an introduction to

Principal Component Analysis

(PCA)

abstract

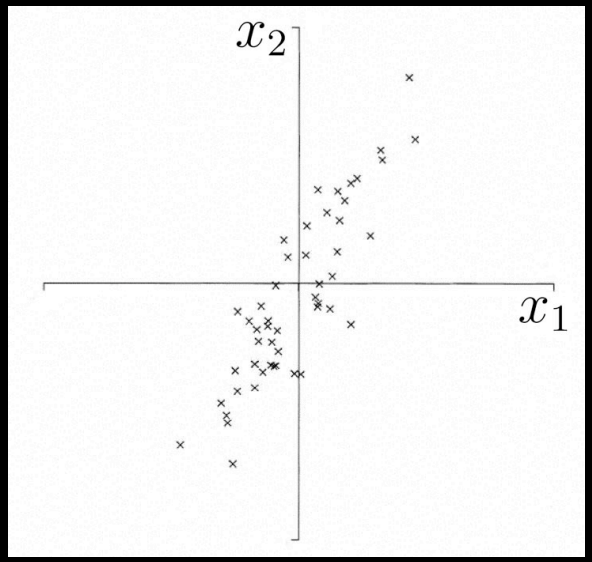
Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables, that nonetheless retains most of the sample's information.

By information we mean the variation present in the sample, given by the correlations between the original variables. The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

overview

- geometric picture of PCs
- algebraic definition and derivation of PCs
- usage of PCA
- astronomical application

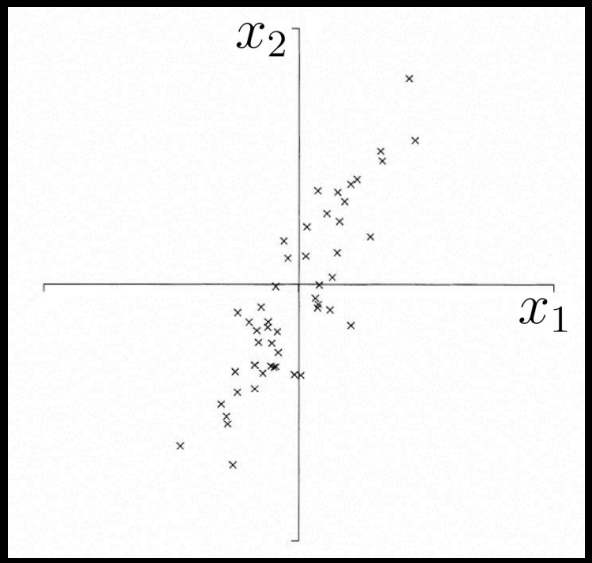
Geometric picture of principal components (PCs)

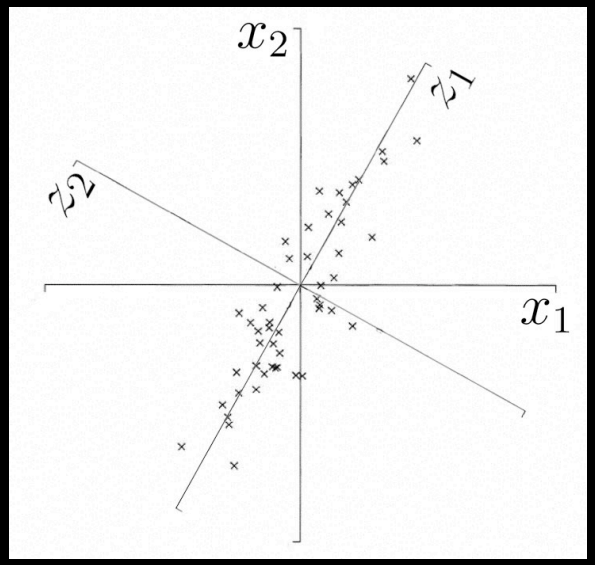


A sample of *n* observations in the 2-D space $\mathbf{x} = (x_1, x_2)$

Goal: to account for the variation in a sample
in as few variables as possible, to some accuracy

Geometric picture of principal components (PCs)





- the 1st PC z_1 is a minimum distance fit to a line in $\mathbf{x}$ space
- the 2nd PC z_2 is a minimum distance fit to a line in the plane perpendicular to the 1st PC

PCs are a series of linear least squares fits to a sample, each orthogonal to all the previous.

