Ridge Regression

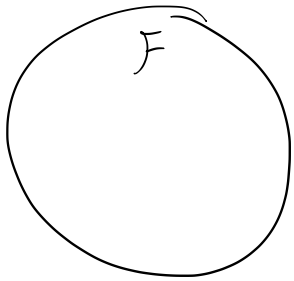
# Generative Models

- Think about the model is generated (prior) and how data are generated from the model (likelihood)
- Bayes Rule: $\underbrace{P(\text{model} \mid \text{data})}_{\text{posterior}} \propto \underbrace{P(\text{data} \mid \text{model})}_{\text{likelihood}} \cdot \underbrace{P(\text{model})}_{\text{prior}}$
- MAP: $\max_{\text{model}} \log p(\text{model} \mid \text{data})$

  ↑ maximum a posteriori

  $= \min_{\text{models}} \underbrace{-\log \text{likelihood}}_{\text{loss function}} \underbrace{- \log \text{prior}}_{\text{regularization}}$

  supervised learning
- Maximum likelihood is a special case (where we don't use a prior, or use uniform prior so log prior is constant)
- Belief? Not necessary




posterior = likelihood · prior

if ↑ gaussian ↑ gaussian then ↑ is gaussian

if ↑ Bernoulli ↑ prior is Beta then posterior is Beta

# Ridge Regression

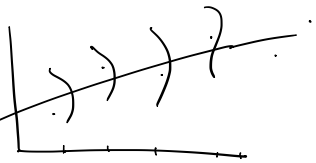
$$\min_{\lambda} \frac{1}{n} \sum_i (y_i - f_\lambda(x_i))^2 + C\|\lambda\|_2^2 \quad \text{where } f_\lambda(x_i) = \sum_j \lambda_j x_{ij}$$

a Bayesian version of it:

$\tau^2[1, \ldots]$

$$\vec{\beta} \sim N(\vec{0}, \tau^2 I)$$

$$\vec{y} \sim N(\bar{X}\vec{\beta}, \sigma^2 I)$$



posterior: $p(\vec{\beta} \mid \vec{y}, \bar{X}) = P(\vec{y} \mid \vec{\beta}, \bar{X}) \cdot p(\vec{\beta}) \cdot \frac{1}{Z}$

$$= \frac{1}{Z} \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma}\|Y - X\beta\|_2^2\right) \frac{1}{(2\pi)^{p/2}\tau^p} \exp\left(-\frac{1}{2\tau^2}\|\beta\|^2\right)$$

$$-\log p(\vec{\beta} \mid \vec{y}, \bar{X}) = -\log(\text{stuff}) + \frac{1}{2\sigma}\|Y - X\beta\|_2^2 + \frac{1}{2\tau^2}\|\beta\|_2^2 \leftarrow \text{Ridge regression}$$

**Fact 1**: Ridge Regression has a generative & frequentist interpretation

**Fact 2**: Least squares regression has a closed form solution.

$$\min_{\lambda} F(\vec{\lambda}) \quad \text{where } F(\vec{\lambda}) = \|\vec{y} - \bar{X}\vec{\lambda}\|^2 = \sum_i (y_i - \bar{x}_i\vec{\lambda})^2$$

$\bar{X}_{\cdot j}$: $j^{th}$ feature vector

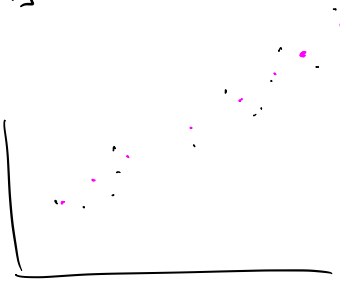
$$\frac{dF(\vec{\lambda})}{d\lambda_j} = -\sum_i 2(y_i - \bar{x}_i\vec{\lambda})x_{ij} = -2\bar{X}_{\cdot j}^T \cdot (\vec{y} - \bar{X}\vec{\lambda})$$

$$\nabla F(\vec{\lambda}) = -2\bar{X}^T(Y - X\lambda) = -2[X^TY - X^TX\lambda] = 0$$

$$X^TY = X^TX\lambda^*$$

$$\boxed{(\bar{X}^T\bar{X})^{-1}\bar{X}\vec{y} = \lambda^*}$$

$$\hat{y} = \bar{X}\cdot\vec{\lambda}^* = \underbrace{X(X^TX)^{-1}X}_{\text{"hat matrix"}}Y = HY = \text{"smoother" of } Y$$



