PCR e PLS

Principal Component Regression

- Costruzione delle componenti principali $Z_1, \ldots, Z_p$
- Utilizzo di tali componenti come esplicative in un modello di regressione lineare.
- Selezione del numero ottimale di componenti

Principal Component Regression

- X: set delle variabili quantitative esplicative.
- $Z = XV$: set delle componenti principali
- V: Matrice dei pesi, costituita dagli autovettori della matrice di covarianza o correlazione.
- Stima dei coefficienti del seguente modello: $Y = Z\gamma + \varepsilon$
- Calcolo dei coefficienti espressi in funzione delle variabili originarie: $Z^{(a)}\gamma^{(a)} = XV^{(a)}\gamma^{(a)} = X\beta^{(a)}$.
- $\beta^{(a)} = V^{(a)}\gamma^{(a)}$

Principal Component Regression

Problema Principale

- Le "direzioni" delle componenti principali sono identificate in modo non supervisionato
- Le componenti con maggiore variabilitá non é detto che siano le migliori per spiegare la variabile risposta.

Alternativa

- Partial Least Sqaures

Partial Least Squares

- $Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$
- $y_i = \theta_0 + \sum_{m=1}^{M} \theta_m + z_{im} + \varepsilon_i$
- $\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$
- Come si ottengono i ϕ_{jm}?

Partial Least Squares

- Standardizzazione dei predittori
- Prima "direzione" (componente) Z_1 calcolata, impostando ogni ϕ_{j1} uguale al coefficiente della regressione lineare semplice di Y su X_j.
- Si dimostra che il coefficiente é proporzionale alla correlazione tra Y e X_j
- Nel calcolo $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$ il PLS attribuisce il massimo peso alle variabili maggiormente legate alla risposta
- Per identificare la seconda componente si "aggiusta" ogni variabile per Z_1 facendo la regressione di ogni variabile su Z_1 e prendendo i residui (che possono essere letti come le informazioni rimanenti che non sono spiegate dalla prima direzione PLS).
- Z_2 é calcolata utilizzando la stessa procedura usata per Z_1, ma sfruttando i residui.
- La procedura viene iterata fino a ottenere m componenti
- Scelta numero componenti
- Regressione minimi quadrati su tali componenti