ASTR509 - 7

Partial Correlation and PCA

ASTR509 © Jasper Wall Fall term 2013

When last we spoke ....

We considered the traps of apparent correlations: **bad data or data grouping;** misleading formal significance; erroneous assumptions; dependence on third variables; and the **error in assuming that a correlation means a causal connection.**

We moved on to formal testing, noting that most formal testing, Bayesian or classical, is based on the **Bivariate Gaussian** model, **NON-PARAMETRIC** ranking tests being the exception.

We looked at Bayesian testing, computing the **probability distribution for (ρ/data),** From this we can assess what we really want to know, e.g. what is the probability that **ρ** is zero? Or that data set A is more highly correlated than B?

We discussed the formalism for the classical tests, all classical tests: set up the **null hypothesis**, calculate a **statistic**, and from the **sampling distribution** for the statistic under the null hypothesis, find at what **level of significance** we can reject it.

For classical correlation testing, we looked the **Fisher test (on r),** and the **Spearman Rank test**. The latter (non-parametric) gets us safely away from the Bivariate Gaussian model .

We noted the DIY method of random **permutations** of the **x**'s and **y**'s to get a sampling distribution for an invented statistic.

We considered 'what to do next' issues; and **Anscombe's Quartet** made an appearance.

Bootstrap and Jackknife testing

In some data-modelling procedures, confidence intervals for the parameters fall out of the procedure. But are these realistic? And what about the procedures where they do not? Computer-power and Monte Carlo (just coming) can provide the answers.

Bootstrap (Efron 1979) is blatant ‘quick-and-dirty Monte Carlo‘. Our sample is of **N** data-points, each consisting of one or more numbers (e.g. single measurements, or [**x,y**] pairs), and we want the error on a parameter estimated from (e.g. the mean). We calculate the parameter using a modelling process (which we’ll come to…..). We then ‘**bootstrap'** to find its uncertainty:

1. **Label (number)** each data-point.

2. Draw at random from the sample **another** sample of **N,** with replacement (simply done by computer with a random-number generator using the point-labels), i.e. your new sample may contain the 10th point more than once.

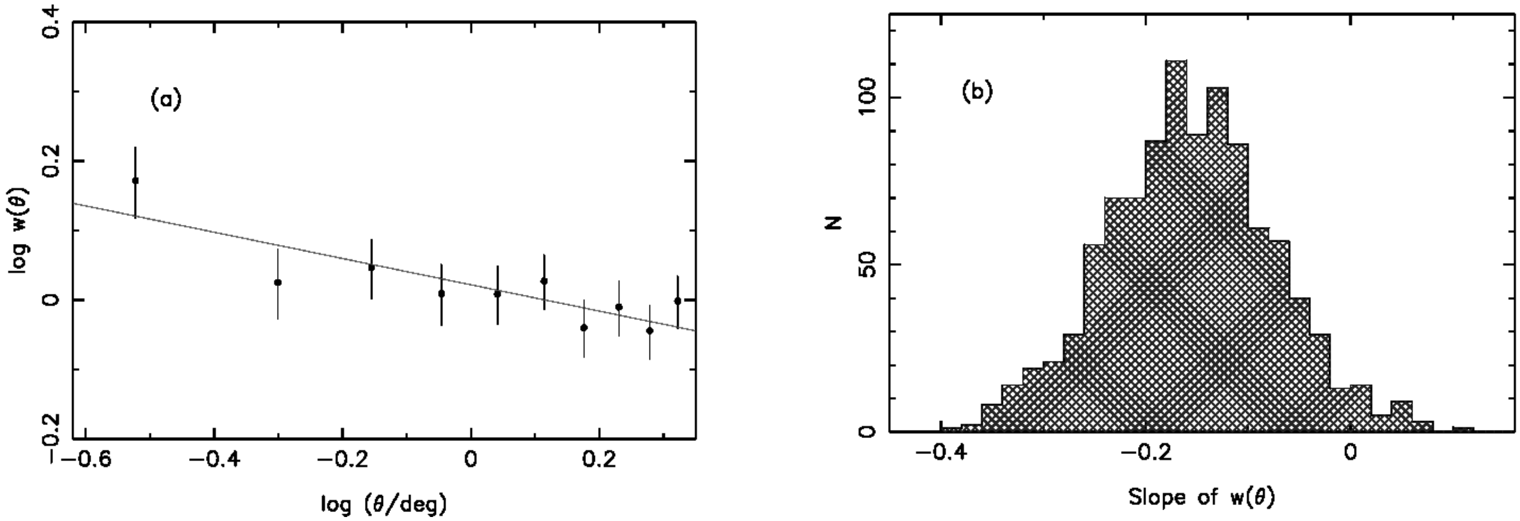
3. **Recalculate** the parameter.

4. **Repeat** this process as many times as possible.

That's it! Provided that the data points are independent (in distribution and in order), the distribution of these recalculated parameters maps the uncertainty in the estimate from the original sample.

Bootstrap - Example

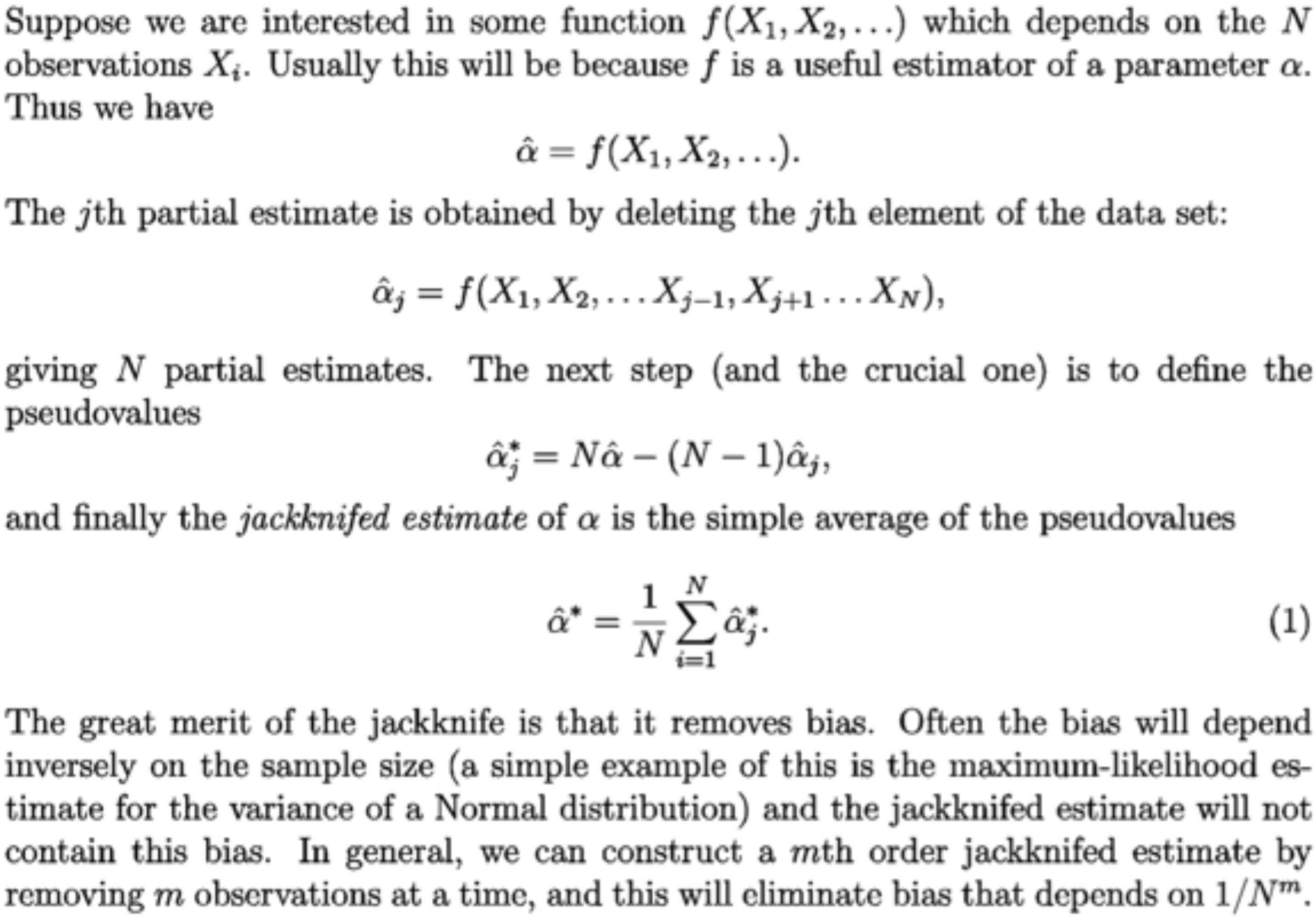
Bhavsar (1990) showed that the bootstrap is well suited to estimating uncertainty in measuring the slope of the angular two-point correlation function for galaxies. This function **w(θ)** measures the excess surface density over that expected from a random distribution at angular scales θ. The data-points are the **(x,y)** pairs of galaxy coordinates.



