Análisis de datos y Estadística Avanzada

Máster Interuniversitario de Astrofísica UCM+UAM

Tema 9: Análisis de componentes principales (PCA)

Javier Gorgas y Nicolás Cardiel
Departamento de Astrofísica y Ciencias de la Atmósfera
Facultad de Ciencias Físicas
Universidad Complutense de Madrid

Esquema

Técnicas multivariantes

Consideremos un conjunto de objetos sobre los que se mide una serie de propiedades diferentes. ¿Estudio óptimo? Uso de técnicas multivariantes, las cuales permiten realizar un **análisis simultáneo** de todos los objetos y sus propiedades (ver Tema 6).

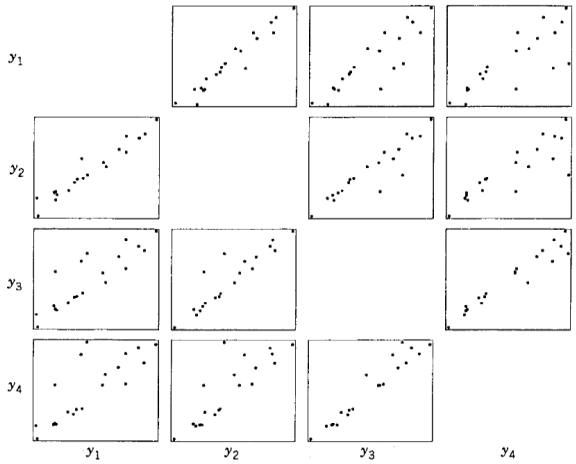
	propiedad #1	propiedad #2	...	...	propiedad #p
objeto #1	y_{11}	y_{12}	...	...	y_{1p}
objeto #2	y_{21}	y_{22}	...	...	y_{2p}
⋮	⋮	⋮	⋮	⋮	⋮
objeto #n	y_{n1}	y_{n2}	...	...	y_{np}

¿Qué hacer?

- Contrastes de hipótesis sobre la matriz de covarianza: testear correlación entre propiedades.
- Análisis de componentes principales: bucar un conjunto reducido de combinaciones lineales de las variables que resuman la variación de los datos.
- Análisis de factores: expresar las variables originales como un conjunto de funciones lineales de factores.
- Análisis de agrupación: determinar agrupaciones entre datos (número de grupos inicialmente desconocido).
- Análisis de clasificación: ubicación de nuevos objetos en distintos grupos predefinidos.
- Regresión lineal múltiple: determinar un modelo que prediga un conjunto de propiedades (variables dependientes) a partir de otro conjunto de propiedades (variables independientes).
- Análisis discriminante: buscar la combinación lineal de las variables que mejor discrimine entre diferentes muestras de objetos.
- ...

¡Simplificar para sobrevivir!

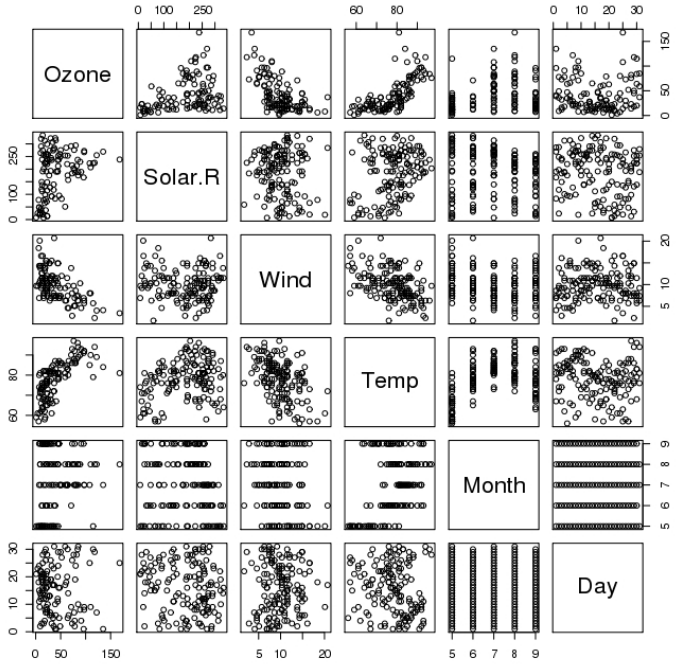
El objetivo principal del análisis de componentes principales es reducir la dimensionalidad de un conjunto (muy) grande de datos.



Tradicionalmente los astrónomos tienden a representar los parámetros medidos uno frente a otro, tratando de inferir conclusiones a partir de las correlaciones observadas. Esta técnica es inviable cuando el número de parámetros representados es superior a 4 ó 5.

R permite, de forma trivial, representar todas los posibles diagramas de dispersión de un conjunto de datos multivariante con la ejecución de un único comando:

```
> plot(airquality)
```
←—los datos están en el paquete `base`, cargado por defecto



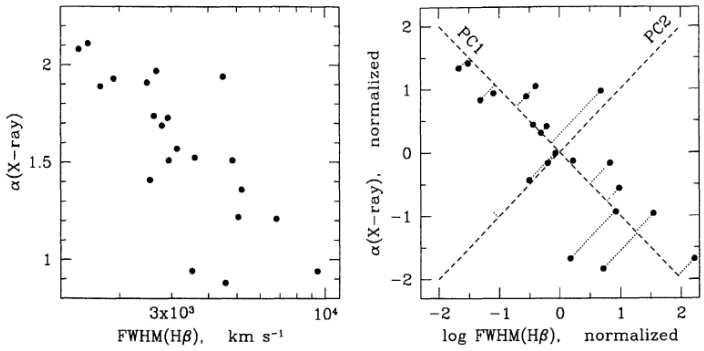
Como ya vimos en su día, en el trabajo dentro del área del análisis multivariante resulta extremadamente útil utilizar álgebra matricial.

	propiedad #1	propiedad #2	...	...	propiedad #p
objeto #1	y_{11}	y_{12}	$\cdots$	$\cdots$	y_{1p}
objeto #2	y_{21}	y_{22}	$\cdots$	$\cdots$	y_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
objeto #i	y_{i1}	y_{i2}	$\cdots$	$\cdots$	y_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
objeto #n	y_{n1}	y_{n2}	$\cdots$	$\cdots$	y_{np}
medias	$\bar{y}_1$	$\bar{y}_2$	$\cdots$	$\cdots$	$\bar{y}_p$

Podemos definir $\boldsymbol{y}$ como un vector aleatorio con p variables (propiedades) medidas en cada objeto. Si tenemos n objetos en la muestra, las observaciones pueden escribirse como $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$, donde

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix}, \qquad \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1' \\ \boldsymbol{y}_2' \\ \vdots \\ \boldsymbol{y}_n' \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \cdots & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & \cdots & y_{np} \end{pmatrix}.$$

Aproximación geométrica al problema



(Francis & Wills 1999)

- Consideremos un conjunto de n observaciones $\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_i}, \ldots, \boldsymbol{y_n}$, que forman una nube de puntos en un espacio p-dimensional (como simplificación, podemos visualizarlo como un elipsoide de puntos). Cada vector $\boldsymbol{y_i}$ es un vector columna con p elementos.
- Si las p propiedades $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_p$ están correlacionadas, la distribución de puntos no estará orientada paralelamente a los ejes definidos por $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_p$.
- A través del PCA vamos a determinar los "ejes naturales" de la distribución de puntos (i.e., los ejes del elipsoide), cuyo origen se encuentra en $\bar{\boldsymbol{y}}$, el vector medio de $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$. Esto se realiza restando $\bar{\boldsymbol{y}}$ y calculando la rotación que minimice la suma de distancias a los ejes (maximice la proyección de los datos sobre los mismos ejes).

Aproximación geométrica al problema

- Podemos rotar los ejes multiplicando cada vector p-dimensional $\boldsymbol{y}_i$ por una matriz ortogonal $\boldsymbol{A}$

$$\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{y}_i.$$

  Como $\boldsymbol{A}$ es ortogonal, $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$, la distancia al origen no cambia

$$\boldsymbol{z}_i'\boldsymbol{z}_i = (\boldsymbol{A}\boldsymbol{y}_i)'(\boldsymbol{A}\boldsymbol{y}_i) = \boldsymbol{y}_i'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{y}_i = \boldsymbol{y}_i'\boldsymbol{y}_i,$$

  y por ello $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{y}_i$ es realmente una rotación.
- Buscamos la matriz ortogonal $\boldsymbol{A}$ que nos proporcione unos nuevos parámetros (componentes principales) $\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_p$ que no estén correlacionadas. Para ello necesitamos que la matriz muestral de covarianzas de $\boldsymbol{z}$, $\boldsymbol{S}_z$, sea diagonal

$$\boldsymbol{S}_z = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}' = \begin{pmatrix} s_{z_1}^2 & 0 & \ldots & 0 \\ 0 & s_{z_2}^2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & s_{z_p}^2 \end{pmatrix},$$

  donde $\boldsymbol{S}$ es la matriz muestral de covarianzas de $\boldsymbol{y}$.

