Exercise 3.6

$$p(\beta \mid y) \propto p(y \mid \beta)\, p(\beta)$$

where

$$p(y\mid\beta) \rightarrow \mathcal{N}(X\beta, \sigma^2 I)$$
$$p(\beta) \rightarrow \mathcal{N}(0, \tau I)$$

$$\frac{1}{\sqrt{2\pi}^p \sigma^p} \quad \text{const}$$

$$p(\beta\mid y) \propto \text{const.}\ \exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right\} \exp\left\{-\frac{1}{2\tau}\beta^T\beta\right\}$$

$$\propto \exp\Big\{ \underbrace{(y-X\beta)^T(y-X\beta) + \underbrace{\frac{\sigma^2}{\tau}}_{\lambda}\beta^T\beta}\Big\} \cdot \left(-\frac{1}{2\sigma^2}\right)$$

Exercise 3.10

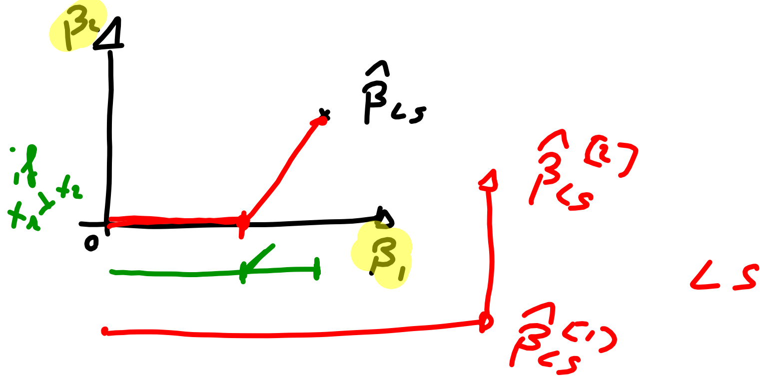
$$F = \frac{(\underline{RSS_0 - RSS_1})/\overbrace{(p_1 - p_0)}^{=1}}{RSS_1/(N-p_1-1)}$$

→ find the β_j s.t. $\beta_j = 0$ lead to the smallest $RSS_0 - RSS_1$

we know (ex 3.1), $F_{1,N-p-1} \overset{d}{=} z_j^2$

⇒ the β_j which, when set equal to 0, increases the least the RSS is that with the smallest z_j^2

Least angle regression



we update $\hat\beta_1$ until
$cor(x_1, r) < cor(x_2, r)$
then
we update both $(\hat\beta_1, \hat\beta_2)$

$r = y - \bar{y} - \sum_{j=1}^{p} x_j \hat\beta_j$

- start with $r = y - \bar{y}$
- check the largest correlation between x_j and r
- update the regression coefficient of x_j ($\hat\beta_j$) until
  $\langle x_j, r \rangle$

Suppose that β_j are ordered by importance, i.e., effect of x_i is larger than effect of x_j, $i<j$

- first we update $\hat\beta_1$
- we reach a point where we add also $\hat\beta_2$ to our solution
  we are updating $\{\hat\beta_1, \hat\beta_2\}$
- we reach a point where the update is related to $\{\hat\beta_1, \hat\beta_2, \hat\beta_3\}$

⋮

once a coefficient enters in the set of active regression coefficients, it stays

FORWARD REGRESSION	LAR
1st step $\hat{y} = \hat\beta_0 + \hat\beta_1^{LS} x_1$	$\hat{y} = \hat\beta_0 + \hat\beta_1^{[2]} x_1$

- lasso can be seen as a special case of a modification of LAR, for which once the estimate of a regression coeff. reach 0, it is excluded by the set of active regression coefficient.

METHODS USING DERIVED INPUT DIRECTIONS

- Principal component regression
- Partial least squares

Principal component regression

IDEA: inputs have different variabilities in different directions
→ directions with largest variability provide more information

↓ principal components

linear combinations of X based on directions of largest variability

$z_m = X v_m$ eigenvectors

z_1 = direction with the largest variability
z_2 = " " the 2^{nd} largest variability s.t. $z_1 \perp z_2$
⋮
z_p = " " the least variability $z_p \perp z_1, \dots, z_{p-1}$

