MA34B – Estadística
ACP

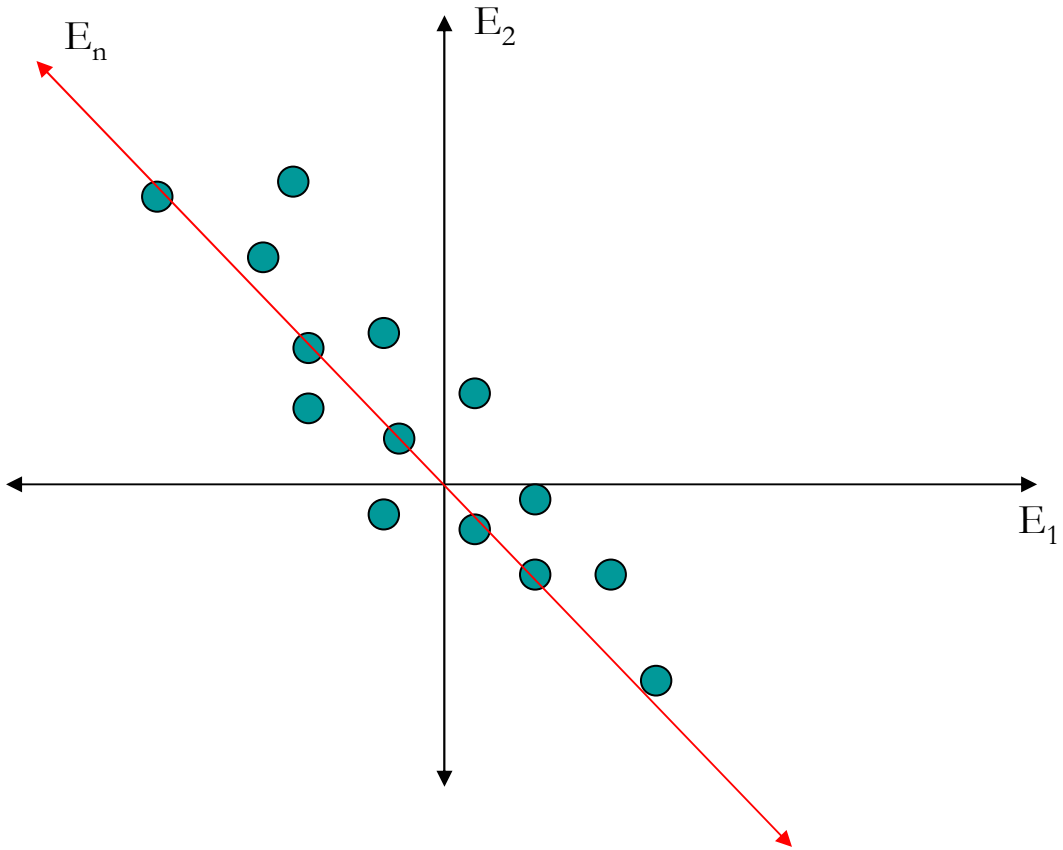
Prof. Rodrigo Abt B.

rabt@dim.uchile.cl

Introducción

- El Análisis de Componentes Principales (ACP) es una técnica que permite describir un conjunto de información que usualmente tiene muchas variables en función de un conjunto menor de variables no correlacionadas, tratando de preservar la mayor cantidad de la variabilidad posible de los datos originales, o sea resumir información sin perder mucha variabilidad.
- Las ventajas del ACP:
  - Reduce el espacio de variables a unas pocas que explican un alto porcentaje de la variabilidad de los datos.
  - Se gana interpretación
  - Las variables resultantes no están correlacionadas.
  - Se hace una reducción de la complejidad del problema
- Dentro de sus principales aplicaciones se encuentran: construcción de índices, reducción de colinealidad en modelos lineales y clasificación entre otras.

Gráficamente



Se busca encontrar un nuevo set de ejes en los cuales se acumule la mayor cantidad de variabilidad.

Planteamiento (1)

- Se tiene un conjunto de datos representado por "p" variables $X_1, X_2, \ldots, X_p$ que previamente estarán centradas, es decir que a cada una se le resta su promedio. Por lo tanto se tendrá una matriz de "n" observaciones (filas), y "p" columnas (variables), en que el promedio por variable es 0 (por el hecho de estar centradas).
- Se busca un nuevo conjunto de variables denomindas "componentes principales", no correlacionadas, de modo que la varianza que acumulen sea la mayor posible.
- Los ejes en los cuales se representen dichas variables se obtienen encontrando las sucesivas direcciones del plano que concentran la máxima varianza.
- Las nuevas variables son resultado de una combinación lineal de las variables originales.

Planteamiento (2)

- Se busca entonces coeficientes a_{ij} tales:

$$Y_1 = Xa_1 = X_1a_{11} + X_2a_{21} + \ldots + X_pa_{p1}$$

$$Y_2 = Xa_2 = X_1a_{12} + X_2a_{22} + \ldots + X_pa_{p2}$$

$$\vdots$$

$$Y_p = Xa_p = X_1a_{1p} + X_2a_{2p} + \ldots + X_pa_{pp}$$

- En que Y_1 representa la variable que acumula la máxima varianza. Luego le sigue Y_2 que es la variable que acumula la mayor varianza del resto que no pudo representar Y_1, y así sucesivamente. El objetivo es quedarse con un conjunto menor de k (k<p) variables que expliquen un alto porcentaje de la varianza total.