Conclusión: El problema se reduce a encontrar $\boldsymbol{A}$ tal que diagonalice $\boldsymbol{S}$.
(Puede demostrarse que una matriz simétrica S puede ser diagonalizada empleando una matriz ortogonal que contenga los autovectores normalizados de S, y la matriz diagonal resultante contiene los autovalores asociados.)

Aproximación geométrica al problema

- La matriz ortogonal $\boldsymbol{A}$ que diagonaliza $\boldsymbol{S}$ puede escribirse como

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \\ \vdots \\ \boldsymbol{a}_p' \end{pmatrix},$$

  donde $\boldsymbol{a}_i$ son los autovectores de $\boldsymbol{S}$ que verifican $\boldsymbol{a}_i'\boldsymbol{a}_j = \delta_{ij}$ (están normalizados y son ortogonales).
- Las componentes principales son las nuevas variables $\mathcal{Z}_i = \boldsymbol{a}_i'\boldsymbol{y}$, por ejemplo

$$\mathcal{Z}_1 = a_{11}\mathcal{Y}_1 + a_{12}\mathcal{Y}_2 + \ldots + a_{1p}\mathcal{Y}_p.$$

- Los autovalores de $\boldsymbol{S}$ serán las varianzas muestrales de las componentes principales

$$\begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{pmatrix} = \begin{pmatrix} s^2_{z_1} & 0 & \ldots & 0 \\ 0 & s^2_{z_2} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & s^2_{z_p} \end{pmatrix},$$

  siendo habitual ordenar las variables de forma que $\lambda_1 > \lambda_2 > \ldots > \lambda_p$.

Reducción de la dimensionalidad

- Como los autovalores son varianzas de las componentes principales, podemos definir la proporción de varianza explicada por las primeras k componentes mediante

$$\text{Proporción de varianza} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\sum_{j=1}^{p} s_{jj}}$$

- Si los parámetros están muy correlacionados, la dimensionalidad efectiva es mucho menor que p. En este caso los primeros autovalores son grandes y la proporción de varianza será próxima a 1 para valores de k pequeños.
- Si las correlaciones entre los parámetros originales son pequeñas, la dimensionalidad efectiva será próxima a p y los autovalores serán parecidos. En este caso las componentes principales esencialmente duplicarán los parámetros originales y no se conseguirá reducir la dimensionalidad.

Aproximación algebraica al problema

- Otra forma de interpretar el PCA es como un método que permita encontrar combinaciones lineales de variables con una varianza máxima. Por ejemplo, dado un conjunto de p parámetros $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_p$, podemos buscar la dirección definida por el vector p-dimensional $\boldsymbol{a}$ en la que un nuevo parámetro definido como

$$\mathcal{Z} = a_1\mathcal{Y}_1 + a_2\mathcal{Y}_2 + \ldots + a_p\mathcal{Y}_p = \boldsymbol{a}'\boldsymbol{y},$$

  presenta una varianza máxima.
- Si tenemos un conjunto de n objetos, la varianza muestral de $\mathcal{Z}_i = \boldsymbol{a}'\boldsymbol{y}_i$, con $i = 1, \ldots, n$, puede calcularse en función de $\boldsymbol{S}$, la matriz muestral de covarianzas de $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$, como

$$s_z^2 = \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}.$$

- No es posible calcular un máximo para la expresión anterior porque su valor crece de forma indefinida para vectores $\boldsymbol{a}$ suficientemente grandes. Una forma de hacerlo es restringir arbitrariamente (pero de forma razonable) el tamaño (norma) de $\boldsymbol{a}$. Por ejemplo, suponiendo que es un vector unitario, i.e., $\boldsymbol{a}'\boldsymbol{a} = 1$.
- Podemos entonces buscar el máximo de s_z^2 con la condición $\boldsymbol{a}'\boldsymbol{a} = 1$. Esto se hace usando la técnica de los multiplicadores de Lagrange a imponiendo que la derivada de $\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a} - \lambda(\boldsymbol{a}'\boldsymbol{a} - 1)$ sea igual a cero, lo que conduce a

$$(\boldsymbol{S} - \lambda\boldsymbol{I})\boldsymbol{a} = 0 \qquad \Leftrightarrow \qquad \boldsymbol{S}\boldsymbol{a} = \lambda\boldsymbol{a}$$

Aproximación algebraica al problema

- El valor óptimo de $\boldsymbol{a}$ (que llamaremos $\boldsymbol{a}_1$) es la solución de

$$\boldsymbol{S}\boldsymbol{a} = \lambda\boldsymbol{a}.$$

  Es decir, $\boldsymbol{a}_1$ es el autovector asociado al autovalor λ_1 de mayor tamaño.
- El segundo eje que maximiza la varianza debe ser perpendicular al primero ya calculado, por lo que tenemos una nueva restricción $\boldsymbol{a}'\boldsymbol{a}_1 = 0$, por lo que la expresión a minimizar es ahora $\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a} - \lambda_2(\boldsymbol{a}'\boldsymbol{a} - 1) - \mu_2(\boldsymbol{a}'\boldsymbol{a}_1)$, donde λ_2 y μ_2 son dos nuevos multiplicadores de Lagrange. Tomando derivadas es fácil mostrar que $\mu_2 = 0$, por lo que la ecuación a resolver vuelve a ser

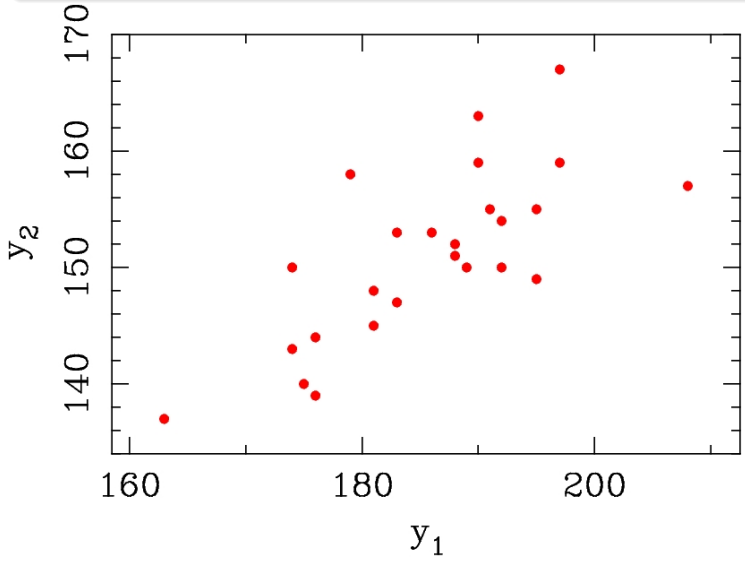
$$\boldsymbol{S}\boldsymbol{a} = \lambda\boldsymbol{a},$$

  siendo λ_2 el segundo autovalor más grande y $\boldsymbol{a}_2$ su autovector asociado.
- De forma similar se razona para el resto de los ejes $\boldsymbol{a}_3, \ldots, \boldsymbol{a}_p$. Es decir, se obtiene el mismo resultado que ya vimos antes en la aproximación geométrica.

Un ejemplo con MATLAB

Consideremos el siguiente conjunto de datos ($n = 25$ objetos y $p = 2$ parámetros)

y_1	191	195	181	183	176	208	189	197	188	192	179	183	174
y_2	155	149	148	153	144	157	150	159	152	150	158	147	150
y_1	190	188	163	195	186	181	175	192	174	176	197	190	
y_2	159	151	137	155	153	145	140	154	143	139	167	163	



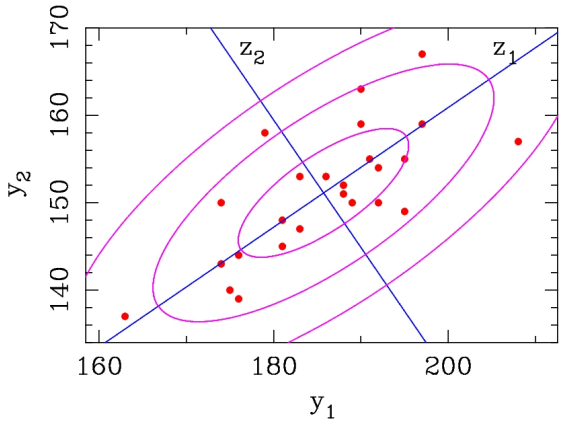

```
» load datos.dat
» y1=datos(:,1);
» y2=datos(:,2);
» plot(y1,y2,'ro');
» xlabel('y1');
» ylabel('y2');
» ymean=mean(datos);
```

$$\bar{\boldsymbol{y}} = \begin{pmatrix} 185.7200 \\ 151.1200 \end{pmatrix}$$

```
» S=cov(datos);
```

$$\boldsymbol{S} = \begin{pmatrix} 95.2933 & 52.8683 \\ 52.8683 & 54.3600 \end{pmatrix}$$

$\boldsymbol{S}$ es la matriz a diagonalizar.