# Fact 3 Ridge Regression has a closed form solution

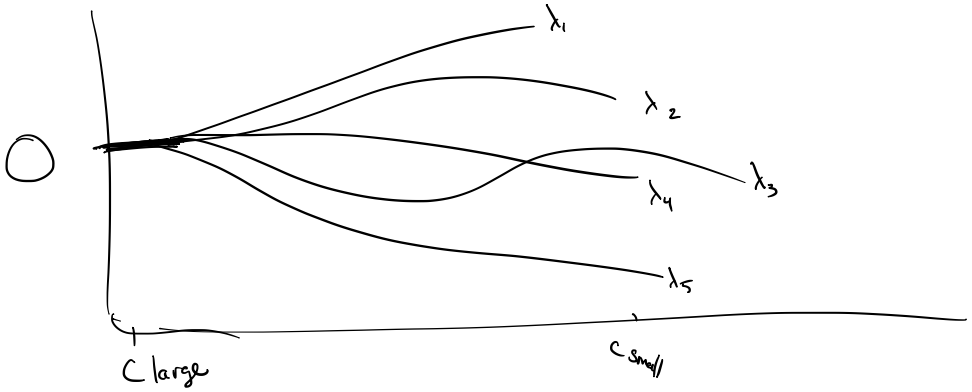
$$F(\bar{\lambda}) = \|\bar{Y} - X\bar{\lambda}\|_2^2 + C\|\lambda\|_2^2$$

$$\frac{dF(\bar{\lambda})}{d\lambda_j} = -2X_{:j}^T(\bar{Y} - X\bar{\lambda}) + 2C\lambda_j$$

$$\nabla F(\bar{\lambda}) = -2X^T(\bar{Y} - X\bar{\lambda}) + 2C\bar{\lambda} = 2(-X^T\bar{Y} + X^TX\bar{\lambda} + C\bar{\lambda}) = 0$$

$$X^TY = (X^TX + CI)\lambda^*$$

$$\boxed{(X^TX + CI)^{-1}X^TY = \bar{\lambda}^*}$$

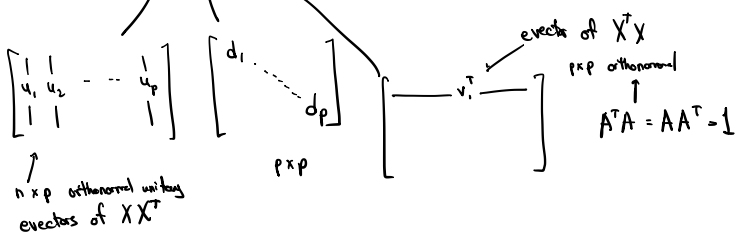


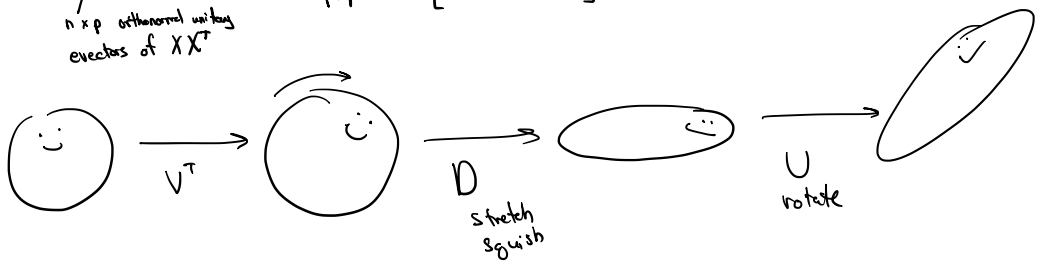

## Computation issues

$$\lambda^* = (X^TX + CI)^{-1}X^TY$$

$O(p^3)$

use SVD: $X = UDV^T$



Note $U^T U = I \qquad V^T V = I$

Simplify $X^T X = (UDV^T)^T (UDV^T) = V D^T U^T U D V^T$

$= VDDV^T = V \begin{pmatrix} d_1^2 & & \\ & d_2^2 & \\ & & \ddots \\ & & & d_p^2 \end{pmatrix} V^T = V D^2 V^T$

sub into $\lambda^*$

$\lambda^* = (VD^2V^T + CI)^{-1} X^T Y$

$CV\begin{pmatrix} 1 & & \\ & 1 & \\ & & \ddots \end{pmatrix}V^T = V\begin{pmatrix} c & & \\ & c & \\ & & \ddots \end{pmatrix}V^T$

$= (VD^2V^T + V\begin{pmatrix} c & & \\ & c & \\ & & \ddots \end{pmatrix}V^T)^{-1} X^T Y$