Left: The two-point correlation function for 2812 radio sources with extended radio structure, from the NRAO 1.4-GHz survey of the northern sky. A least-squares fit gives a slope of -0.19. Right: The distribution of slopes obtained in bootstrapping the sample with 1000 trials; the slope is less than zero (i.e. signal is present) for 96.8 per cent of the trials.

Jackknife

Similar to the bootstrap, first described by Tukey in 1958.



Suppose we are interested in some function $f(X_1, X_2, \ldots)$ which depends on the N observations X_i. Usually this will be because f is a useful estimator of a parameter α. Thus we have

$$\hat{\alpha} = f(X_1, X_2, \ldots).$$

The jth partial estimate is obtained by deleting the jth element of the data set:

$$\hat{\alpha}_j = f(X_1, X_2, \ldots X_{j-1}, X_{j+1} \ldots X_N),$$

giving N partial estimates. The next step (and the crucial one) is to define the pseudovalues

$$\hat{\alpha}_j^* = N\hat{\alpha} - (N-1)\hat{\alpha}_j,$$

and finally the *jackknifed estimate* of α is the simple average of the pseudovalues

$$\hat{\alpha}^* = \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_j^*. \tag{1}$$

The great merit of the jackknife is that it removes bias. Often the bias will depend inversely on the sample size (a simple example of this is the maximum-likelihood estimate for the variance of a Normal distribution) and the jackknifed estimate will not contain this bias. In general, we can construct a mth order jackknifed estimate by removing m observations at a time, and this will eliminate bias that depends on $1/N^m$.

If the bootstrap can be used, it is computationally much cheaper.
If both can be used, so much the better.

Jackknife - Example

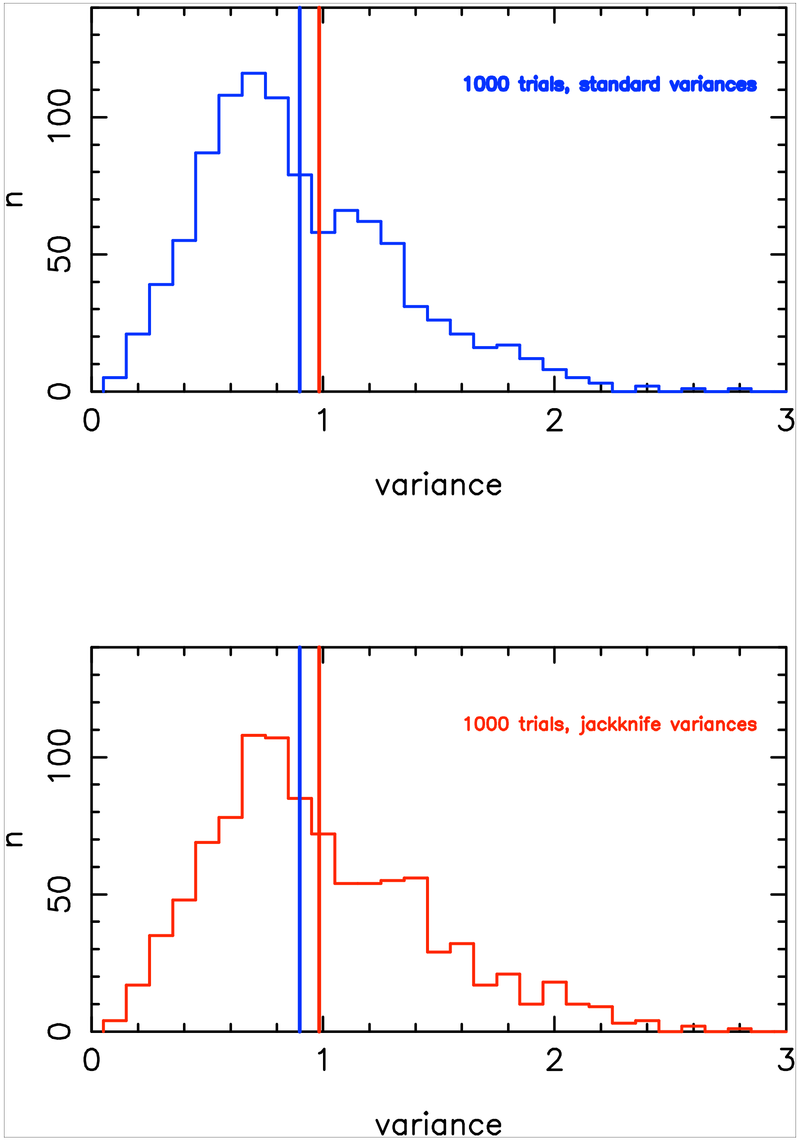
Draw 10 samples from a Gaussian distribution of μ = 0 and σ = 1.0. Calculate the variance $\Sigma(x_i - \langle x \rangle)^2/10$. Do this for 1000 trials, and the result is the blue histogram of the upper diagram.

The mean variance (blue vertical) is 0.8999, less than 1.0; the infamous 1/N vs 1/(N-1) issue rears its ugly head.

Calculate the variances using the jackknife method for each of 1000 similar trials - the red histogram of the lower diagram. The peak has shifted to larger values; in fact the mean is 0.9834, very close to 1.0.

The jackknife variance is **larger** than the standard variance.

Bias has been removed.



Partial correlation

Partial correlation between two variables is taken into account by nullifying the effects of the third (or fourth, or more) variable upon the variables being considered. A science in itself; there are books …

Parametric form - consider a sample of **N** objects for which **3** parameters $\mathbf{X_1}$, $\mathbf{X_2}$, and $\mathbf{X_3}$ have been measured. The **first-order partial correlation coefficient** between variables $\mathbf{X_1}$ and $\mathbf{X_2}$ is

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

where the **r** are the product moment coefficients.
If there are **4** variables, then the **second-order partial correlation coefficient** is

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}}$$

where the correlation is being examined between $\mathbf{X_1}$ and $\mathbf{X_2}$ with $\mathbf{X_3}$ and $\mathbf{X_4}$ held constant. Manipulate the subscripts for the rest…. The **standard error** of the partial correlation coefficients is

$$\sigma_{r_{12.34\ldots m}} = \frac{1 - r_{12.34\ldots m}^2}{\sqrt{N - m}}$$

where **m** is the number of variables involved. The significance then comes from the **t test,** as we've done before.

Partial Correlation - Example

Sample of **N=10** young males aged 12 to 19.

Correlation between **height (X_1)**and **weight (X_2)** will be high because **older = taller** on average, and **older = heavier** on average. So we recognize the presence of a **third variable, age (X_3)**.

But with **age (X_3)** held constant, the correlation would still be significantly positive because at all ages, **taller = heavier** (on average), common sense would tell us?

We want to know if there is **a real correlation between height (X_1) and weight (X_2).**

Correlation coefficient between height and weight **r_{12} = 0.78**
between height and age **r_{13} = 0.52**
and between weight and age **r_{23} = 0.54**

The first-order partial coefficient of correlation is thus **$r_{12,3}$ = 0.69**; and **$\sigma_{12.3}$ = 0.198**

The correlation is **significant at the level of 0.2%** (2 chances in 1000 that it could arise by chance if NO true correlation were present.)

Back to the Bivariate Gaussian / Covariance Matrix

Let’s do this again.

