2.9 Regularized fixed design linear regression

We've considered a model of which dimention is fixed. Now we show that when d is comparable to n, reglurization can help promote the accuracy.

1. James-Stein estimator

Given $\{x^{(1)}, \cdots, x^{(n)}\} \sim N(\theta^*, \sigma^2 I)$, i.i.d., we defined

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

and we have $E[\hat{\theta}-\theta^*] = \frac{d\sigma^2}{n}$. Can we do better?

James-Stein estimator: $\hat{\theta}_{JS} := \left(1-\frac{(d-2)\sigma^2}{n\|\hat{\theta}\|_2^2}\right)\hat{\theta}$

whose shrinkage is governed by $\|\theta\|_2^2$. For example, if $\theta^*=0$, $\|\hat{\theta}\|_2^2 \sim \frac{d\sigma^2}{n}$, which provides a massive shrinkage factor of $2/d$.

Short: standard estimators are often not optimal.

2. Fixed design linear regression

(i) $\{x^{(1)}, \cdots, x^{(n)}\}$ i.i.d., $x^{(i)} \in \mathbb{R}^d$.

(ii) Design matrix $X \in \mathbb{R}^{n\times d}$

(iii) Responses: $Y \in \mathbb{R}^d$

(iv) Noise $\varepsilon \in \mathbb{R}^d$, $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2$

(v) Assume data satisfies: $Y = X\theta^* + \varepsilon$

Regularized least squares estimator:

$$\hat{\theta} = \operatorname{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}\|X\theta - Y\|_2^2 + \lambda\|\theta\|_2^2$$

$$= \frac{1}{n}\Sigma_\lambda^{-1} X^T Y$$ can be verified by differentiating.

where $\Sigma_\lambda = \frac{1}{n}X^TX + \lambda I$, $\lambda \geq 0$ is the regularization strength

Insight:

(i) Bias variance tradeoff

(i) Bias-variance tradeoff, the excess risk:

$$\begin{aligned}
E[\|\hat\theta-\theta^*\|_\Sigma^2] &= E[\|\hat\theta - E[\hat\theta] + E[\hat\theta] - \theta^*\|_\Sigma^2] \\
&= E[\|\hat\theta - E[\hat\theta]\|_\Sigma^2] + \|E[\hat\theta]-\theta^*\|_\Sigma^2 \\
&:= \mathrm{Var} + \mathrm{Bias}^2
\end{aligned}$$

In the unregularized case, $E[\hat\theta] = (X^TX)^{-1}X^T(X\theta^* + E[\varepsilon]) = \theta^*$

Bias $= 0$. When $\lambda > 0$, bias will be non-zero.

(ii) Rotation will not influence the regularized least squares.

Suppose $R \in \mathbb{R}^{d\times d}$ is an orthogonal matrix, so $X \to XR$, $\theta^* \to R^T\theta^*$. We still consider the excess risk:

$$E[XR(R^TX^TXR + n\lambda I)^{-1}R^TX^T(XRR^T\theta^* + \varepsilon) - XRR^T\theta^*\|_2^2]$$

which comes from $E[\|X\hat\theta - X\theta^*\|_2^2]$.

If we take SVD: $X = USV^T$ and set $R = V$, we have $X^TX \mapsto (V^TVSU^T)(USV^TV) = S^2$, which is diagonal.

Therefore we could assume that Σ is diagonal:

$$\Sigma := \mathrm{diag}(\tau_1, \cdots, \tau_d)$$

Now we do some computation:

- Compute the mean of estimator:

$$\begin{aligned}
\bar\theta_j &:= E[\hat\theta_j] \\
&= E[\Sigma_\lambda^{-1}\tfrac{1}{n}X^T(X\theta^* + \varepsilon)]_j \\
&= E[\Sigma_\lambda^{-1}\Sigma\theta^* + \Sigma_\lambda^{-1}\tfrac{1}{n}X^T\varepsilon]_j \qquad (\Sigma = \tfrac{1}{n}X^TX) \\
&= \frac{\tau_j}{\tau_j + \lambda}\theta_j^* \qquad (\text{since } E[\varepsilon] = 0.)
\end{aligned}$$

Note: $\hat\theta$ is shrunk by λ.

