Regression Notes

① Basics

Given $x_1, \dots, x_n \in \mathbb{R}^p$ & $y_1, \dots, y_n \in \mathbb{R}$, produce "good" β

Fitting: $\min_{\beta'} \|X\beta' - y\|^2$ $\quad$ "Ridge" $\min_{\beta} \|y - X\beta\|^2 + \lambda\|\beta\|^2$

- Fitting data
- Prediction
- Parameter Estimation

$\beta_{OLS} = (X^TX)^{-1}X^Ty$ $\quad$ $\beta_\lambda = (X^TX + \lambda I_p)^{-1}X^Ty$

↑ Say: "assume invertible"

Fun Fact #1

Hat matrix $H = X(X^TX)^{-1}X^T$

Hat ["projection" or "influence"]

$\hat{y} = X\beta_{OLS}$
$= X(X^TX)^{-1}X^Ty$
$= Hy$

$H_{ij} = x_i^T(X^TX)^{-1}x_j$

$H_{ii} \triangleq h_i$ the *leverage of i

$\sum_{i=1}^n h_i = \text{rank}$

Fun Fact #2

Influence of point i. Claim $h_i = \frac{\partial \hat{y}_i}{\partial y_i}$

Even stronger

LOO

$\beta_{-i} = (X^TX - x_ix_i^T)^{-1}(X^Ty - x_iy_i)$

$= \left((X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1 - x_i^T(X^TX)^{-1}x_i}\right)(X^Ty - x_iy_i)$ ← Sherman–Morrison formula

$= \beta_{OLS} + \frac{(X^TX)^{-1}x_i}{1 - h_i}(y_i - \langle x_i, \beta_{OLS}\rangle)$

$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$

$\langle x_i, \beta_{-i}\rangle = \langle x_i, \beta_{OLS}\rangle + \frac{H_{ij}}{1 - h_i}e_i$

Regression Notes

(2) Modeling ~~Assumptions~~ & Inference

Fixed design: X not random

$$y_i = \langle x_i, \beta \rangle + \eta_i, \quad \eta_i \sim N(0, \sigma^2)$$

$$y = X\beta + \eta, \quad \eta \sim N(0, \sigma^2 I)$$

Random design $\{x_i\} \overset{iid}{\sim} N(0, \Sigma)$

$$y = X\beta + \eta$$

Fixed design is easy:

Claim In fixed design setting, Let $\hat{\Sigma} = \frac{1}{n} X^T X$, $\beta_{OLS} = \beta + N(0, \frac{\sigma^2}{n} \hat{\Sigma}^{-1})$

Pf $\beta_{OLS} = (X^T X)^{-1} X^T y$

$$= (X^T X)^{-1} X^T (X\beta + \eta)$$

$$= (X^T X)^{-1} X^T X \beta + (X^T X)^{-1} X^T \eta$$

$A N(0, I) \overset{d}{=} N(0, AA^T)$

$$= \beta + N(0, (X^T X)^{-1} X^T X (X^T X)^{-1} \cdot \sigma^2)$$

$\square$

Regression Notes

Ask: everyone know about MLE & MAP?

For Gaussian data: $\operatorname{argmax}_{\beta'} p(y, X \mid \beta')$ ~~$= \operatorname{argmax}_{\beta'} \dots$~~

$= \operatorname{argmax}_{\beta'} p(y \mid X, \beta')\, p(X \mid \beta')$ → indep of β

$= \operatorname{argmax}_{\beta'} \mathcal{N}(X\beta', \sigma^2 I)$

$= \operatorname{argmax}_{\beta'} \exp\left\{-\frac{1}{2\sigma^2}\|y - X\beta'\|^2\right\}$

$= \operatorname{argmin} \|y - X\beta'\|^2$

Maximum a posteriori prior $p(\beta) = \mathcal{N}(0, \tau^2 I)$

$$\operatorname*{argmax}_{\beta'} p(\beta \mid y, X) = \operatorname*{argmax} p(y \mid X, \beta')\, p(X \mid \beta')\, p(\beta')$$

$$= \operatorname{argmax} \exp\left\{\frac{-1}{2\sigma^2}\|y - X\beta'\|^2 - \frac{1}{2\tau^2}\|\beta'\|^2\right\}$$

$$= \operatorname{argmin} \|y - X\beta'\|^2 + \frac{\sigma^2}{\tau^2}\|\beta'\|^2$$

