PCA

Pricipal Component Analysis

(PCA) Análisis de Componentes Principales

- Técnica multivariante que trata de reducir el número de variables originales (X1, X2,...,Xn) a un número menor de variables (CP1, CP2,...,CPp), denominadas *componentes principales*
- Son una combinación lineal de las variables iniciales y sintetizan la mayor parte de la información contenida en los datos originales

$$CP1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1n}X_n$$

…

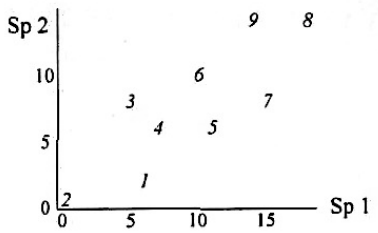
$$CPn = a_{n1}X_1 + a_{n2}X_2 + \ldots + a_{nn}X_n$$

donde aij es el coeficiente correspondiente a la especie j en el componente i.

(PCA) Análisis de Componentes Principales

- Ejemplo:

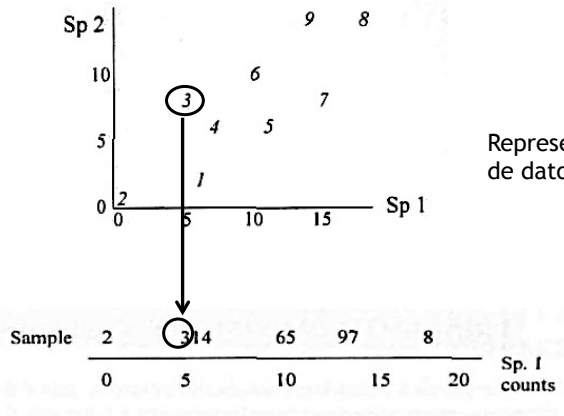
Muestra	1	2	3	4	5	6	7	8	9
Sp.1	6	0	5	7	11	10	15	18	14
Sp.2	2	0	8	6	6	10	8	14	14



Representación 2D de datos 2D

(PCA) Análisis de Componentes Principales

- Proyección de 2D a 1D

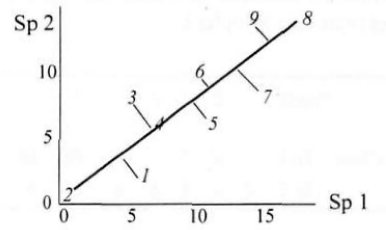


Representación 2D de datos 2D

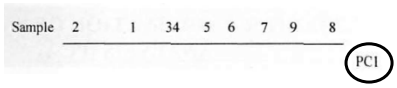
Representación 1D de datos 2D

(PCA) Análisis de Componentes Principales

- Mejor" ajuste



Representación 2D
de datos 2D

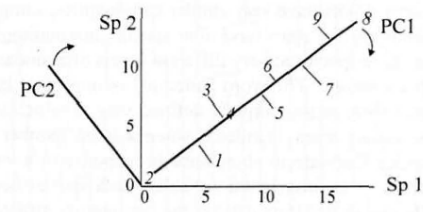


Evidentemente es más realista!!

Se denomina: Primer Componente Principal, PC1 (ó CP1)

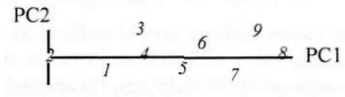
(PCA) Análisis de Componentes Principales

- "Mejor" ajuste



El 2º Componente Principal, PC2 (ó CP2) se define como el eje perpendicular (en el plano, 2D) al PC1.

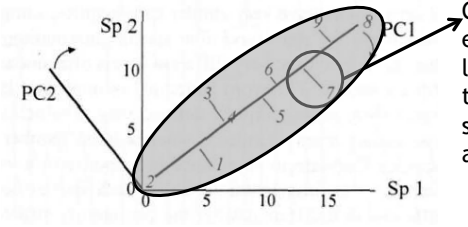
El análisis de componentes principales consiste en la rotación del gráfico 2D.



Para este caso, si no representamos el PC2 se pierde poca información y se reduce la dimensionalidad

(PCA) Análisis de Componentes Principales

- "Mejor" ajuste



Cuidado!! el "mejor" ajuste es SIMILAR al concepto de línea de regresión pero NO se toman distancias VERTICALES sino ajustes perpendiculares a la recta

Métodos de cálculo del Primer Componente Principal (PC1)
- Método de mínimos cuadrados
- Método de máximas diferencias entre las muestras (visual)

(PCA) Análisis de Componentes Principales

- Ejemplo 3D:

Muestra	1	2	3	4	5	6	7	8	9
Sp.1	6	0	5	7	11	10	15	18	14
Sp.2	2	0	8	6	6	10	8	14	14
Sp.3	3	1	6	6	9	11	10	16	15