Algebraic definition of PCs

Given a sample of *n* observations on a vector of *p* variables

$$\mathbf{x} = (x_1, x_2, \ldots, x_p)$$

define the first principal component of the sample
by the linear transformation

$$z_1 \equiv \mathbf{a}_1^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{p} \mathrm{a}_{i1} x_i$$

where the vector $\mathbf{a}_1 = (\mathrm{a}_{11}, \mathrm{a}_{21}, \ldots, \mathrm{a}_{p1})$

is chosen such that $\mathrm{var}[z_1]$ is maximum

Algebraic definition of PCs

Likewise, define the k^{th} PC of the sample by the linear transformation

$$z_k \equiv \mathbf{a}_k^{\mathrm{T}} \mathbf{x} \qquad k = 1, \ldots, p$$

where the vector

$$\mathbf{a}_k = (\mathrm{a}_{1k}, \mathrm{a}_{2k}, \ldots, \mathrm{a}_{pk})$$

is chosen such that

$$\mathrm{var}[z_k] \quad \text{is maximum}$$

subject to

$$\mathrm{cov}[z_k, z_l] = 0 \quad \text{for } k > l \geq 1$$

and to

$$\mathbf{a}_k^{\mathrm{T}} \mathbf{a}_k = 1$$

Algebraic derivation of coefficient vectors $\mathbf{a}_k$

To find $\mathbf{a}_1$ first note that

$$\mathrm{var}[z_1] = \langle z_1^2 \rangle - \langle z_1 \rangle^2$$

$$= \sum_{i,j=1}^{p} \mathrm{a}_{i1}\mathrm{a}_{j1}\langle x_i x_j \rangle - \sum_{i,j=1}^{p} \mathrm{a}_{i1}\mathrm{a}_{j1}\langle x_i \rangle \langle x_j \rangle$$

$$= \sum_{i,j=1}^{p} \mathrm{a}_{i1}\mathrm{a}_{j1} S_{ij} \quad \text{where} \quad S_{ij} \equiv \sigma_{x_i x_j} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

$$= \mathbf{a}_1^{\mathrm{T}} \mathbf{S} \mathbf{a}_1$$

$\mathbf{S}$ is the covariance matrix for the variables $\mathbf{x} = (x_1, x_2, \ldots, x_p)$

Algebraic derivation of coefficient vectors $\mathbf{a}_k$

To find $\mathbf{a}_1$ maximize $\mathrm{var}[z_1]$ subject to $\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$

Let λ be a Lagrange multiplier

then maximize $\mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1 - \lambda(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 - 1)$

by differentiating… $\mathbf{S}\mathbf{a}_1 - \lambda\mathbf{a}_1 = 0$

$\Rightarrow (\mathbf{S} - \lambda\mathbf{I}_p)\mathbf{a}_1 = 0$

therefore

$\mathbf{a}_1$ is an eigenvector of $\mathbf{S}$

corresponding to eigenvalue $\lambda \equiv \lambda_1$

Algebraic derivation of $\mathbf{a}_k$

We have maximized

$$\mathrm{var}[z_1] = \mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_1 = \mathbf{a}_1^{\mathrm{T}}\lambda_1\mathbf{a}_1 = \lambda_1$$

So λ_1 is the largest eigenvalue of $\mathbf{S}$

The first PC z_1 retains the greatest amount of variation in the sample.