We have a pair of experimental measurements **($X_{i,}Y_i$),** each with uncertainty **σ_x, σ_y** associated with it, Gaussian ‘experimental resolution’. For each:

$$P(x) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma_x^2}\right), \qquad P(y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{y^2}{\sigma_y^2}\right).$$

If **X_i** and **Y_i** are independent,

$$\begin{aligned} P(x,y) &= P(x)P(y) \\ &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)\right). \end{aligned}$$

Back to the Bivariate Gaussian / Covariance Matrix

This will be down on the origin value by √**e** when

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = 1.$$

Why choose this value? Because it corresponds to the **1D** case when **x = ± σ.**

We can rewrite the original probability equation as

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 1.$$

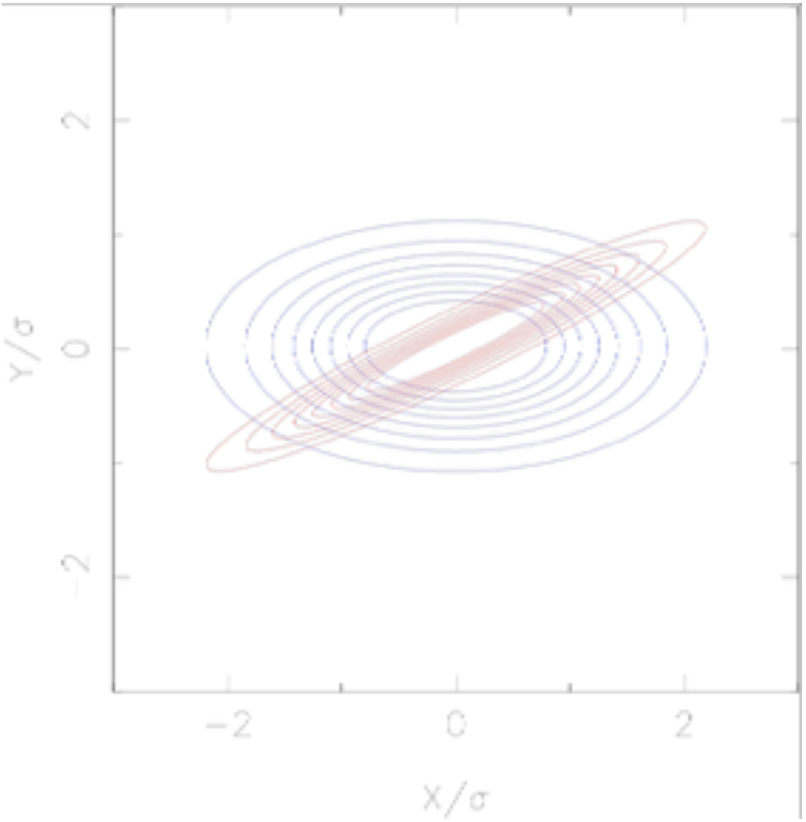
and we can invert the 2x2 matrix to get the matrix: $\begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$

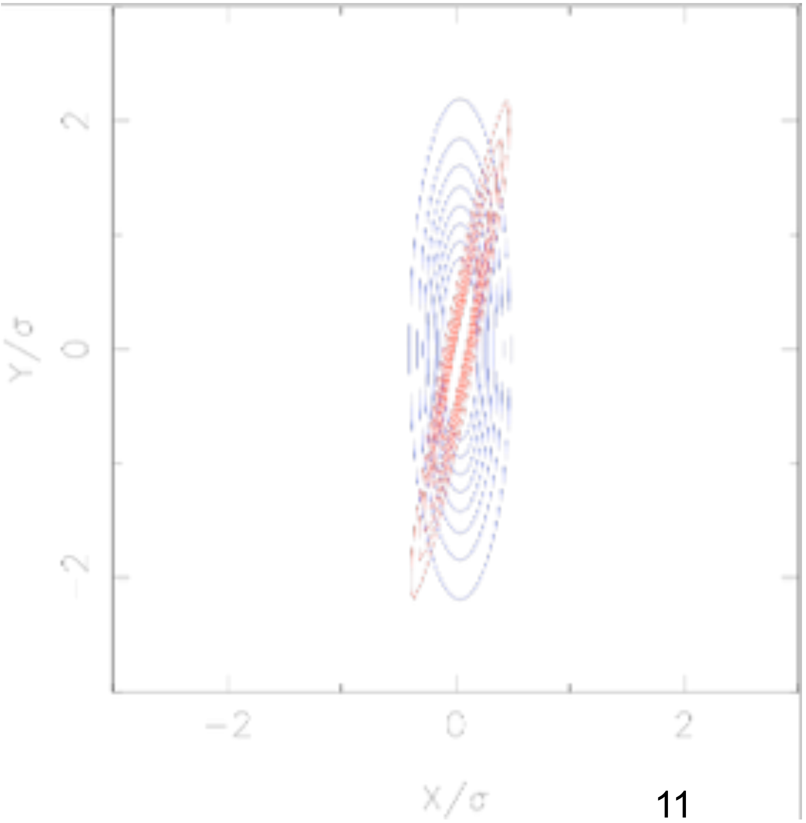
This is known as the error (covariance) matrix for **x** and **y.**
Off-diagonal zeros show that x and y are uncorrelated.

Error (covariance) matrices, continued

In general each **error-matrix element** of a set of variables $\mathbf{x_1,x_2,x_3\ldots x_n}$ is defined as the **expectation value** $\mathbf{<(x_i - <x_i>)(x_j - <x_j>)>}$

The error matrix is **symmetrical,** and **off-diagonal elements** are $\mathbf{cov(x_i,x_j)}$**.** In the previous example the diagonal terms are zero – errors of **x** and **y** are uncorrelated.





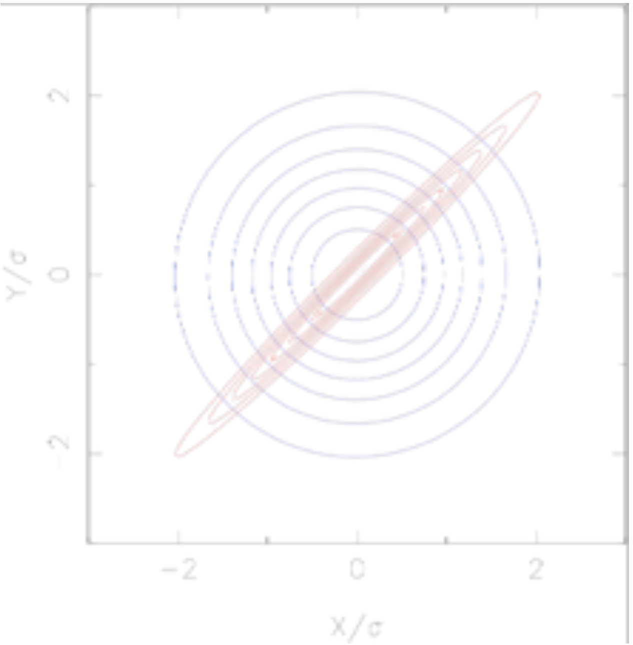
Left: σ_x = 1.0, σ_y = 0.5, ρ = 0 and 0.95. Right: σ_x = 0.2, σ_y = 1.0, ρ = 0 and 0.95

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**:

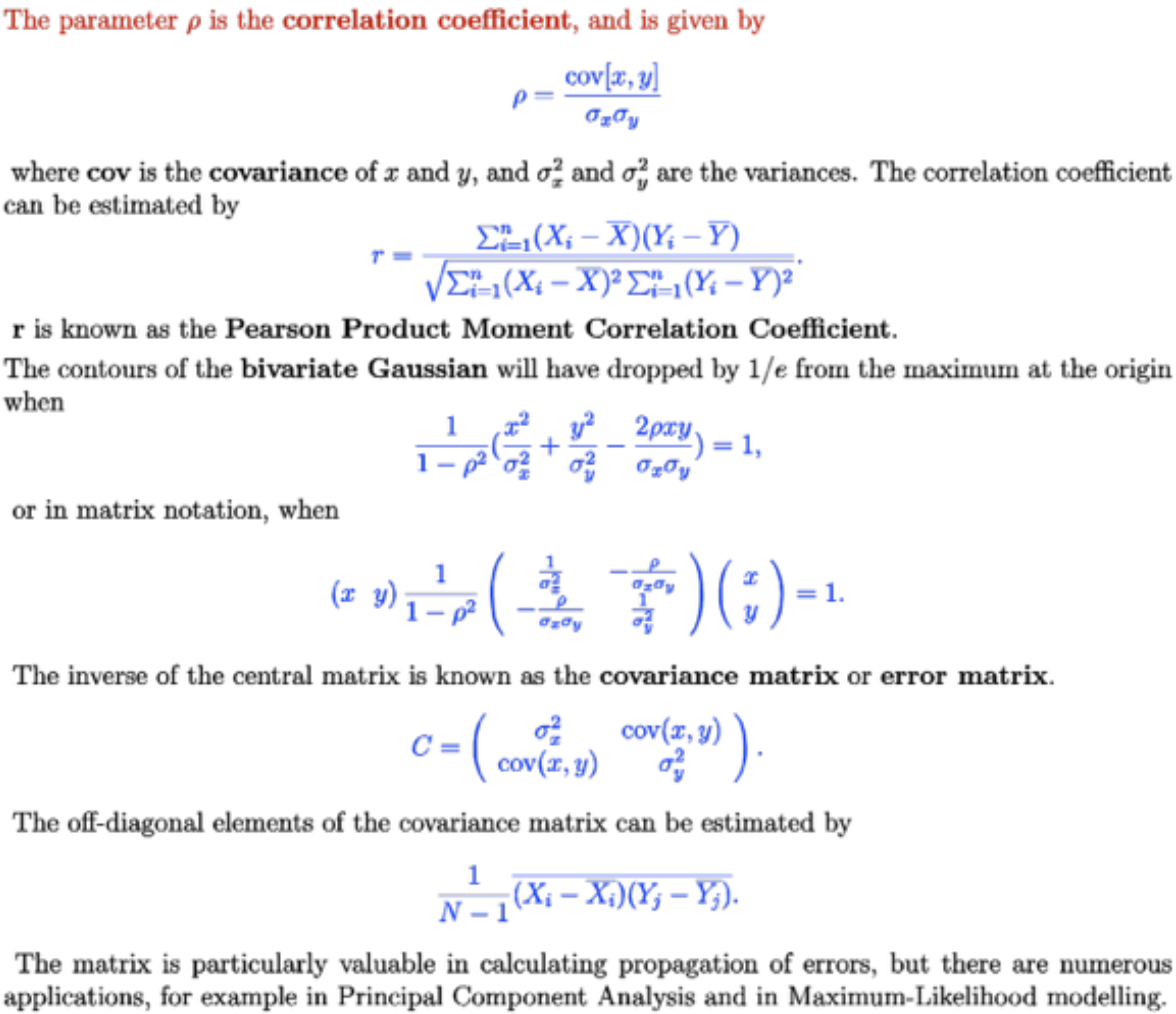
$$\begin{aligned}\mathrm{prob}(x, y \mid \sigma_x, \sigma_y, \rho) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\times \exp\left(\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y}\right)\right).\end{aligned}$$

