselves so that covering them with the thumb destroys the correlation to the eye, then we should doubt it. **Selection effects, data errors, or some other form of statistical conspiracy?**



Dodgy correlations: in each case formal calculation will indicate that a correlation exists to a high degree of significance.

The fishing trip continues…

If **still** confident, remember that

a correlation does not prove a causal connection. Examples:

¤ **The price of fish in Billingsgate Market and the size of feet in China.**

¤ **Number of violent crimes in cities versus number of churches.**

¤ **The quality of student handwriting versus their height.**

¤ **Stock market prices and the sunspot cycle.**

¤ **In World War II, bombing accuracy was far greater when enemy fighter planes were present.**

¤ **Cigarette smoking versus lung cancer.**

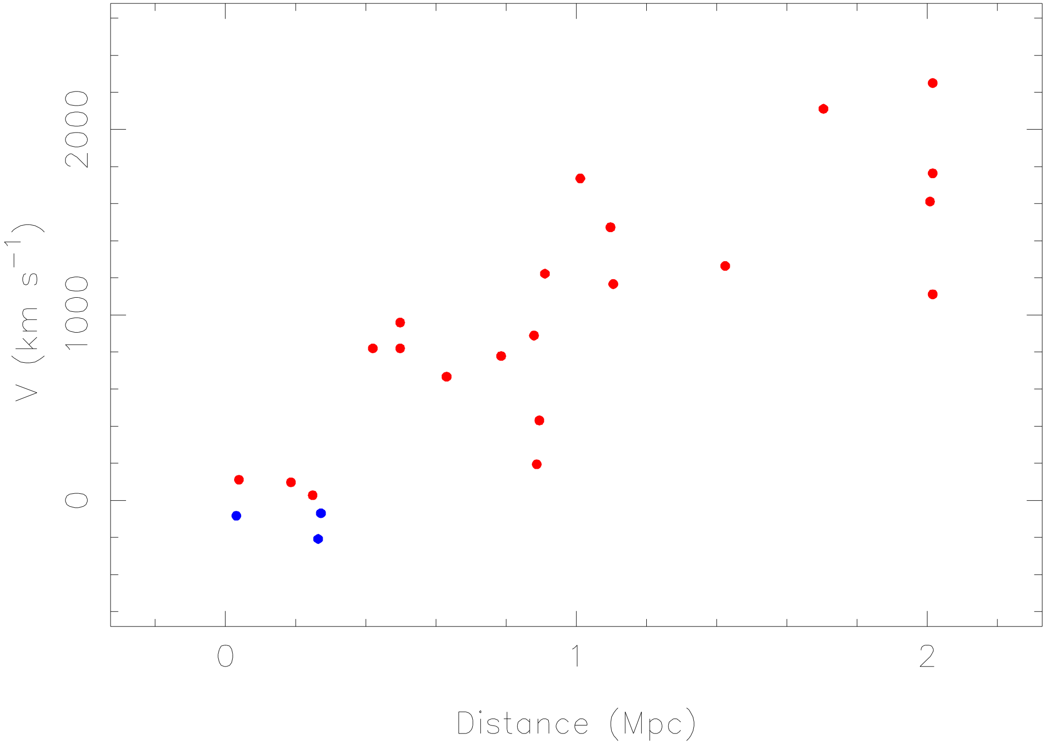
¤ **Health versus alcohol intake.**

1. Lurking third variables
2. Similar time scales
3. Causal connection

There are ways of searching for intrinsic correlation between variables when they are known to depend mutually upon a third variable. But 'known'?????

The fishing trip ends – big fish are out there

Don't get too discouraged by all the foregoing. Consider the example figure, a ragged correlation if ever there was one, although there are no nasty groupings of the type rejected by the Rule of Thumb.



An early Hubble diagram (Hubble 1936); recession velocities of a sample of 24 galaxies versus distance measure.

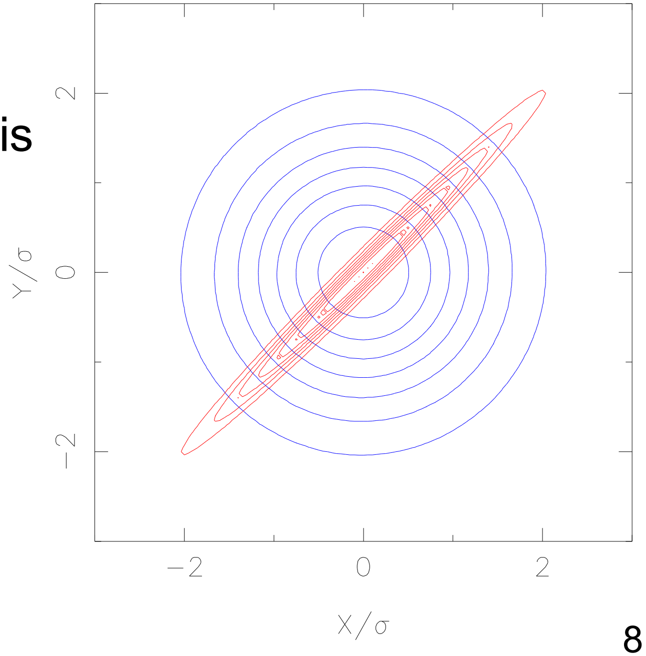
Correlation: standard model

We have a set of measurements **(X_i, Y_i)** and we ask (formally) if they are related to each other. What does 'related' mean? In general we model our data as a **bivariate** or **joint Gaussian** of **correlation coefficient ρ**:

$$\mathrm{prob}(x, y \mid \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left(\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho x y}{\sigma_x\sigma_y}\right)\right).$$

This model is so well developed that **'correlation'** and **'ρ ≠ 0'** are nearly synonymous; if **ρ** → **0** there is little correlation, while if **ρ** → **1** the correlation is perfect.

Left: linear contours of the bGpd. Near circular: **ρ = 0.01**, little connection between **x** and **y**; highly elliptical: **ρ = 0.99**, strong correlation between **x** and **y**. Negative values of **ρ** reverse the tilt: **'anticorrelation'**.



Correlation: standard model, continued

The parameter ρ is the **correlation coefficient**, and is given by

$$\rho = \frac{\mathrm{cov}[x,y]}{\sigma_x \sigma_y}$$

where **cov** is the **covariance** of x and y, and σ_x^2 and σ_y^2 are the variances. The correlation coefficient can be estimated by

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

r is known as the **Pearson Product Moment Correlation Coefficient**.

