Then select a random permutation of *folds.

ex) n = 26, K = 5 folds:

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 2 | 3 | 5 | 3 | 3 | 1 | 5 | 2 | 2 | 5 | 1 | 2 | 1 | 3 | 4 | 2 | 1 | 5 | 5 | 1 | 4 | 1 | 4 | 4 | 3 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |

↑ You add the case numbers

Cases 7, 12, 14, 18, 21 and 23 are in fold 1.

The other folds have 5 cases.

see hw6 go to 24) ex)

12) variable selection: put standardized predictors into $W$ and center the response $\underline{z} = \underline{y} - \bar{y}$.

Do RR on $\underline{z} = W\underline{\beta} + \underline{e}$

where there is no intercept ($\underline{\beta}$ and $(\beta_2, \ldots, \beta_p)^T$ are closely related).

Keep the $k_i$ variables with the largest $|\hat{\beta}_i|$, $i = 1, \ldots, J$. Use CV to select $k_i^*$.

PhD topic

13) Another ridge regression estimator $\hat{\beta}_{RR}$ minimizes the criterion $Q_{RR}(\beta) = RSS(\beta) + \lambda \sum_{j=2}^{p} \beta_j^2$.

$\beta \to 0$ then $Q_{RR} \to \bar{Y}$

§3.6 and 3.7 lasso

14) The lasso MLR estimator $\hat{\beta}_L$ minimizes $Q_L(\beta) = \frac{RSS(\beta)}{a} + \frac{\lambda_n}{a} \sum_{j=2}^{p} |\beta_j|$.

There is not a simple formula for $\hat{\beta}_L$, but the criterion is convex, so fast algorithms exist.

15) Convex criterion often have an equivalent dual problem.

Suppose $\underset{n\times 1}{Z} = \underset{n\times p}{W}\underset{p\times 1}{\beta} + \underset{n\times 1}{e}$ where $Z = Y - \bar{Y}$ and the model has no constant.

Then $Q_L(\beta) = RSS(\beta) + \lambda_n \sum_{j=1}^{p} |\beta_j|$,

and the dual problem is

$\hat{\beta}_L$ minimizes

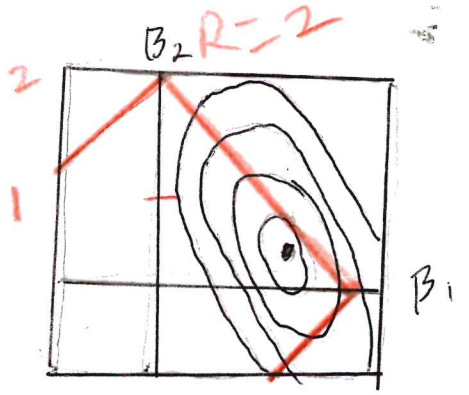
$Q_{L,R}(\beta) = \|z - W\beta\|^2$ subject to

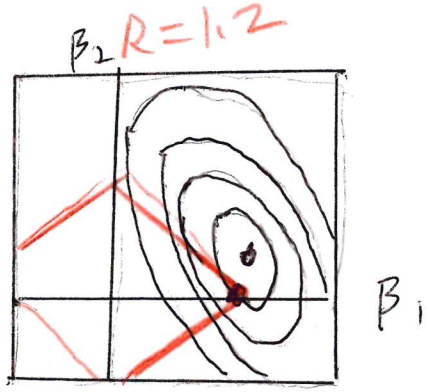
$\{\beta \in \mathbb{R}^p$ with $\sum_{j=1}^{p} |\beta_j| \leq R\}$ where

each value of $R > 0$ corresponds to a $\lambda = \lambda_m > 0$.

Hyperellipsoids correspond to

$RSS(\beta) = \text{constant}$. The $\bullet$ to the OLS estimator.




Grand p 93

$\hat{\beta}_L = \hat{\beta}_{OLS}$ minimizes $Q_{L,R=2}(\beta)$

$\hat{\beta}_L = \begin{pmatrix} 1.2 \\ 0 \end{pmatrix}$ minimizes $Q_{L,R=1.2}(\beta)$

16) The lasso estimator does variable selection: as $\lambda \uparrow$ from 0, some $\hat{\beta}_{iL}$ values $= 0$.

