Análisis de componentes principales

Rocío Maehara Aliaga, PhD

Departamento de Ciencias
Pontifícia Universidad Católica del Perú

19 de octubre de 2018

Outline

Introducción

- Cuando se recoge la información de una muestra de datos, lo más frecuente es tomar el mayor número posible de variables. Evidentemente, en este caso es difícil visualizar relaciones entre las variables.
- Otro problema que se presenta es la fuerte correlación que muchas veces se presenta entre las variables: si tomamos demasiadas variables (cosa que en general sucede cuando no se sabe demasiado sobre los datos o sólo se tiene ánimo exploratorio).
- Se hace necesario, pues, reducir el número de variables. Es importante resaltar el hecho de que el concepto de mayor información se relaciona con el de mayor variabilidad o varianza.

Ejemplos

El análisis de componentes principales se aplica al cruce de tablas con individuos en fila y variables cuantitativas en columnas.

Dominio	**Individuos**	**Variables**	x_{ik}
Ecología	Río	Concentración de contaminantes	Concentración del contaminante k en el río i
Economía	Año	Indicadores económicos	Valor del indicador k en el año i
Genética	Paciente	Genes	Expresión del gen k para el paciente i
Marketing	Marca	Índices de satisfacción	Valor del índice k para la marca i
Pedología	Suelo	Composición granulométrica	Índice del componente k para el suelo i
Biología	Animal	Medidas	Medida k para el animal i

Componentes Principales

- Para estudiar las relaciones que se presentan entre p variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorreladas entre sí (que no tenga repetición o redundancia en la información) llamado conjunto de componentes principales.
- Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.
- De modo ideal, se buscan $m < p$ variables que sean combinaciones lineales de las p originales y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.

Componentes Principales

- El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos.
- Habitualmente, se calculan los componentes sobre variables originales estandarizadas, es decir, variables con media 0 y varianza 1. Esto equivale a tomar los componentes principales, no de la matriz de covarianzas sino de la matriz de correlaciones (en las variables estandarizadas coinciden las covarianzas y las correlaciones).

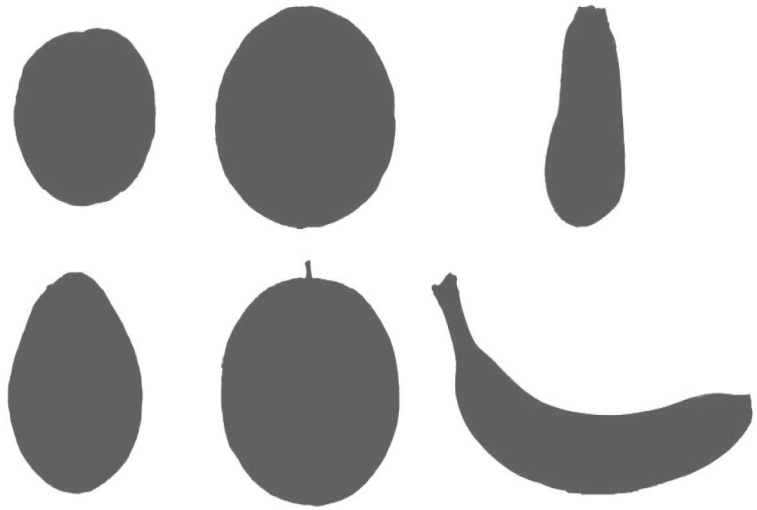
Componentes Principales

- El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos.
- Habitualmente, se calculan los componentes sobre variables originales estandarizadas, es decir, variables con media 0 y varianza 1. Esto equivale a tomar los componentes principales, no de la matriz de covarianzas sino de la matriz de correlaciones (en las variables estandarizadas coinciden las covarianzas y las correlaciones).

Componentes Principales

- Si las variables originales $x_1, ..., x_p$ están incorreladas, entonces carece de sentido calcular unos componentes principales. Si se hiciera, se obtendrían las mismas variables pero reordenadas de mayor a menor varianza.
- El cálculo de los componentes principales de una serie de variables $x_1, ..., x_p$ depende normalmente de las unidades de medida empleadas. Si transformamos las unidades de medida, lo más probable es que cambien a su vez los componentes obtenidos.
- Una solución frecuente es usar variables $x_1, ..., x_p$ tipificadas. Con ello, se eliminan las diferentes unidades de medida y se consideran todas las variables implícitamente equivalentes en cuanto a la información recogida

Componentes Principales



- ¿Qué es lo que diferencia los enfoques de la misma fruta entre la primera fila y la segunda? Las distancias están menos deformadas en los segundos enfoques y las representaciones ocupan mejor el espacio en la fotografía.
- El ACP vuelve a buscar el mejor espacio de representación (de dimensión reducida) que permite visualizar lo mejor posible la forma de una nube de K dimensiones.