---

(3) Prediction (see: Hsu Kakade Zhang 11 & Mourtada Rosasco 22)

In random design setting:

$$\mathbb{E}_{x,y}\left(\langle x, \hat{\beta}\rangle - y\right)^2 = \mathbb{E}\left(\langle x, \hat{\beta}\rangle - \langle x, \beta\rangle - \eta\right)^2$$

$$= \mathbb{E}(\hat{\beta} - \beta)^T x x^T (\hat{\beta} - \beta) + \mathbb{E}\eta^2$$

$$= (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta) + \sigma^2$$

$$= \|\Sigma^{1/2}(\hat{\beta} - \beta)\|^2 + \sigma^2$$

Regression Notes

Main Question What is $\|\Sigma^{1/2}(\beta_\lambda - \beta)\|^2$?

Hope: as in fixed design & OLS

$$\|\Sigma^{1/2}(\beta_{OLS} - \beta)\|^2 = \|\Sigma^{1/2} \cdot \mathcal{N}(0, \tfrac{\sigma^2}{n}\hat{\Sigma}^{-1})\|^2$$

$$= \|\Sigma^{1/2} \tfrac{\sigma}{\sqrt{n}} \cdot \hat{\Sigma}^{1/2} \cdot \mathcal{N}(0, I)\|^2$$

$$\approx \tfrac{\sigma^2}{n} \cdot \|\mathcal{N}(0, I)\|^2 \approx \tfrac{\sigma^2 d}{n}$$

Assumptions on data-generating distribution

(A1) $\exists\, \beta \in \mathbb{R}^d$ s.t. $\mathbb{E}[y|x] = \langle \beta, x \rangle$ and $\mathrm{Var}(y|x) \le \sigma^2$

(A2) $\exists\, R$ s.t. $\|x\| \le R$ almost surely

Thm [Montanda & Rosasco, 22(?)] Let $\Sigma = \mathbb{E} x x^T$

Under A1 & A2, if $\lambda \gtrsim R^2/n$

$$\mathbb{E}\left[\|\Sigma^{1/2}(\beta_\lambda - \beta)\|^2\right] \lesssim \lambda \|\beta\|^2 + \frac{\sigma^2 \mathrm{Tr}\left((\Sigma + \lambda I)^{-1}\Sigma\right)}{n}$$

More generally, $\forall \lambda > 0$

$$\mathbb{E}[\ ] \le \left(1 + \frac{R^2}{\lambda n}\right)^2 \lambda \|\beta\|^2 + \left(1 + \frac{R^2}{\lambda n}\right) \frac{\sigma^2 \mathrm{Tr}\left((\Sigma + \lambda I)^{-1}\Sigma\right)}{n}$$

↳ or $\inf_{\beta \in \mathbb{R}^d} \{L(\beta) + \lambda \|\beta\|^2 - L(\beta)\}$

Regression Notes

Lemma 1 Under A1 & A2,

$$\mathbb{E}\left[\|\Sigma^{1/2}(\beta_\lambda - \beta)\|^2\right] \le \lambda^2 \mathbb{E}\left[\langle(\hat{\Sigma}+\lambda I)^{-1}\Sigma(\hat{\Sigma}+\lambda I)^{-1}\beta, \beta\rangle\right] \quad \text{(bias)}$$

$$+ \frac{\sigma^2}{n}\mathbb{E}\left[\operatorname{tr}\left((\hat{\Sigma}+\lambda I)^{-1}\Sigma\right)\right] \quad \text{(variance)}$$

Proof Idea

~~Key step~~

$$\beta_\lambda = (\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n x_i y_i$$

$$= (\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n x_i\left(x_i^T\beta + \eta_i\right)$$

$$= (\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}\beta + (\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n x_i\eta_i$$

$$\beta_\lambda - \beta = (\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma}\beta + (\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n x_i\eta_i - (\hat{\Sigma}+\lambda I)^{-1}(\hat{\Sigma}+\lambda I)\beta$$

$$= -\lambda(\hat{\Sigma}+\lambda I)^{-1}\beta + (\hat{\Sigma}+\lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n x_i\eta_i. \qquad \text{"}\square\text{"}$$

Lemma 2 Under A2,

$$\mathbb{E}\operatorname{tr}\left[(\hat{\Sigma}+\lambda I)^{-1}\Sigma\right] \le \left(1+\frac{R^2}{\lambda n}\right)\operatorname{tr}\left[(\Sigma+\lambda I)^{-1}\Sigma\right]$$