The contours of the **bivariate Gaussian** will have dropped by $1/e$ from the maximum at the origin when

$$\frac{1}{1-\rho^2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho x y}{\sigma_x \sigma_y}\right) = 1,$$

or in matrix notation, when

$$(x \quad y)\,\frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x\sigma_y} \\ -\frac{\rho}{\sigma_x\sigma_y} & \frac{1}{\sigma_y^2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 1.$$

The inverse of the central matrix is known as the **covariance matrix** or **error matrix**.

$$C = \begin{pmatrix} \sigma_x^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \sigma_y^2 \end{pmatrix}.$$

The off-diagonal elements of the covariance matrix can be estimated by

$$\frac{1}{N-1}\overline{(X_i - \overline{X_i})(Y_j - \overline{Y_j})}.$$

The matrix is particularly valuable in calculating propagation of errors, but there are numerous applications, for example in Principal Component Analysis and in Maximum-Likelihood modelling.

Formal testing – what are we doing?

1. For bivariate data, what we really want to know is whether or not **ρ = 0.**

2. Using the bivariate Gaussian is a very specific model.

3. A Gaussian is assumed - it allows only two variances, and assumes that both **x** and **y** are random variables.

4. $\boldsymbol{\sigma}_x$ and $\boldsymbol{\sigma}_y$ include both the errors in the data, and their intrinsic scatter -- all presumed Gaussian.

5. Does not apply, for example, to data where the **x**-values are well-defined and there are 'errors' only in **y**, perhaps different at different **x**. In such cases we would use model-fitting, perhaps of a straight line. This is a **different issue** – This is model-fitting, or parameter-estimation.

CAUTION! Are your data right for this testing process?

Correlation testing - comments

- The **non-parametric tests** circumvent some of the issues involved in the non-Bayesian approach, but they have no bearing on the fundamental issue – **what was the real question**?

2. But as ever, the Bayesian approach, strong in answering the real question, forces reliance on a **model**.

3. In practice there is little difference between the **Fisher test** and results from **Jeffreys distribution**. We can show this with some random Gaussian data with a correlation of zero. In the standard way, we can use the **r-distribution** to find the probability of **r** being as large, or larger, than we observe, on the hypothesis that **ρ=0**. If this probability is small, the test is hinting at the possibility that the correlation is actually positive. Therefore we compare with the probability, from the Jeffreys distribution, that **ρ** is positive. If the probability from Fisher's **r-distribution** is small we expect the probability from **ρ** to be large; and in fact we can see, either from simulations or from the algebraic form of the distributions, that **the sum of these two probabilities is always → 1**.

Interpreting the standard Fisher test (illegally!) to be telling us the chance that **ρ** is positive, actually works very well!

Correlation found! Now what?

First question: what's the law relating the variables?

We rush off and fit 'regression lines', often by Least Squares.

But recognize that we're now model-fitting. There is a crucial distinction.

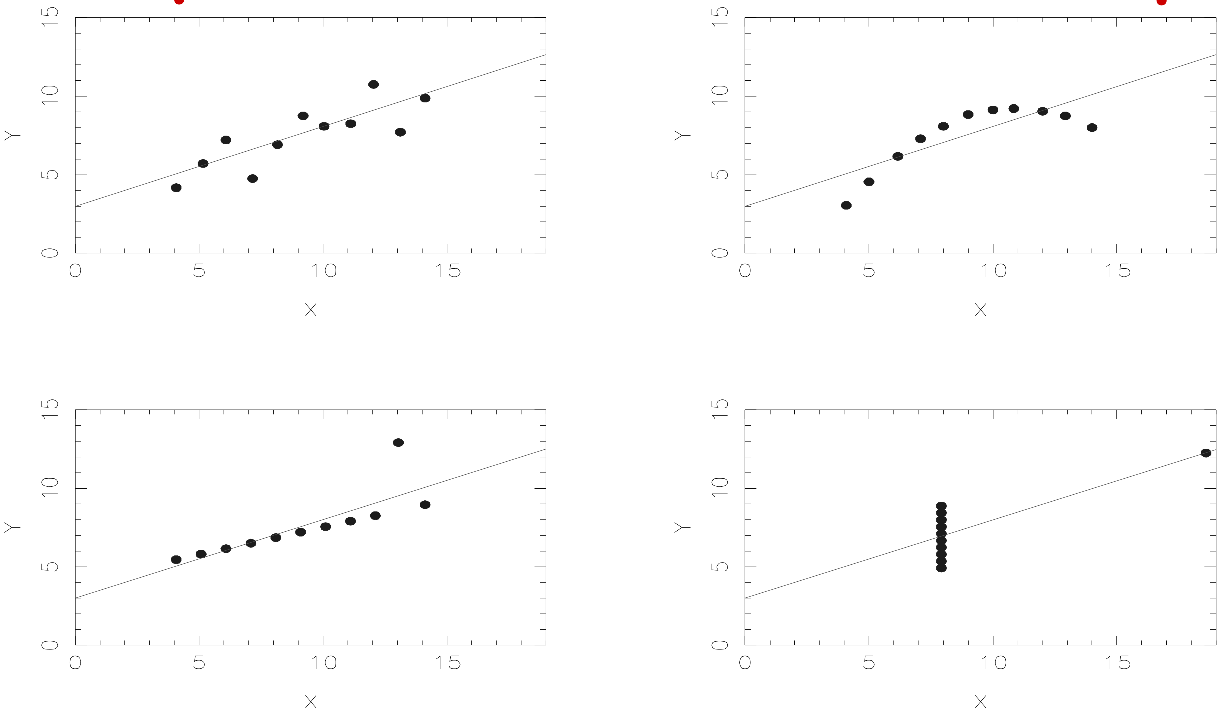
In the model fitting (coming later) :

- Are there better quantities to minimize than the squares of deviations?
- What errors result on the regression-line parameters?
- Why should the relation be linear?
- What are we trying to find out?

Example: If we have found a correlation between **x** and **y**, which variable is dependent; do we want to know (**x** on **y**) or (**y** on **x**)? The coefficients are generally completely different.

As an argument against blind application of correlation testing, consider the example of Anscombe's (1973) famous quartet:

Anscombe's quartet: correlation vs independence



Anscombe's quartet: 4 fictitious sets of 11 (Xi,Yi), each with the same <X>, <Y>, identical coefficients of correlation, regression lines, residuals in Y, estimated standard errors on slope, and covariance matrices.

1. In ¾ cases, the points are clearly related; they are far from independent but still show only indifferent quality of correlation. At upper right, choice of the 'right' relation would result in a perfect fit.

2. **X** independent of **Y** means **prob(X,Y)=prob(X)prob(Y)**, or **prob (X|Y)=prob(X).**

X correlated with **Y** means **prob(X,Y) ≠ prob(X)prob(Y)** in a way such as to give **r ≠ 0.**

We can have **prob(X,Y) ≠ prob(X)prob(Y)** AND **r = 0**, example: Union Jack.