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

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

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.

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 **x_n**, find a **new set of ξ_n variables** that are **orthogonal (independent),** each one a **linear** combination of the original variables:

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

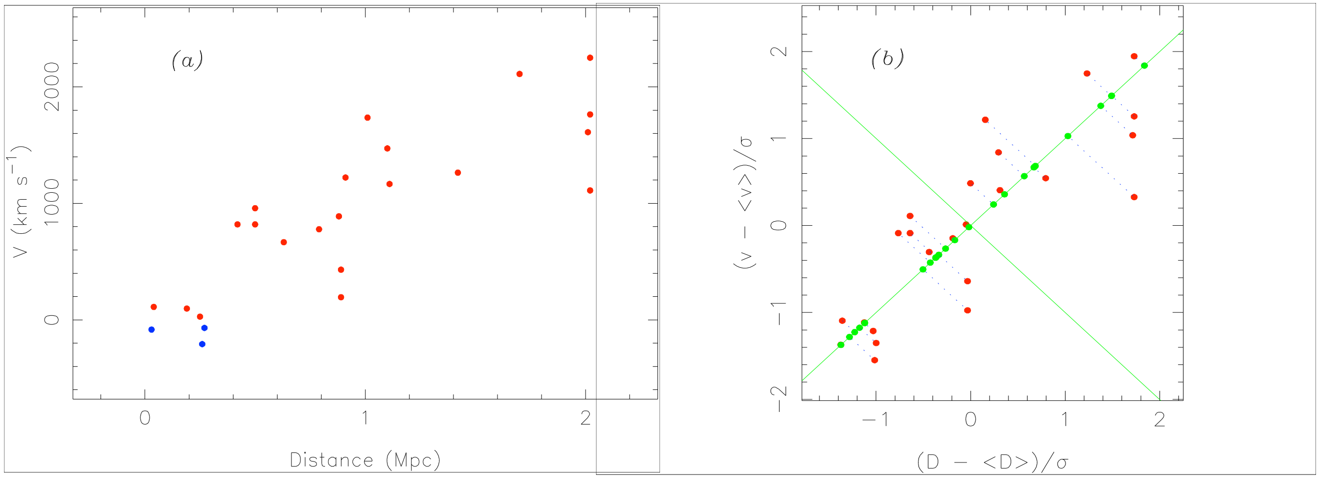
with values of **a_{ij}** such that the **smallest number** of new variables account for as much of the variance as possible. The **ξ_i** are the **principal components**.
If most of the variance involves **just a few** of the **n** new variables, we have found **a simplified description of the data.**

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

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

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



b. 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.

PCA - Example 1 continued

- **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**.

- 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.

	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

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

PCA - Example I continued

(2) Matrix approach.

(a) Construct the error matrix. i.e. for the two-variable case of the example, **a(1,1) = Σd'²**, **a(2,2) = Σv'²** , **a(1,2) = a(2,1) = Σv'd'.**

(b) 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.

This is simple in matrix notation! (1) 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 figure.)

Then (2) we 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 data are 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 are drawn from a **multivariate Gaussian** or in general when a simple quadratic form, using the covariance matrix, can describe the distribution of the data.

- But the clouds of points in most n-variate hyperspaces will NOT 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 – **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 $\rho = 0$.

- It may be apparent how to reject outliers or to transform coordinates to reduce the problem to a linear analysis. In large datasets such processes can reveal **unusual objects.**

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 - Example 2 continued (2)

(1) – subtract mean and divide by std dev:

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 - Example 2 continued (3)

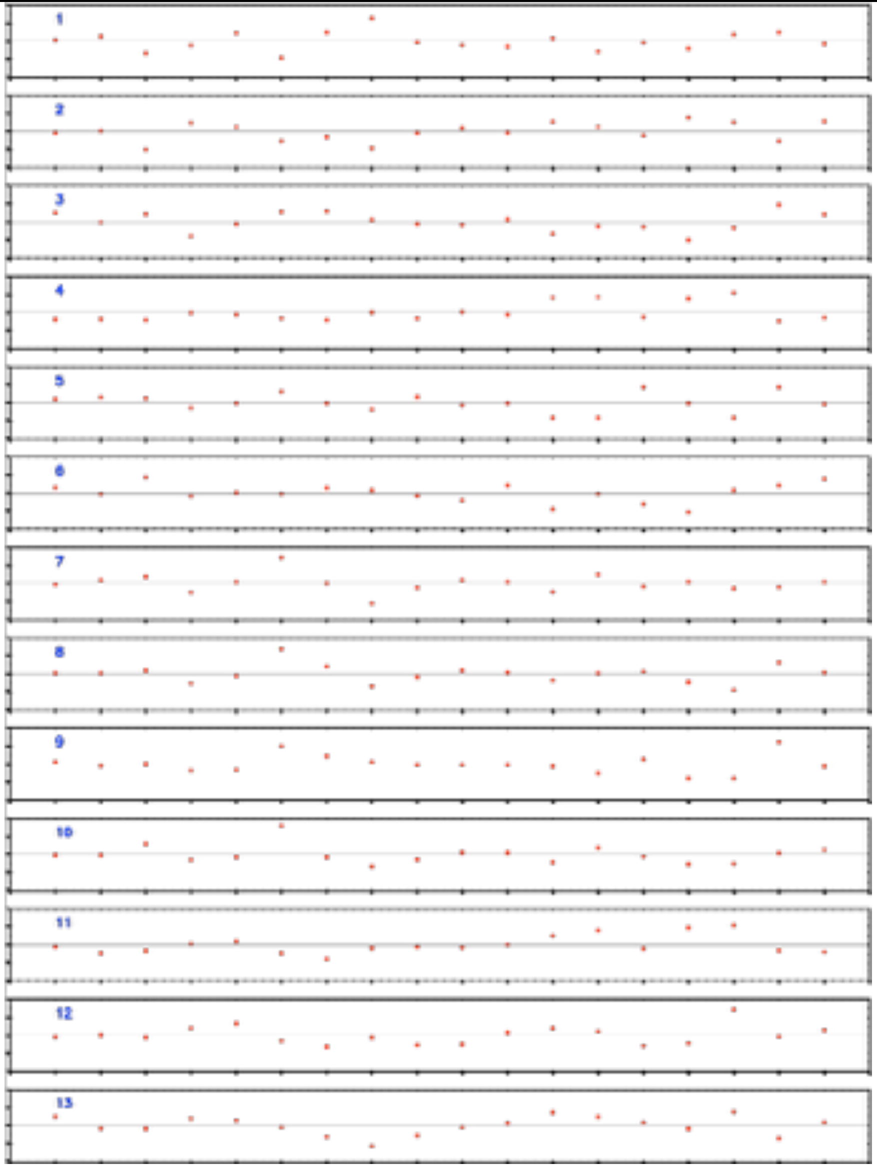
‘Data adjustment', weighting, normalizing, is critical to the outcome!

Do we understand the significance of the results? Has the error/covariance matrix done the job we expect of it?

NB **many ways of ‘adjusting’:** logs of the data, weighting by factors other than std devs, use of prior knowledge, etc.

For the present data: 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 points are there; only one deviation **>3σ** in 234 points, close to expected for Gaussian stats, and the distributions look reasonable. Results will be understandable.