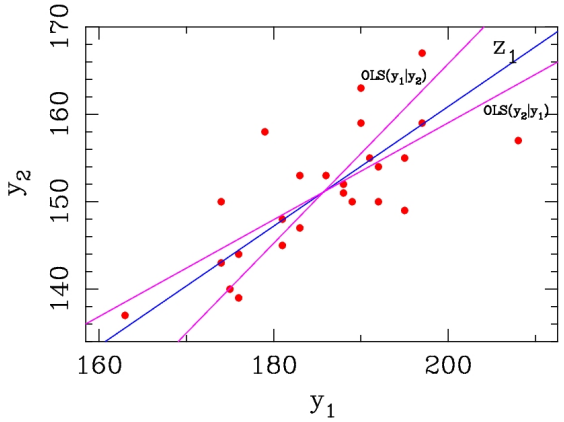

```
» [A,score,lambda,tsquare]=princomp(datos);
```

$$\boldsymbol{A} = \begin{pmatrix} -0.8249 & -0.5652 \\ -0.5652 & 0.8249 \end{pmatrix} \qquad \boldsymbol{\lambda} = \begin{pmatrix} 131.5183 \\ 18.1350 \end{pmatrix}$$

y los autovectores son

$$\boldsymbol{a}_1 = \begin{pmatrix} -0.8249 \\ -0.5652 \end{pmatrix} \quad \text{y} \quad \boldsymbol{a}_2 = \begin{pmatrix} -0.5652 \\ 0.8249 \end{pmatrix}$$

Las elipses tienen semiejes proporcionales a $\sqrt{\lambda_1} = 11.47$ y $\sqrt{\lambda_2} = 4.26$ (calculadas como $y_1 \propto \sqrt{\lambda_1}\cos t$, $y_2 \propto \sqrt{\lambda_2}\sin t$, con $t \in [0, 2\pi]$, rotadas por $\boldsymbol{A}$ y con origen en $\bar{\boldsymbol{y}}$).



La proporción de varianza explicada por la primera componente será

$$\text{Proporción de varianza} = \frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.879 \sim 88\%$$

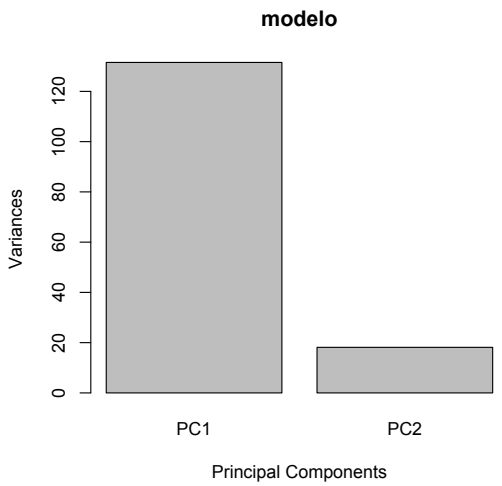
¿Significado de la primera componente?

Al ser el PCA una rotación de ejes, la primera componente principal minimiza la suma cuadrática de distancias entre los puntos y la dirección principal (distancia perpendicular). Es, por tanto, equivalente a la regresión ortogonal (ver Tema 4). De hecho, la dirección de la primera componente principal se encuentra ubicada entre la regresión ordinaria de y_1 sobre y_2 y la regresión ordinaria de y_2 sobre y_1.

El mismo ejemplo con R

Volvemos a considerar el mismo conjunto de datos ($n = 25$ objetos y $p = 2$ parámetros)

y_1	191	195	181	183	176	208	189	197	188	192	179	183	174
y_2	155	149	148	153	144	157	150	159	152	150	158	147	150
y_1	190	188	163	195	186	181	175	192	174	176	197	190	
y_2	159	151	137	155	153	145	140	154	143	139	167	163	




```
> datos <- read.table('datos.dat',header=FALSE)
> modelo <- prcomp(datos)
> barplot(modelo$sdev^2,names.arg=c("PC1","PC2"),
+ xlab="Principal Components",
+ ylab="Variances",main="modelo")
> print(modelo)
Standard deviations:
[1] 11.468144 4.258521

Rotation:
           PC1        PC2
V1 -0.8249295 -0.5652357
V2 -0.5652357 +0.8249295
> summary(modelo)
Importance of components:
                            PC1    PC2
Standard deviation      11.4681 4.2585
Proportion of Variance   0.8788 0.1212
Cumulative Proportion    0.8788 1.0000
```


¡El PCA no es invariante de escala!

- Dado que el PCA se basa en la diagonalización de la matriz muestral de covarianzas $\boldsymbol{S}$, es sensible a un cambio de escala en alguno de los parámetros $\mathcal{Y}$. Por tanto, las componentes principales no son invariantes bajo cambios de escala.
- Siempre que sea posible, las variables bajo estudio deben expresarse en unidades comparables.
- Si las variables tienen escalas muy distintas, pueden estandarizarse antes del cálculo de la componentes principales. Esto es equivalente a calcular las componentes principales de la matriz muestral de correlación $\boldsymbol{R}$.

  Es importante resaltar que los resultados que se obtienen al calcular las componentes principales a partir de la matriz muestral de covarianzas $\boldsymbol{S}$ o a partir de la matriz muestral de correlación $\boldsymbol{R}$ son diferentes:

$$\boldsymbol{S} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix} \quad \text{vs.} \quad \boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix},$$

  donde $\boldsymbol{R} = \boldsymbol{D}_s^{-1}\boldsymbol{S}\boldsymbol{D}_s^{-1}$, con $\boldsymbol{D}_s = \text{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \dots, \sqrt{s_{pp}})$.

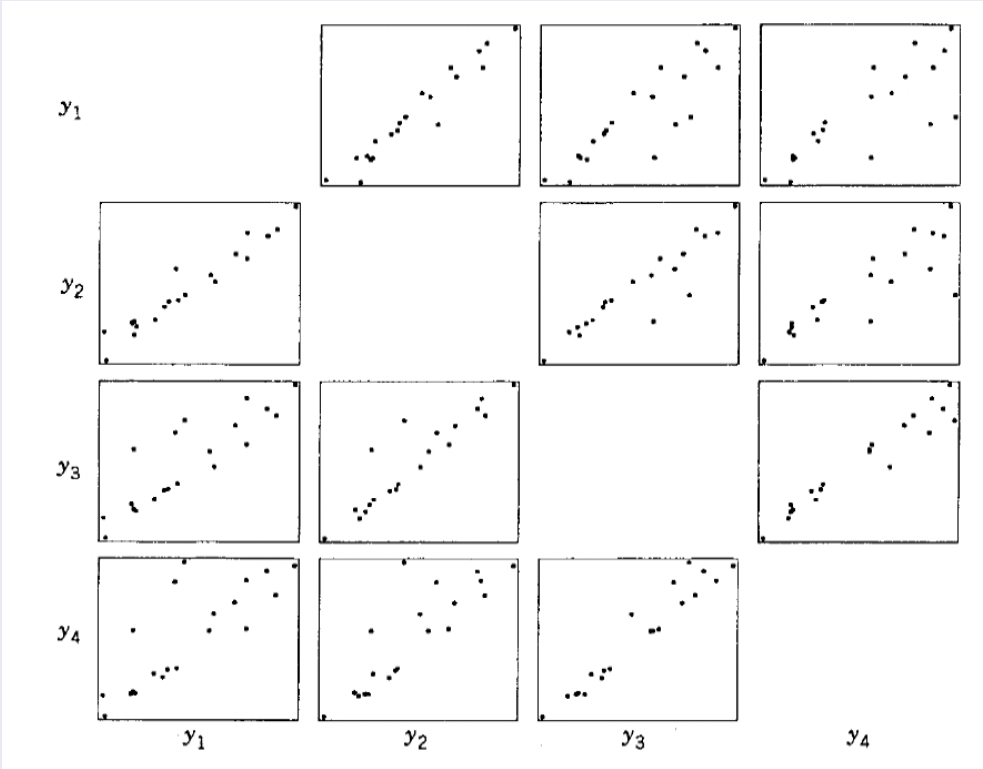
Utilidad del PCA

- En algunas aplicaciones el PCA constituye un objetivo en sí mismo y es objeto de interpretación.
- Otras veces es simplemente una herramienta que permite reducir la dimensionalidad de un conjunto de datos que posteriormente puede ser objeto de un análisis estadístico posterior.