} see

17] LassoCLT: Assume $p$ is fixed and 7] holds, $\underline{Y} = \underline{X}\underline{\beta} + \underline{e}$.

a) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\underline{\beta}}_L - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, \sigma^2 V).$$

b) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$ and $s_n \xrightarrow{P} \underline{s}$ ← see 18]

then $\sqrt{n}(\hat{\underline{\beta}}_L - \underline{\beta}) \xrightarrow{D} N_p(-\frac{\tau}{2} V \underline{s}, \sigma^2 V)$.

c) If $\frac{\hat{\lambda}_n}{n} \xrightarrow{P} 0$, $\hat{\underline{\beta}}_L \xrightarrow{P} \underline{\beta} = \underline{\beta}_{OLS} = \underline{\beta}_F$.

18] Here $s_{in} \in [-1, 1]$, $s_{in} = \text{sign}(\hat{\beta}_{iL})$ if $\hat{\beta}_{iL} \neq 0$. $= \hat{\beta}_{iLn}$

$\text{Sign}(\beta_i) = 1$ if $\beta_i > 0$

$-1$ if $\beta_i < 0$.

19] Since $\hat{\underline{\beta}}_L \xrightarrow{P} \underline{\beta} = \underline{\beta}_{OLS,F}$, $P(S \subseteq I_{min}) \to 1$

where $I_{min}$ is $\hat{\underline{\beta}}_L$ corresponds to chosen by $k$-fold CV on grid $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_J$.

Thus the lasso variable selection estimator is $\sqrt{n}$ consistent.

20] Lasso variable selection is poor if $\frac{\lambda}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$. Typically (glmnet) $\lambda_1 \propto n^{3/4}$. Hence $\hat{\beta}_L$ from K-Fold CV is at best $n^{1/4}$ consistent: far worse than $\sqrt{n}$ consistent.

21] If $p > n$, "lasso variable selection" uses $p$ predictors for $\lambda$ very close to 0, but at most $n+1$ predictors, including a constant, once some $\hat{\beta}_i = 0$. The value $\hat{\lambda}_J$ is the smallest value of $\lambda$ such that $\hat{\beta}_2 = \cdots = \hat{\beta}_p = 0$.

(If $Y = \alpha + x^T\beta + e$, MLR2, the smallest value of $\lambda \Rightarrow \beta = 0$. Hence $\hat{\beta}_{\lambda_i} \neq 0$ for $i = 1, 2, \ldots, J-1$.)

22] In low dimensions, $\lambda_1 \leq \cdots \leq \lambda_J$, behavior near $\lambda_1$ is important, and $J=100$ is fine. In high dimensions, using larger $J$ gives more models, making it more likely to get a model good for prediction (multitude of models).

23] simulations: To compare variable selection estimators and MLR estimators, we will count the number of times $S \subseteq I_{min}$ and make prediction intervals for $Y_t$.

24] know for E2. Let $Z_1, \ldots, Z_n, Z_t$ be iid. (training $Z_i$, test)

The Shorth(c) interval is the shortest closed interval containing at least c of the $Z_i$. For small ordered data sets, compute the lengths of the intervals containing c cases.

$[Z_{(1)}, Z_{(c)}]$, $[Z_{(2)}, Z_{(c+1)}]$, ..., $[Z_{(n-c+1)}, Z_{(n)}]$.

$Z_{(c)} - Z_{(1)}$ $\quad Z_{(c+1)} - Z_{(2)}$ $\quad Z_{(n)} - Z_{(n-c+1)}$.

Then $\text{Shorth}(c) = [Z_{(s)}, Z_{(s+c-1)}]$ is the interval with shortest length.

ex] Find Shorth(4).

0, 1, 3, 6, 9, 10, 11

$6-0=6$

$9-1=8$

$10-3=7$

$11-6=5$

$\boxed{\text{Shorth}(4) = [6, 11]}$ see HW6

25] Let $\lceil x \rceil$ = smallest integer $\geq x$ = ceiling(x).

So $\lceil 7.7 \rceil = \lceil 8 \rceil = 8$. If $\frac{c_n}{n} \to 1-\delta$,

the Shorth($c_n$) interval is a large sample $100(1-\delta)\%$ PI for $Z_t$. $c_n = k_n = \lceil n(1-\delta) \rceil$

contains $\approx 100(1-\delta)\%$ of the training data.

The maximum undercoverage $\wedge \approx 1.12\sqrt{\frac{\delta}{n}}$ for $n$ large

at uniform $(a,b)$ dist's

so use $c_n = \min\left(n, \left\lceil n\left[1-\delta + 1.12\sqrt{\frac{\delta}{n}}\right]\right\rceil\right)$.

26] This large sample $100(1-\delta)\%$ shorth$(c_n)$ PI is for iid RVs.