PCA - Example 2 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 - Example 2 continued (5)

(3) Solve 13 - 13th order equations in 13 unknowns to get the eigenvalues of this matrix! Jacobi rotations: each plane rotation or transformation gets 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	-0.135

PCA - Example 2 continued (6)

The Francis-Wills results table:

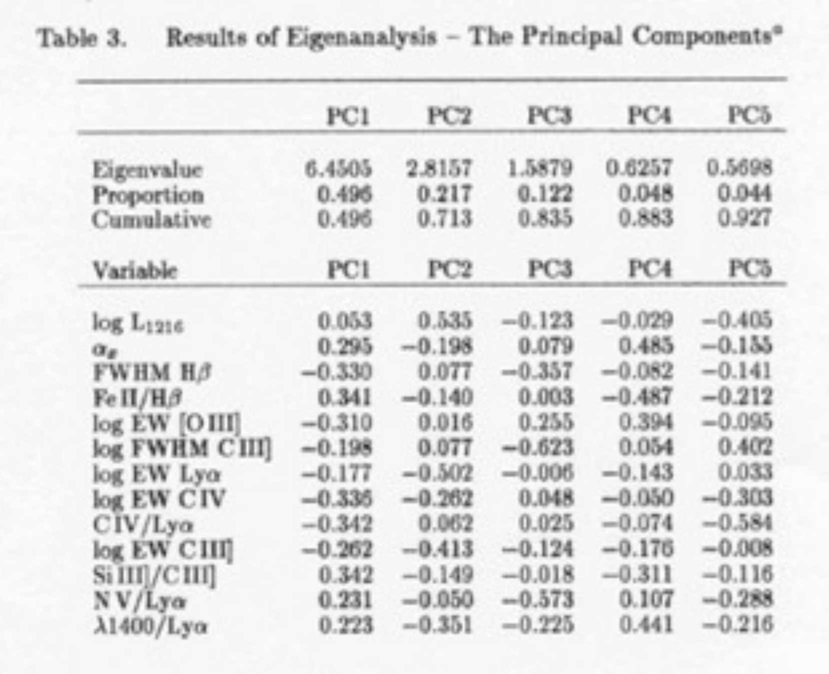


Table 3. Results of Eigenanalysis – The Principal Components[a]

	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

(1) Cols (2)-(6) show the first 5 principal components.

(2) First row gives **variances (eigenvalues)** of the data along the direction of the corresponding principal component.

(3) The sums of the variances add up to the sums of the variances of the input variables, here13.

(4) Principal components are given in order of contribution to the total variance: ’**Proportion’** on the 2nd line, **‘Cumulative proportion’** on 3rd.

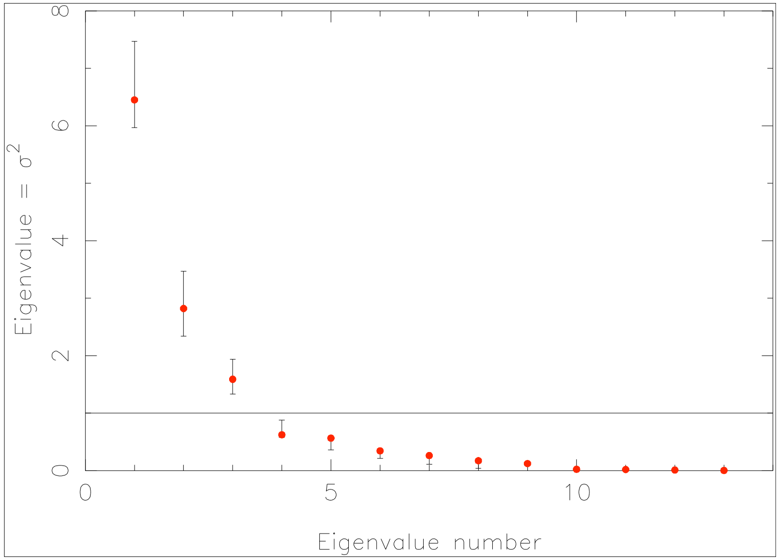
(5) The first principal component contributes 50% of the spectrum-to-spectrum variance, the 2nd **22%,** 3rd **12%.** The first 3 components contribute 84% of the variance.

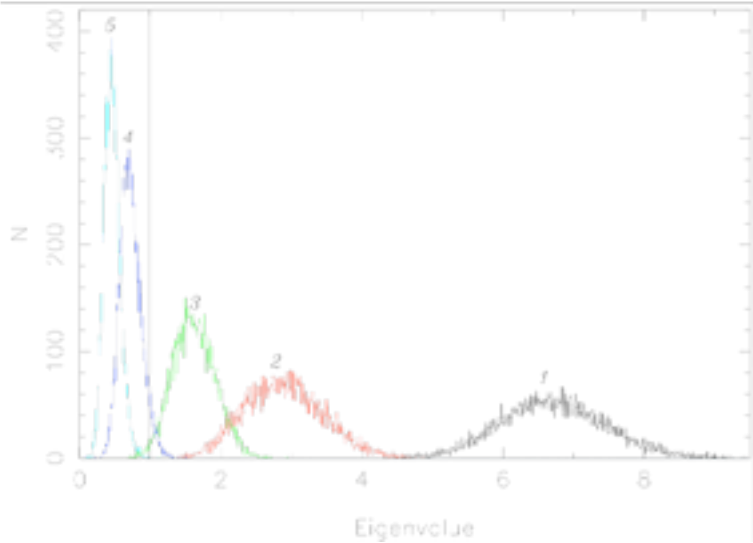
(6) The **cols of numbers for each principal component represent the weights assigned to each input variable. Thus PC1 = 0.053x_1 + 0.295x_2 - 0.330x_3…, where x_1, x_2, x_3 are the values of the normalized variables** corresponding to log L_{1216}, α_x, FWHM Hβ, etc. By convention these weights are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

PCA - Example 2 continued (7)

(4) Check it out. Simple test 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 - Example 2 continued (8)

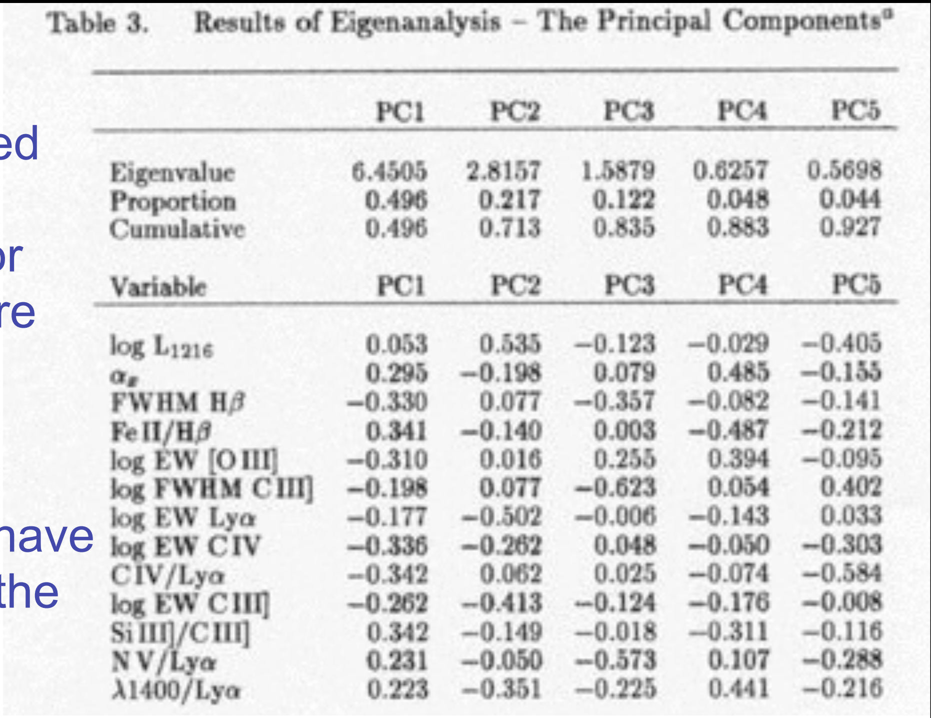


Table 3. Results of Eigenanalysis – The Principal Components[a]

	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

1. The **first principal component** is elongated with variance about 6.5 times that of any individual measurements, and accounts for ~50% of the total variance. This is therefore likely to be highly significant.

2. If all measured, normalized quantities contributed equally to **PC1**, they would all have weight 0.277 (**1√13 for 13 variables**), but the variables contribute more or less than this.

3. To test the significance of the contribution of any one measured variable, **perform the PCA without that variable, then check the significance of the correlation between that variable and the scores of the new principal component.**

4. This procedure shows that all measured variables except **L_{1216}, log FWHM CIII],** and **log EW Lyα**, correlate with **PC1**, but correlations involving **NV/Lyα** and **λ1400/Lyα** are not very strong.

