

Resumen de Análisis de Componentes Principales

FRANCISCO SILVA

Supongamos que tenemos un vector aleatorio $X = (X_1, X_2, \dots, X_p)$ donde cada componente mide una característica de una población. Si se toman n muestras de X podemos formar la matriz Σ de $m \times p$ donde la componente Σ_{ij} indica el valor observado de la característica j en la i -ésima observación. La idea del análisis de componentes principales es reducir la dimensión p , que cuando p es grande implica una difícil interpretación de los datos, mediante combinaciones lineales de las p variables originales. A éstas nuevas variables, que tendrán correlación nula, se les llamará componentes principales.

Evidentemente es deseable que esta reducción no tenga como consecuencia una pérdida de información considerable, por lo tanto se busca una reducción tal que las nuevas variables concentren gran parte de la varianza de los datos originales. De este modo, si $C = \text{Var}(X)$ denota la matriz de covarianza de X (que es simétrica), entonces, si su diagonalización está dada por $C = G\Lambda G^t$ donde en la matriz Λ consideramos los valores propios ordenados de mayor a menor, se prueba que el vector de componentes principales está dado por $Y = G^t X$. Es decir, la i -ésima componente principal está dada por una combinación lineal de las variables originales, donde los coeficientes están dados por las coordenadas del i -ésimo vector propio.

Hasta el momento lo que se ha hecho es encontrar nada más que un cambio de coordenadas, pero no hemos reducido la dimensionalidad del problema. Sin embargo, un sencillo cálculo muestra que $\lambda_i = \text{Var}Y_i$ donde λ_i es el i -ésimo valor propio más grande. También es fácil probar que $\sum_{i=1}^p \text{Var}Y_i = \sum_{i=1}^p \text{Var}X_i = \sum_{i=1}^p \lambda_i$ y por lo tanto los valores propios concentran la varianza de las variables originales y la de las nuevas variables. Así pues, supongamos que λ_1 y λ_2 concentran la mayor parte de la varianza, esto es $\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^p \lambda_i} \times 100\% \approx 90\%$, entonces sabemos que las dos primeras componentes principales, Y_1 e Y_2 , acumulan más o menos el 90% de la varianza de **todas** las variables originales y por lo tanto una reducción razonable sería graficar los datos en estas nuevas coordenadas (no se perdería una cantidad importante de información).

En la práctica no se conoce la distribución del vector X y por lo tanto hay que trabajar con los datos, es decir se considera S_X , la matriz de covarianza empírica (haciendo el papel de C) y se diagonaliza esta matriz.

En la interpretación de los datos un rol importante lo juega el círculo de correlación. Se consideran las dos primeras componentes principales (donde en la práctica se visualizan los datos) Y_1 e Y_2 y se calculan los siguientes valores $\text{Corr}(X_i, Y_1) = g_{i1} \sqrt{\frac{\lambda_1}{s_{X_i X_i}}} \forall i \in 1, \dots, n$ y $\text{Corr}(X_i, Y_2) = g_{i2} \sqrt{\frac{\lambda_2}{s_{X_i X_i}}} \forall i \in 1, \dots, n$ y se grafican los puntos $(\text{Corr}(X_i, Y_1), \text{Corr}(X_i, Y_2))$ dentro del círculo unidad. Estos valores muestran el tipo de relación de las n nuevas con respecto a alguna de las antiguas. Por ejemplo si 2 de las antiguas se encuentran en la frontera del círculo esto indica una relación lineal entre las variables nuevas y esas 2 antiguas en especial.

Básicamente lo anterior es la materia de componentes principales, hay que notar que en el análisis anterior se supuso implícitamente que las variables se miden en una escala similar. Si este no fuese el caso es **indispensable** estandarizar los datos y trabajar con la matriz de correlación, en lugar de la de covarianza.