MSE can help us with the selection of Model Complexity

→ how can we automate this?

→ regularization
- Best subset selection
- Ridge regression
- LASSO

---

take $h_\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_D x_D$

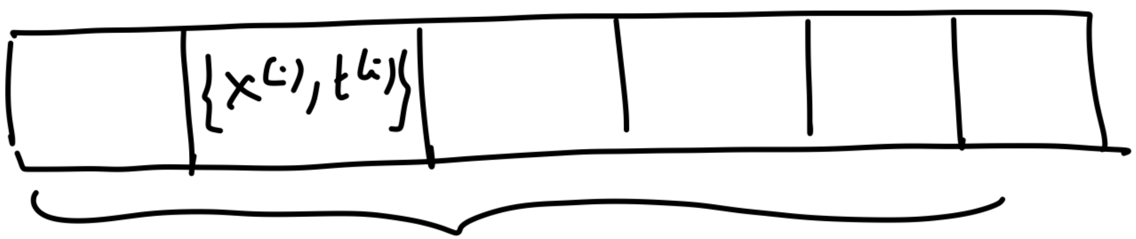
Best subset selection: find the subset of $\{\beta_1, \beta_2, \dots, \beta_D\}$ that achieves the best prediction

Best Subset Selection

1) if enough data, split dataset between training part and a test part, then for each subset of β_j learn the corresponding linear model on training set (through least squares) and evaluate it on test set.

$$\sum_{k=1}^{D} \binom{D}{k}$$

2) if small dataset $\rightarrow$ use cross-validation
(K fold cross validation)



learn model on all bins but the k^{th} one then evaluate this model on the k^{th} bin

$$\text{error}_{CV} = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_{\beta}^{-k(i)}\left(x^{(i)}\right) \right)^2$$

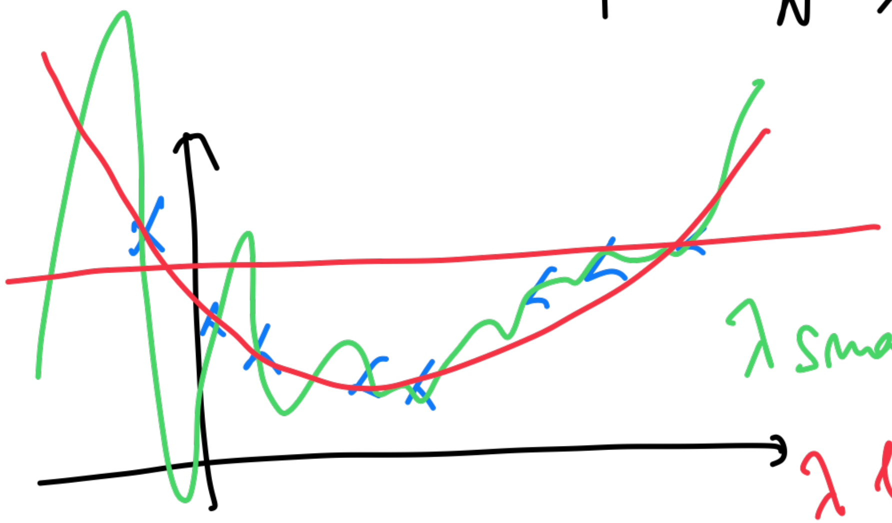
→ to solve the best subset selection problem, compute error_{CV} for each subset and pick subset with best error_{CV}.

$$\sum_{k=1}^{D} \binom{D}{k} \rightarrow$$

Ridge regression

$$\ell(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2$$

$$+ \lambda \sum_{j=1}^{D} |\beta_j|^2$$



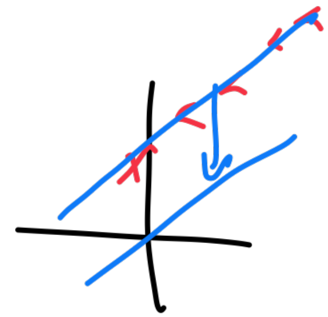
λ small

λ large

Ridge Regression

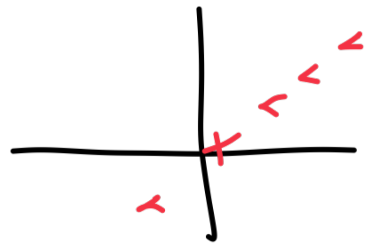
$$\ell_{Ridge}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2 + \lambda\sum_{j=1}^{D}|\beta_j|^2$$

$$\beta^* = \min_{\beta} \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2 + \lambda\sum_{j=1}^{D}|\beta_j|^2$$