Algebraic derivation of coefficient vectors $\mathbf{a}_k$

To find the next coefficient vector $\mathbf{a}_2$ maximize $\mathrm{var}[z_2]$

subject to $\mathrm{cov}[z_2, z_1] = 0$

and to $\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = 1$

First note that

$$\mathrm{cov}[z_2, z_1] = \mathbf{a}_1^{\mathrm{T}}\mathbf{S}\mathbf{a}_2 = \lambda_1\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_2$$

then let λ and φ be Lagrange multipliers, and maximize

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{S}\mathbf{a}_2 - \lambda(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - 1) - \phi\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1$$

Algebraic derivation of coefficient vectors $\mathbf{a}_k$

We find that $\mathbf{a}_2$ is also an eigenvector of $\mathbf{S}$ whose eigenvalue $\lambda \equiv \lambda_2$ is the second largest.

In general

$$\mathrm{var}[z_k] = \mathbf{a}_k^{\mathrm{T}} \mathbf{S} \mathbf{a}_k = \lambda_k$$

- The k^{th} largest eigenvalue of $\mathbf{S}$ is the variance of the k^{th} PC.
- The k^{th} PC z_k retains the k^{th} greatest fraction of the variation in the sample.

Algebraic formulation of PCA

Given a sample of n observations on a vector of p variables $\mathbf{x} = (x_1, x_2, \ldots, x_p)$

define a vector of p PCs $\mathbf{z} = (z_1, z_2, \ldots, z_p)$

according to $\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x}$

where $\mathbf{A}$ is an orthogonal p x p matrix
whose k^{th} column is the k^{th} eigenvector $\mathbf{a}_k$ of $\mathbf{S}$

Then $\mathbf{\Lambda} = \mathbf{A}^{\mathrm{T}}\mathbf{S}\mathbf{A}$ is the covariance matrix of the PCs,

being diagonal with elements $\Lambda_{ij} = \lambda_i \delta_{ij}$

usage of PCA: Probability distribution for sample PCs

If (i) the *n* observations of $\mathbf{x}$ in the sample are independent &

(ii) $\mathbf{x}$ is drawn from an underlying population that follows a *p*-variate normal (Gaussian) distribution with known covariance matrix $\tilde{\mathbf{S}}$

then $$(n-1)\mathbf{S} \sim W_p(\tilde{\mathbf{S}}, n-1)$$

where W_p is the Wishart distribution

else utilize a bootstrap approximation

usage of PCA: Probability distribution for sample PCs

If (i) $(n-1)\mathbf{S}$ follows a Wishart distribution &

(ii) the population eigenvalues $\tilde{\lambda}_k$ are all distinct

then the following results hold as $n \longrightarrow \infty$

- all the λ_k are independent of all the $\mathbf{a}_k$

  $\lambda \equiv (\lambda_1, \lambda_2, \ldots, \lambda_p)$ $\mathbf{a}_k$

  are jointly normally distributed

$$\langle \lambda \rangle = \tilde{\lambda}$$

$$\langle \mathbf{a}_k \rangle = \tilde{\mathbf{a}}_k$$

(a tilde denotes a population quantity)

usage of PCA: Probability distribution for sample PCs

and

- $$\mathrm{cov}[\lambda_k, \lambda_{k'}] = \begin{cases} \frac{2\tilde{\lambda}_k^2}{n-1} & k = k' \\ 0 & k \neq k' \end{cases}$$

$$\mathrm{cov}[\mathrm{a}_{jk}, \mathrm{a}_{j'k'}] = \begin{cases} \frac{\tilde{\lambda}_k}{(n-1)} \sum_{l \neq k}^{p} \frac{\tilde{\lambda}_l \tilde{\mathrm{a}}_{jl} \tilde{\mathrm{a}}_{j'l}}{(\tilde{\lambda}_l - \tilde{\lambda}_k)^2} & k = k' \\ -\frac{\tilde{\lambda}_k \tilde{\lambda}_{k'} \tilde{\mathrm{a}}_{jk} \tilde{\mathrm{a}}_{j'k'}}{(n-1)(\tilde{\lambda}_k - \tilde{\lambda}_{k'})^2} & k \neq k' \end{cases}$$