Solución (1)

- Veamos la varianza de Y_1:

$$Var(Y_1) = Var(Xa_1) = a_1^t \left(\frac{1}{n} X^t X\right) a_1 = a_1^t \Sigma a_1$$

- Se resuelve entonces el problema de maximización:

$$\max \quad a_1^t \Sigma a_1$$

$$\text{s.a. } ||a_1||^2 = a_1^t a_1 = 1$$

- La condición impuesta sobre a_1 es para obtener unicidad de la solución.
- Usando el método de los lagrangeanos:

NOTA: La varianza calculada aquí es la MUESTRAL, no confundir con la teórica

Solución (2)

- Sea: $Q = a_1^t \Sigma a_1 - 2\lambda(a_1^t a_1 - 1)$
- Derivando con respecto a a_1, se tiene: $\Sigma a_1 - \lambda a_1 = 0$
- Que no es otra cosa que el método para encontrar los vectores propios de Σ y sus valores propios asociados.
- Luego el vector director a_1 del primer eje principal corresponde al primer vector propio de Σ.
- Los siguientes vectores directores se obtienen resolviendo:

$$\max \quad a_2^t \Sigma a_2$$

$$\text{s.a. } ||a_2||^2 = a_2^t a_2 = 1 \quad \text{y} \quad a_2^t a_1 = 0$$

Observaciones Al Método

- Debe considerarse que Σ como se define es la matriz de varianzas y covarianzas de muestrales de X, es decir:

$$\Sigma = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 & \frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) & \cdots & \cdots \\ \frac{1}{n}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1) & \frac{1}{n}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2 & \cdots & \cdots \\ \vdots & \vdots & \cdots & \vdots \\ \cdots & \cdots & \cdots & \frac{1}{n}\sum_{i=1}^{n}(x_{ip}-\bar{x}_p)^2 \end{pmatrix}$$

Propiedades (1)

- Con este mecanismo podemos generar "p" duplas valor-vector propio: (λ_1,e_1), (λ_2,e_2), …, (λ_p,e_p), en que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.
- Los valores propios son todos reales no negativos, ya que la matriz Σ es simétrica y semidefinida positiva.
- Se tiene además que $Var(Y_j) = \lambda_j$ (j=1,…,p).
- Con esta solución se conserva la varianza generalizada:

$$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} Var(Y_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

Propiedades (2)

- Como las nuevas variables no son correlacionadas, la matriz de varianzas y covarianzas de las componentes principales es diagonal, cuyos términos son los valores propios.
- Se puede construir un índice que explica la cantidad de variabilidad que aporta cada componente como:

$$\phi_h = \frac{\lambda_h}{\sum_{i=1}^{p} \lambda_i}$$

- Y se puede calcular la variabilidad acumulada hasta la componente "h" como:

$$\Phi_h = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_h}{\sum_{i=1}^{p} \lambda_p}$$

Propiedades (3)

- La matriz de componentes principales se puede escribir como:

$$Y = XA$$

en que las columnas de A son los vectores propios obtenidos (que definen los llamados "ejes principales"), y las columnas de Y contienen las componentes principales. Además $A^tA=I$.

- La covarianza entre las componentes principales y las variables originales centradas es:

$$Cov(Y_i, X_j) = \lambda_i a_{ij}$$

Propiedades (4)

- La correlación entre las componentes principales y las variables originales centradas es:

$$Corr(Y_i, X_j) = \frac{Cov(Y_i, X_j}{\sqrt{Var(Y_i)Var(X_j)}} = a_{ij}\frac{\sqrt{\lambda_i}}{S_j}$$

- Estas correlaciones permiten construir el denominado "Círculo de Correlaciones", el que se utiliza para encontrar el aporte y contribución de cada variable en la construcción de cada componente.

Método Alternativo (1)

- Otra manera usual de resolver el problema de las componentes principales, es no solamente centrar las variables originales, sino que estandarizarlas, es decir, dividir cada variable por su desviación estándar, en cuyo caso la matriz de varianzas y covarianzas se convierte en la Matriz de Correlaciones de X, o sea R.
- La ventaja de este enfoque es que nos independizamos de las unidades, y con ello se elimina el problema de la influencia de las variables con mayor varianza en la determinación de las componentes. Adicionalmente, se simplifican los cálculos, y se mejora la interpretación.

Método Alternativo (2)

- En este caso se resuelve:

$$Ra = \lambda^* a$$

- Se cumple entonces:

$$\sum_{i=1}^{p} Var(Z_i) = \text{Traza}(R) = \sum_{i=1}^{p} \lambda_i^* = p$$

en que Z_i es la variable X_i estandarizada.

Varianza Explicada

- La proporción de varianza explicada por cada componente es:

$$\phi_h^* = \frac{\lambda_h^*}{p}$$

- Y la varianza acumulada explicada hasta la componente "h" es:

$$\Phi_h^* = \frac{\lambda_1^* + \lambda_2^* + \ldots + \lambda_h^*}{p}$$

Selección De Componentes

- Para determinar cuántas componentes se retendrán, se recurre a los siguientes criterios:
  - Graficar λ_i^* versus i, y buscar retener componentes hasta que los cambios entre cada punto sean más suaves.
  - Fijarse una cota de variabilidad, y elegir el número de componentes en orden hasta alcanzar o sobrepasar dicha cota.
  - Eliminar las componentes que sean inferiores a una determinada cota. Usualmente esta cota es la unidad.