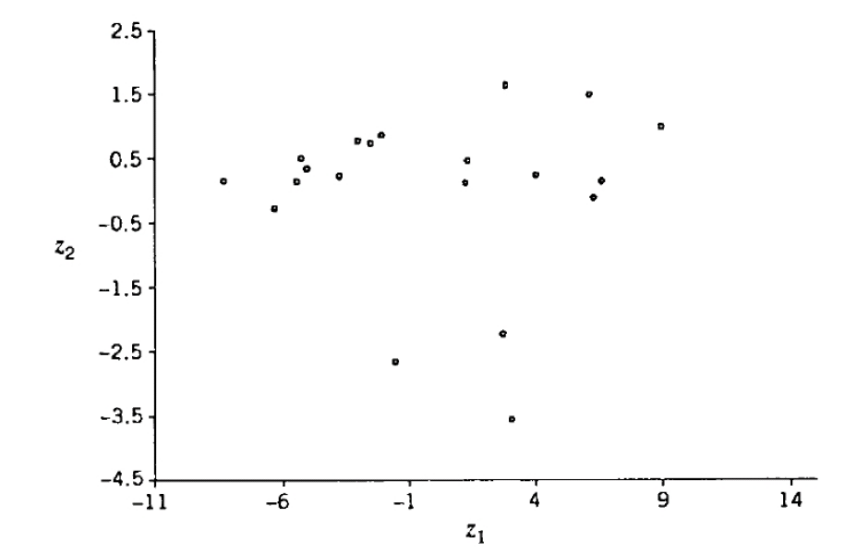
Ejemplos de reducción de la dimensionalidad

- Análisis de regresión: cuando el número de variales es muy grande comparado con el número de observaciones (los tets pueden ser ineficientes o imposibles de realizar).
- Análisis de regresión: cuando las variables independientes están muy correlacionadas (las estimaciones de los coeficientes de regresión son inestables).
- Chequeo de normalidad multivariada, presencia de “outliers”,. . . (a través de diagramas de dispersión de las primeras dos componentes).

Detección de “outliers” (diagrama de dispersión)

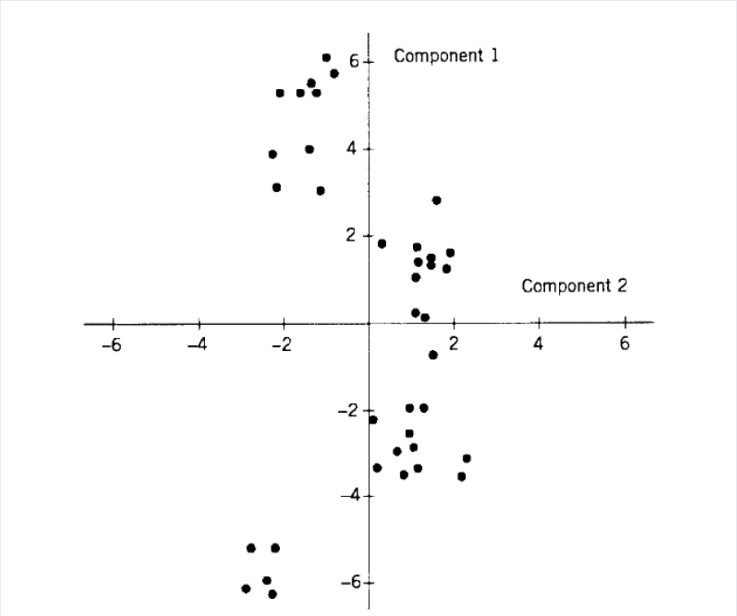


Detección de “outliers” (primeras 2 componentes principales)



¡Los “outliers” no aparecen al examinar las cuatro variables por separado!

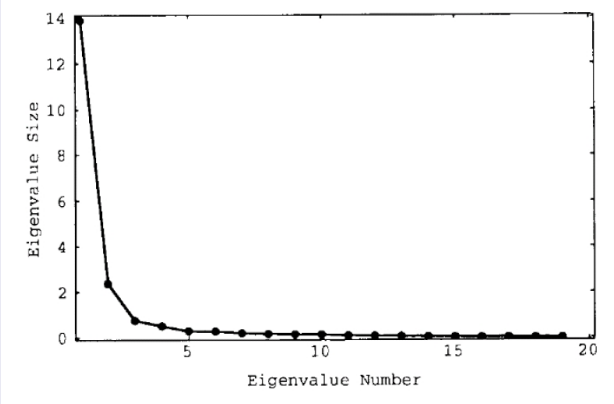
Detección de agrupaciones



Detección de 4 grupos en el estudio de 19 propiedades en 40 objetos. En este ejemplo la proporción de varianza explicada por las dos primeras componentes es del 85%, por lo que la representación gráfica de z_1 y z_2 muestra la información existente en los datos con poca distorsión.

Indicaciones generales

1. Retener suficientes componentes para garantizar un porcentaje predefinido de la varianza total, por ejemplo el 80%.
2. Retener aquellas componentes cuyos autovalores superen el promedio de todos los autovalores, $\sum_{i=1}^{p} \lambda_i/p$. Para la matriz de correlación este promedio es 1.0.
3. Utilizar una representación gráfica de λ_i frente a i, y determinar el "codo" en el que se produce la transición entre los autovalores grandes y los pequeños.



Indicaciones generales (continuación)

4. Utilizar tests de significación.

   Un test preliminar que resulta útil es testear la completa independencia de las variables, por ejemplo en la matriz poblacional de covarianzas

   $$H_0 : \boldsymbol{\Sigma} = \text{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}),$$

   o lo que es equivalente, en la matriz poblacional de correlación

   $$H_0 : \boldsymbol{P}_\rho = \boldsymbol{I}.$$

   En este caso, el estadístico definido por

   $$u' = -[(n-1) - \frac{1}{6}(2p+5)] \ln(u),$$

   donde

   $$u = \frac{|\boldsymbol{S}|}{s_{11}s_{22}\ldots s_{pp}} = |\boldsymbol{R}|,$$

   sigue aproximadamente una distribución χ^2_f, con $f = \frac{1}{2}p(p-1)$.
   Se rechaza H_0 **si** $u' > \chi^2_{\alpha,f}$.

   Si el test indica que las variables son independientes, no tiene sentido realizar un análisis de componentes principales.

Indicaciones generales (continuación)

4. Utilizar tests de significación.

   Para testear la significación de las componentes principales, se realiza la hipótesis nula de que los últimos k autovalores son pequeños e iguales, $H_{0k} : \gamma_{p-k+1} = \gamma_{p-k+2} = \ldots = \gamma_p$, donde $\gamma_1, \gamma_2, \ldots, \gamma_p$ son los autovalores poblacionales, es decir, los autovalores de $\mathbf{\Sigma}$.

   Para testear H_{0k} se calcula el promedio de los últimos k autovalores

   $$\overline{\lambda} = \sum_{i=p-k+1}^{p} \frac{\lambda_i}{k},$$

   y se calcula el estadístico

   $$u = \left(n - \frac{2p+11}{6}\right)\left(k\ln(\overline{\lambda}) - \sum_{i=p-k+1}^{p} \ln(\lambda_i)\right),$$

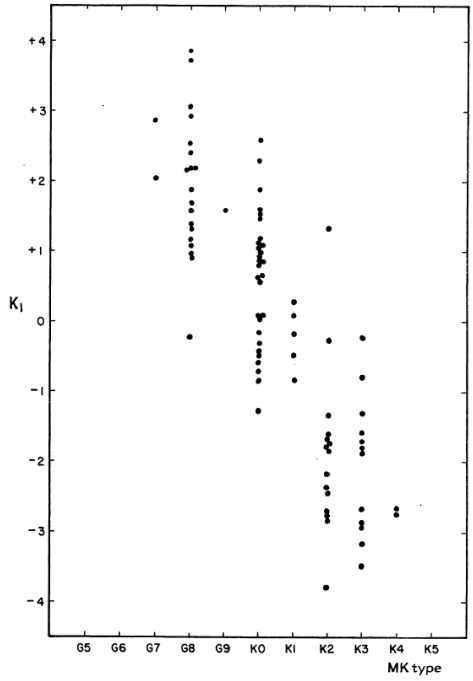
   que sigue aproximadamente una distribución χ^2_ν, con $\nu = \frac{1}{2}(k-1)(k+2)$.
   Se rechaza H_0 si $u \geq \chi^2_{\alpha,\nu}$.

   Normalmente se empieza con $H_{02} : \gamma_{p-1} = \gamma_p$. Si se acepta, se sigue con $H_{03} : \gamma_{p-2} = \gamma_{p-1} = \gamma_p$, y se sigue testeando hasta que H_{0k} se rechaza para algún valor de k.

¿Significado de las componentes?

- Las componentes principales se obtienen por rotación de ejes en el espacio de parámetros, proporcionando unas nuevas variables que no están correlacionadas y que reflejan las direcciones de máxima varianza. Estas direcciones no tienen por qué tener una interpretación evidente.
- Si las componentes resultantes no pueden interpretarse fácilmente, pueden rotarse buscando anular el mayor número de coeficientes de las combinaciones lineales para simplificar la intepretación. Sin embargo, las nuevas componentes rotadas volverán a estar correlacionadas y ya no suministrarán direcciones de máxima varianza (ya no serán componentes principales).
- Cuando la interpretación de los datos sea el objetivo fundamental (y no la reducción de su dimensionalidad), el análisis de factores es una técnica alternativa más útil.

- *Stellar Spectral Classification. I. Application of Component Analysis*, T.J. Deeming (1963).