$$\ell_{Ridge} = \frac{1}{N}(t - \underline{\underline{X}}\beta)^T(t - \underline{\underline{X}}\beta) + \lambda\beta^T\beta$$

$$= \frac{1}{N}\left(t^Tt + \beta^TX^TX\beta - 2\beta^TX^Tt\right) + \lambda\beta^T\beta$$



take $\frac{\partial \ell_{Ridge}}{\partial \beta} \frac{1}{N}\left(2X^TX\beta - 2X^Tt\right) + 2\lambda\beta = 0$

$$\boxed{\begin{aligned} \left(\frac{1}{N}X^T X + \lambda I\right)\beta &= X^T t \frac{1}{N} \\ \beta &= \underbrace{\left(\frac{1}{N}X^T X + \lambda I\right)}_{\frac{1}{\lambda}}^{-1} X^T t \frac{1}{N} \end{aligned}}$$

LASSO

$$\ell_{LASSO}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2 + \underbrace{\lambda \sum_{j=1}^{D} |\beta_j|}$$

3 approaches to reduce variance (mitigate overfitting)

→ Best subset selection

→ Ridge regression

$$\min_{\beta} \ l_{Ridge}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2 + \underline{\underline{\lambda}} \sum_{j=1}^{D} |\beta_j|^2$$

→ LASSO regression

$$\min_{\beta} \ l_{LASSO}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2 + \lambda \sum_{j=1}^{D} |\beta_j|$$

