Lecture 12: October 10

Lecturer: Alessandro Rinaldo *Scribes: Biswajit Paria*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

12.1 Linear Model

Consider a matrix $X \in \mathbb{R}^{n\times d}$, and unknown parameter $\theta^* \in \mathbb{R}^d$, and observations $y \in \mathbb{R}^n$. A linear model is the one where y is a perturbed linear function of X. More precisely

$$y = X\beta^* + \epsilon \tag{12.1}$$

where $\epsilon \in \mathbb{R}^n$ is a random vector. For simplicity we assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n$ and independent. The rows of X are often populated from a set of functions $\{f_j\}_{j=1}^d$. That is,

$$y_i = \sum_{j=1}^{d} \beta_j^* f_j(t_i) + \epsilon_i \tag{12.2}$$

where $\{t_1, \ldots, t_n\}$ is a set of design points, and $(f_1(t_i), \ldots, f_d(t_d)) =: X_i$ constitutes the i th row of X. For simplicity of analysis we assume that the design points and consequently the design matrix is fixed.

In reality however

1. The model is not linear.
2. The design matrix X is not fixed.
3. The variance is not constant.

12.1.1 Prediction and Estimation

We are interested in the following two problems:

1. **Prediction or mean estimation:** To predict the value of $y = x\beta^* + \epsilon$ for some given x. More formally, we want to minimize the mean square error for estimating $\mathbb{E}[y] = X\beta^*$

$$\frac{1}{n}\mathbb{E}[\|\tilde{Y} - X\widehat{\beta}\|^2] = \frac{1}{n}\mathbb{E}[\|X\beta^* - X\widehat{\beta}\|^2] + \frac{1}{n}\mathbb{E}[\|\tilde{\epsilon}\|^2] \tag{12.3}$$

   where $\tilde{y} = X\beta + \tilde{\epsilon}$, and the expectation is with respect to $\tilde{y}$ and y.

2. **Parameter estimation:** To estimate the unknown parameter β^*. That is to minimize $\mathbb{E}[\|\widehat{\beta} - \beta^*\|^2]$. In general, this problem is harder than prediction.

12.2 Ordinary Least Squares (OLS)

Let $(Y_1, X_1), \dots, (Y_n, X_n), (X, Y) \in \mathbb{R}^{d+1}$. The true parameter β^* is given as the solution minimizing the squared loss
$$\beta^* = \arg\min_{\beta} \mathbb{E}[(Y - X^T\beta)^2] = \Sigma^{-1}\alpha, \tag{12.4}$$
for $\alpha = \mathbb{E}[XY]$ and $\Sigma = \mathrm{cov}[X]$. For given samples $(Y_1, X_1), \dots, (Y_n, X_n)$ replacing the expectation by the mean of the sample square losses and minimizing yields the OLS estimator given by
$$\widehat{\beta} = (X^TX)^{-1}X^TY \tag{12.5}$$
whenever $(X^TX)^{-1}$ exists. The inverse may not always exist, for instance when $d > n$.

However, $\beta \mapsto \|Y - X\beta\|^2$ is a convex function, and can always be minimized. By the first order optimality condition, we have
$$X^TX\beta = X^TY. \tag{12.6}$$
Any β that satisfies the above equation minimizes the squared loss. If X is rank deficient for some β satisfying the above equation and $\Delta \in \mathrm{kernel}(X)$, $\beta + \Delta$ also satisfies the above equation as $X\Delta = 0$. One solution is given by
$$\widehat{\beta} = (X^TX)^+X^TY. \tag{12.7}$$
$(X^TX)^+$ is the Moore Penrose pseudo-inverse of $(X^TX)^+$, as described below.

12.2.1 Pseudo-Inverse

Pseudo-inverse of a matrix $A \in \mathbb{R}^{m\times n}$ is given by the unique matrix $A^+ \in \mathbb{R}^{n\times m}$ satisfying
$$AA^+A = A, \quad A^+AA^+ = A^T, \quad A^+A \text{ and } AA^+ \text{ are symmetric} \tag{12.8}$$
If A is independent columns $A^+ = (A^TA)^{-1}A^T$, and if A has independent rows $A^+ = A^T(AA^T)^{-1}$. More generally if $\mathrm{rank}(A) = r \le \min\{n, m\}$ then if $A = UDV^T$ be its SVD with D being a diagonal matrix of size r, then $A^+ = VD^{-1}U^T$.