(a tilde denotes a population quantity)

usage of PCA: Inference about population PCs

If $\mathbf{x}$ follows a p-variate normal distribution

then analytic expressions exist* for

MLE's of $\tilde{\lambda}_k$, $\tilde{\mathbf{a}}_k$, and $\tilde{\mathbf{S}}$

confidence intervals for $\tilde{\lambda}_k$ and $\tilde{\mathbf{a}}_k$

hypothesis testing for $\tilde{\lambda}_k$ and $\tilde{\mathbf{a}}_k$

else bootstrap and jackknife approximations exist

*see references, esp. Jolliffe

usage of PCA: Practical computation of PCs

In general it is useful to define standardized variables by

$$\mathbf{x} \longrightarrow \mathbf{x}^* = \left(\frac{x_1}{\sqrt{\sigma_1^2}}, \frac{x_2}{\sqrt{\sigma_2^2}}, \ldots, \frac{x_p}{\sqrt{\sigma_p^2}} \right)$$

If the x_k are each measured about their sample mean

then the covariance matrix $\mathbf{S}^*$ of $\mathbf{x}^*$

will be equal to the correlation matrix of $\mathbf{x}$

and the PCs $\mathbf{z}^* = \mathbf{A}^{*\mathrm{T}}\mathbf{x}^*$ will be dimensionless

usage of PCA: Practical computation of PCs

Given a sample of *n* observations on a vector $\mathbf{x}$ of *p* variables x_k (each measured about its sample mean)

compute the covariance matrix $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$

where $\mathbf{X}$ is the *n* x *p* matrix

whose i^{th} row is the i^{th} obsv. $\mathbf{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{ip})$

Then compute the *n* x *p* matrix $\mathbf{Z} = \mathbf{X}\mathbf{A}$

whose i^{th} row is the PC score $\mathbf{z}_i \equiv (z_{i1}, z_{i2}, \ldots, z_{ip})$

for the i^{th} observation.

usage of PCA: Practical computation of PCs

Write $\mathbf{X} = \mathbf{Z}\mathbf{A}^{\mathrm{T}}$ to decompose each observation into PCs

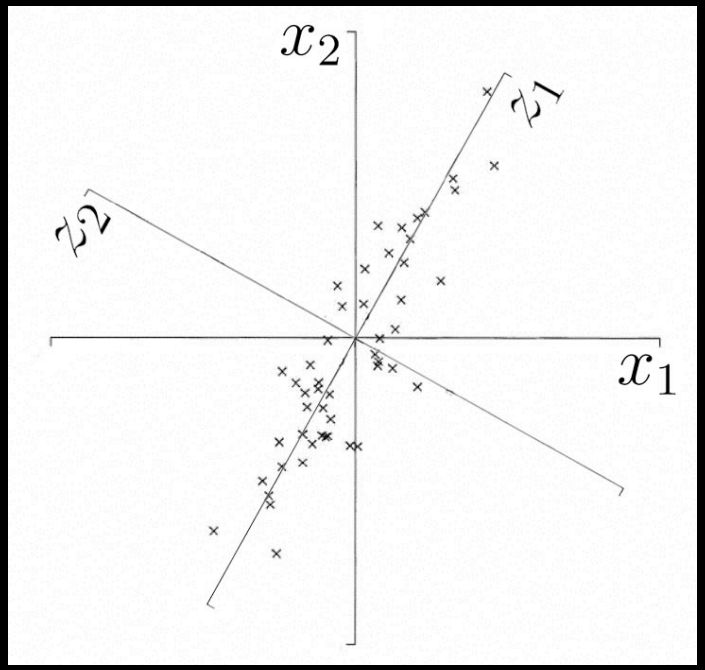
$$\mathbf{x}_i = \mathbf{A}\mathbf{z}_i = \sum_{k=1}^{p} z_{ik}\mathbf{a}_k$$