Principal Component Analysis (PCA)

PCA is the ultimate correlation searcher when many variables are present.

Given a sample of **N** objects with **n** parameters measured for each, what is correlated with what?

What variables produce primary correlations, and what produce secondary, via the lurking **third** (or indeed **n-2**) variables?

PCA is one of a family of algorithms (known as multivariate statistics) designed for this situation. Its task: given a sample of **N** objects with **n** measured variables $\mathbf{x}_n$, find a **new set of ξ_n variables** that are **orthogonal (independent),** each one a **linear** combination of the original variables:

$$\xi_i = \sum_{j=1}^{n} a_{ij} x_j$$

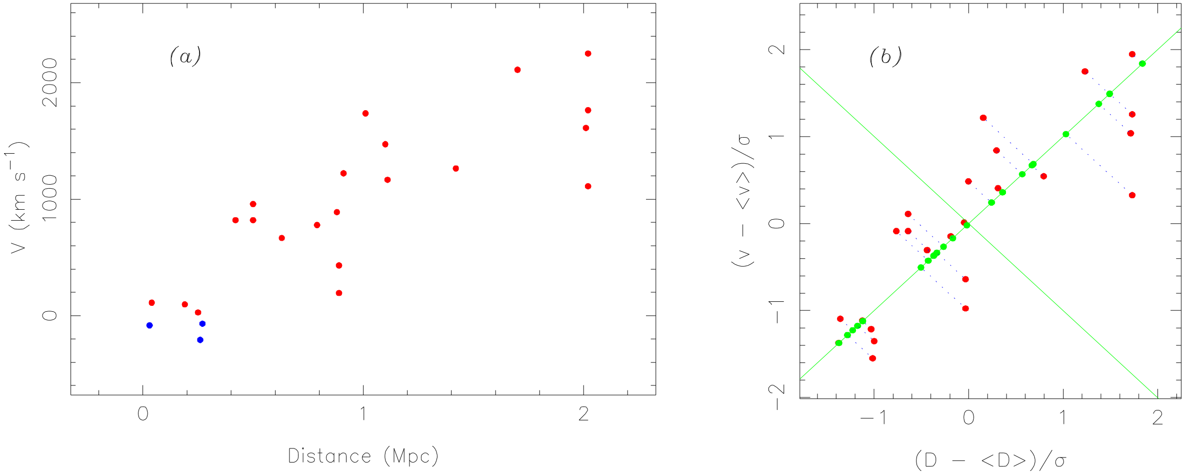
with values of $\mathbf{a}_{ij}$ such that the **smallest number** of new variables account for as much of the variance as possible. The ξ_i are the **principal components**.
If most of the variance involves **just a few** of the **n** new variables, we have found **a simplified description of the data.**

Finding which of the variables correlate (and how) may lead to success on our fishing expedition - we may have caught new physical insight.

PCA - Example 1

- ***Geometric approach***: Back to the early Hubble diagram, 24 galaxies with two measured variables, **recession velocity v** and **distance d**. Procedure:

 - Normalize by subtracting the means from each variable and divide by the std dev, i.e. plot **$v_i' = (v_i - <v>) / \sigma_v$** vs **$d_i' = (d_i - <v>) / \sigma_d$**



 - Find the **first principal component** by rotating the axis through the origin to align with max elongation, the direction of apparent correlation, using least-squares.

 - Maximizing the variance along PC1 is **equivalent to minimizing the sums of the squares** of the distances of the points from this line through the origin.

 - The distance of a point from the direction PC1 (dotted verticals) represents the value (**score**) of PC1 for that point.

 - PC1 is clearly a linear combination of the two original variables; in fact it is **v' = d'.**

 - Because the new coordinate system was found by simple rotation, distances from origin are unchanged; the **total variance** of **v'** and **d'** remains **2.0**.

PCA - Example 1 continued

- The variance of PC1, the normalized distances squared from PC2, is 1.837.
- The remaining variance of the sample must be accounted for by the projection of data points onto the axis PC1, perpendicular to PC2; lengths of these are scores of the second principal component PC2, and this is verified as 0.163; sum = 2.0.

The table sets out the the results in the standard way of PCA.

	PC1	PC2
Eigenvalue	1.837	0.163
Proportion	0.918	0.082
Cumulative	0.918	1.000

Variable	PC1	PC2
d (Mpc)	1.0	1.0
v (km s^{-1})	1.0	-1.0

2. The matrix approach. Procedure:

- Construct the error matrix. i.e. for the two-variable case of the example, **a(1,1) = $\Sigma d'^2$, a(2,2) = $\Sigma v'^2$, a(1,2) = a(2,1) = $\Sigma v'd'$.**

.- Seek a principal axis transformation that makes the cross-terms vanish, an axis transformation to rotate the ellipses of our BVGD so that the axes of the ellipses coincide with the principal axes of the coordinate system.

- Simple in matrix notation! We determine the **eigenvalues of the error matrix** and form its **eigenvectors** (for the example, **v' = d'** and **v' = -d'** as seen in the Hubble fig.)
- Use these eigenvectors to form the **transpose matrix T**, for variable transformation and axis rotation. The axis rotation **diagonalizes the matrix**, i.e. in the new axis system, the cross terms are zero; we have rotated the axes until there is **no (v',d') covariance**.

PCA notes

- Our set of data has been reduced from 48 numbers for the 24 galaxies to 4 numbers, a 2 x 2 matrix. How? **PCA assumes that the covariance (error) matrix describes the data**

- This is the case if data drawn from a **multivariate Gaussian** or in general when a simple quadratic form, using the covariance matrix, can describe the distribution of the data.

- Far from generally true that the clouds of points in most n-variate hyperspaces will be so simply distributed.

- In multivariate data sets, the disparate units are taken care of by normalizing: **subtracting mean values and dividing by variances.**

- This is not a prescription. The variance for any particular variable might be dominated by an **outlier** which there are good grounds to reject.

- The choice of weights does therefore depend on familiarity with the data and preferences – there is **plenty of room for subjectivity.**

- PCA is a linear analysis and **tests need to be performed on the linearity of the principal components**. For example, plotting the scores of PC1 vs PC2 should show a Gaussian distribution consistent with ρ = 0.