- Compute the squared bias term:

$$\begin{aligned}
\mathrm{Bias}^2 &= \|\bar\theta - \theta^*\|_\Sigma^2 \\
&= \sum_{j=1}^d \tau_j\left(\frac{\tau_j}{\tau_j+\lambda}\theta_j^* - \theta_j^*\right)^2
\end{aligned}$$

$$= \sum_{j=1}^{d} \tau_j \lambda^2 (\theta_j^*)^2 / (\tau_j + \lambda)^2$$

Note: as $\lambda \to \infty$, Bias $\to \|\theta^*\|_\Sigma^2$, which implies $\bar{\theta} \to 0$.

- Compute the variance term:

$$\begin{aligned}
\mathrm{Var} &= E[\|\hat{\theta} - \bar{\theta}\|_\Sigma^2] \\
&= E[\| \Sigma_\lambda^{-1} n^{-1} X^T (X\theta^* + \varepsilon) - \Sigma_\lambda^{-1} n^{-1} X^T X \theta^* \|_\Sigma^2] \\
&= E[\| \Sigma_\lambda^{-1} n^{-1} X^T \varepsilon \|_\Sigma^2] \\
&= \frac{1}{n^2} E[\varepsilon^T X \Sigma_\lambda^{-1} \Sigma \Sigma_\lambda^{-1} X^T \varepsilon] \\
&= \frac{1}{n^2} \mathrm{tr}(\Sigma_\lambda^{-1} \Sigma \Sigma_\lambda^{-1} X^T E[\varepsilon\varepsilon^T] X) \\
&= \frac{\sigma^2}{n} \mathrm{tr}(\Sigma_\lambda^{-1} \Sigma \Sigma_\lambda^{-1} \Sigma) \\
&= \frac{\sigma^2}{n} \sum_{j=1}^{d} \left(\frac{\tau_j}{\tau_j + \lambda}\right)^2
\end{aligned}$$

Note: Regularization reduce the variance since $\lambda > 0$.

Now we should balance bias and variance.

Goal: minimise the sum of upper bounds of Bias^2 and Var.

① $$\mathrm{Bias}^2 = \sum_{j=1}^{d} \tau_j \lambda^2 (\theta_j^*)^2 / (\tau_j + \lambda)^2 \le \sum_{j=1}^{d} \tau_j \lambda^2 (\theta_j^*)^2 / (2\tau_j \lambda) = \lambda \|\theta^*\|_2^2 / 2$$

② $$\mathrm{Var} = \frac{\sigma^2}{n} \sum_{j=1}^{d} \tau_j^2 / (\tau_j + \lambda)^2 \le \frac{\sigma^2}{n} \sum_{j=1}^{d} \tau_j^2 / (2\tau_j \lambda) = \mathrm{tr}(\Sigma) \sigma^2 / (2n\lambda)$$

$$\Rightarrow \quad \min \quad \lambda \|\theta^*\|_2^2 / 2 + \mathrm{tr}(\Sigma) \sigma^2 / (2n\lambda)$$

We have $\lambda = \sqrt{\mathrm{tr}(\Sigma)\sigma^2 / n\|\theta^*\|_2^2}$ and

$$E[L(\hat{\theta}) - L(\theta^*)] \le \sqrt{\|\theta^*\|_2^2 \mathrm{tr}(\Sigma) \sigma^2 / n} \quad \cdots (*)$$

Note: $(*)$ no longer depends on d.

$(*) \sim O(\sqrt{\frac{1}{n}})$, which is slower than previous $O(\frac{1}{n})$.