PCA - Example 2 (9) F&W commentary con’t

5. **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 Lyα** contributes predominantly to **PC2.**

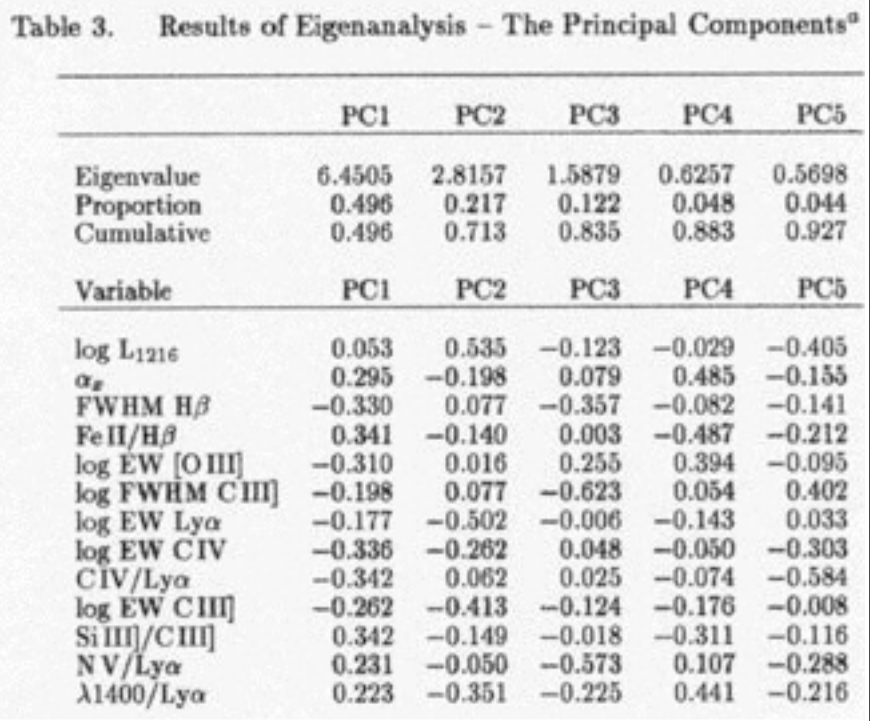


Table 3. Results of Eigenanalysis – The Principal Components[a]

	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

6. Is **PC2** a significant component? A similar correlation test shows that individually the **EW**s do anti-correlate with **L_{1216}**, but this result depends on the **lowest EW**s for the highest luminosity QSO PG1226+023 (3C273) and the **highest EW**s 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.

7. 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 is to check correlations between observed measurements for those measurements that contribute to only one significant principal component – for example, **CIV/Lyα vs. FeII/Hβ**...”

PCA - more stuff

PCA represents **the ultimate 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 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.

PCA - Example 3

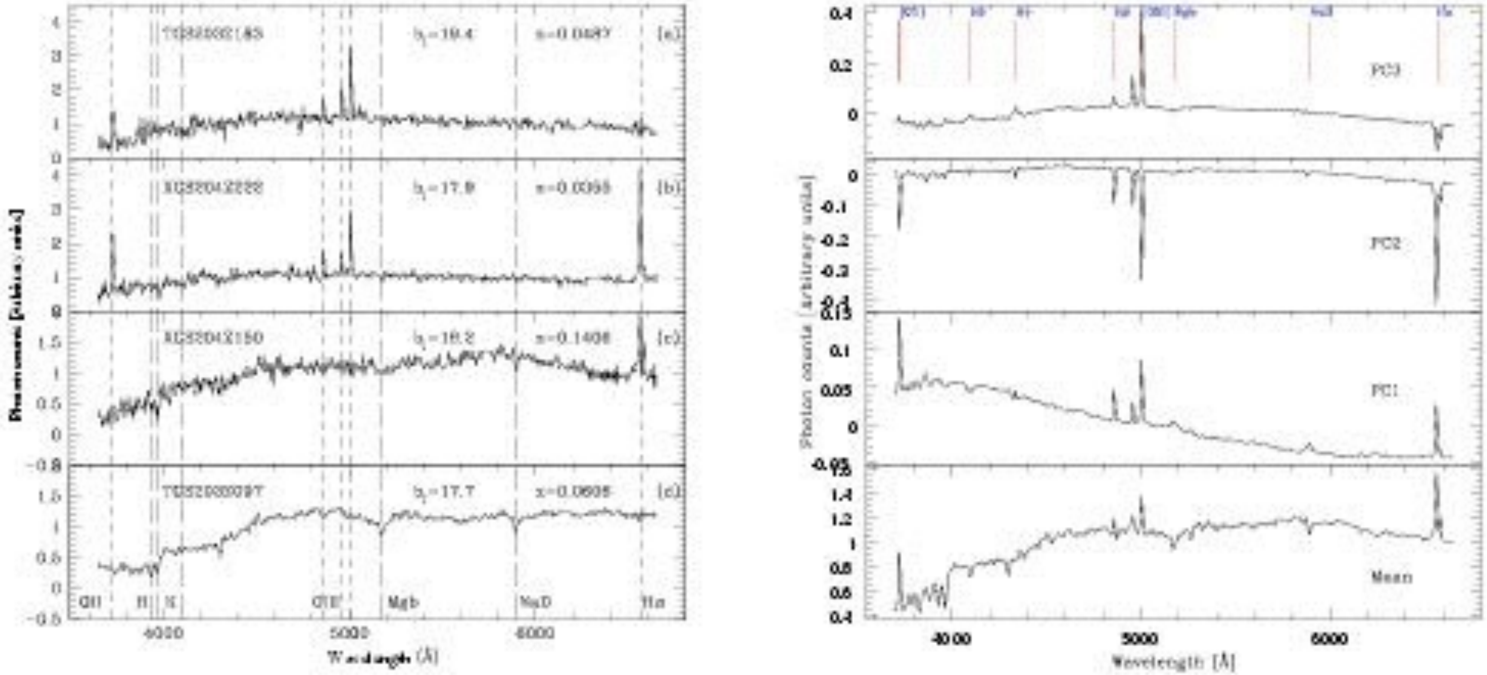
♣ Some PCA problems have a larger number of variables than input observables, **m > N**, resulting in singular matrices requiring modifications to standard techniques to solve the eigenvector equations.

♣ This situation occurs in **spectral PCA** for which the **m** variables are fluxes in wavelength or frequency bins. **Singular Value Decomposition….**

♣ The technique is ideal for dealing with a huge sample and was therefore adopted in the **2dF survey** which aimed to measure 250,000 galaxy spectra to provide a detailed picture of the galaxy distribution out to a redshift of 0.25.

♣ The PCA approach to **2dF galaxy classification** is discussed in detail by Folkes et al. (1999).

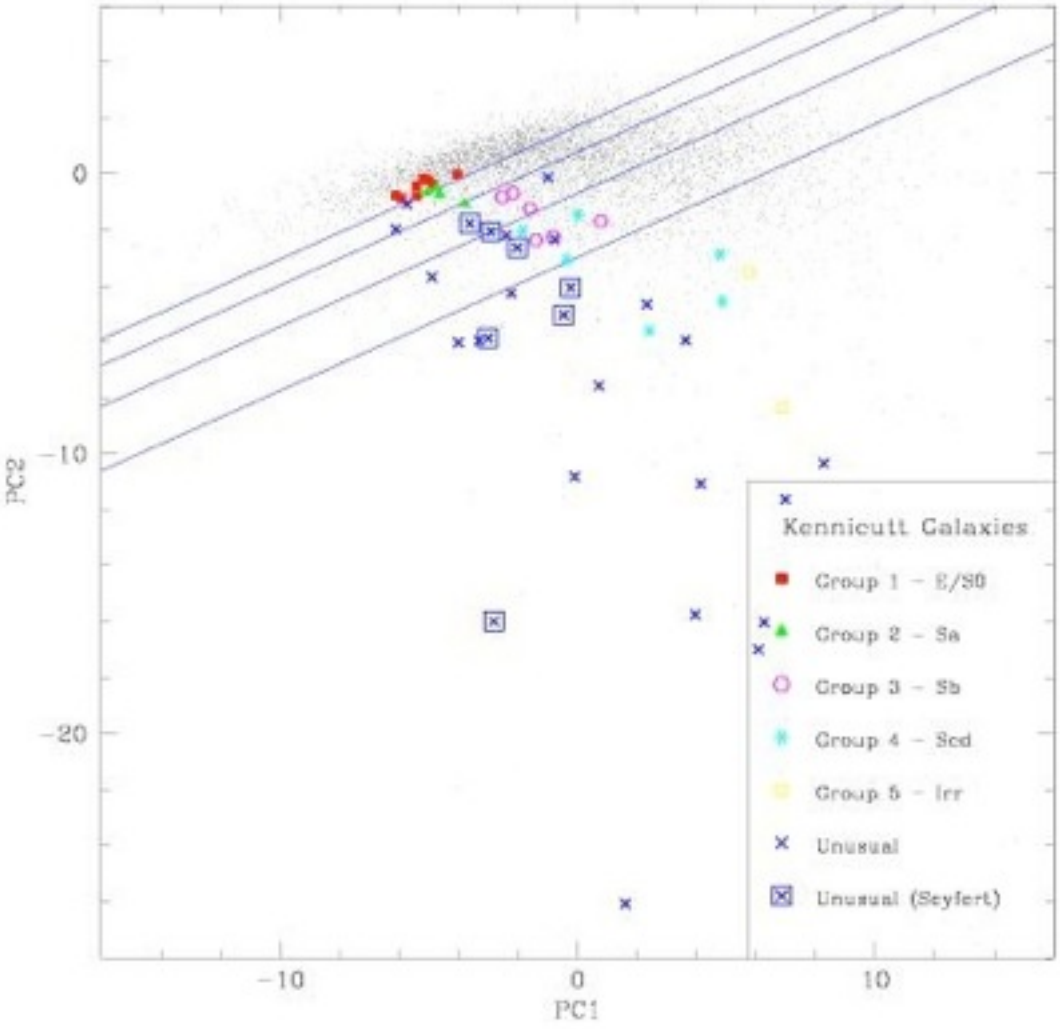
PCA - Example 3 -2dF galaxy classification