PCA – Example 2

The Francis and Wills sample of QSOs, 1999

	log		logFWHM	FeII/	logEW	logFWHM	logEW	logEW	CIV/	logEW	SiIII/	NV/	λ1400/
PG name	L_{1216}	α_x	Hβ	Hβ	[OIII]	CIII]	Lyα	CIV	Lyα	CIII]	CIII]	Lyα	Lyα
0947+396	45.66	1.51	3.684	0.23	1.18	3.520	2.08	1.78	0.45	1.24	0.306	0.179	0.143
0953+414	45.83	1.57	3.496	0.25	1.26	3.432	2.19	1.78	0.40	1.24	0.164	0.189	0.093
1114+445	44.99	0.88	3.660	0.20	1.23	3.654	2.27	1.85	0.42	1.48	0.222	0.175	0.092
1115+407	45.41	1.89	3.236	0.54	0.78	3.403	1.90	1.51	0.33	1.14	0.385	0.228	0.134
1116+215	46.00	1.73	3.465	0.47	1.00	3.446	2.14	1.71	0.34	1.20	0.440	0.254	0.126
1202+281	44.77	1.22	3.703	0.29	1.56	3.434	2.72	2.41	0.69	1.87	0.164	0.154	0.098
1216+069	46.03	1.36	3.715	0.20	1.00	3.514	2.12	1.95	0.54	1.20	0.037	0.121	0.056
1226+023	46.74	0.94	3.547	0.57	0.70	3.477	1.64	1.44	0.45	1.00	0.280	0.174	0.018
1309+355	45.55	1.51	3.468	0.28	1.28	3.406	2.01	1.68	0.41	1.15	0.303	0.131	0.064
1322+659	45.42	1.69	3.446	0.59	0.90	3.351	2.19	1.85	0.41	1.30	0.291	0.135	0.097
1352+183	45.34	1.52	3.556	0.46	1.00	3.548	2.14	1.80	0.41	1.29	0.357	0.203	0.116
1402+261	45.74	1.93	3.281	1.23	0.30	3.229	1.91	1.59	0.39	1.09	0.568	0.227	0.161
1415+451	45.08	1.74	3.418	1.25	0.30	3.434	2.32	1.78	0.29	1.40	0.688	0.210	0.142
1427+480	45.54	1.41	3.405	0.36	1.76	3.300	2.03	1.82	0.49	1.21	0.265	0.126	0.117
1440+356	45.23	2.08	3.161	1.19	1.00	3.192	2.14	1.54	0.21	1.05	0.747	0.141	0.092
1444+407	45.92	1.91	3.394	1.45	0.30	3.479	1.99	1.34	0.21	1.06	0.809	0.335	0.164
1512+370	46.04	1.21	3.833	0.16	1.76	3.546	2.02	2.05	0.75	1.28	0.228	0.182	0.050
1626+554	45.48	1.94	3.652	0.32	0.95	3.631	2.14	1.80	0.39	1.36	0.197	0.217	0.118

PCA – Francis & Wills, continued (2)

(1) – subtract mean and divide by variance:

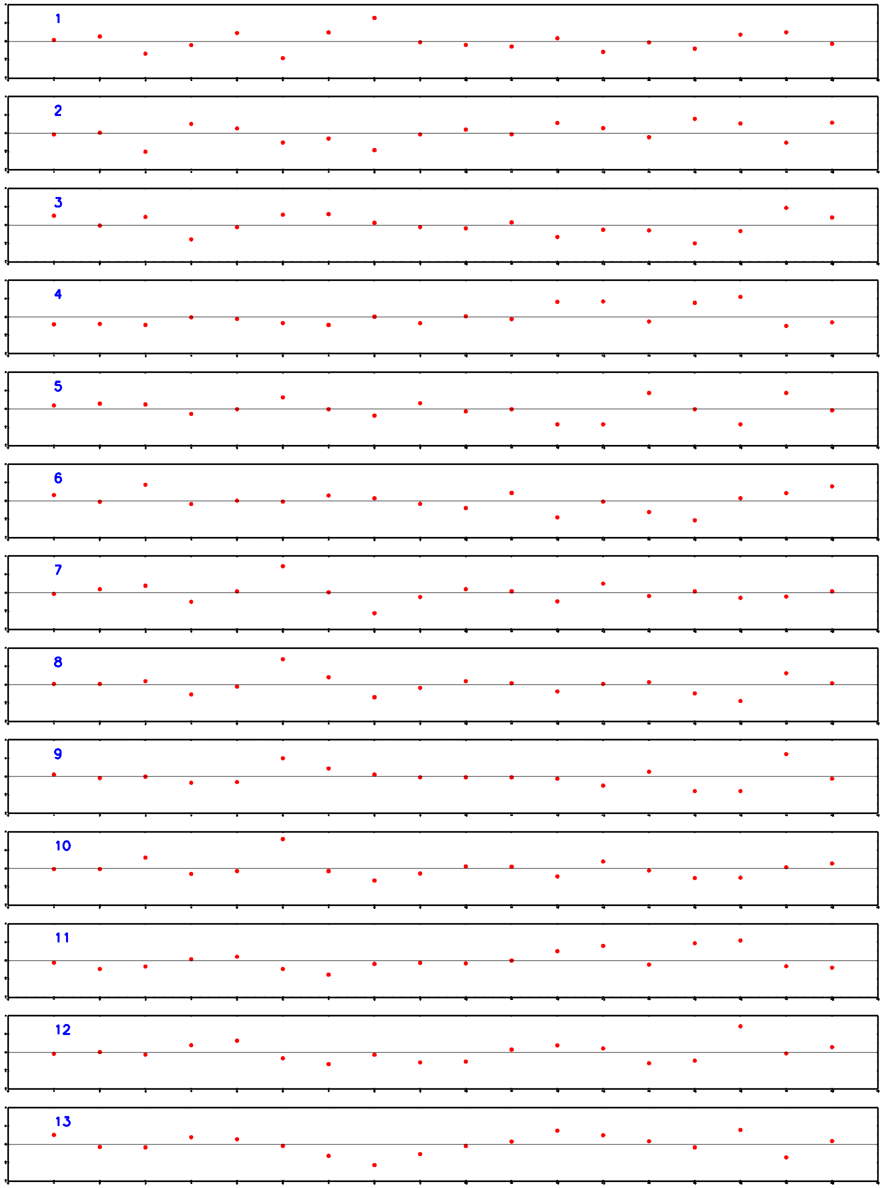
qso\data:	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.14	-0.14	1.014	-0.80	0.39	0.636	-0.13	0.09	0.215	-0.069	-0.252	-0.170	1.005
2	0.52	0.04	-0.061	-0.75	0.57	-0.103	0.39	0.09	-0.157	-0.069	-0.934	0.022	-0.300
3	-1.35	-2.03	0.877	-0.88	0.50	1.760	0.77	0.38	-0.008	1.175	-0.655	-0.246	-0.326
4	-0.42	0.99	-1.548	-0.04	-0.55	-0.346	-0.99	-1.06	-0.678	-0.588	0.128	0.771	0.770
5	0.89	0.51	-0.238	-0.21	-0.03	0.015	0.15	-0.21	-0.604	-0.276	0.392	1.270	0.561
6	-1.84	-1.01	1.123	-0.66	1.27	-0.086	2.90	2.77	2.001	3.197	-0.934	-0.649	-0.170
7	0.96	-0.59	1.192	-0.88	-0.03	0.585	0.06	0.81	0.885	-0.276	-1.544	-1.283	-1.266
8	2.54	-1.85	0.231	0.03	-0.73	0.275	-2.22	-1.36	0.215	-1.313	-0.377	-0.265	-2.258
9	-0.11	-0.14	-0.221	-0.68	0.62	-0.321	-0.47	-0.34	-0.083	-0.536	-0.266	-1.091	-1.057
10	-0.40	0.40	-0.347	0.08	-0.27	-0.782	0.39	0.38	-0.083	0.242	-0.324	-1.014	-0.196
11	-0.57	-0.11	0.282	-0.24	-0.03	0.870	0.15	0.17	-0.083	0.190	-0.007	0.291	0.300
12	0.31	1.11	-1.291	1.64	-1.66	-1.805	-0.94	-0.72	-0.232	-0.847	1.007	0.752	1.475
13	-1.15	0.54	-0.507	1.69	-1.66	-0.086	1.00	0.09	-0.976	0.760	1.584	0.425	0.979
14	-0.13	-0.44	-0.582	-0.48	1.74	-1.210	-0.37	0.26	0.513	-0.225	-0.449	-1.187	0.326
15	-0.82	1.56	-1.977	1.55	-0.03	-2.116	0.15	-0.94	-1.571	-1.054	1.867	-0.899	-0.326
16	0.72	1.05	-0.644	2.18	-1.66	0.292	-0.56	-1.79	-1.571	-1.002	2.165	2.824	1.553
17	0.98	-1.04	1.867	-0.97	1.74	0.854	-0.42	1.23	2.448	0.138	-0.626	-0.112	-1.423
18	-0.26	1.14	0.831	-0.58	-0.15	1.567	0.15	0.17	-0.232	0.553	-0.775	0.560	0.352

