Regression and SVD

Kaiwen Zhou

Contents

1 Topic: Regression and SVD

Problem 1.1. *Show that* $\text{rank}(X) = \text{rank}\left(X^\top X\right)$ *for any matrix* X *with real entries.*

Solution:

Method 1: Suppose $X \in \mathbb{R}^{m\times n}$. Then, by SVD, we have $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ having non-negative singular values on its diagonal and zeroes elsewhere.

Since U, V are orthogonal, they are invertible. Because multiplying an invertible matrix does not change the rank of a matrix, we get $\text{rank}(X) = \text{rank}(U\Sigma V^\top) = \text{rank}(\Sigma)$.

On the other hand, we have

$$\text{rank}(X^\top X) = \text{rank}(V\Sigma^\top U^\top U\Sigma V^\top) = \text{rank}(V\Sigma^\top\Sigma V^\top) = \text{rank}(\Sigma^\top\Sigma) = \text{rank}(\Sigma)$$

Hence, we have $\text{rank}(X) = \text{rank}(X^\top X)$.

Method 2: We will prove $Im(X^\top) = Im(X^\top X)$. First, we have

$$Im(X^\top X) = \{X^\top X v \mid v \in \mathbb{R}^n\} = \{X^\top u \mid u \in Im(X)\}$$

By the rank-nullity theorem, we get $dim(Im(X)) + dim(Ker(X^\top)) = m$. Moreover, $Im(X)$ and $Ker(X^\top)$ are orthogonal. Therefore $Im(X)$ and $Ker(X^\top)$ together span the whole space $\mathbb{R}^m$.

It follows that $\forall u \in \mathbb{R}^m$, we can uniquely decompose it as the sum $u = u_r + u_n$ where $u_r \in Im(X)$ and $u_n \in Ker(X^\top)$.

Therefore, we obtain

$$Im(X^\top) = \{X^\top u, u \in \mathbb{R}^m\} = \{X^\top(u_r + u_n) \mid u_r \in Im(X), u_n \in Ker(X^\top)\} = \{X^\top u_r \mid u_r \in Im(X)\} = Im(X^\top X)$$

Hence, we conclude

$$\text{rank}(X) = dim(Im(X)) = dim(Im(X^\top)) = dim(Im(X^\top X)) = \text{rank}(X^\top X)$$

Problem 1.2. *Write a numpy function to compute the pseudo-inverse of a real matrix, building upon the numpy SVD function discussed in class. Test that your function works by generating a sequence of random invertible matrices, and check that for each one, your pseudo-inverse equals the actual inverse up to numerical precision.*

Solution: WLOG, suppose $X \in \mathbb{R}^{m\times n}$, $m \le n$. Then, by SVD, we have $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ has non-negative singular values $\sigma_1 \ge \cdots \ge \sigma_p \ge \sigma_{p+1} = \cdots = \sigma_m = 0$, $p \le \min\{m, n\}$, on its diagonal and zeroes elsewhere.

Then, we have

$$X^\dagger = \left(X^\top X\right)^{-1} X^\top = \left(V\Sigma^\top U^\top U\Sigma V^\top\right)^{-1} V\Sigma^\top U^\top = \left(V\left(\Sigma^\top\Sigma\right)V^\top\right)^{-1} V\Sigma^\top U^\top = V\left(\Sigma^\top\Sigma\right)^{-1}\Sigma^\top U^\top = V\Sigma^\dagger U^\top$$

That is, the pseudo-inverse given by the SVD is $X^\dagger = V\Sigma^\dagger U^\top$.

For the code, see the attached jupyter notebook.

Problem 1.3. *Prove that the Moore-Penrose pseudo-inverse is given as a one-sided limit of ridge regression problems, i.e. prove*

$$X^\dagger y = \lim_{\lambda\to 0^+} \left(X^\top X + \lambda I\right)^{-1} X^\top y$$

Hint: use the SVD.

Solution: Recall that an alternate definition of $X^\dagger$, see [1] Albert (1972), is

Definition 1.4. *The Moore-Penrose pseudo-inverse* $X^\dagger$ *is defined so that* $X^\dagger y$ *is the minimum-norm vector among all minimizers of* $\min_\beta \|y - X\beta\|^2$.

Therefore, we want to show the limit we get from the RHS is a minimizer of $\min_\beta \|y - X\beta\|^2$, and it's the one with the smallest norm among all minimizers. We apply the SVD to help us show this.

WLOG, suppose $X \in \mathbb{R}^{m\times n}$, $m \le n$. Then, by SVD, we have $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ has non-negative singular values $\sigma_1 \ge \cdots \ge \sigma_p \ge \sigma_{p+1} = \cdots = \sigma_m = 0$, $p \le \min\{m, n\}$, on its diagonal and zeroes elsewhere.