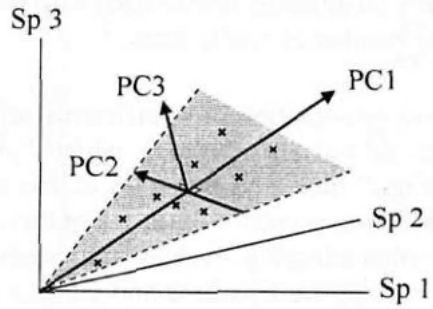
La definición de los Componentes Principales es:

PC1: es el eje que maximiza la varianza de los puntos proyectados perpendicularmente hacia él
PC2: está restringido (debe ser perpendicular a PC1) pero es elegido también como la dirección en la cual se maximiza la varianza de los puntos proyectados perpendicularmente hacia él
PC3: No hay elección, solo puede ser uno: perpendicular a PC1 y PC2

(PCA) Análisis de Componentes Principales

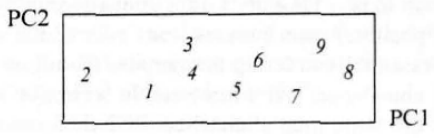
Podemos "visualizar" lo anterior en términos de "mejor ajuste":

PC1 es la LÍNEA de "mejor ajuste"
PC2 y PC1 forman un PLANO de "mejor ajuste"



(PCA) Análisis de Componentes Principales

PLANO de ordenación (Si podemos obviar el PC3!!)



Están en un plano (prácticamente en una línea)

En nuestro ejemplo, entre PC1 y PC2 explican el 99% de la varianza muestral, Por lo que obviar PC3 no afecta prácticamente al análisis final.

Podemos usar 2 Componentes principales (o incluso uno solo)

(PCA) Análisis de Componentes Principales

N- Dimensiones

- Para casos n-dimensionales la regla “no escrita” es intentar explicar al menos el 75% de la varianza total
- No es necesario usar la aproximación geométrica explicada, se recomienda el uso de tablas
- La perpendicularidad (ortogonalidad) se explicará en el modelo algebraico

(PCA) Análisis de Componentes Principales

Definición algebraica:

$$PC1 = 0.62 \times Sp.1 + 0.52 \times Sp.2 + 0.58 \times Sp.3$$
$$PC2 = -0.73 \times Sp.1 + 0.65 \times Sp.2 + 0.20 \times Sp.3$$
$$PC3 = 0.28 \times Sp.1 + 0.55 \times Sp.2 - 0.79 \times Sp.3$$

PC1 es la suma de las contribuciones (aproximadamente) iguales de las Sp. Además produce una ordenación de menor a mayor de las abundancias

PC2: a un nivel mas “fino” para muestras con una misma abundancia, valores altos de Sp.2 y bajos de Sp.1 (y viceversa)

PC3: similar al anterior pero con Sp.2 y Sp.3

(PCA) Análisis de Componentes Principales

Definición algebraica:

$$PC1 = 0.62 \times Sp.1 + 0.52 \times Sp.2 + 0.58 \times Sp.3$$
$$PC2 = -0.73 \times Sp.1 + 0.65 \times Sp.2 + 0.20 \times Sp.3$$
$$PC3 = 0.28 \times Sp.1 + 0.55 \times Sp.2 - 0.79 \times Sp.3$$

Ortogonalidad

$$(0.62) \times (-0.73) + (0.52) \times (0.65) + (0.58) \times (0.20) = 0$$

$$(0.62) \times (0.28) + (0.52) \times (0.55) + (0.58) \times (-0.79) = 0$$

etc.,

(PCA) Análisis de Componentes Principales

Notas adicionales

- Excluir variables menos comunes (especies con menor presencia o especies raras, etc) (Si fuera necesario incluirlas en el análisis se puede utilizar el Clustering)
- Transformar (raiz, log, etc) permite suavizar el efecto que pueden tener variables con mayor presencia en los datos
- Normalizar para igualar la participación de las variables en el cálculo de las componentes principales (dado que la varianza = 1) Se denomina *Correlation-based PCA* en contraste con el *covariance-based PCA*
- Para variables con magnitudes distintas es recomendable *Correlation-based PCA*

(PCA) Análisis de Componentes Principales

Fortalezas

- Es conceptualmente simple (si relacionamos la analogía geométrica de 2D)
- Es computacionalmente muy fácil (cálculo de matrices que actualmente están implementados en todos los procesadores matemáticos)
- Los ejes de ordenación son interpretables (a diferencia de MDS) dado que son combinaciones de valores de las variables

Debilidades

- Poca flexibilidad para definir la disimilaridad (distancia Euclídea) (transformaciones de variables son la única opción)
- La propiedad de preservar distancias es muy pobre (dado que son proyecciones sobre los ejes principales)
- No explica bien las relaciones no lineales (curvilíneas, etc.) entre variables
- Es sensible a la falta de normalidad