$= \left(V \begin{pmatrix} d_1^2+c & & \\ & d_2^2+c & \\ & & \ddots \end{pmatrix} V^T\right)^{-1} X^T Y$

$= \left((V^T)^{-1}\begin{pmatrix} d_1^2+c & & \\ & d_2^2+c & \\ & & \ddots \end{pmatrix}^{-1} V^{-1}\right) X^T Y$

$= V \begin{pmatrix} 1/d_1^2+c & & \\ & 1/d_2^2+c & \\ & & \ddots \end{pmatrix} V^{-1} X^T Y$ — $(UDV^T)^T = V^{T^T} D^T U^T = VDU^T$

$= V \begin{pmatrix} 1/d_1^2+c & & \\ & \ddots & \\ & & \ddots \end{pmatrix} V^{-1} VDU^T Y$

$= V \begin{pmatrix} 1/d_1^2+c & & \\ & \ddots & \\ & & \ddots \end{pmatrix} \begin{pmatrix} d_1 & & \\ & d_2 & \\ & & \ddots \end{pmatrix} U^T Y$

$= V \begin{pmatrix} d_1/d_1^2+c & & \\ & \ddots & \\ & & \ddots \end{pmatrix} U^T Y = V \operatorname{diag}\left(\frac{d_j}{d_j^2+c}\right) U^T Y = \lambda^*$

$\hat{Y}^{ridge} = X\lambda^* = UDV^T V \operatorname{diag}\left(\frac{d_j}{d_j^2+c}\right) U^T Y$

$= U \operatorname{diag}\left(\frac{d_j^2}{d_j^2+c}\right) U^T Y$

$= \sum_j \underset{n\times 1}{u_j} \frac{d_j^2}{d_j^2+c} \underset{1\times n}{u_j^T} \cdot \underset{n\times 1}{Y}$

Fact 4 This ↑ is not too difficult to compute.

# Fact 5: There's a relationship of ridge regression to principal components analysis (PCA)

PCA: finds orthogonal directions in which variance is maximized along first component, second component, etc.


loading vector

$$\omega_{(1)} = \underset{\omega:\|\omega\|=1}{\operatorname{argmax}} \left(X_i \cdot \omega\right)^2 = \underset{\omega}{\operatorname{argmax}} \underbrace{\|\vec{X}\vec{\omega}\|^2}$$

$$\langle X\omega, X\omega \rangle$$
$$(X\omega)^T X\omega$$
$$\omega^T \underbrace{X^T X}\, \omega$$

↑ occurs at largest evalue of $X^TX$ & when $\vec{\omega}$ is corresponding evector $v_1$

$$\underline{pca}_1 = X \cdot v_1$$

matrix of all principal components $= \underline{\underline{PCA}} = UDV^T V = UD$

$$pca_j = \vec{u}_j \cdot d_j$$

---

Back to ridge

$$\hat{Y} = X \cdot \lambda^* = \underbrace{X}\, V \operatorname{diag}\left(\frac{d_j}{d_j^2 + C}\right) U^T Y$$

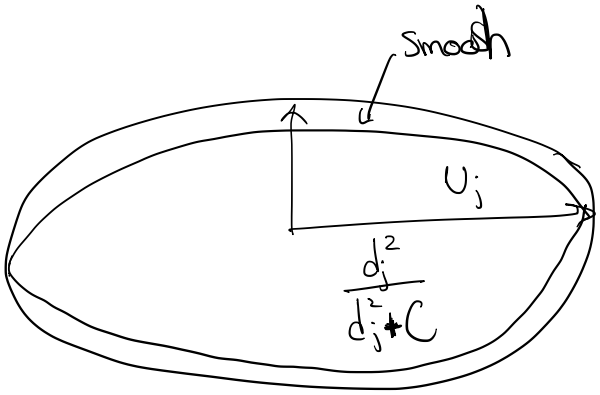
$$XV = UDV^T V = \underbrace{UD}$$

call $UD$ "new data" $\tilde{X}$

$$= UD \operatorname{diag}\left(\frac{d_j}{d_j^2 + C}\right) U^T Y$$

$$= \underbrace{U}_{\tilde{X}} \underbrace{\operatorname{diag}\left(\frac{d_j^2}{d_j^2 + C}\right)}_{\tilde{\lambda}^*} U^T Y$$

$$= U \operatorname{diag}\left(\frac{1}{\quad}\right) DU^T Y$$




$$C = 0 \Rightarrow \lambda_j^* = U_j^T Y$$

$$C > 0 \Rightarrow \lambda_j^* = \frac{d_j^2}{d_j^2 + C} U_j^T Y$$