Then, we have

$$
\begin{aligned}
\left(X^\top X + \lambda I\right)^{-1} X^\top y &= \left(V\Sigma^\top U^\top U\Sigma V^\top + \lambda I\right)^{-1} V\Sigma^\top U^\top y \\
&= \left(V(\Sigma^\top\Sigma + \lambda I)\, V^\top\right)^{-1} V\Sigma^\top U^\top y \\
&= V\left(\Sigma^\top\Sigma + \lambda I\right)^{-1} \Sigma^\top U^\top y \\
&= V \begin{bmatrix} \frac{\sigma_1}{\lambda+\sigma_1^2} & & & \\ & \ddots & & \\ & & \frac{\sigma_p}{\lambda+\sigma_p^2} & \\ & & & \\ & & & \end{bmatrix} U^\top y
\end{aligned}
$$

We then take the limit and obtain

$$
X^\dagger y = \lim_{\lambda\to 0^+} \left(X^\top X + \lambda I\right)^{-1} X^\top y = \lim_{\lambda\to 0^+} V \begin{bmatrix} \frac{\sigma_1}{\lambda+\sigma_1^2} & & & \\ & \ddots & & \\ & & \frac{\sigma_p}{\lambda+\sigma_p^2} & \\ & & & \\ & & & \end{bmatrix} U^\top y = V \begin{bmatrix} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_p} & \\ & & & \\ & & & \end{bmatrix} U^\top y := VSU^\top y, \quad S \in \mathbb{R}^{n\times m}
$$

To be a minimizer of $\min_\beta \|y - X\beta\|^2$, we need $\beta^* = X^\dagger y = VSU^\top y$ to satisfy the first-order necessary condition

$$
\frac{\partial L(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta}\frac{1}{2}\|y - X\beta\|^2 = \frac{\partial}{\partial \beta}\frac{1}{2}(y - X\beta)^\top (y - X\beta) = X^\top X\beta - X^\top y = 0 \tag{1}
$$

Plug $\beta^* = VSU^\top y$ in Equation (1) and apply SVD on X, we get

$$
\begin{aligned}
X^\top X\beta^* - X^\top y &= V\Sigma^\top U^\top U\Sigma V^\top VSU^\top y - V\Sigma^\top U^\top y \\
U, V\text{orthogonal} \implies &= V\Sigma^\top\Sigma SU^\top y - V\Sigma^\top U^\top y \\
&= V \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_p & \\ & & & \\ & & & \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_p & & \\ & & & & \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_p} & \\ & & & \\ & & & \end{bmatrix} U^\top y - V\Sigma^\top U^\top y \\
&= V\Sigma^\top U^\top y - V\Sigma^\top U^\top y \\
&= 0
\end{aligned}
$$

Therefore, $\beta^* = X^\dagger y = VSU^\top y$ is a minimizer of $\min_\beta \|y - X\beta\|^2$.

We now assume that there exists another minimizer $\beta_1 \in \mathbb{R}^n$ other than β^* and that $y = X\beta_1 + \varepsilon$ where $\varepsilon \in \mathbb{R}^m$.

Then, by Equation (1), we have

$$
0 = X^\top X\beta_1 - X^\top y = X^\top X\beta_1 - X^\top (X\beta_1 + \epsilon) \implies 0 = X^\top \varepsilon = V\Sigma^\top U^\top \varepsilon \implies 0 = \Sigma^\top U^\top \varepsilon = \begin{bmatrix} \sigma_1 u_1^\top \varepsilon \\ \vdots \\ \sigma_p u_p^\top \varepsilon \\ 0 \\ \vdots \\ 0 \end{bmatrix} \implies u_i^\top \varepsilon = 0, \text{ for } i = 1, \ldots, p
$$

It follows that

$$
SU^\top \varepsilon = \begin{bmatrix} \frac{1}{\sigma_1} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma_p} & \\ & & & \\ & & & \end{bmatrix} U^\top \varepsilon = \begin{bmatrix} \frac{1}{\sigma_1} u_1^\top \varepsilon \\ \vdots \\ \frac{1}{\sigma_p} u_p^\top \varepsilon \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0
$$

Finally, we consider ${\beta^*}^\top \beta^*$ and obtain

$$
\begin{aligned}
\|\beta^*\|^2 = {\beta^*}^\top \beta^* &= (VSU^\top y)^\top VSU^\top y \\
&= (VSU^\top X\beta_1 + VSU^\top \varepsilon)^\top (VSU^\top X\beta_1 + VSU^\top \varepsilon) \\
SU^\top \varepsilon = 0 \implies &= (VS\Sigma V^\top \beta_1)^\top (VS\Sigma V^\top \beta_1) \\
&= \beta_1^\top V \Sigma^\top S^\top V^\top V S \Sigma V^\top \beta_1 \\
V \text{ orthogonal } \implies &= \beta_1^\top V \Sigma^\top S^\top S \Sigma V^\top \beta_1 \\
&= \beta_1^\top V \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}^\top \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} V^\top \beta_1 \\
&= \beta_1^\top \left(\sum_{i=1}^{p} v_i v_i^\top \right) \beta_1 \\
&= \beta_1^\top \left(\sum_{i=1}^{n} v_i v_i^\top \right) \beta_1 - \beta_1^\top \left(\sum_{i=p+1}^{n} v_i v_i^\top \right) \beta_1 \\
&= \beta_1^\top V V^\top \beta_1 - \sum_{i=p+1}^{n} \beta_1^\top v_i v_i^\top \beta_1 \\
&= \|\beta_1\|^2 - \sum_{i=p+1}^{n} \|v_i^\top \beta_1\|^2 \\
&\le \|\beta_1\|^2
\end{aligned}
$$

where the equality is attained when $p = n$, i.e., X must be full column ranked.

Since β_1 is chosen arbitrarily, we can now conclude that $\beta^* = X^\dagger y = VSU^\top y$ is a minimizer of $\min_\beta \|y - X\beta\|^2$, and it's the one with the smallest norm among all minimizers.

Hence, we conclude that the limit given by $X^\dagger y = \lim_{\lambda \to 0^+} \left(X^\top X + \lambda I\right)^{-1} X^\top y$ is, by definition, the Moore-Penrose pseudo-inverse.

References

[1] Albert, Arthur (1972). *Regression and the Moore-Penrose pseudo-inverse*. Academic Press.