27] For MLR with $\hat{\underset{\sim}{\beta}}_E$ (E an estimator like OLS, OPLS, MMLE, RR, lasso, forward selection, lassoVS), and $\underset{\sim}{Y} = \mathbb{X}\underset{\sim}{\beta} + \underset{\sim}{e}$, find the residuals $\hat{\underset{\sim}{e}} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}} = \underset{\sim}{Y} - \mathbb{X}\hat{\underset{\sim}{\beta}}_E$.

Roughly, find shorth$(c_n)$ of the $r_i$, say $[r_{(s)}, r_{(s+c-1)}]$, which is a PI for $e_f$. Then $[\hat{Y}_f + r_{(s)}, \hat{Y}_f + r_{(s+c-1)}]$ is a large sample $100(1-\delta)\%$ PI for $Y_f$ where $Y_f = \underset{\sim}{x}_f^T \hat{\underset{\sim}{\beta}}_E$ if $\hat{\underset{\sim}{\beta}}_E$ is a consistent estimator of $\underset{\sim}{\beta}_E$ and $Y = \underset{\sim}{x}_E^T \underset{\sim}{\beta}_E + e$ where $e$ depends on $E$.

28] $Y = X\beta + \varepsilon$, $\underset{\sim}{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$.

The elastic net estimator $\hat{\beta}_{EN}$

minimizes the criterion

$$Q_{EN}(\beta) = \frac{1}{2} RSS(\beta) + \lambda_n \left[ \frac{1}{2}(1-\alpha) \|\beta_S\|_2^2 + \alpha \|\beta_S\|_1 \right]$$

or $Q_2(\beta) = RSS(\beta) + \lambda_1 \|\beta_S\|_2^2 + \lambda_2 \|\beta_S\|_1$

where $0 \le \alpha \le 1$, $\lambda_1 = (1-\alpha)\lambda_n$, $\lambda_2 = 2\alpha\lambda_n$.

Then $\alpha = 1$, $\lambda_1 = 0$ "corresponds to lasso" (using $a = 0.5$ or $a = 1$)

$\alpha = 0$, $\lambda_2 = 0$ corresponds to $\tilde{\beta}_{RR}$

$\lambda_n = 0$, $\lambda_1 = \lambda_2 = 0$ corresponds to $\beta_{OLS}$.

29] EN does variable selection, can be used in HD, and has large sample theory similar to that of RR and lasso.

§ 3.11 K-component estimators

30] Let $Y = \alpha + \underline{x}^T \underline{\beta} + e$.

The K-component MLR estimators use $p$ linear combinations $\underline{\eta}_1^T \underline{x}, \underline{\eta}_2^T \underline{x}, \ldots, \underline{\eta}_p^T \underline{x}$.

Then there are $p$ conditional distributions

$$Y | \underline{\eta}_1^T \underline{x}$$
$$Y | (\underline{\eta}_1^T \underline{x}, \underline{\eta}_2^T \underline{x})$$
$$\vdots$$
$$Y | (\underline{\eta}_1^T \underline{x}, \ldots, \underline{\eta}_p^T \underline{x}).$$

↑ want $K$ small for dimension reduction

Estimating the $\underline{\eta}_i$ and performing the OLS reg of $Y$ on $(\underline{\hat{\eta}}_1^T \underline{x}, \ldots \underline{\hat{\eta}}_K^T \underline{x})$ gives the $K$-component estimator $\underline{\hat{\beta}}_{KE}$, eg the $K$-component PLS estimator $\underline{\hat{\beta}}_{K\,PLS}$ or the $K$-component PCR estimator $\underline{\hat{\beta}}_{K\,PCR}$.

31] Let $Y = \alpha + \underline{x}^T \underline{\beta} + e$ and $\underline{X} = (\underline{1}\ \underline{X}_1)$.

Let $\underset{K\times 1}{\underline{w}_i} = \underset{K\times p}{\hat{A}_{Kn}} \underset{p\times 1}{\underline{x}_i} = \begin{pmatrix} \underline{\hat{\eta}}_1^T \underline{x}_i \\ \vdots \\ \underline{\hat{\eta}}_K^T \underline{x}_i \end{pmatrix}$, $\hat{A}_{Kn} = \begin{pmatrix} \underline{\hat{\eta}}_1^T \\ \vdots \\ \underline{\hat{\eta}}_K^T \end{pmatrix}$.

not $p\times 1$

Let $\underset{\sim}{c}_i = \Sigma_1 \hat{\underset{\sim}{n}}_i = \begin{pmatrix} \underset{\sim}{x}_1^T \hat{\underset{\sim}{n}}_i \\ \vdots \\ \underset{\sim}{x}_n^T \hat{\underset{\sim}{n}}_i \end{pmatrix}$ be the $i$th component vector for $i=1,\dots,p$. Let

$$\underset{n\times k}{V_k} = (\underset{\sim}{c}_1,\dots,\underset{\sim}{c}_k) = \begin{pmatrix} \underset{\sim}{v}_1^T \\ \vdots \\ \underset{\sim}{v}_n^T \end{pmatrix} = \Sigma_1 \hat{A}_{kn}^T$$

for $k=1,\dots,p$.

