1 Setup

1.1 Normal means problem

- Data: $b = (b^{(1)}, \ldots, b^{(n)})$ are n real-valued random variables corresponding to n "units" that we are concerned about
- Goal: estimate the unknown means $\mu = (\mu^{(1)}, \ldots, \mu^{(n)}) \in \mathbb{R}^n$, where
$$\mu^{(i)} = \mathbb{E}(b^{(i)})$$
- "Normal means" assumption:
$$b \sim \mathrm{N}(\mu, \sigma^2 I_n)$$
for some $\sigma^2 > 0$
  - Maximum likelihood estimate of μ is $\hat{\mu}_{\mathrm{mle}} := b$
  - $\mathbb{E}\|\hat{\mu}_{\mathrm{mle}} - \mu\|^2 = \sigma^2 n$
    (Holds even under "weak" assumption: $\mathbb{E}(b) = \mu$ and $\mathrm{cov}(b) = \sigma^2 I_n$, but not necessarily normally distributed)

1.2 Fixed design matrix

- Suppose we have inputs $a^{(i)}, \ldots, a^{(i)} \in \mathbb{R}^d$, which we regard as "fixed" (i.e., non-random) feature vectors describing the n units
- Suppose we have the following belief about how $\mu^{(i)}$'s relate to $a^{(i)}$'s: there is a linear function $L \colon \mathbb{R}^d \to \mathbb{R}$ such that
$$\mu^{(i)} \approx L(a^{(i)}), \quad \forall i \in \{1, \ldots, n\}$$
- This is equivalent to belief that $\mu = (\mu_1, \ldots, \mu_n)$ is "close to" $\mathrm{CS}(A)$, where A is the $n \times d$ matrix with the $(a^{(i)})^\intercal$'s as rows:
$$A = \begin{bmatrix} \longleftarrow & (a^{(i)})^\intercal & \longrightarrow \\ & \vdots & \\ \longleftarrow & (a^{(n)})^\intercal & \longrightarrow \end{bmatrix}$$
- Note: "normal linear regression model" *assumes* $\mu \in \mathrm{CS}(A)$ (but here, we are not considering this model)

1.3 How well can we do with estimates from $\mathrm{CS}(A)$?

- Let Π be $n \times n$ orthogonal projection matrix for $\mathrm{CS}(A)$
  - This is determined by A alone
- The closest vector in $\mathrm{CS}(A)$ to μ (in Euclidean distance) is $\Pi\mu$
  - Can write $\Pi\mu = A\bar{w}$ for some $\bar{w} \in \mathbb{R}^d$ satisfying $A^\intercal(\mu - A\bar{w}) = 0$
  - But this depends on the unknown μ
- Goal: estimate μ by a vector from $\mathrm{CS}(A)$, obtained using A and b
- Note: Consider plane containing μ, $\Pi\mu$, and any other $u \in \mathrm{CS}(A)$
  - These points form a right triangle
  - By Pythagorean theorem:

$$\|u - \mu\|^2 = \|\Pi\mu - \mu\|^2 + \|u - \Pi\mu\|^2$$

  - Every $u \in \mathrm{CS}(A)$ has $\|u - \mu\|^2 \geq \|\Pi\mu - \mu\|^2$

2 Ordinary least squares

2.1 Estimator

- OLS estimate of μ is $\hat{\mu} := \Pi b$
  - Can write $\hat{\mu} = A\hat{w}$ for some $\hat{w} \in \mathbb{R}^d$ satisfying $A^\intercal(b - A\hat{w}) = 0$
- By Pythagorean theorem:

$$\|\hat{\mu} - \mu\|^2 = \|\Pi\mu - \mu\|^2 + \|\hat{\mu} - \Pi\mu\|^2,$$

- By linearity of expectation and bias-variance decomposition:

$$\begin{aligned}\mathbb{E}\|\hat{\mu} - \mu\|^2 &= \|\Pi\mu - \mu\|^2 + \mathbb{E}\|\hat{\mu} - \Pi\mu\|^2 \\ &= \|\Pi\mu - \mu\|^2 + \|\mathbb{E}(\hat{\mu}) - \Pi\mu\|^2 + \mathbb{E}\|\hat{\mu} - \mathbb{E}(\hat{\mu})\|^2\end{aligned}$$