Thus we get $X\widehat{\beta} = \sum_{J=1}^{r} U_J(U_JY)$ where $U_1, \dots, U_r$ are the columns of U in the SVD of $X = UDV^T$. $X\widehat{\beta}$ is the orthogonal projection of Y in the range of X. We next study the behaviour of this estimator $X\widehat{\beta}$.

12.2.2 Bounding the prediction risk

Theorem 12.1 *Suppose $\epsilon \in \mathrm{SG}_n(\sigma^2)$ be a vector of independent sub-gaussian random variables. Then for some universal constant $c > 0$*
$$\mathbb{P}\left(\frac{1}{n}\|X(\widehat{\beta} - \beta^*)\|^2 \ge c\sigma^2\left(\frac{r + \log(1/\delta)}{n}\right)\right) \le \delta \tag{12.9}$$
for $\forall\delta \in (0, 1),\ r = \mathrm{rank}(X^TX)$.

Proof: We start with an inequality following from the definition of $\widehat{\beta}$.
$$\|Y - X\widehat{\beta}\|^2 \le \|Y - X\beta^*\|^2 = \|\epsilon\|^2. \tag{12.10}$$

Next we have

$$
\begin{aligned}
&\|Y - X\widehat{\beta}\|^2 = \|X\beta^* + \epsilon - X\widehat{\beta}\|^2 = \|X(\beta^* - \widehat{\beta})\|^2 + \|\epsilon\|^2 - 2\epsilon^T X(\widehat{\beta} - \beta^*) \\
&\implies \|X(\beta^* - \widehat{\beta})\|^2 \le 2\epsilon^T X(\widehat{\beta} - \beta^*) \\
&\implies \|X(\beta^* - \widehat{\beta})\| \le 2 \underbrace{\epsilon^T}_{\text{a random quantity}} \underbrace{\frac{X(\widehat{\beta} - \beta^*)}{\|X(\widehat{\beta} - \beta^*)\|}}_{\text{also a random quantity depending on } \epsilon} \\
&\implies \|X(\beta^* - \widehat{\beta})\| \le 2 \sup_{v \in \mathbb{S}^{n-1}} \epsilon^T v
\end{aligned}
$$

The last step is a common trick known as *sup-out*. We have seen in the previous lectures that this expression can be bounded using an appropriate ϵ-net. The above however gives us a weaker bound with a factor of n. It is possible to have a dependence on r instead.

Note that when X is rank deficient, in the above expression ϵ is projected to a subspace of rank r. Intuitively, this should lead to a factor of r. We make this more formal as follows. Let $\phi_{n\times r}$ be a orthogonal matrix spanning the column space of X. Then $X(\widehat{\beta} - \beta^*) = \phi v$ for some $v \in \mathbb{R}^r$. Substituting this we get

$$
\|X(\beta^* - \widehat{\beta})\| \le 2\epsilon^T \frac{X(\widehat{\beta} - \beta^*)}{\|X(\widehat{\beta} - \beta^*)\|} = 2\epsilon^T \frac{\phi v}{\|\phi v\|} = 2\frac{\tilde{\epsilon}^T v}{\|v\|} \le 2 \sup_{x \in \mathbb{S}^{r-1}} \tilde{\epsilon}^T x,
$$

where $\tilde{\epsilon} = \phi^T \epsilon \in SG_r(\sigma^2)$. Squaring we get

$$
\begin{aligned}
&\mathbb{E}[\|X(\beta^* - \widehat{\beta})\|^2] \le 4\mathbb{E}\left[(\sup_{x \in \mathbb{S}^{r-1}} \tilde{\epsilon}^T x)^2\right] = 4\sum_{i=1}^{r} \mathbb{E}[\tilde{\epsilon}_i^2] \le 4r\sigma^2 \\
&\implies \frac{1}{n}\mathbb{E}[\|X(\beta^* - \widehat{\beta})\|^2] \le \frac{4r\sigma^2}{n}
\end{aligned}
$$

To get a high probability bound, we can use the fact that $(\tilde{\epsilon}^T x)^2 \in \mathrm{SE}((11\sigma^2)^2, 11\sigma^2)$ for all $x \in \mathbb{S}^{r-1}$, and then use a discretization argument to bound the sup. ■