Utilización de 5 índices de intensidad de línea en 84 estrellas gigantes de tipos espectrales G y K para la realización de una clasificación espectral:

1. línea K de Ca II
2. banda CNλ4200
3. Mgb
4. Hα
5. el triplete de Ca Iλ6102–6162)

Se obtiene que λ_1 es claramente mayor que $\lambda_2, \lambda_3, \lambda_4$ y λ_5. Como las estrellas se restringen a un intervalo en magnitud absoluta, es razonable interpretar la primera componente (K_1) como un parámetro indicativo del tipo espectral.

TABLE IV

Correlation matrix

1·000	0·285	0·546	0·463	0·497
0·285	1·000	0·700	0·603	0·812
0·546	0·700	1·000	0·558	0·761
0·463	0·603	0·558	1·000	0·615
0·497	0·812	0·761	0·615	1·000

TABLE V

Eigenvalues and eigenvectors

$k=$		1	2	3	4	5
λ_k		3·371	0·752	0·479	0·246	0·152
a_{ki}	$i=1$	0·354	0·850	−0·110	−0·277	−0·251
	2	0·463	−0·496	−0·085	−0·288	−0·670
	3	0·479	0·013	−0·403	0·779	0·037
	4	0·429	0·022	0·879	0·187	0·088
	5	0·497	−0·175	−0·212	−0·446	0·692

- *Variations in spectral-energy distributions and absorption-line strengths among elliptical galaxies*, S.M. Faber (1973).

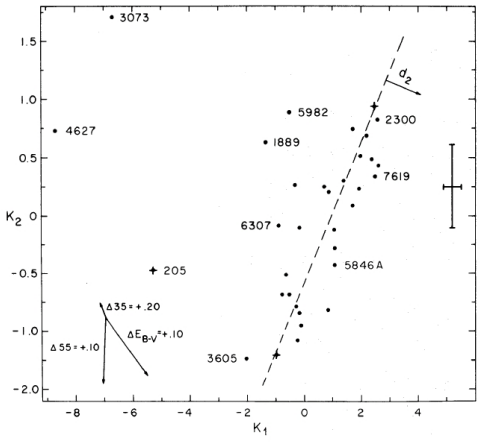


TABLE 9

COORDINATES FOR THE ANALYSIS INTO PRINCIPAL COMPONENTS

Coordinate	Index	Coordinate	Index
x_{j1}	$(35-55)_0$	x_{j6}	$(Mg)_0$
x_{j2}	$(LB)_0$	x_{j7}	$(TiO)_0$
x_{j3}	$(CN)_0$	x_{j8}	$(67-55)_0$
x_{j4}	$(G)_0$	x_{j9}	$(74-55)_0$
x_{j5}	$(45-55)_0$		

λ_1, λ_2, λ_3 y λ_4 son significativamente mayores que las varianzas esperadas por los errores (Q_i). Pero λ_3 y λ_4 son significativos sólo si los colores de M31, M32 y NGC205 se incluyen. Como estos últimos son inciertos, entonces sólo parecen significativas las 2 primeras componentes.

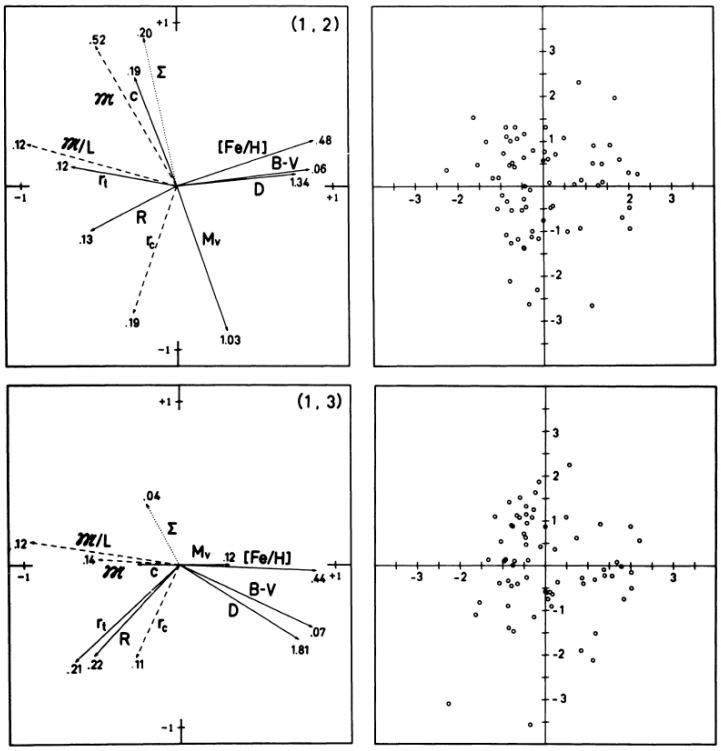
⇒ Sólo hacen falta dos parámetros para especificar completamente los colores de las galaxias elípticas estudiadas.

TABLE 10

ANALYSIS INTO PRINCIPAL COMPONENTS: ALL GALAXIES

i	λ_i	Q_i	λ_i/Q_i	v_{i1}	v_{i2}	v_{i3}	v_{i4}	v_{i5}	v_{i6}	v_{i7}	v_{i8}	v_{i9}
1.....	6.560	0.033	197.00	+0.383	−0.327	+0.301	+0.374	+0.372	−0.344	+0.021	−0.361	−0.355
2.....	0.515	0.172	2.98	−0.018	+0.088	+0.849	−0.231	−0.306	−0.250	−0.023	+0.058	+0.237
3.....	0.434	0.137	3.16	−0.028	+0.692	+0.028	−0.244	+0.044	+0.005	−0.280	−0.408	−0.460
4.....	0.122	0.043	2.80	+0.461	+0.512	+0.067	+0.303	+0.376	+0.135	−0.032	+0.177	+0.484
5.....	0.011	0.010	1.03	−0.754	+0.070	+0.100	+0.319	+0.343	−0.082	−0.006	−0.313	+0.301
6.....	0.022	0.027	0.81	+0.156	−0.220	−0.113	−0.695	+0.396	−0.048	+0.008	−0.362	+0.373
7.....	0.215	0.149	1.43	+0.008	−0.223	+0.345	+0.031	+0.067	+0.878	−0.141	−0.162	−0.086
8.....	1.064	0.641	1.67	+0.009	+0.200	+0.064	−0.044	−0.024	+0.139	+0.945	−0.174	−0.095
9.....	0.060	0.047	1.25	−0.211	+0.014	+0.188	−0.257	+0.585	−0.013	+0.069	+0.617	−0.351

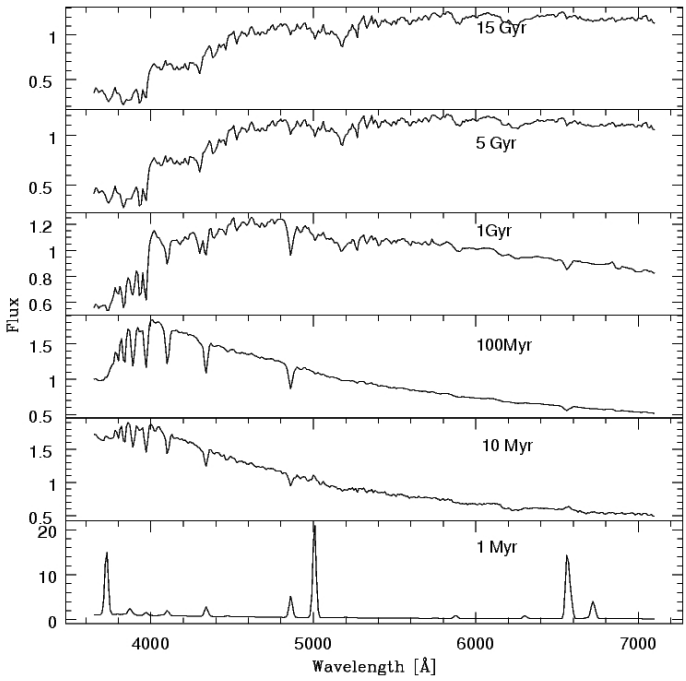
- *The manifold of globular clusters*, Brosche & Lentes (1984).



Our sample consists of the data of Harris and Racine (1979) for those globular clusters for which each of the following quantities is known (94 of a total of 150 clusters):
(1) The galactocentric distance R in kpc
(2) The tidal radius r_t in pc
(3) The concentration index $c = \log(r_t/r_c)$
(4) The total absolute visual magnitude M_V
(5) The colour $B-V$ corrected for interstellar reddening
(6) The metallicity [Fe/H]
(7) A horizontal branch type number D. This quantity is from Straižys (1982) and limits the sample to $N = 67$.

El número de parámetros significativos es $p = 2$ (el tercer autovalor es sólo marginalmente significativo).
No se detectan agrupaciones en el plano PC1, PC2.