PCA – Francis and Wills, continued (3)

This process of `data adjustment', weighting, normalizing, is critical to the outcome, in particular to whether we understand the significance of the results, and whether the error/covariance matrix really does the job we expect of it. There are many ways of doing this: we can take logs of the data, we can weight by factors other than standard deviations based on prior knowledge, etc.

What about the present weighting system? The figure plots the run of the 18 points, one from each QSO, for each of the 13 data, i.e. 13 mini-plots.

Looks OK! All the points are there; there's only one deviation **>3σ** in 234 points, not far off expectation for Gaussian stats, and the distributions look reasonable. We may be confident that the results will be understandable.



PCA – Francis and Wills continued (4)

(2) Construct the covariance or error matrix. This is a 13 x 13 symmetric matrix:

$$\mathbf{C} = \begin{vmatrix} < x_1^2 > & < x_1 x_2 > & \\ < x_1 x_2 > & < x_2^2 > & \\ & & \ddots \end{vmatrix}$$

```
 1.0000 -0.1530  0.1135 -0.0414 -0.1420  0.0627 -0.7656 -0.4387  0.0620 -0.6803 -0.0962  0.1764 -0.3794
-0.1530  1.0000 -0.6775  0.6117 -0.5009 -0.4853 -0.0647 -0.4348 -0.6603 -0.3460  0.6255  0.4159  0.6514
 0.1135 -0.6775  1.0000 -0.7000  0.5029  0.7748  0.2860  0.6694  0.7656  0.5151 -0.7008 -0.2118 -0.4287
-0.0414  0.6117 -0.7000  1.0000 -0.7829 -0.5204 -0.1602 -0.5852 -0.6826 -0.3701  0.9295  0.5139  0.5182
-0.1420 -0.5009  0.5029 -0.7829  1.0000  0.1549  0.3013  0.6476  0.6979  0.3944 -0.6505 -0.5894 -0.4519
 0.0627 -0.4853  0.7748 -0.5204  0.1549  1.0000  0.1207  0.2595  0.2923  0.3465 -0.4627  0.1881 -0.1898
-0.7656 -0.0647  0.2860 -0.1602  0.3013  0.1207  1.0000  0.7653  0.2489  0.8897 -0.1574 -0.1864  0.1630
-0.4387 -0.4348  0.6694 -0.5852  0.6476  0.2595  0.7653  1.0000  0.7925  0.8609 -0.6196 -0.4830 -0.2307
 0.0620 -0.6603  0.7656 -0.6826  0.6979  0.2923  0.2489  0.7925  1.0000  0.5117 -0.7328 -0.4608 -0.5046
-0.6803 -0.3460  0.5151 -0.3701  0.3944  0.3465  0.8897  0.8609  0.5117  1.0000 -0.3930 -0.2054  0.0287
-0.0962  0.6255 -0.7008  0.9295 -0.6505 -0.4627 -0.1574 -0.6196 -0.7328 -0.3930  1.0000  0.5622  0.5626
 0.1764  0.4159 -0.2118  0.5139 -0.5894  0.1881 -0.1864 -0.4830 -0.4608 -0.2054  0.5622  1.0000  0.6198
-0.3794  0.6514 -0.4287  0.5182 -0.4519 -0.1898  0.1630 -0.2307 -0.5046  0.0287  0.5626  0.6198  1.0000
```

Here it is – nb diagonal elements are all 1.0, as they must be.

PCA - Francis and Wills continued (5)

(3) Solve 13 - 13th order equations in 13 unknowns to get the eigenvalues of this matrix! This is 150-year old technology; for symmetric matrices, {\it Jacobi rotations} do the trick, each plane rotation or transformation designed to get rid of one off-diagonal matrix element. ``absolutely foolproof for all real symmetric matrices" – NumRec. The NumRec routine **jacobi**, when supplied with the covariance matrix, returns the **eigenvalues**, the **array of eigenvectors**, and the **number of rotations required**, which turns out to be about 3x13^2 = 500. The cpu time required is insignificant. My results:

Rotations: 459

Eigenvalues: 6.451 2.820 1.589 0.624 0.565 0.343 0.261 0.172 0.122 0.023 0.019 0.010 0.002

Eigenvectors:

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
-0.055	0.534	0.126	-0.018	0.408	0.193	-0.128	-0.322	-0.418	0.075	0.250	0.280	-0.226
-0.294	-0.197	-0.082	0.490	0.151	0.511	-0.456	0.146	0.282	0.090	0.147	-0.018	-0.071
0.330	0.077	0.357	-0.081	0.149	0.133	-0.213	0.422	-0.296	0.111	0.150	-0.480	0.366
-0.342	-0.139	-0.006	-0.484	0.222	-0.001	-0.074	0.184	0.013	0.656	-0.297	0.146	0.015
0.310	0.016	-0.252	0.396	0.093	-0.619	-0.389	-0.017	-0.064	0.352	-0.019	0.105	0.018
0.198	0.075	0.624	0.044	-0.399	0.007	-0.183	0.234	0.132	0.064	-0.129	0.394	-0.351
0.177	-0.503	0.005	-0.138	-0.026	0.127	-0.312	-0.352	-0.396	-0.101	-0.283	-0.242	-0.391
0.336	-0.262	-0.051	-0.046	0.302	0.196	-0.049	0.046	-0.041	-0.276	-0.214	0.601	0.441
0.342	0.064	-0.031	-0.067	0.581	-0.034	0.180	0.215	0.411	-0.112	-0.128	-0.171	-0.479
0.261	-0.414	0.124	-0.177	0.012	0.016	0.146	-0.257	0.203	0.294	0.698	0.101	-0.016
-0.342	-0.149	0.015	-0.310	0.125	-0.399	-0.362	0.301	-0.056	-0.469	0.348	0.106	-0.113
-0.231	-0.053	0.571	0.112	0.288	-0.258	-0.088	-0.465	0.291	-0.083	-0.190	-0.159	0.279
-0.223	-0.351	0.225	0.441	0.207	-0.136	0.499	0.251	-0.424	0.054	0.019	0.087	-0.125

PCA – Francis and Wills (6)

The Francis-Wills results table:

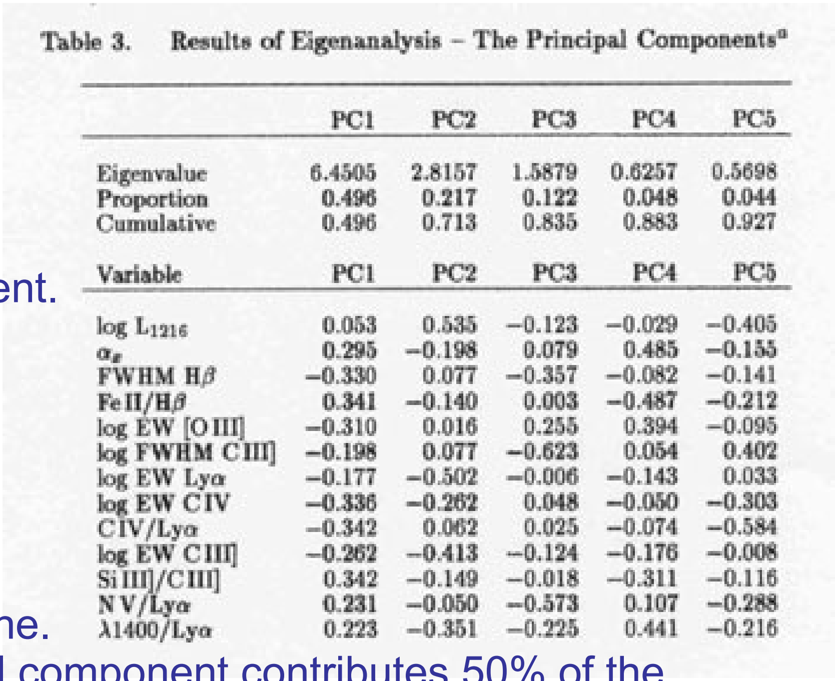


Table 3. Results of Eigenanalysis – The Principal Components[a]

	PC1	PC2	PC3	PC4	PC5
Eigenvalue	6.4505	2.8157	1.5879	0.6257	0.5698
Proportion	0.496	0.217	0.122	0.048	0.044
Cumulative	0.496	0.713	0.835	0.883	0.927
Variable	PC1	PC2	PC3	PC4	PC5
log L_{1216}	0.053	0.535	−0.123	−0.029	−0.405
α_x	0.295	−0.198	0.079	0.485	−0.155
FWHM Hβ	−0.330	0.077	−0.357	−0.082	−0.141
Fe II/Hβ	0.341	−0.140	0.003	−0.487	−0.212
log EW [O III]	−0.310	0.016	0.255	0.394	−0.095
log FWHM C III]	−0.198	0.077	−0.623	0.054	0.402
log EW Lyα	−0.177	−0.502	−0.006	−0.143	0.033
log EW C IV	−0.336	−0.262	0.048	−0.050	−0.303
C IV/Lyα	−0.342	0.062	0.025	−0.074	−0.584
log EW C III]	−0.262	−0.413	−0.124	−0.176	−0.008
Si III]/C III]	0.342	−0.149	−0.018	−0.311	−0.116
N V/Lyα	0.231	−0.050	−0.573	0.107	−0.288
λ1400/Lyα	0.223	−0.351	−0.225	0.441	−0.216

Columns (2)-(6) show the first 5 out of a total of 13 principal components. The first row gives **the variances (eigenvalues)** of the data along the direction of the corresponding principal component.

The sums of all the variances add up to the sums of the variances of the input variables, in this case, 13.

By convention, the principal components are given in order of their contribution to the total variance. This is given as **'Proportion'** in the second line, and the **'Cumulative proportion'** on the third line.

Thus, among the parameters used, the first principal component contributes 50% of the spectrum-to-spectrum variance, the second **22%,** the third, **12%.**The first two principal components together contribute 71% of the variance, the first 3, 84%, the first 4, nearly 90%.

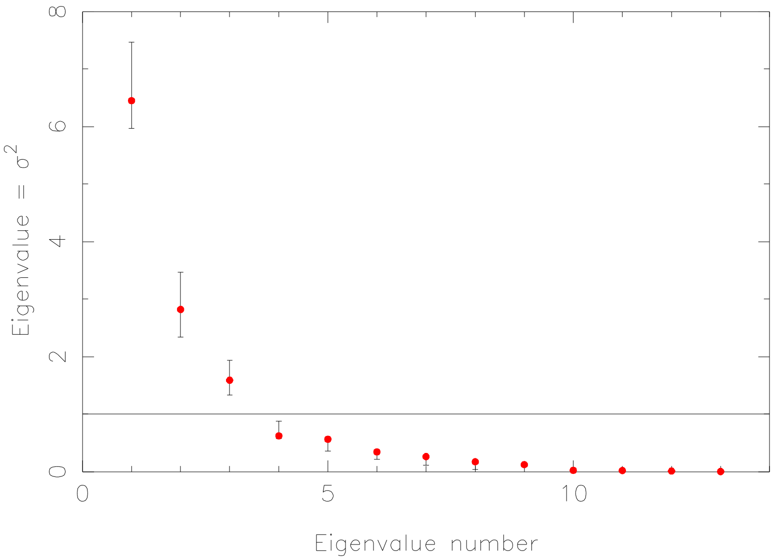
The **columns of numbers for each principal component represent the weights assigned to each input variable. Thus PC1 = 0.053x_1 + 0.295x_2 - 0.330x_3…, where x_1, x_2, x_3 are the values of the normalized variables** corresponding to log L_{1216}, α_x, FWHM Hβ, etc. By convention these weights are chosen so that the sum of their squares = 1. This arbitrariy fixes the scale of the new variable. The sign of the new variable is arbitrary as well.