usage of PCA: Data compression

Because the k^{th} PC retains the k^{th} greatest fraction of the variation

we can approximate each observation

by truncating the sum at the first $m < p$ PCs



$$\mathbf{x}_i \cong \mathbf{x}_i^m = \sum_{k=1}^{m} z_{ik}\mathbf{a}_k$$

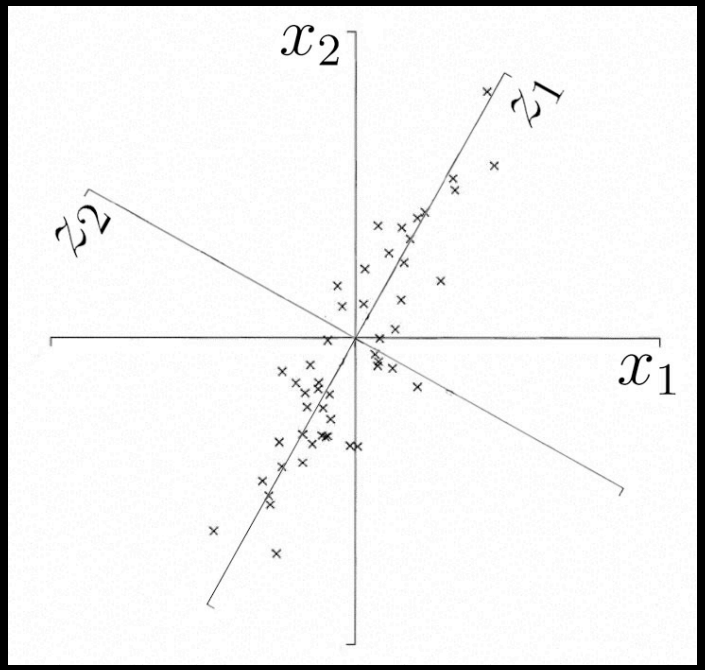
usage of PCA: Data compression

Reduce the dimensionality of the data

from p to $m < p$ by approximating $\mathbf{X} \cong \mathbf{X}^m = \mathbf{Z}^m \mathbf{A}^{m\mathrm{T}}$