- Expected value of $\hat{\mu}$:
$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(\Pi b) = \Pi\mu$$
- "Variance" of $\hat{\mu}$:
$$\mathbb{E}\|\hat{\mu} - \mathbb{E}(\hat{\mu})\|^2 = \operatorname{tr}(\operatorname{cov}(\hat{\mu}))$$
where covariance matrix of $\hat{\mu}$ is
$$\operatorname{cov}(\hat{\mu}) = \mathbb{E}(\Pi b - \mathbb{E}(\Pi b))(\Pi b - \mathbb{E}(\Pi b))^\intercal = \Pi \operatorname{cov}(b)\Pi^\intercal$$

2.2 Analysis under normal means assumption

- Under (weak) "normal means" assumption, $\operatorname{cov}(b) = \sigma^2 I_n$
$$\operatorname{cov}(\hat{\mu}) = \sigma^2\Pi$$
and
$$\operatorname{tr}(\operatorname{cov}(\hat{\mu}) = \sigma^2 \operatorname{tr}(\Pi) = \sigma^2 \operatorname{rank}(A)$$
- Conclusion:
$$\mathbb{E}\|\hat{\mu} - \mu\|^2 = \|\Pi\mu - \mu\|^2 + \sigma^2 \operatorname{rank}(A)$$
  - Recall: $\mathbb{E}\|\hat{\mu}_{\text{mle}} - \mu\|^2 = \sigma^2 n$
  - $\hat{\mu}$ is improvement over $\hat{\mu}_{\text{mle}}$ when $\operatorname{rank}(A) < n$ and $\|\Pi\mu - \mu\|$ is small enough

3 Ridge regression

3.1 Estimator

- Regularization parameter $\lambda > 0$
- Define $\hat{w}_\lambda$ as (unique) solution w to
$$(A^\intercal A + \lambda I_d)w = A^\intercal b$$
- Ridge regression estimate of μ is
$$\hat{\mu}_\lambda := A\hat{w}_\lambda = A(A^\intercal A + \lambda I_d)^{-1}A^\intercal b$$

- By Pythagorean theorem:

$$\|\hat{\mu}_\lambda - \mu\|^2 = \|\Pi\mu - \mu\|^2 + \|\hat{\mu}_\lambda - \Pi\mu\|^2$$

- By bias-variance decomposition:

$$\mathbb{E}\|\hat{\mu}_\lambda - \Pi\mu\|^2 = \|\mathbb{E}(\hat{\mu}_\lambda) - \Pi\mu\|^2 + \|\hat{\mu}_\lambda - \mathbb{E}(\hat{\mu}_\lambda)\|^2$$

- Expected value of $\hat{\mu}_\lambda$:

$$\mathbb{E}(\hat{\mu}_\lambda) = A(A^\intercal A + \lambda I_d)^{-1} A^\intercal \mu$$

- Covariance of $\hat{\mu}_\lambda$:

$$\operatorname{cov}(\hat{\mu}_\lambda) = A(A^\intercal A + \lambda I_d)^{-1} A^\intercal \operatorname{cov}(b) A(A^\intercal A + \lambda I_d)^{-1} A^\intercal$$

3.2 Analysis under normal means assumption

- Under (weak) "normal means" assumption, $\operatorname{cov}(b) = \sigma^2 I_n$; in this case,

$$\operatorname{cov}(\hat{\mu}_\lambda) = \sigma^2 A(A^\intercal A + \lambda I_d)^{-1} A^\intercal A(A^\intercal A + \lambda I_d)^{-1} A^\intercal = \sigma^2 (A(A^\intercal A + \lambda I_d)^{-1} A^\intercal)^2$$

- Need eigendecomposition of

$$A(A^\intercal A + \lambda I_d)^{-1} A^\intercal$$

- Write singular value decomposition of A as

$$A = \sum_{i=1}^{r} s_i \, u_i v_i^\intercal = \sum_{i=1}^{d} s_i \, u_i v_i^\intercal$$

(If $r := \operatorname{rank}(A) < d$, then $s_{r+1} = \cdots = s_d = 0$, and we extend right singular vectors $(v_i)_{i=1}^r$ to obtain a complete orthonormal basis $(v_i)_{i=1}^d$ for $\mathbb{R}^d$)