Model

$$y = \theta_0 + \sum_{m=1}^{M} \theta_m z_m + \varepsilon$$

$$\hat{\theta}_0 = \bar{y}$$

$$\hat{\theta}_m = \frac{\langle z_m, y \rangle}{\langle z_m, z_m \rangle}$$

NOTE
principal component analysis is scale dependent
⇓
IMPORTANT to first standardize X

for each X_j

$$x_{ij}^{st} = \frac{(x_{ij} - \bar{x}_j)}{sd(X_j)}$$

$$\hat{y} = \hat{\theta}_0 + \sum_{m=1}^{M} \hat{\theta}_m z_m \qquad z_m = X v_m$$

$$= \hat{\theta}_0 + \sum_{m=1}^{M} \hat{\theta}_m X v_m$$

$$= \hat{\theta}_0 + X \sum_{m=1}^{M} \hat{\theta}_m v_m$$

$$= \hat{\theta}_0 + X \hat{\beta}_{PCR}$$

if $M = p$, principal component regression provides the least square estimates, because all p principal component span the space of X

- idea: use $M < p$ principal components, to exclude directions with less information

↓

we obtain results similar to ridge regression

$$\hat{\beta}_{ridge} = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y \qquad \hat{\beta}_{LS} = \sum_{j=1}^{p} u_j u_j^T y$$

$$\hat{\beta}_{PCR} = \sum_{j=1}^{p} u_j \underbrace{\mathbb{1}[j \leq M]}\, u_j^T y = \sum_{j=1}^{M} u_j u_j^T y$$

Partial least squares

- in the construction of principal components we do not take into account y

$\downarrow$

- in the construction of derived input directions we also consider y

as for principal component regression it is important to first standardize X

1st pls direction

1st step: $\hat{\varphi}_{1j} = \dfrac{\langle x_j, y\rangle}{\underbrace{\langle x_j, x_j\rangle}_{=1}}$ $\longrightarrow z_1 = \sum_{j=1}^{p} \hat{\varphi}_{1j} x_j$

$\hat{\theta}_1 = \dfrac{\langle z_1, y\rangle}{\langle z_1, z_1\rangle}$

$$x_j^{(2)} = x_j - \frac{\langle z_1, x_j\rangle}{\langle z_1, z_1\rangle} \qquad j = 1, \dots, p$$

$$\hat{\varphi}_{2j} = \frac{\langle x_j^{(2)}, y\rangle}{\langle x_j^{(2)}, x_j^{(2)}\rangle}$$

$$z_2 = \sum_{j=1}^{p} \hat{\varphi}_{2j} x_j^{(2)} \qquad \hat{\theta}_2 = \frac{\langle z_2, y\rangle}{\langle z_2, z_2\rangle}$$

Differences:

pcr : derived input directions are the principal components of X, constructed by looking at the variability within X

pls : directions take into consideration both the variability and the correlation with y

Mathematically

PCR $\quad \max_{\alpha} \mathrm{Var}(X\alpha) \quad$ s.t. $\|\alpha\| = 1$

$\alpha^T S v_\ell = 0 \quad \ell = 1, \dots, m-1$

sample covariance matrix

uncorrelated directions

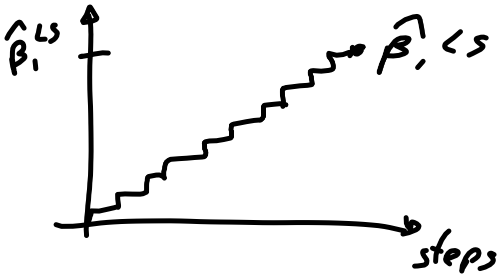
PLS $\quad \max_{\alpha} \mathrm{Cor}^2(y, X\alpha)\, \mathrm{Var}(X\alpha)$

s.t. $\|\alpha\| = 1$

$\alpha^T S \hat{\varphi}_\ell = 0 \quad \ell = 1, \dots, m-1$

In practice, the variance component tends to dominate the correlation one $\longrightarrow$ similar results !

Incremental forward stagewise regression



similar to LAR, but each update involves only one parameter each time

$$\hat{\beta}_j^{[m]} = \hat{\beta}_j^{[m-1]} + \underbrace{\delta}_{\text{step}}$$

→ we will go back when we will talk about boosting

→ grouped LASSO

different construction of penalty to take into account structures in the data

- dummy variables related to the same categorical variable
- working with genetic data, group genes belonging to the same pathway

$$\hat{\beta}_{gL} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{\ell=1}^{L} X_\ell \beta_\ell \right) + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} \, \|\beta_\ell\|_2 \right\}$$

group size

$\| \ \|_2$ = Euclidean distance

it is 0 <=> all components are 0

- adaptive LASSO
- relaxed LASSO

} → 2-step procedure

- SCAD $\qquad L(\beta) + J(\beta)$

$$\frac{dJ(\beta)}{d\beta} = \lambda \operatorname{sign}(\beta) \left[\mathbb{1}[|\beta| \le \lambda] + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} \mathbb{1}(|\beta| > \lambda) \right]$$