where $\mathbf{Z}^m$ is the *n* x *m* portion of $\mathbf{Z}$

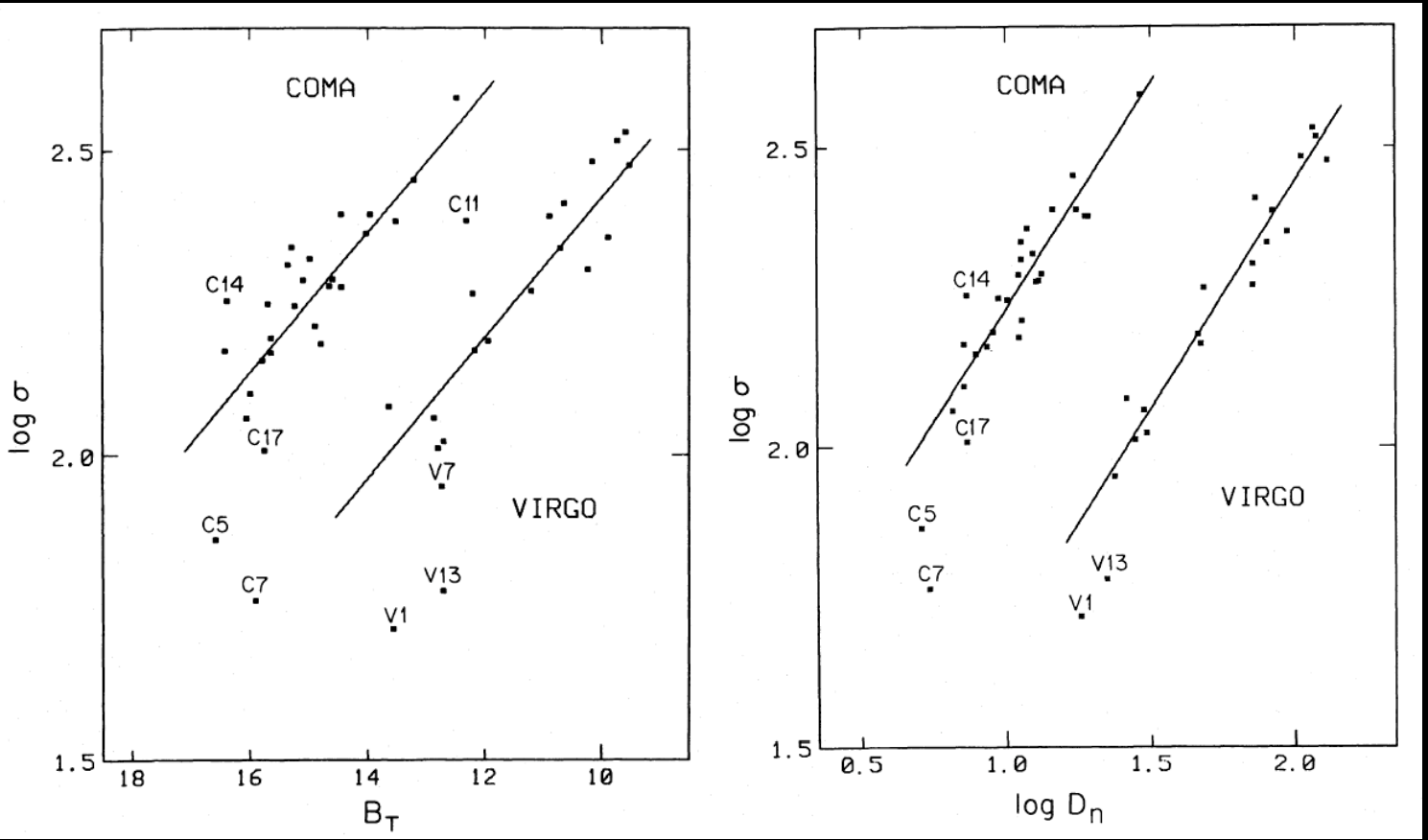
and $\mathbf{A}^m$ is the *p* x *m* portion of $\mathbf{A}$