Left: four examples of 2dF spectra prepared for PCA. Right: the mean spectrum, and first three principal components. These three components represent the eigenvectors of the covariance matrix of these prepared spectra. In this example, the first PC accounts for 49.6 per cent of the variance; the first three components account for 65.8 per cent of the variance. Much of the remainder is due to noise.

PCA - Example 3 - 2dF galaxy classification (2)

Right: distribution of 2dF galaxy spectra in the PC1 - PC2 plane. Slanted lines divide the plane into the five spectral classes adopted by Folkes et al.; the positions of galaxies typed by Kennicut (1992) are shown.



Note how asymmetrical the distribution looks. This need not invalidate the analysis -- here primarily one of classification -- but the effectiveness must in general be reduced. Asymmetrical shapes in the PC planes must result in unquantifiable errors in the classification.

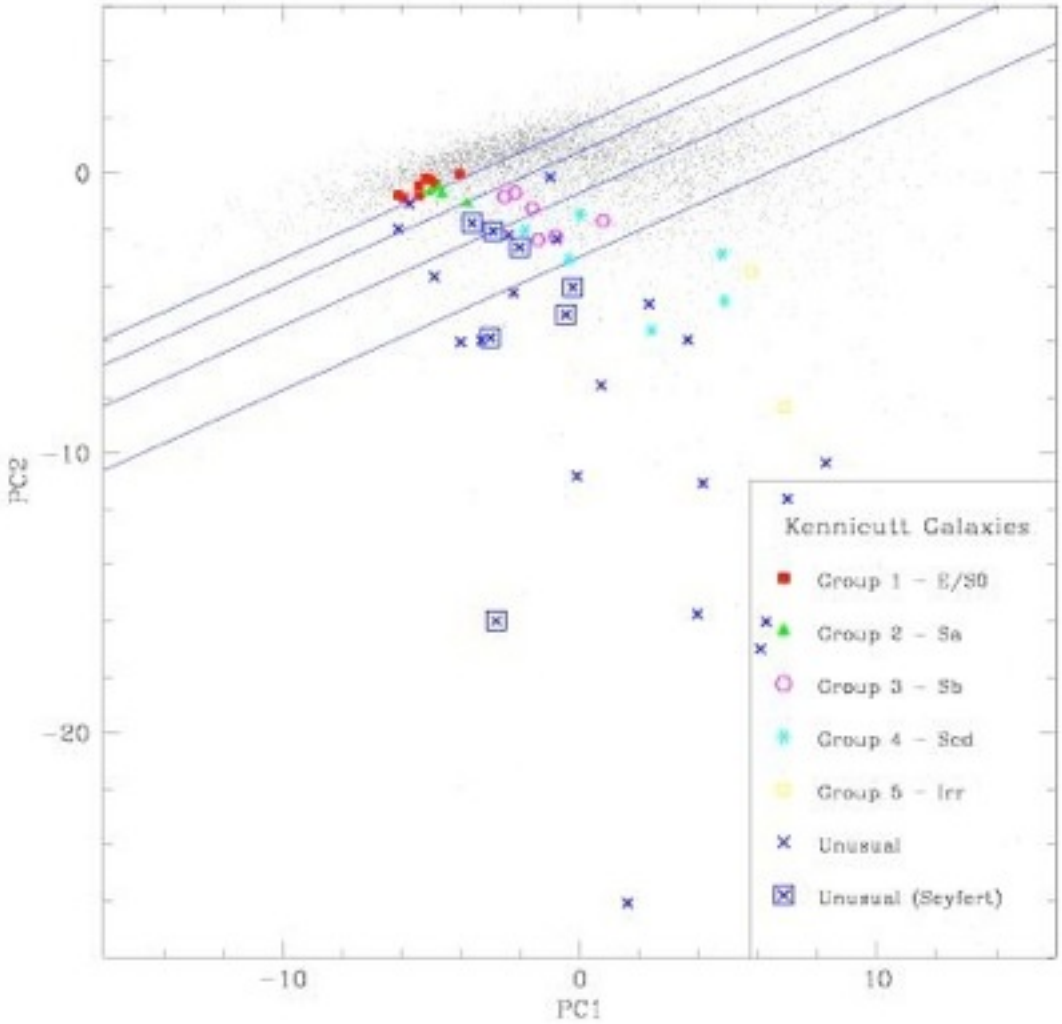
PCA - Example 3 - 2dF galaxy classification (3)

The key aspect Folkes et al wished to address was how luminosity function depends on galaxy type.

The objects in the PC1 -- PC2 plane form a single cluster, blue emission-line objects to the right, red objects with absorption lines to the left, and strong-emission-line objects straggling downward.

Five spectral classes were then adopted, shown by the slanted lines.

Confirmation that these spectral classes correspond to morphological classification came from placing the 55 Kennicut (1992) standard galaxies into this plot; the 5 classes are roughly E/SO, Sa, Sb, Scd and Irr.

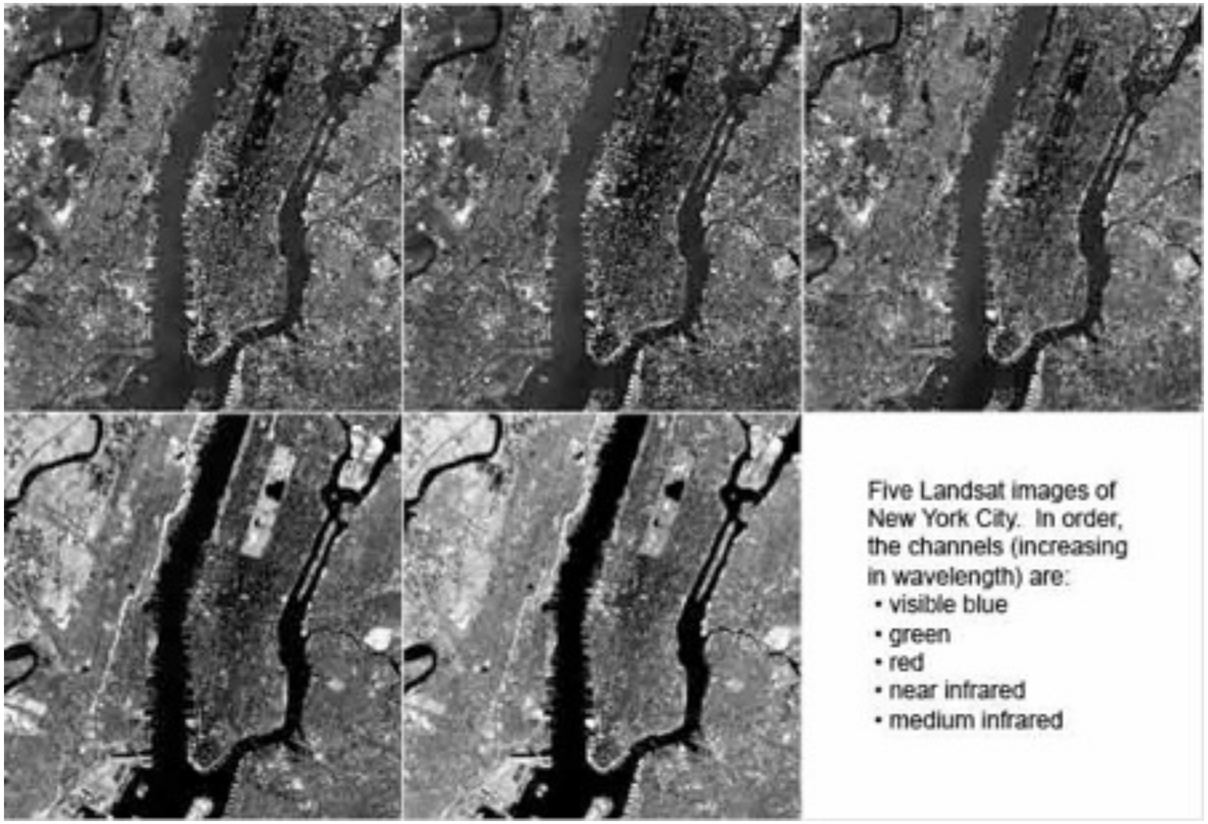


The way ahead to use the PCA classes to work out luminosity functions for each is clear, and the punch line is that **significantly different Schechter functions emerged for each class.**

