31.1 High-Dimensional Linear Regression

We recall some basic facts about linear regression. Assume we have n independent observations of the form

$$Y_j = \langle \lambda, X_j \rangle + \varepsilon_j,$$

where λ is a p-dimensional unknown vector (regression coefficients), X_j is a p-dimensinoal known vector (design), and ε_j is $\mathcal{N}(0, \sigma^2)$.

Seeking

$$\widehat{\lambda} = \underset{v \in \mathcal{R}^p}{\operatorname{argmin}} \left|\left| Y - \mathbb{X}v \right|\right|_2^2 ,$$

where $\mathbb{X} = [X_1; \cdots ; X_n] \in \mathbb{R}^{n \times p}$. The Moore-Penrose pseudo-inverse gives

$$\widehat{\lambda} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y.$$

Then

$$\begin{aligned}
\left|\left|\widehat{\lambda} - \lambda\right|\right|^2 &= \left|\left|(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{X}\lambda + \varepsilon)\right|\right|^2 \\
&= \left|\left|(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\varepsilon\right|\right|^2 \\
\mathbb{E}\left|\left|\widehat{\lambda} - \lambda\right|\right|^2 &= \langle(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\varepsilon, (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\varepsilon\rangle \\
&= \mathbb{E}\langle\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-2}\mathbb{X}^T\varepsilon, \varepsilon\rangle \\
&= \mathbb{E}\left|\left|\mathbb{X}^T\varepsilon\right|\right|^2 \\
&= p\sigma^2.
\end{aligned}$$

if $\mathbb{X}^T\mathbb{X} = I$. This bound is not satisfactory when p is large.

Assume that we have additional prior information about λ, expressed as $\lambda \in K$ where K is known, e.g.

$$K = \{\lambda \in \mathbb{R}^p : |\lambda|_0 \leq s\},$$

where $|\lambda|_0 := |\{j : \lambda_j \neq 0\}|$ and s stands for "sparsity" since usually $s << K$. First consider the scenario when the measurement vectors $x_1, \ldots, x_n$ are random. More specifically, assume that $x_j \sim \mathcal{N}(0, I_p)$ are iid and that $\Sigma^2 = 0$.

Let $\lambda \in K \subseteq \mathbb{R}^p$ be unknown and let $Y = \mathbb{X}\lambda$, where $\mathbb{X} \in \mathbb{R}^{n \times p}$ and the rows are $X_j \sim \mathcal{N}(0, \sigma^2 I_p)$. We wish to estimate λ. We know that $\lambda \in E$ an affine *random* subspace of dimension $p-\text{rank}(X) \geq p-n$. Assume that K is bounded and let $\eta \in \mathbb{R}^p$ be a unit vector.

The width of K in direction η is

$$w_\eta(K) = \sup_{u,v \in K} \langle \eta, u - v \rangle.$$

The mean width of K is

$$\hat{w}(K) = \mathbb{E} w_\eta(K),$$

where $\eta \sim U(S^{p-1})$.

The Gaussian mean width of K is

$$w(K) = \mathbb{E} w_g(K),$$

where $g \sim \mathcal{N}(0, I_p)$. It is equivalent to the usual mean width since

$$w(K) = \mathbb{E} \sup_{u,v \in K} \langle g, u - v \rangle = \mathbb{E} \left|\left|g\right|\right|_2 \sup_{u,v \in K} \langle \frac{g}{||g||}, u - v \rangle$$

where $g/\left|\left|g\right|\right|_2 \in S^{p-1}$ and $||g||$ are independent. It's true that

$$\mathbb{E} \left|\left|g\right|\right|_2 = \sqrt{2} \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)}$$

and $p/\sqrt{p+1} \leq \mathbb{E} \left|\left|g\right|\right|_2 \leq \sqrt{p}$.