Pf x_{n+1} ~~$\sim N(0,\Sigma)$~~, indep of iid from $x_1, \dots, x_n$

$$\mathop{\mathbb{E}}_{x_1,\dots,x_n}\operatorname{tr}\left[(\hat{\Sigma}+\lambda I)^{-1}\Sigma\right] = \mathop{\mathbb{E}}_{\substack{x_1,\dots,x_n \\ x_{n+1}}}\operatorname{tr}\left[(\hat{\Sigma}+\lambda I)^{-1}x_{n+1}x_{n+1}^T\right]$$

$$= n \mathop{\mathbb{E}}_{\substack{x_1,\dots,x_n \\ x_{n+1}}}\left[\langle(n\hat{\Sigma}+\lambda n I)^{-1}x_{n+1}, x_{n+1}\rangle\right]$$

$\Rightarrow$

Sherman-Morrison again

$$x_{n+1}^\top \left(n\hat{\Sigma} + x_{n+1}x_{n+1}^\top + \lambda n I\right)^{-1} x_{n+1}$$

$$= x_{n+1}^\top \left(n\hat{\Sigma} + \lambda n I\right)^{-1} x_{n+1} - \frac{x_i^\top \left(n\hat{\Sigma}+\lambda n I\right)^{-1} x_i x_i^\top \left(n\hat{\Sigma}+\lambda n I\right)^{-1} x_i}{1 + x_i^\top \left(n\hat{\Sigma}+\lambda n I\right)^{-1} x_i}$$

$$= x_{n+1}^\top \left(n\hat{\Sigma}+\lambda n I\right)^{-1} x_{n+1}\left(1 - \frac{x_i^\top \left(n\hat{\Sigma}+\lambda n I\right)^{-1} x_i}{1 + x_i^\top \left(n\hat{\Sigma}+\lambda n I\right)^{-1} x_i}\right)$$

So

$$n \mathop{\mathbb{E}}_{\substack{x_1,\dots,x_n \\ x_{n+1}}} \left[x_{n+1}^\top \left(n\hat{\Sigma} + n\lambda I\right)^{-1} x_{n+1}\right]$$

$$= n \mathbb{E}\left[\left(1 + x_{n+1}^\top \left(n\hat{\Sigma} + n\lambda I\right)^{-1} x_{n+1}\right) x_{n+1}^\top \left((n+1)\hat{\Sigma}_{n+1} + \lambda n I\right)^{-1} x_{n+1}\right]$$

$$\hookrightarrow \le \left(1 + \frac{R^2}{\lambda n}\right)$$

$$\le n\left(1 + \frac{R^2}{\lambda n}\right) \mathbb{E}\left[x_{n+1}^\top \left((n+1)\hat{\Sigma}_{n+1} + \lambda n I\right)^{-1} x_{n+1}\right]$$

keep

$$= n\left(1 + \frac{R^2}{\lambda n}\right) \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}\left[x_i^\top \left((n+1)\hat{\Sigma}_{n+1} + \lambda n I\right)^{-1} x_i\right]$$

$$= n\left(1 + \frac{R^2}{\lambda n}\right) \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}\left[\mathrm{tr}\left[\left((n+1)\hat{\Sigma}_{n+1} + \lambda n I\right)^{-1} x_i x_i^\top\right]\right]$$

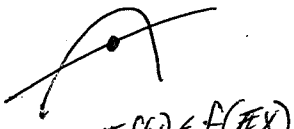
$$= n\left(1 + \frac{R^2}{\lambda n}\right) \mathbb{E}\left[\mathrm{tr}\left[\left((n+1)\hat{\Sigma}_{n+1} + \lambda n I\right)^{-1} \hat{\Sigma}_{n+1}\right]\right]$$

Observe: $\hat{\Sigma}_{n+1}$ & $\hat{\Sigma}_{n+1}$

$1 + \frac{1}{n}$ $\quad \lambda I$

Regression Notes

$$\left(\left(1+\tfrac{1}{n}\right)\hat{\Sigma}_{n+1} + \lambda I\right)^{-1} \preceq \left(\hat{\Sigma}_{n+1} + \lambda I\right)^{-1}$$



Claim: $f: A \mapsto \mathrm{tr}\left[(A+\lambda I)^{-1} A\right]$ is concave.

Pf idea: $\mathrm{tr}\left[(A+\lambda I)^{-1} A\right] = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \lambda}$