PCA - Example 4 - a colour transform

Reindeer Graphics: FoveaPro Principal Component Analysis: **The ultimate color transform, from 2 to 56 channels.**

In optical science, a prism is used to break up white light into its componenty colors. The Reindeer Graphics PCA plugins do something similar for color and multichannel images. They find the principal colors that make up the image, and in so doing isolate the information in new and very useful ways. This is not just a matter of converting from RGB to Lab, CMYK, HSL, or some other standard color space. Rather, the method defines a fundamental set of basic color coordinates unique to each image. The technique used is a standard statistical method, rarely applied to image data, called **Principal Component Analysis.**

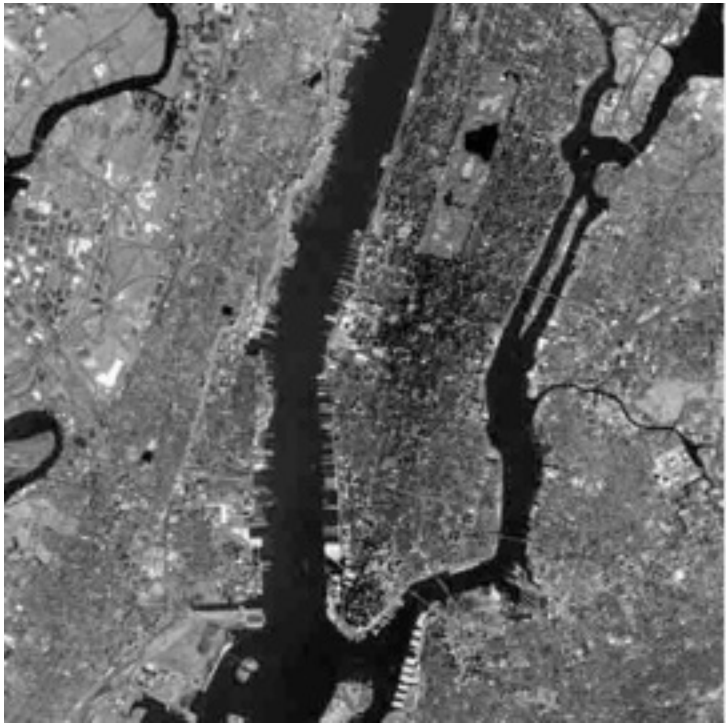


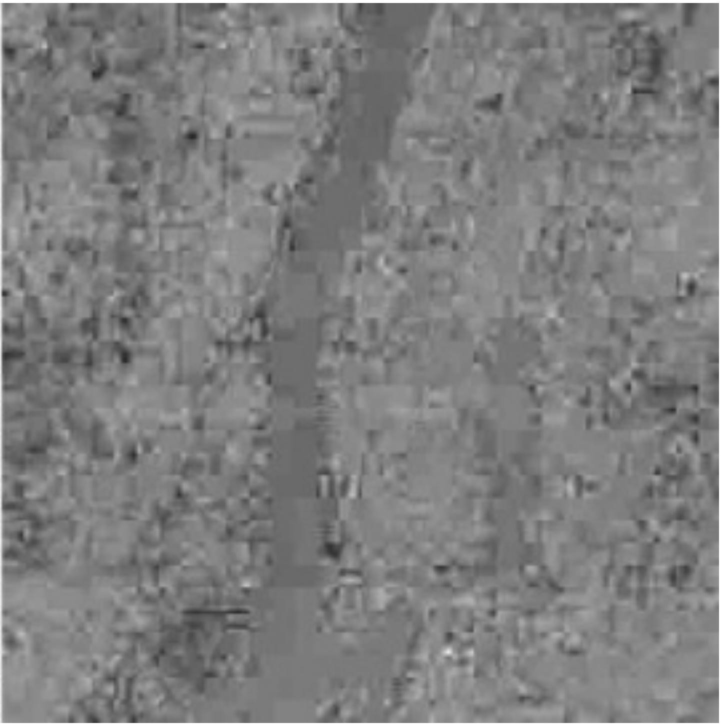
Five Landsat images of New York City. In order, the channels (increasing in wavelength) are:

- visible blue
- green
- red
- near infrared
- medium infrared

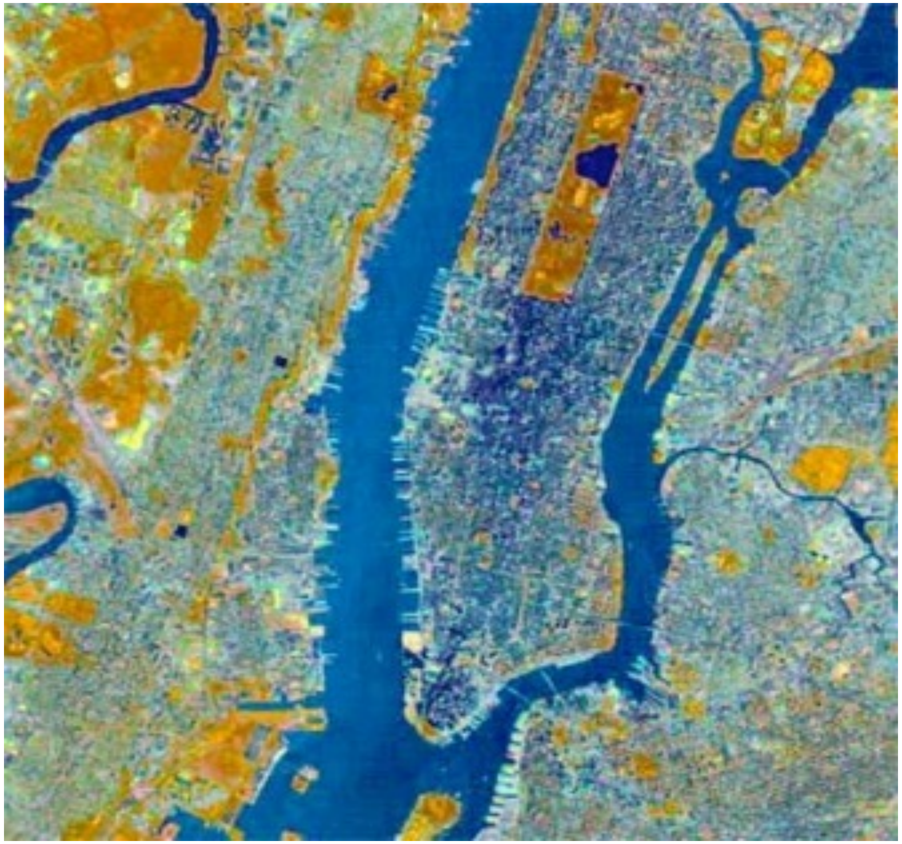
PCA - Example 4 - a colour transform (2)

In the process of computing the Principal Components of this set of images, a rotation in 5-space (because there are five images) with a set of new images created. Below are the *most* significant (73.69% - left) and the *least* significant (0.27% - right) channels after the transform. Notice that the JPEG artifacts from the set of images have all appeared in the bottom channel (right). This characteristic of PCA will prove useful for removing both pattern noise (such as JPEG and DV encoding) and random noise within images, below.





PCA - Example 4 - a colour transform (3)



Another useful trick with PCA is to place the most significant three channels into the Red, Green, and Blue channels of a standard RGB image. In this image, Red contains the primary component (73.69%), Green holds the second component (21.67%), and Blue holds the third component (3.31%). Only a tiny fraction remains unaccounted for (1.34%) and most of that is the JPEG signature.