- Then:

$$(A^\intercal A + \lambda I_d)^{-1} = \left(\sum_{i=1}^{d}(s_i^2+\lambda)\, v_i v_i^\intercal\right)^{-1} = \sum_{i=1}^{d}\frac{1}{s_i^2+\lambda}\, v_i v_i^\intercal$$

$$A(A^\intercal A + \lambda I_d)^{-1}A^\intercal = \sum_{i=1}^{r}\frac{s_i^2}{s_i^2+\lambda}\, u_i u_i^\intercal$$

$$\left(A(A^\intercal A + \lambda I_d)^{-1}A^\intercal\right)^2 = \sum_{i=1}^{r}\left(\frac{s_i^2}{s_i^2+\lambda}\right)^2 u_i u_i^\intercal$$

- Part of expected squared distance $\mathbb{E}\|\hat{\mu}_\lambda - \Pi\mu\|^2$ due to "bias":

$$\|\Pi\mu - \mathbb{E}(\hat{\mu}_\lambda)\|^2 = \left\|\sum_{i=1}^{r} u_i u_i^\intercal \mu - \sum_{i=1}^{r}\frac{s_i^2}{s_i^2+\lambda}\, u_i u_i^\intercal \mu\right\|^2 = \sum_{i=1}^{r}\left(\left(1-\frac{s_i^2}{s_i^2+\lambda}\right)u_i^\intercal\mu\right)^2$$

- Part of expected squared distance $\mathbb{E}\|\hat{\mu}_\lambda - \Pi\mu\|^2$ due to "variance":

$$\mathbb{E}\|\hat{\mu}_\lambda - \mathbb{E}(\hat{\mu}_\lambda)\|^2 = \mathrm{tr}(\mathrm{cov}(\hat{\mu}_\lambda)) = \sigma^2\sum_{i=1}^{r}\left(\frac{s_i^2}{s_i^2+\lambda}\right)^2$$

- So overall expected squared distance is:

$$\mathbb{E}\|\hat{\mu}_\lambda - \Pi\mu\|^2 = \sum_{i=1}^{r}\left(1-\frac{s_i^2}{s_i^2+\lambda}\right)^2 (u_i^\intercal\mu)^2 + \sigma^2\sum_{i=1}^{r}\left(\frac{s_i^2}{s_i^2+\lambda}\right)^2$$

  One term goes up with λ and the other goes down with λ

  - Recall: for OLS estimate $\hat{\mu}$,

$$\mathbb{E}\|\hat{\mu} - \Pi\mu\|^2 = \sigma^2\,\mathrm{rank}(A)$$

  - Ridge regression has potential to improve over $\hat{\mu}_{\mathrm{mle}}$ even when $\mathrm{rank}(A) = n$, but OLS cannot

4 Ridge regression with orthogonal design

- Suppose $A = I_n$, so $r = \text{rank}(A) = n$ and

$$\hat{\mu}_\lambda = \frac{1}{1+\lambda} b$$

and

$$\mathbb{E}\|\hat{\mu}_\lambda - \mu\|^2 = \|\mu\|^2 \left(1 - \frac{1}{1+\lambda}\right)^2 + \sigma^2 n \left(\frac{1}{1+\lambda}\right)^2$$

- Choosing $\lambda = \frac{\sigma^2 n}{\|\mu\|^2}$ gives

$$\mathbb{E}\|\hat{\mu}_\lambda - \mu\|^2 = \left(1 - \frac{\sigma^2 n}{\|\mu\|^2 + \sigma^2 n}\right) \sigma^2 n$$

in which case $\hat{\mu}_\lambda$ strictly improves over $\hat{\mu}_{\text{mle}}$

  - The catch: choice of λ depends on $\|\mu\|^2$ (and σ^2)
    (Not really a big deal; choose λ by cross-validation anyway)
  - Amazing: there is a choice of $\lambda = \lambda(b, \sigma^2)$ that strictly improves over $\hat{\mu}_{\text{mle}}$ (James-Stein estimator; requires normal means assumption)
    * Key: it is "easier" to estimate $\|\mu\|^2$ than it is to estimate μ