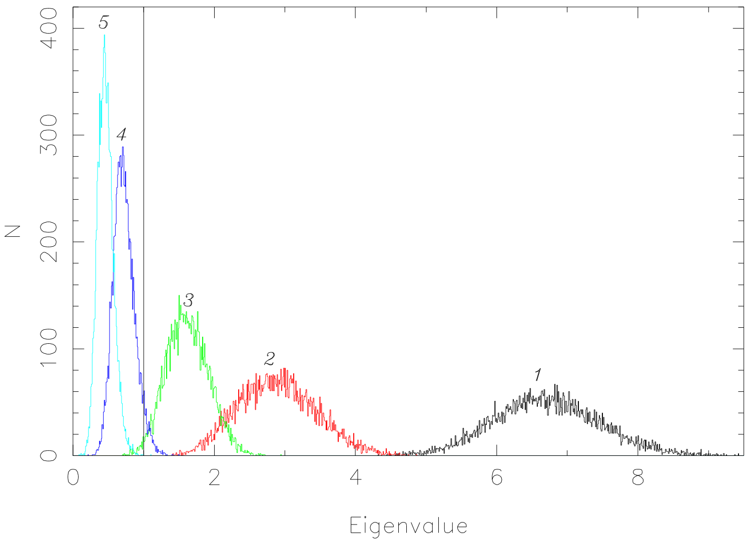
PCA – Francis and Wills continued (7)

(4) Check it out. One simple check of this step: the eigenvalues must add up to the trace of the array, the sum of the diagonal elements, =13 here.

For eigenvalues to be significant, they must be greater than 1.0. How to test this?

(a) Remove any variable, and recompute, to assess how much it contributes to any particular eigenvalue.





(b) Find the **errors (uncertainties)** on the eigenvalues. **Bootstrap** is perfect for this – see the right figure, 10000 trials. The widths of the distributions are reflected in the error bars in the left figure.

Eigenvalues 1, 2 and maybe 3 are significant. The rest – garbage.

PCA – Francis and Wills concluded

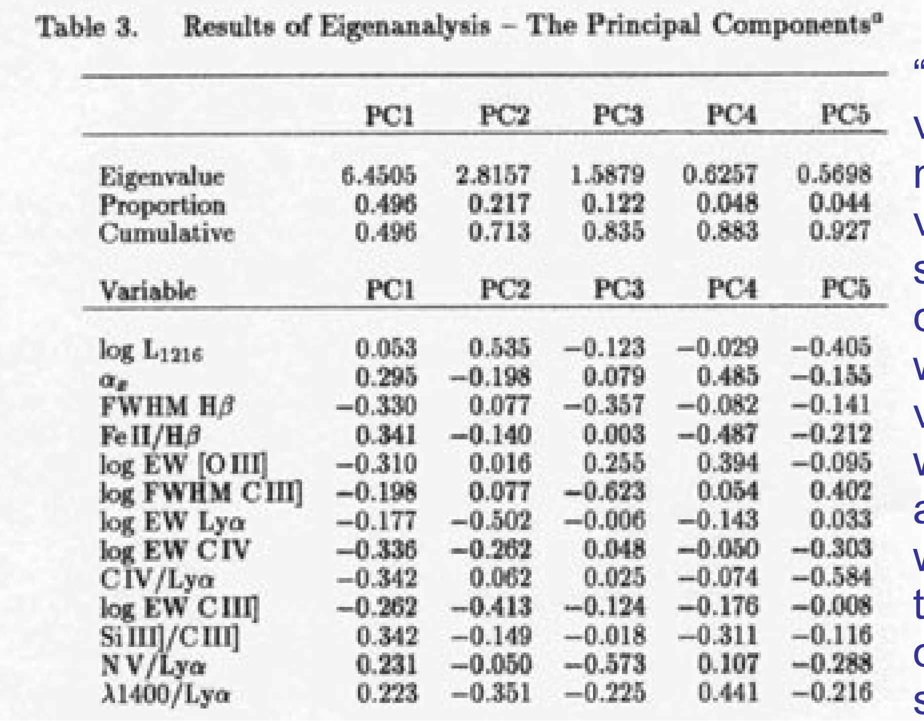


Table 3. Results of Eigenanalysis – The Principal Components[a]

	PC1	PC2	PC3	PC4	PC5
Eigenvalue	6.4505	2.8157	1.5879	0.6257	0.5698
Proportion	0.496	0.217	0.122	0.048	0.044
Cumulative	0.496	0.713	0.835	0.883	0.927
Variable	PC1	PC2	PC3	PC4	PC5
log L_{1216}	0.053	0.535	−0.123	−0.029	−0.405
α_g	0.295	−0.198	0.079	0.485	−0.155
FWHM Hβ	−0.330	0.077	−0.357	−0.082	−0.141
Fe II/Hβ	0.341	−0.140	0.003	−0.487	−0.212
log EW [O III]	−0.310	0.016	0.255	0.394	−0.095
log FWHM C III]	−0.198	0.077	−0.623	0.054	0.402
log EW Lyα	−0.177	−0.502	−0.006	−0.143	0.033
log EW C IV	−0.336	−0.262	0.048	−0.050	−0.303
C IV/Lyα	−0.342	0.062	0.025	−0.074	−0.584
log EW C III]	−0.262	−0.413	−0.124	−0.176	−0.008
Si III]/C III]	0.342	−0.149	−0.018	−0.311	−0.116
N V/Lyα	0.231	−0.050	−0.573	0.107	−0.288
λ1400/Lyα	0.223	−0.351	−0.225	0.441	−0.216

