11

La riduzione delle dimensioni

11.1 Analisi in componenti principali (PC): introduzione

Date p variabili quantitative X_1, X_2, . . . , X_p, l'obiettivo dell'analisi in componenti principali (PC, dall'inglese "Principal Components") è quello di sostituire alle variabili originali, che possono essere correlate, un nuovo insieme di variabili, dette componenti principali che hanno le seguenti proprietà (Jolliffe, 2002):

- sono incorrelate (ortogonali);
- sono elencate in ordine decrescente rispetto alla loro varianza.

La prima componente principale Y_1 è la combinazione lineare delle p variabili di partenza avente la massima varianza. La seconda componente principale Y_2 è la combinazione lineare delle p variabili con la varianza immediatamente inferiore alla varianza di Y_1 e ad essa incorrelata, ecc. fino alla p-esima componente. Se le p variabili originali sono molto correlate, un numero $q < p$ tiene conto di una quota elevata di varianza totale, per cui le prime q componenti forniscono una buona approssimazione di dimensione ridotta della struttura dei dati.

Lo scopo primario di questa tecnica è la riduzione di un numero elevato di variabili (rappresentanti altrettante caratteristiche del fenomeno analizzato) in alcune variabili latenti. Ciò avviene tramite una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano nel quale le variabili vengono ordinate in modo decrescente di varianza: pertanto, la variabile con maggiore varianza viene proiettata sul primo asse, la seconda sul secondo asse e così via. La riduzione della complessità avviene limitandosi ad analizzare le principali (per varianza) tra le nuove variabili.

Ci sono 3 approcci per arrivare a determinare le componenti principali:

- Proiezioni di punti in un sottospazio.
- Rappresentazione di una matrice di rango p con una matrice di rango ridotto.