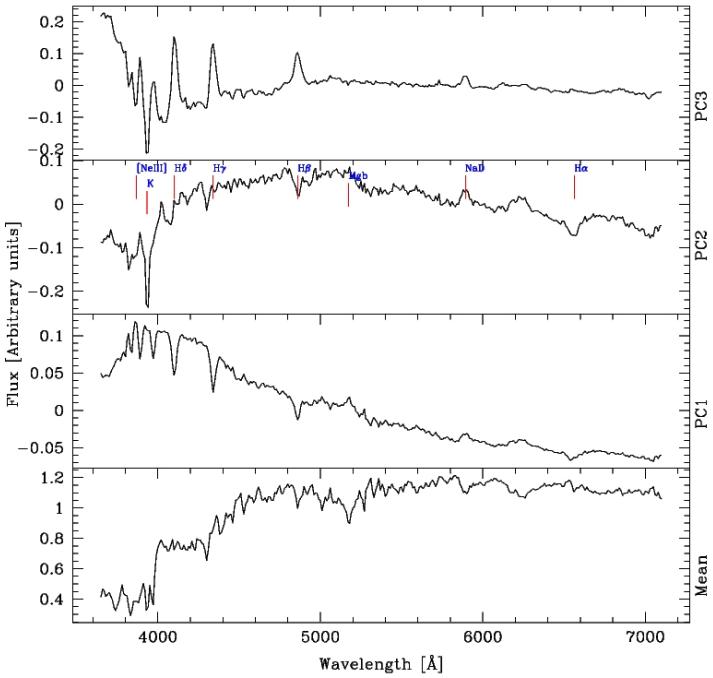
- *Analysis of synthetic galaxy spectra*, Ronen, Aragón-Salamanca, & Lahav (1999).



Estudio de 1850 espectros simulados con PEGASE, considerando brotes instantáneos que ocurren a $t = 0$ y con edades comprendidas entre 0.01 y 18.5 Gaños.

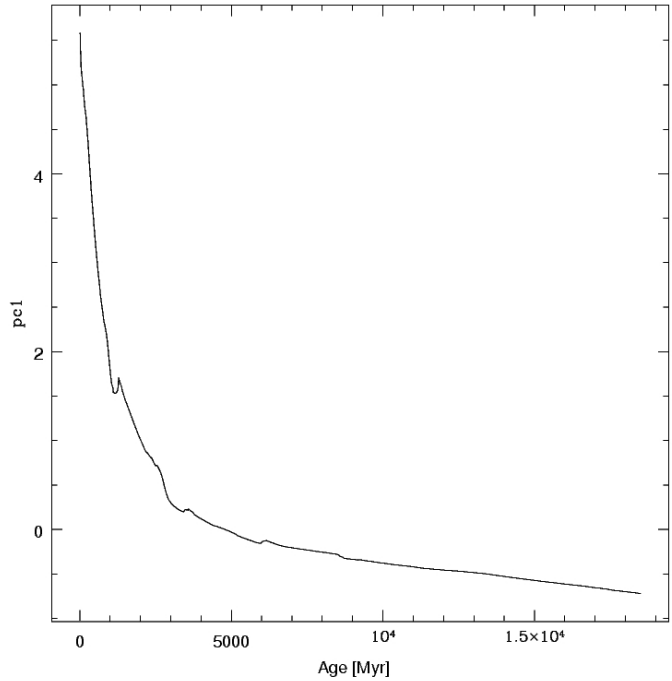
← Ejemplos de espectros simulados para diferentes edades.

- *Analysis of synthetic galaxy spectra*, Ronen, Aragón-Salamanca, & Lahav (1999).



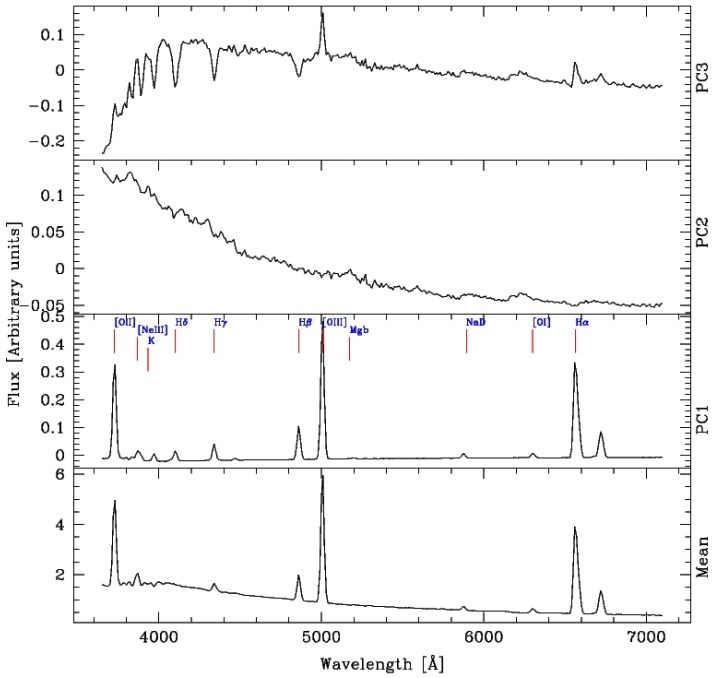
- PC1: explica el 98.5% de la variación en los espectros. Correlaciona el continuo azul con las absorciones de Balmer.
- PC2: sólo explica el 0.9% de la variación, por lo que las características espectrales reconocibles (como la línea K del Ca) tan sólo añaden información de segundo orden con respecto a PC1.
- PC3: explica el 0.5% de la variación. Correlaciona las líneas de Balmer con las absorciones por debajo de 4000 Å.

- *Analysis of synthetic galaxy spectra*, Ronen, Aragón-Salamanca, & Lahav (1999).



La proyección de PC1 frente a la edad muestra el claro enrojecimiento de las galaxias al hacerse éstas más viejas.

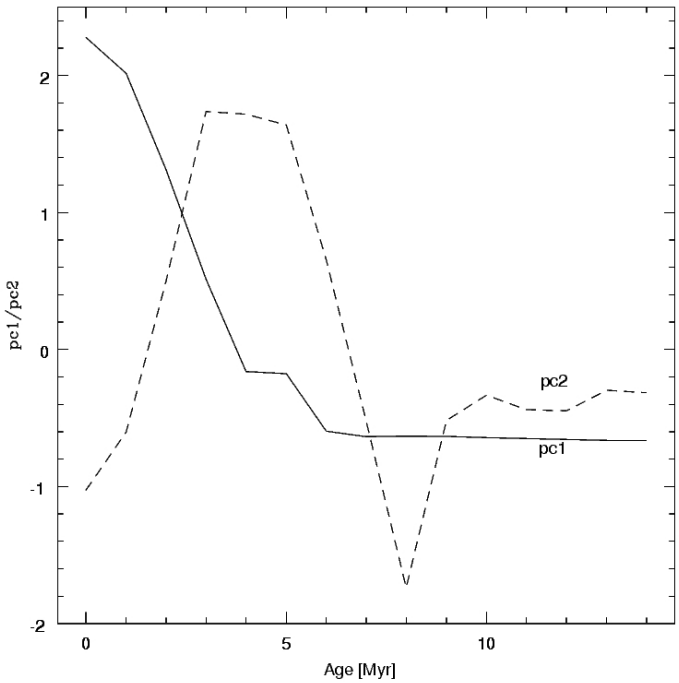
- *Analysis of synthetic galaxy spectra*, Ronen, Aragón-Salamanca, & Lahav (1999).



Brote joven: espectros simulados con edades comprendidas entre 0 y 14 Maños

- PC1: explica el 99.5% de la variación en los espectros. La información que contiene es básicamente las líneas de emisión.
- PC2: sólo explica el 0.7% de la variación, y se reduce de forma casi exclusiva a un continuo azul.
- PC3: explica el 0.05% de la variación. Muestra la correlación entre líneas de absorción y la disminución en el continuo por debajo de 4000 Å.

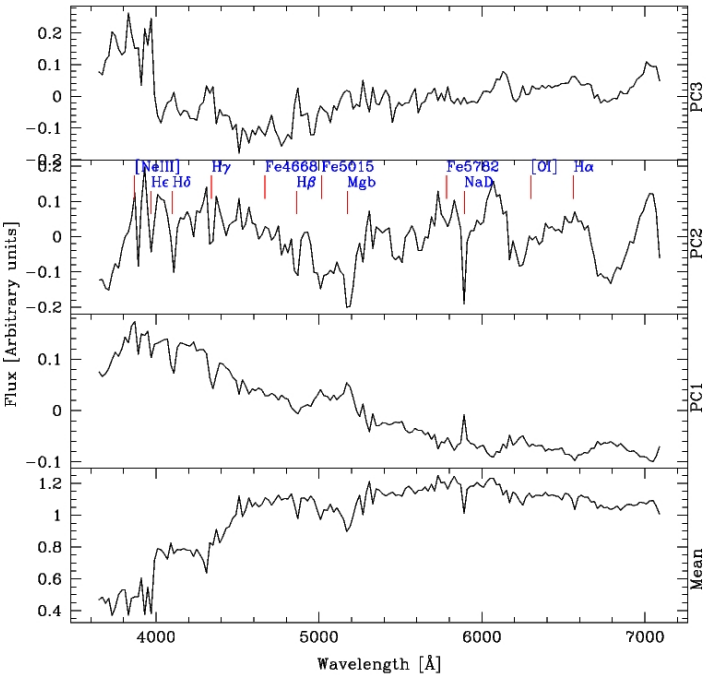
- *Analysis of synthetic galaxy spectra*, Ronen, Aragón-Salamanca, & Lahav (1999).