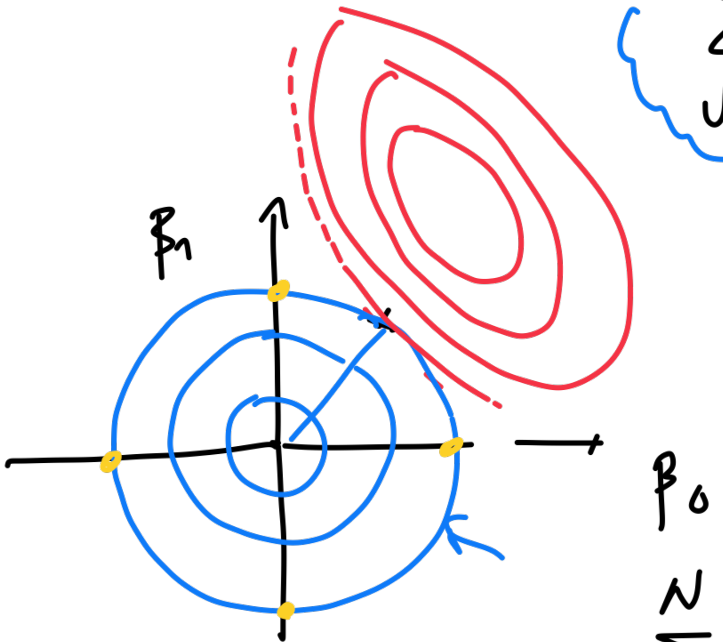
→ geometric intuition

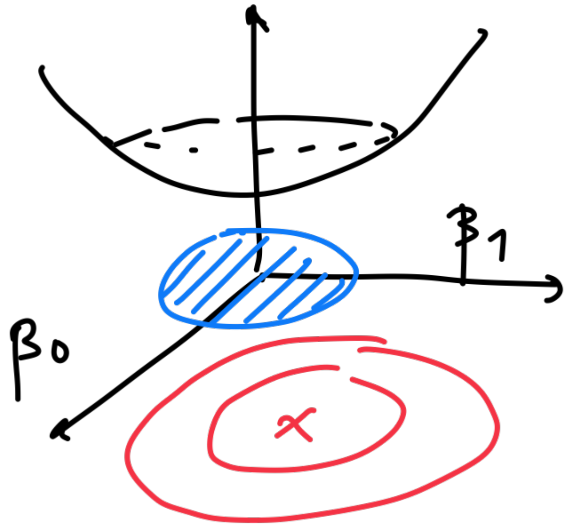
→ statistical motivation

Ridge

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta(x^{(i)})\right)^2$$

$$\sum_{j=1}^{D} |\beta_j|^2 < t$$

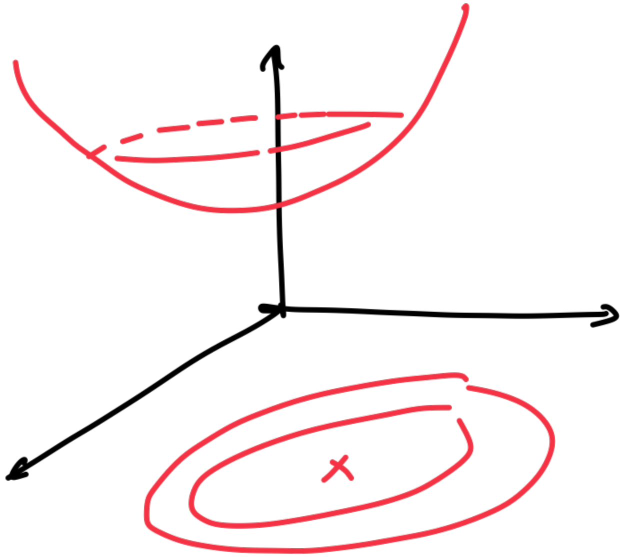




LASSO

$$\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta(x^{(i)})\right)^2$$

$$\sum_{j=1}^{D} |\beta_j| < t$$

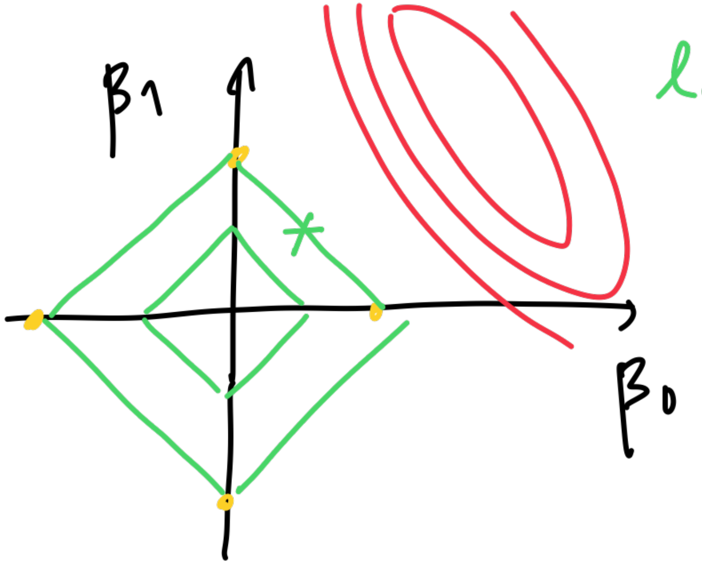


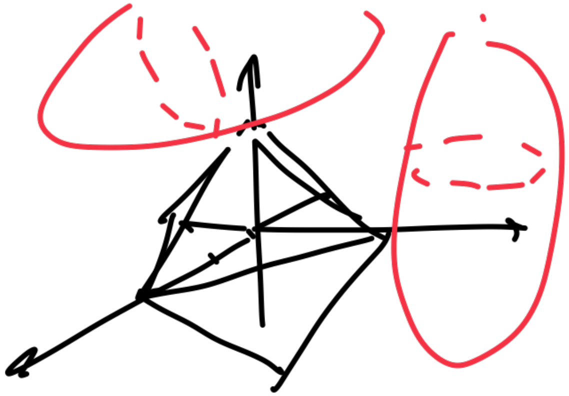
$$\frac{1}{\sqrt{2}} \frac{1}{\sqrt{t}} = \beta_0$$

$$\frac{1}{\sqrt{2}} \frac{1}{\sqrt{t}} = \beta_1$$

$$\frac{1}{\sqrt{2}} \frac{1}{\sqrt{t}} + \frac{1}{\sqrt{2}} \frac{1}{\sqrt{t}}$$

ℓ_1 norm ball





Statistical motivation

recall that RSS corresponds to MLE estimator when considering linearly generated data with Gaussian

$$t^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} + \varepsilon^{(i)} \qquad \varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

$\varepsilon^{(i)}$ independent

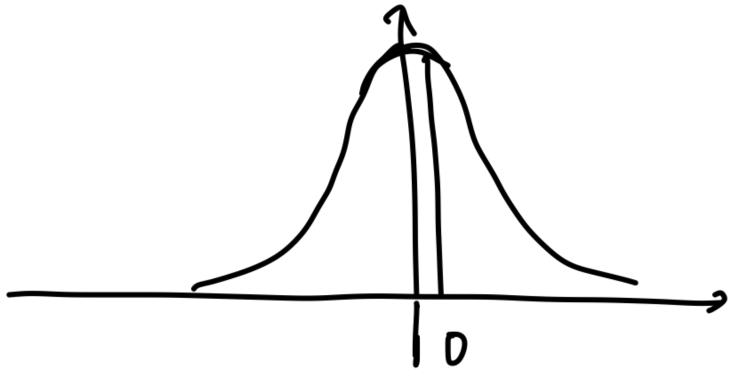
$$\hat{\beta}_{RSS} = \text{Maximum likelihood estimator}$$