Let the working OLS model for regressing $y$ on $\underset{\sim}{v} = \hat{A}_{kn}\underset{\sim}{x}$ be

$$\underset{\sim}{y} = \alpha_k \underset{\sim}{1} + V_k \underset{\sim}{\gamma}_k + \underset{\sim}{\varepsilon} \quad \text{where}$$

$\underset{\sim}{\varepsilon}$ depends on the model.

The OLS regression of $y$ on $\underset{\sim}{v} = \hat{A}_{kn}\underset{\sim}{x}$ gives $\hat{\underset{\sim}{\gamma}}_k = \hat{\Sigma}_{\underset{\sim}{v}}\, \hat{\Sigma}_{\underset{\sim}{v}\underset{\sim}{y}} = (\hat{A}_{kn}\hat{\Sigma}_{\underset{\sim}{x}}\hat{A}_{kn}^T)^{-1}\hat{A}_{kn}\hat{\Sigma}_{\underset{\sim}{x}\underset{\sim}{y}}$.

$$\therefore\ \hat{\underset{\sim}{\beta}}_{kE} = \hat{A}_{kn}^T \hat{\underset{\sim}{\gamma}}_k =$$

$$\hat{A}_{kn}^T (\hat{A}_{kn}\hat{\Sigma}_{\underset{\sim}{x}}\hat{A}_{kn}^T)^{-1}\hat{A}_{kn}\hat{\Sigma}_{\underset{\sim}{x}\underset{\sim}{y}} = \hat{\Lambda}_k \hat{\Sigma}_{\underset{\sim}{x}\underset{\sim}{y}}$$

$$= \hat{A}_{kn}^T (\hat{A}_{kn}\hat{\Sigma}_{\underset{\sim}{x}}\hat{A}_{kn}^T)^{-1}\hat{A}_{kn}\hat{\Sigma}_{\underset{\sim}{x}}\hat{\underset{\sim}{\beta}}_{OLS} = \hat{\Lambda}_k \hat{\Sigma}_{\underset{\sim}{x}}\hat{\underset{\sim}{\beta}}_{OLS}.$$

($P\times P$, $P\times 1$)

if $\hat{A}_{kn}^T \simeq A_k$,

$$y_i = \alpha_k + \underset{\sim}{x}_i^T \hat{A}_{kn}^T \underset{\sim}{\gamma}_k + \varepsilon_i \simeq y_i = \alpha_k + \underset{\sim}{x}_i^T \underset{\sim}{\beta}_{kE} + \varepsilon_i$$

If $\hat{\underline{n}}_i \xrightarrow{P} \underline{n}_i$ and $\hat{A}_{kn} \xrightarrow{P} A_k = \begin{pmatrix} \underline{n}_1^T \\ \vdots \\ \underline{n}_k^T \end{pmatrix}$,

then $\hat{\underline{B}}_{kE} \xrightarrow{P} \underline{B}_{kE} =$

$$A_k^T (A_k \Sigma_x A_k^T)^{-1} A_k \Sigma_x \underline{B}_{OLS}(x,y) = I_k \Sigma_x \underline{B}_{OLS}(x,y).$$

32] $\hat{\underline{B}}_{pE} = \hat{\underline{B}}_{OLS}(x,y)$ (if $k=p$) if the inverse matrices exist (also if $p=1$). see HW7

33] If $\underline{B}_{OLS} = \sum_{j=1}^{m} c_{ij} \underline{n}_{ij}$ where $1 \le m \le p$, and $c_{ij} \neq 0$, then $\underline{B}_{kE} = \underline{B}_{OLS}$ if $k$ is large enough to contain the $m$ vectors $\underline{n}_{ij}$. This condition can hold to a good approx in low dimensions.

34] In high dimensions, want $k < n-1$ and likely want $n \ge Jk$ with $J \ge 5, 10$ etc.

35] use model selection, such as 10-fold CV, to select $k^*$. Call the estimator $\hat{\underline{B}}_{MS,E}$.