Brote joven: espectros simulados con edades comprendidas entre 0 y 14 Maños

La proyección de PC1 y PC2 frente a la edad muestra que las líneas de emisión dominan sólo para edades muy jóvenes, disminuyendo drásticamente por encima de $t = 6$ Maños. PC2 indica que sin embargo el continuo tiene un máximo entre 3 y 5 Maños después del brote, para caer bruscamente en $t = 8$ Maños.

- *Analysis of synthetic galaxy spectra*, Ronen, Aragón-Salamanca, & Lahav (1999).

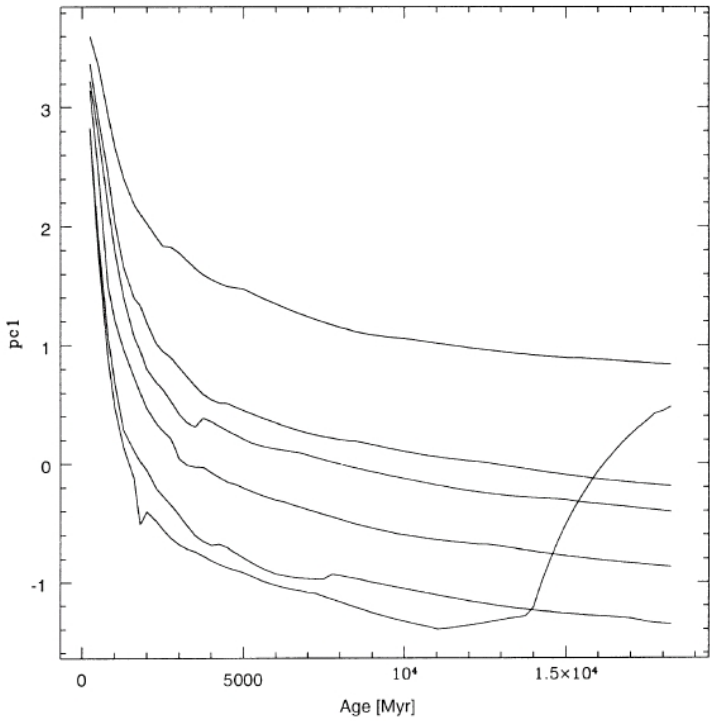


El efecto de la metalicidad

Simulación de espectros con distintas metalicidades, y edades comprendidas entre 100 y 1850 Maños.

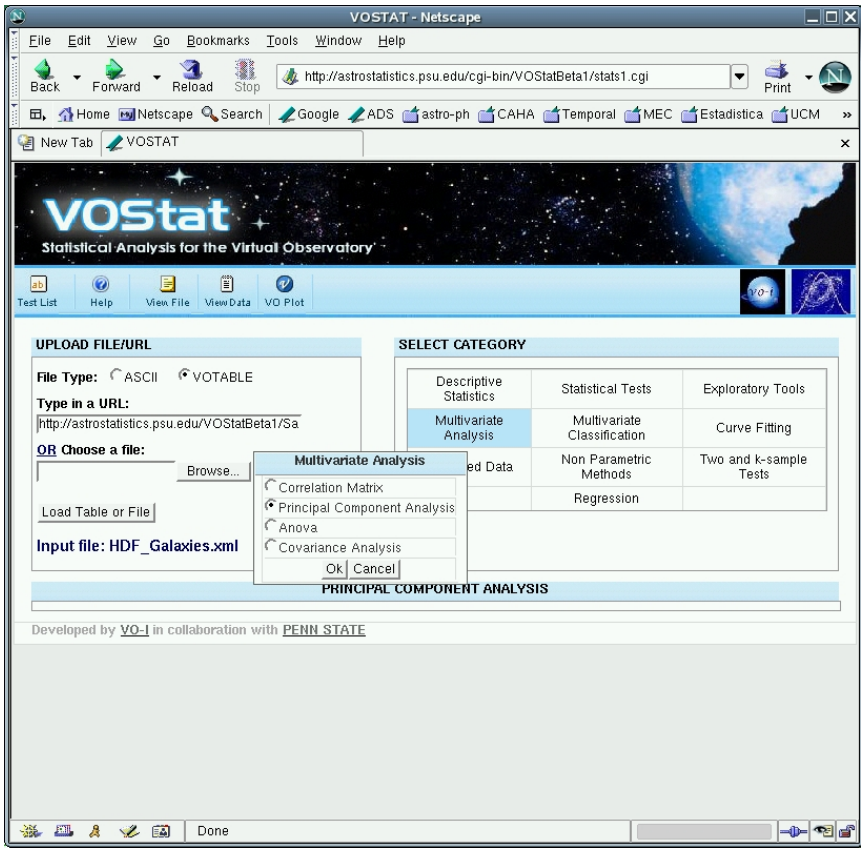
- PC1: explica el 96.41% de la variación en los espectros. Alguna de las líneas metálicas aparecen en "emisión" (e.g. NaD, Mgb, líneas de Fe). Son líneas de abosorción pero están anticorrelacionadas con las absorciones de Balmer y con los colores azules.
- PC2: sólo explica el 2.03% de la variación, y resulta muy diferente de lo que vimos anteriormente para metalicidad solar constante (tenemos mayor varianza debido precisamente a los efectos de la metalicidad). En este caso tenemos correlación positiva (mismo signo) para las absorciones metálicas (NaD, Mgb, Fe5015, Fe5782) y las absorciones de Balmer.
- PC3: explica el 0.34% de la variación y es más difícil de interpretar que en el caso de metalicidad solar.

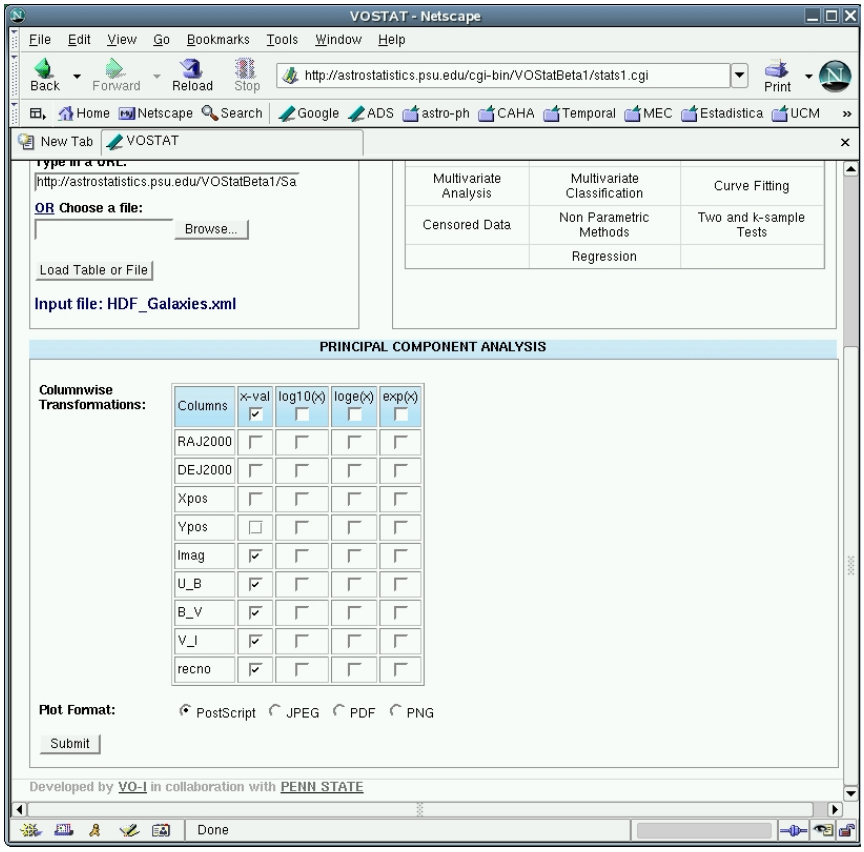
- *Analysis of synthetic galaxy spectra*, Ronen, Aragón-Salamanca, & Lahav (1999).