"The **first principal component** is elongated with variance about **6.5 times** that of any individual measurements, and accounts for **about half** the total variance. This is therefore likely to be highly significant. If all measured, normalized quantities contributed equally to PC1, they would all have weight **0.277 (1/√13 for 13 variables),** but each variable contributes more or less than this. One way to test the significance of the contribution of any one measured variable, is to perform the PCA without that variable, then check the significance of the correlation between that variable and the scores of the new principal component. This procedure shows that all measured variables except L_{1216}, logFWHM CIII], and log EW Lyα, correlate with PC1, but correlations involving NV/Ly α and λ1400/Ly α are not very strong. **PC2**, accounting for **22% of the variance** in this dataset, appears to link the EW Ly α, EW CIV, and EW CIII] with L_{1216}, so EW CIV and EW CIII] appear to contribute to both PC1 and PC2, but EW Ly α contributes predominantly to PC2. Is PC2 a significant component? A similar correlation test shows that individually the EWs do anti-correlate with L_{1216}, but this result depends on the lowest EWs for the highest luminosity QSO PG1226+023 and the highest EWs for the low luminosity QSO PG1202+281. However L_{1216} correlates significantly (Pearson's ordinary correlation coefficient = -0.77) with PC2 formed when L_{1216} is excluded. Thus **there is a significant overall correlation between EW and L_{1216}**, although a larger sample is clearly needed to investigate the individual EW correlations. Another test may be to check correlations between observed measurements for those measurements that contribute to only one significant principal component - for example, CIV/Lyα vs. FeII/Hβ..."

…And Principal Component Analysis concluded

گ PCA represents the ultimate powerful way of searching for correlations in a stack of data. It is so simple to perform and **no special numerical skills are required.** There are a few buts:

گ The distribution of points in the multi-dimension space must be essentially unimodal. Consider the two-blob case…

گ Thus the data need to be of **quadratic** form; they need to cluster continuously around the PC, but they need not do this necessarily in a Gaussian manner. In fact the method is **immensely forgiving in terms of distribution,** provided the **unimodal** condition is met.

گ Check at the start what the form of the data scatter will be. Look at plots! It may well be worth considering other methods of central location for zero-pointing, such as the median; and normalizing other than via an rms std dev.

گ PCA software is available in widely used software packages - SPSS, SAS, Minitab
It is also available at Francis's web site **http://msowww.anu.edu.au/~pfrancis**
If using this, please observe the acknowledgement requested by Paul.

Challenge 4

Distribute points at random according to a bivariate Gaussian distribution

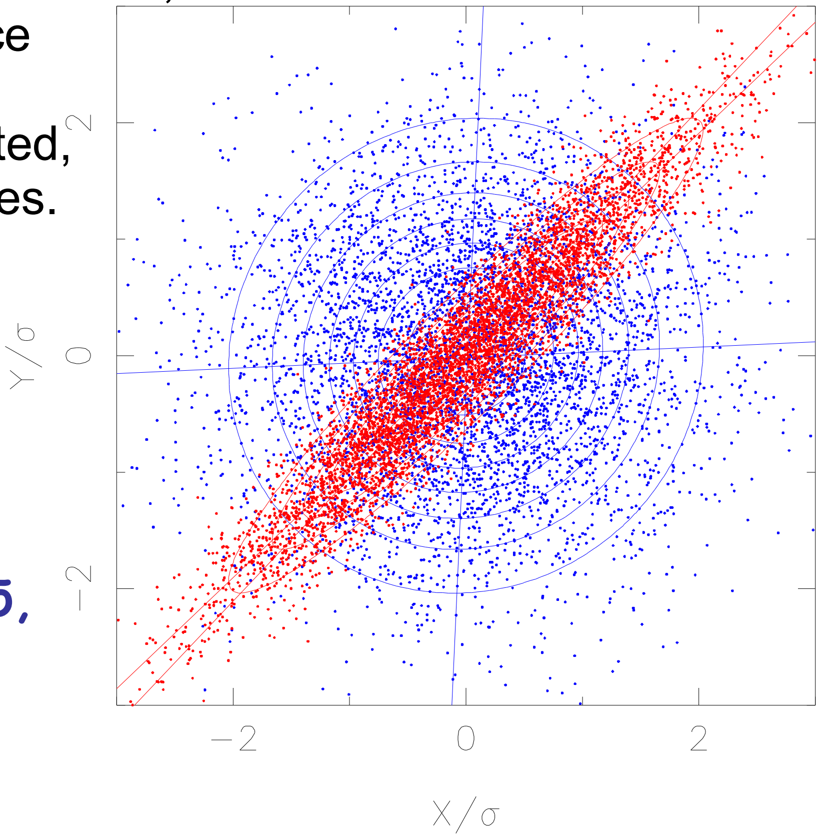
Back to the random numbers game

To generate numbers obeying a bivariate (or even multivariate) Gaussian, with given $\boldsymbol{\sigma}_i$ and $\boldsymbol{\rho}_i$? Following the discussion of error matrices, it's quite simple to formulate:

1. Set up the error matrix and determine the covariance matrix from it. (For the bivariate case, the error matrix is $\mathbf{e}_{1,1} = \boldsymbol{\sigma}_x{}^2$, $\mathbf{e}_{2,1} = \mathbf{e}_{1,2} = \mathbf{cov[x,y]} = \boldsymbol{\rho}\,\boldsymbol{\sigma}_x\,\boldsymbol{\sigma}_y$, $\mathbf{e}_{2,2} = \boldsymbol{\sigma}_y{}^2$, as we have seen.)
2. Find the **eigenvalues** and **eigenvectors** of the covariance matrix.
3. Combine the eigenvectors, the column vectors, into the transformation matrix **T**, the matrix that diagonalizes the covariance matrix.
4. Draw **(X',Y')** Gaussian pairs, uncorrelated, with variances equal to the two eigenvalues. Compute the **(X,Y)** pairs according to

$$\begin{pmatrix} X \\ Y \end{pmatrix} = [T] \begin{pmatrix} X' \\ Y' \end{pmatrix}$$

The points in the figure were obtained in this manner, with $\boldsymbol{\rho}$ = **0.05** and **0.95**, **5000** points each.



Challenge 5

Assign luminosities at random to your 1000000 points in your random universe, according to a power-law luminosity function with a given slope, say -2.

Assign a second set of random luminosities to your galaxies, according to a different power-law luminosity function, with a slope of say -3.

Now each galaxy of yours shines on the earth (at the centre) with a flux of La_i/R_i^2 and Lb_i/R_i^2. Make a 'survey' of the whole sky; because your telescope has a fixed sensitivity, you will only be able to see your pet objects if their flux is above the survey flux llimit. (Choose your flux units and your survey limits with carfe, so that R-1, the edge of the Universe, does not come into play!) Make a second survey for the other set of luminosities.

Your luminosities were chosen independently and at random, right?

So for your two samples plot flux1 against flux 2, and **consider carefully what you find**......

End Bertinoro 3