astronomical application: PCs for elliptical galaxies

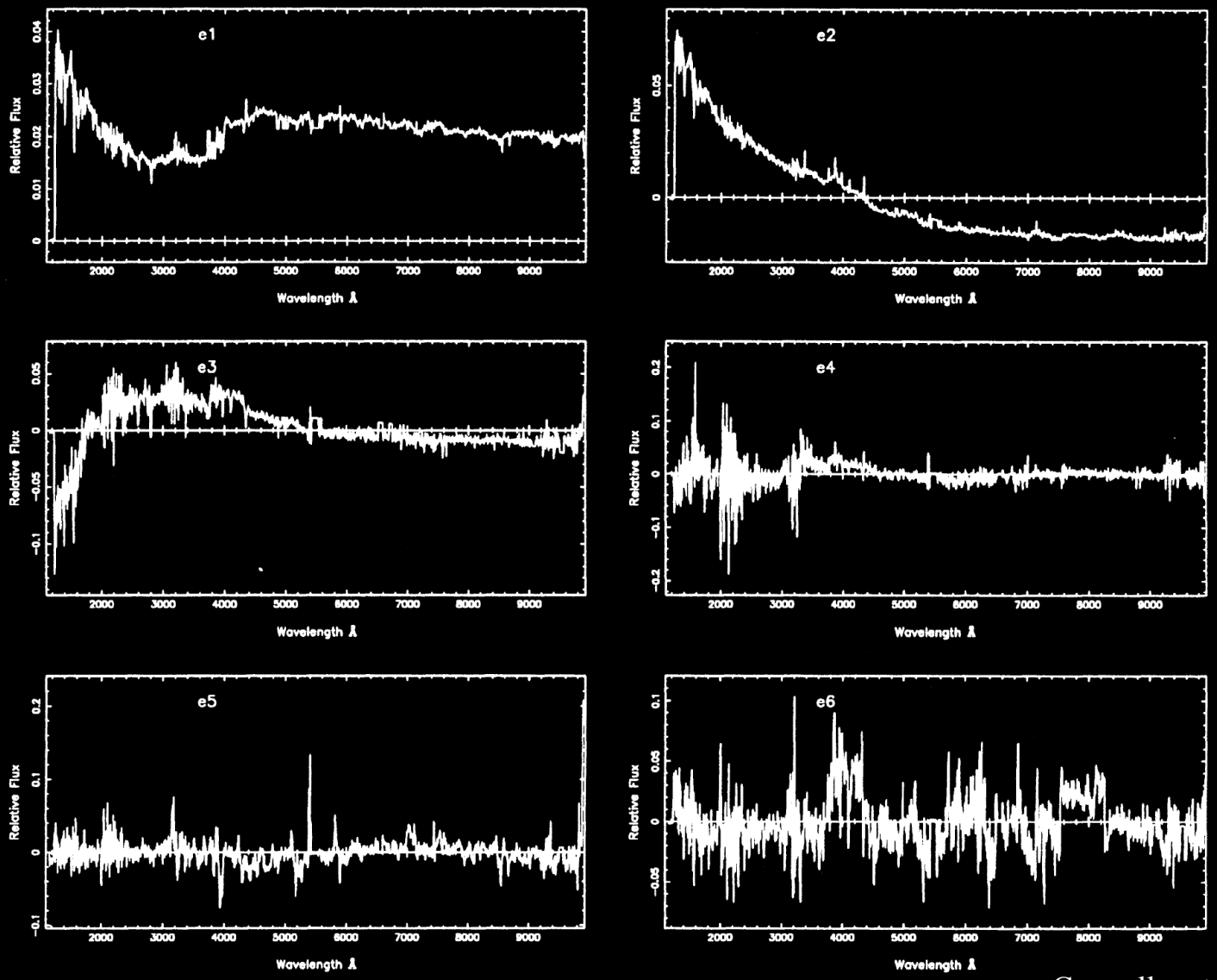
Rotating to PC in $B_T - \Sigma$ space improves Faber-Jackson relation as a distance indicator

$$L \propto \sigma^4$$



Dressler, et al. 1987

astronomical application: Eigenspectra (KL transform)



Connolly, et al. 1995

references

Connolly, and Szalay, et al., "Spectral Classification of Galaxies: An Orthogonal Approach", AJ, **110**, 1071-1082, 1995.
Dressler, et al., "Spectroscopy and Photometry of Elliptical Galaxies. I. A New Distance Estimator", ApJ, **313**, 42-58, 1987.
Efstathiou, G., and Fall, S.M., "Multivariate analysis of elliptical galaxies", MNRAS, **206**, 453-464, 1984.
Johnston, D.E., et al., "SDSS J0903+5028: A New Gravitational Lens", AJ, **126**, 2281-2290, 2003.
Jolliffe, Ian T., 2002, *Principal Component Analysis* (Springer-Verlag New York, Secaucus, NJ).
Lupton, R., 1993, *Statistics In Theory and Practice* (Princeton University Press, Princeton, NJ).
Murtagh, F., and Heck, A., *Multivariate Data Analysis* (D. Reidel Publishing Company, Dordrecht, Holland).
Yip, C.W., and Szalay, A.S., et al., "Distributions of Galaxy Spectral Types in the SDSS", AJ, **128**, 585-609, 2004.

That’s no quasar.

It’s a space station.