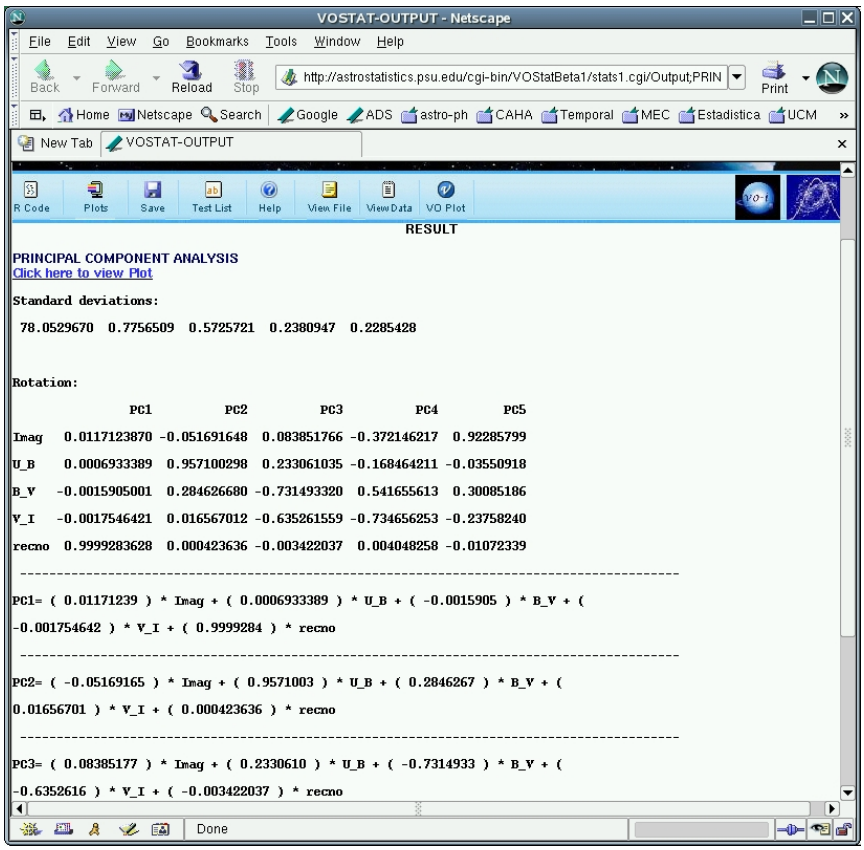


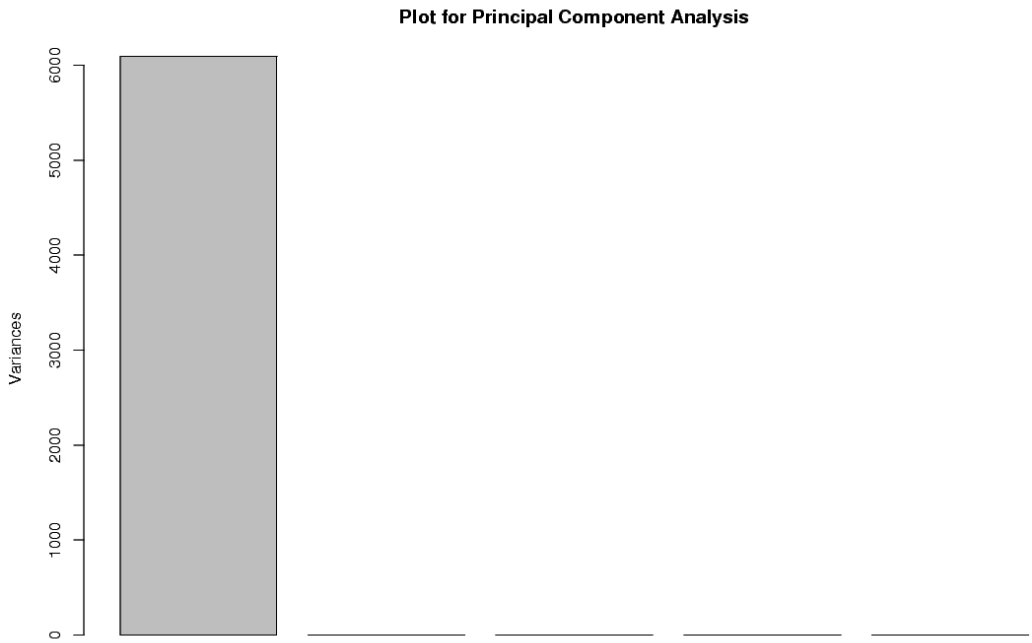
El efecto de la metalicidad

La proyección de PC1 frente a la edad para diferentes metalicidades ($Z = 0.1$, 0.05, 0.02, 0.008 y 0.004, de arriba a abajo). Como PC1 contiene un continuo azul, su valor disminuye (galaxias más rojas) a medidad que la edad aumenta. Sin embargo, en los modelos más metálicos se hacen más azules de nuevo a partir de $t = 14$ Gaños. Esto se explica asumiendo que en este caso las estrellas esquivan la fase AGB y se mueven rápidamente a la rama horizonal azul debido a la existencia de unos fuertes vientos estelares.


VOSTAT - Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://astrostatistics.psu.edu/cgi-bin/VOStatBeta1/stats1.cgi
Print
Home Netscape Search Google ADS astro-ph CAHA Temporal MEC Estadistica UCM
New Tab VOSTAT
VOStat
Statistical Analysis for the Virtual Observatory
Test List Help View File View Data VO Plot
UPLOAD FILE/URL
File Type: ASCII VOTABLE
Type in a URL:
http://astrostatistics.psu.edu/VOStatBeta1/Sa
OR Choose a file:
Browse...
Load Table or File
Input file: HDF_Galaxies.xml
SELECT CATEGORY
Descriptive Statistics
Statistical Tests
Exploratory Tools
Multivariate Analysis
Multivariate Classification
Curve Fitting
Non Parametric Methods
Two and k-sample Tests
Regression
Multivariate Analysis
Correlation Matrix
Principal Component Analysis
Anova
Covariance Analysis
Ok Cancel
PRINCIPAL COMPONENT ANALYSIS
Developed by VO-I in collaboration with PENN STATE
Done


VOSTAT - Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://astrostatistics.psu.edu/cgi-bin/VOStatBeta1/stats1.cgi
Print
Home Netscape Search Google ADS astro-ph CAHA Temporal MEC Estadistica UCM
New Tab VOSTAT
http://astrostatistics.psu.edu/VOStatBeta1/Sa
OR Choose a file:
Browse...
Load Table or File
Input file: HDF_Galaxies.xml
Multivariate Analysis
Multivariate Classification
Curve Fitting
Censored Data
Non Parametric Methods
Two and k-sample Tests
Regression
PRINCIPAL COMPONENT ANALYSIS
Columnwise Transformations:
Columns
x-val
log10(x)
loge(x)
exp(x)
RAJ2000
DEJ2000
Xpos
Ypos
Imag
U_B
B_V
V_I
recno
Plot Format:
PostScript JPEG PDF PNG
Submit
Developed by VO-I in collaboration with PENN STATE
Done


VOSTAT-OUTPUT - Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://astrostatistics.psu.edu/cgi-bin/VOStatBeta1/stats1.cgi/Output;PRIN
Print
Home Netscape Search Google ADS astro-ph CAHA Temporal MEC Estadistica UCM
New Tab VOSTAT-OUTPUT
R Code Plots Save Test List Help View File View Data VO Plot
RESULT
PRINCIPAL COMPONENT ANALYSIS
Click here to view Plot
Standard deviations:
78.0529670 0.7756509 0.5725721 0.2380947 0.2285428
Rotation:
PC1 PC2 PC3 PC4 PC5
Imag 0.0117123870 -0.051691648 0.083851766 -0.372146217 0.92285799
U_B 0.0006933389 0.957100298 0.233061035 -0.168464211 -0.03550918
B_V -0.0015905001 0.284626680 -0.731493320 0.541655613 0.30085186
V_I -0.0017546421 0.016567012 -0.635261559 -0.734656253 -0.23758240
recno 0.9999283628 0.000423636 -0.003422037 0.004048258 -0.01072339
PC1= (0.01171239) * Imag + (0.0006933389) * U_B + (-0.0015905) * B_V + (
-0.001754642) * V_I + (0.9999284) * recno
PC2= (-0.05169165) * Imag + (0.9571003) * U_B + (0.2846267) * B_V + (
0.01656701) * V_I + (0.000423636) * recno
PC3= (0.08385177) * Imag + (0.2330610) * U_B + (-0.7314933) * B_V + (
-0.6352616) * V_I + (-0.003422037) * recno
Done


Plot for Principal Component Analysis
Variances
0
1000
2000
3000
4000
5000
6000

Referencias

- Babu G.J., Feigelson E.D., 1996, *Astrostatistics*, Chapman & Hall, London
- Brosche P., Lentes F.-T., 1984, *The manifold of globular clusters*, A&A, 139, 474
- Deeming T.J., 1964, *Stellar spectral classification*, MNRAS, 127, 493
- Faber S.M., 1973, *Variations in spectral-energy distributions and absorption-line strengths among elliptical galaxies*, ApJ, 179, 731
- Francis P.J., Wills B.J., 1999, *Introduction to Principal Components Analysis*, ASP Conference Series, 162, 363
- Rencher A.C., 2002, *Methods of multivariate analysis*, 2nd edition, John Wiley & Sons
- Ronen S., Aragón-Salamanca A., Lahav O., 1999, *Principal component analysis of synthetic galaxy spectra*, MNRAS, 303, 284
- Wall J.V., Jenkins C.R., 2003, *Practical statistics for astronomers*, Cambridge University Press

Página WEB de VOStat: `http://vo.iucaa.ernet.in/~voi/VOStat.html`