Question 1: (1)

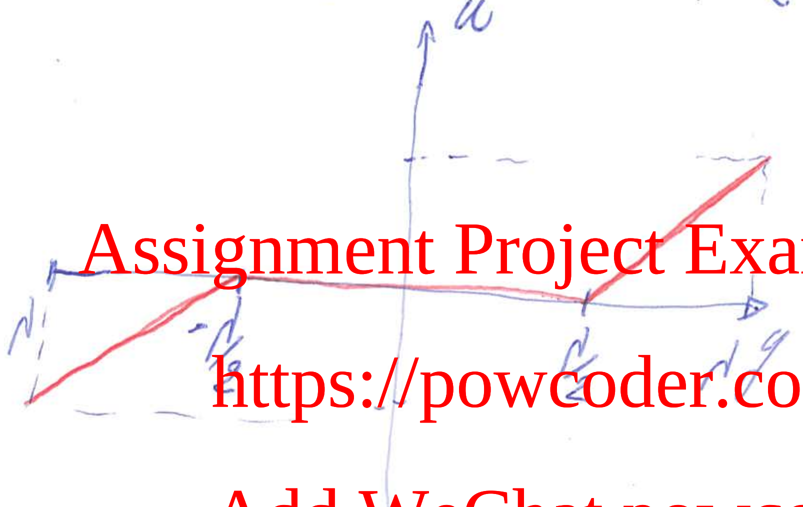
1. there are two cases to consider $u>0$ and $u<0$. Let $L=(y-u)^2+\lambda|u|$

$$\frac{\partial L}{\partial u} = -2(y-u)+\lambda\, sgn(u)$$

if $u>0$: $-(y-u)+\frac{\lambda}{2}=0 \Rightarrow u=y-\frac{\lambda}{2}$ if $y>\frac{\lambda}{2}$.

if $u<0$: $-(y-u)-\frac{\lambda}{2}=0 \Rightarrow u=y+\frac{\lambda}{2}$ if $y<-\frac{\lambda}{2}$.

2.



$u=sgn(y)(|y|-\frac{\lambda}{2})_+$

3. The Lasso cost is

$$\|y-X\beta\|^2+\lambda|\beta|$$

we compute the solution for each coordinate j of β and iterate to convergence.

$$L=\|y-X_{-j}\beta_{-j}-x_j\beta_j\|^2+\lambda|\beta_j| = \|\tilde{y}-x_j\beta_j\|^2+\lambda|\beta_j|$$

$$\frac{\partial L}{\partial\beta_j} = -x_j^T(\tilde{y}-x_j\beta_j)+\frac{\lambda}{2}sgn(\beta_j)$$

if $\beta_j>0$: $-x_j^T\tilde{y}+x_j^Tx_j\beta_j+\frac{\lambda}{2}=0$

$$\beta_j=\frac{x_j^T\tilde{y}}{x_j^Tx_j}-\frac{\lambda}{2x_j^Tx_j}=\hat{\beta}_j-\frac{\lambda}{2x_j^Tx_j} \text{ if } \hat{\beta}_j>\frac{\lambda}{2x_j^Tx_j}$$

if $\beta_j < 0$: $-x_j^T\tilde{y} + x_j^T x_j \beta_j - \frac{\lambda}{2} = 0$

$$\beta_j = \frac{x_j^T\tilde{y}}{x_j^T x_j} + \frac{\lambda}{2x_j^T x_j} = \hat{\beta}_j + \frac{\lambda}{2x_j^T x_j} \text{ if } \hat{\beta}_j < -\frac{\lambda}{2x_j^T x_j}$$

$$\beta_j = \text{sgn}(\hat{\beta}_j)\left(|\hat{\beta}_j| - \frac{\lambda}{2x_j^T x_j}\right)_+ .$$

The updates β_j, $j=1,\dots,p$ are used to update $\tilde{y}$ and therefore $\hat{\beta}_j$.

Question 2:

1. $\hat{Y} = X\hat{B}$ where $\hat{B} = (X'X)^{-1}X'Y$

$$\hat{\Sigma}_k = \frac{RSS_k}{n} \text{ where } RSS_k = (Y-\hat{Y})'(Y-\hat{Y}).$$

$$\log(f) = -\frac{nq}{2}\log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}_k| - \frac{1}{2}tr\{(Y-X\hat{B})\hat{\Sigma}_k^{-1}(Y-X\hat{B})'\}$$

$$-2\log(\text{likelihood}) = nq\log(2\pi) + n\log|\hat{\Sigma}_k| + nq$$

2. There are kq parameters for B and $\frac{1}{2}q(q+1)$ parameters for the covariance matrix Σ_k

$$AIC = nq\log(2\pi) + n\log|\hat{\Sigma}_k| + 2kq + q(q+1) + nq$$

The constants $nq\log(2\pi) + nq$ play no practical role in model selection and can be ignored

$$AIC = n\log|\hat{\Sigma}_k| + 2kq + q(q+1)$$

or

$$AIC_k = \log|\hat{\Sigma}_k| + \frac{2kq + q(q+1)}{n}$$

$$BIC_k = \log|\hat{\Sigma}_k| + \frac{\log(n)\,kq}{n}$$

Question 3:

1. $-2\log(\text{likelihood}) = n\log(2\pi) + n\log\sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$

Using the maximum likelihood

$$-2\log(\text{likelihood}) = n\log(2\pi) + n\log(\hat{\sigma}_k^2) + n$$

The number of parameters is k for β and 1 for σ^2

$$AIC = n\log(2\pi) + n\log\hat{\sigma}_k^2 + n + 2(k+1)$$

The constants $n\log(2\pi) + n$ play no practical role in model selection and can be ignored

$$AIC = n\log\hat{\sigma}^2 + 2(k+1)$$

$$AIC_k = \log(\hat{\sigma}_k^2) + \frac{2(k+1)}{n}$$

2. Suppose the true model order is k_0 and we fit a candidate model of order $k_0 + L$ where $L > 0$

AIC overfits if $AIC_{k_0+L} < AIC_{k_0}$

- $P\{AIC_{k_0+L} < AIC_{k_0}\} = P\left\{\log(\hat{\sigma}^2_{k_0+L}) + \frac{2(k_0+L+1)}{n} < \log\hat{\sigma}^2_{k_0} + \frac{2(k_0+1)}{n}\right\}$

- $\hat{\sigma}_k = \frac{RSS_k}{n} \Rightarrow \log\hat{\sigma}_k = \log RSS_k - \log n$

$$P\left\{ \log(RSS_{k_0+L}) - \log(n) + \frac{2(k_0+L+1)}{n} < \log RSS_{k} - \log n + \frac{2(k_0+1)}{n} \right\}$$

$$= P\left\{ \log\left(\frac{RSS_{k_0+L}}{RSS_{k_0}}\right) < -\frac{2L}{n} \right\}$$

$$= P\left\{ \log\left(\frac{RSS_{k_0}}{RSS_{k_0+L}}\right) > \frac{2L}{n} \right\}$$

$$= P\left\{ \frac{RSS_{k_0}}{RSS_{k_0+L}} - 1 > \exp\left(\frac{2L}{n}\right) - 1 \right\}$$

$$= P\left\{ \frac{RSS_{k_0} - RSS_{k_0+L}}{RSS_{k_0+L}} > \exp\left(\frac{2L}{n}\right) - 1 \right\}$$

$$= P\left\{ \frac{\chi^2_L}{\chi^2_{n-k_0-L}} > \exp\left(\frac{2L}{n}\right) - 1 \right\}$$

$$= P\left\{ F_{L,n-k_0-L} > \frac{n-k_0-L}{L} \exp\left(\frac{2L}{n}\right) - 1 \right\}.$$

Extra for Question 2.1.

In this case we assume that the n observations $y \in R^q$ follow $N(x\hat{B}, \hat{\Sigma}_k)$

$$f(y) = \frac{1}{(2\pi)^{\frac{1}{2}q} |\hat{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(y - x\hat{b})' \hat{\Sigma}_k^{-1} (y - x\hat{b})\right\}$$

and likelihood function is

$$L = \prod_{i=1}^{n} f(y_i)$$

$$= \frac{1}{(2\pi)^{\frac{qn}{2}} |\hat{\Sigma}_k|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - x\hat{b}_i)' \hat{\Sigma}_k^{-1} (y_i - x\hat{b}_i)\right\}$$

where $\hat{b}_i$ is the column i of $\hat{B}$.

$$\sum_{i=1}^{n}(y_i - x\hat{b}_i)' \hat{\Sigma}_k^{-1}(y_i - x\hat{b}_i) = \mathrm{tr}\left\{(Y - x\hat{B}) \hat{\Sigma}_k^{-1} (Y - x\hat{B})'\right\}$$

$$= \mathrm{tr}\left\{\hat{\Sigma}_k^{-1}(Y - x\hat{B})'(Y - x\hat{B})\right\} = n\,\mathrm{tr}\left[\hat{\Sigma}_k^{-1}\hat{\Sigma}_k\right]$$

$$= n\,\mathrm{tr}[I_q] = nq$$

where Y is taken $n \times q$.