Data Science II
(P8106)

Department of Biostatistics
Mailman School of Public Health
Columbia University

Methods using derived input directions

- A large number of inputs, often correlated
- Use a small number of linear combinations of the original inputs X_j, $j = 1, \ldots, p$
  - New predictors Z_m, $m = 1, \ldots, M$
  - Z_m are linear combinations of X_j
- Two approaches
  - Principal components regression (PCR)
  - Partial least squares (PLS)

Details

- Two steps: dimension reduction + regression
- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ linear combinations of our original p predictors

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j$$

  for some constants $\phi_{m1}, \phi_{m2}, \ldots, \phi_{mp}$
- We can then fit the linear regression model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, i = 1, \ldots, n \tag{1}$$

  using ordinary least squares

Details

Notice that

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{j=1}^{p} \beta_j x_{ij},$$

where

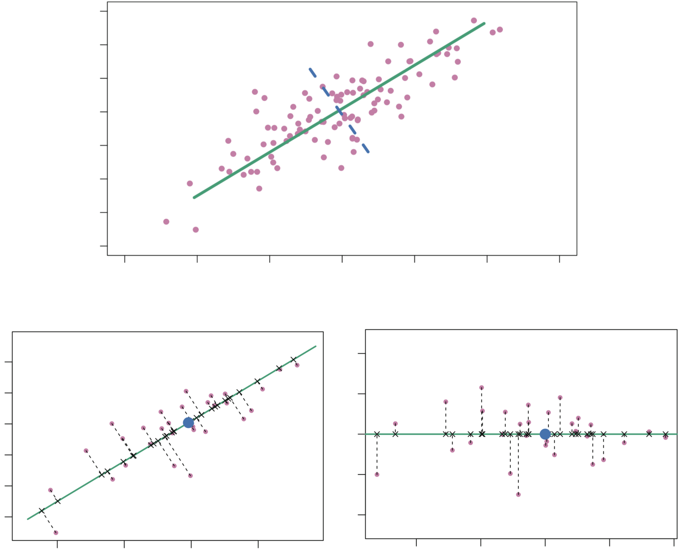
$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj} \tag{2}$$

- Model (1) is a special case of the original linear regression
- Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the form (2)

Principal components regression (PCR)

- We apply principal components analysis (PCA) to define the linear combinations of the predictors
- The first PC is the linear combination of the X variables that captures as much of the information as possible
- The second PC is the linear combination of X that captures as much of the information as possible and is uncorrelated with the first PC
- ...
- We replace correlated original variables with the first M PCs that capture the joint variation
- There is no sample covariance between different PCs over the dataset

Example of PCA



Details of PCA

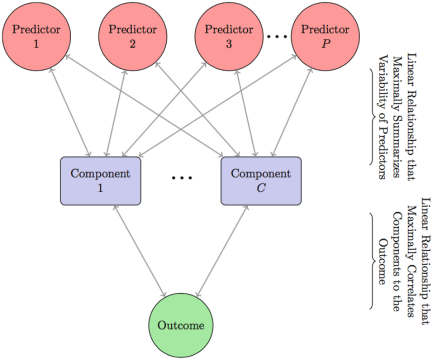
$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}$$

PCR: continued

- Choice of M: cross-validaton
- PCR identifies linear combinations that best represent the predictor $X_1, \ldots, X_p$
- The response does not supervise the identification of the principal components
- Drawback: there is no guarantee that Z_m are the best linear combinations of X_j in predicting the response

Partial least squares (PLS)

- PLS identifies these new features in a supervised way
- Make use of the response Y in order to identify new features that not only approximate the old features well, but also **are related to the response**



Applied Predictive Modeling, Fig 6.9

PLS

- ϕ_{mj} are selected so that Z_m have high variance and high correlation with response
- S is the sample covariance matrix
- PCR: $\phi_m = (\phi_{m1}, \phi_{m2}, \ldots, \phi_{mp})$ solves

$$\max_{\phi} \operatorname{Var}(\boldsymbol{X}\phi)$$
$$\text{subject to } \|\phi\| = 1, \phi^\intercal S \phi_l = 0, l = 1, \ldots, m-1$$

- PLS: $\phi_m = (\phi_{m1}, \phi_{m2}, \ldots, \phi_{mp})$ solves

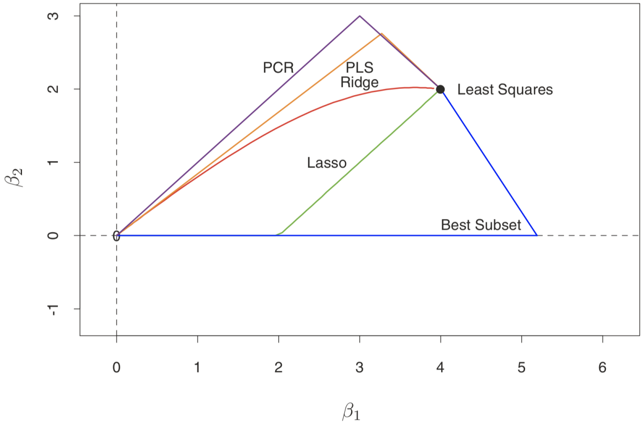
$$\max_{\phi} \operatorname{Var}(\boldsymbol{X}\phi)\operatorname{Corr}^2(\boldsymbol{y}, \boldsymbol{X}\phi)$$
$$\text{subject to } \|\phi\| = 1, \phi^\intercal S \phi_l = 0, l = 1, \ldots, m-1$$

Details

- Dimension reduction in PLS can be viewed as a supervised dimension reduction procedure; dimension reduction in PCR is an unsupervised procedure
- Choice of M?

An example

- True coefficients $(4, 2)$
- $\rho = 0.5$



Summary

- Linear regression and its cousins
  - Ridge regression
  - The lasso
  - Principal components regression
  - Partial least squares
- Next lecture: nonlinear models