$$= \max_{\beta} \; p(t^{(i)} \mid \beta)$$

Maximum a posteriori

$$\max_{\beta} \; p(\underbrace{\beta | t^{(i)}}) = \frac{p(t^{(i)}|\beta)\, p(\beta)}{p(t^{(i)})}$$

$$p(t^{(i)}) = \sum_{\text{all } \beta} p(t^{(i)}|\beta)\, p(\beta)$$



$$\max_{\beta} \; \underbrace{p(t^{(i)}|\beta)}\, \overbrace{p(\beta)}^{\text{prior on } \beta}$$

since $t^{(i)} = \overbrace{\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}} + \varepsilon^{(i)}$

$$t^{(i)} \sim \exp\left(-\frac{(t^{(i)} - h_\beta(x^{(i)}))^2}{2\sigma^2}\right)$$

For gaussian prior

$$\max_{\beta} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - h_\beta(x^{(i)}))^2}{2\sigma^2}\right) \cdot \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\beta_j^2}{2\sigma^2}\right)$$

if $\{\varepsilon^{(i)}\}_{i=1}^{N}$ independent

$$p(\{t^{(i)}\}_{i=1}^{N} | \beta) p(\beta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - h_\beta(x^{(i)}))^2}{2\sigma^2}\right) \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi} b} \exp\left(\frac{-\beta_j^2}{2b^2}\right)$$

$$\arg\max_{\beta} p(\{t^{(i)}\}_{i=1}^{N} | \beta) p(\beta) = ?$$

log

$$\arg\max_{\beta} \log\left(p(\{t^{(i)}\}_{i=1}^{N} | \beta) p(\beta)\right)$$

$$= \underset{\beta}{\arg\max} \sum_{i=1}^{N} \cancel{\log\left(\frac{1}{\sqrt{2\pi\sigma}}\right)} + \sum_{i=1}^{N} -\frac{\left(t^{(i)} - h_\beta(x^{(i)})\right)^2}{2\sigma^2} + \sum_{j=1}^{\Delta} -\frac{\beta_j^2}{2b^2} + \sum_{j=1}^{\Delta} \cancel{\log\left(\frac{1}{\sqrt{2\pi\sigma}}\right)}$$

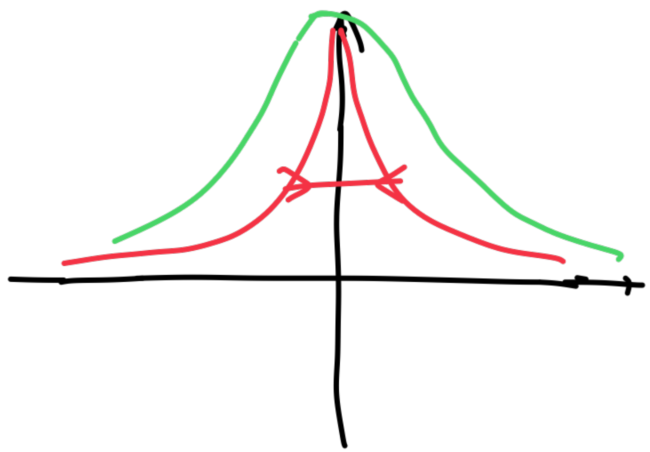
do not depend on β

$$= \underset{\beta}{\arg\max} - \sum_{i=1}^{N} \frac{\left(t^{(i)} - h_\beta(x^{(i)})\right)^2}{2\sigma^2} - \sum_{j=1}^{\Delta} \frac{\beta_j^2}{2b^2}$$

$$\boxed{= \underset{\beta}{\arg\min} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta(x^{(i)})\right)^2 + \underbrace{\cancel{\frac{2\sigma^2}{2b^2}}}_{\lambda =} \sum_{j=1}^{\Delta} \beta_j^2}$$

$$p(x) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$$

How about using a laplace prior on β_j?



$$\beta_{MAP,\,laplace} = \arg\max\; p(t^{(i)}|\beta)\, p(\beta)$$

$$\times \prod_{j=1}^{D} \frac{1}{2b}\exp\left(-\frac{|\beta_j|}{b}\right)$$

$$= \arg\max \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(t^{(i)}-h_\beta(x^{(i)}))^2}{2\sigma^2}\right)$$

take the log

$$\arg\max\; -\sum_{i=1}^{N} \frac{(t^{(i)}-h_\beta(x^{(i)}))^2}{2\sigma^2} - \sum_{j=1}^{D} \frac{1}{b}|\beta_j|$$