SD-TSIA204
Properties and non uniqueness of Ordinary Least Squares

Ekhine Irurozki
Télécom Paris

Non uniqueness of the OLS solution

Singularity in High-Dimensional Design Matrices

Motivation

- Let $X \in \mathbb{R}^{n\times(p+1)}$ be a design matrix.
- Super-collinearity occurs if the columns of X are linearly dependent.
- Consequence :
$$\text{rank}(X^\top X) < p+1 \quad \Rightarrow \quad X^\top X \text{ is singular (non-invertible).}$$

Spectral Decomposition of Symmetric Matrices

Notations and preliminaries

- A square matrix A is singular iff $\det(A) = 0$.
- For symmetric A, $\det(A) = \prod_j \lambda_j$, with real eigenvalues λ_j.
- Hence, A is singular iff at least one $\lambda_j = 0$.
- Spectral theorem : if $A \in \mathbb{R}^{p \times p}$ is symmetric, then
$$A = V \Lambda V^\top, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p),$$
  with $V = [v_1 \ldots v_p]$ orthogonal ($V^\top V = I$).
- Equivalently :
$$A = \sum_{j=1}^{p} \lambda_j v_j v_j^\top.$$
- This expresses A as a sum of rank-1 matrices.

Inverse via Spectral Decomposition

- If all $\lambda_j \neq 0$, the inverse of A is
$$A^{-1} = \sum_{j=1}^{p} \lambda_j^{-1} v_j v_j^\top$$
- If any $\lambda_j = 0$, the inverse is undefined.

Moore-Penrose Inverse via Spectral Decomposition

- For symmetric A :

$$A^+ = \sum_{j:\lambda_j \neq 0} \lambda_j^{-1} v_j v_j^\top$$

- v_j are eigenvectors of A.
- Properties :

$$A^+ A A^+ = A^+, \quad A A^+ A = A$$

- Provides the minimum-norm solution for rank-deficient systems.

Exercise: Show that $A^+ A A^+ = A^+$

Solutions for the OLS using the normal equations and the generalized inverse

- **A** solution of the normal equations :
$$\widehat{\boldsymbol{\theta}} = (X^\top X)^+ X^\top \mathbf{y}$$
- Let $\ker(X) = \{v \in \mathbb{R}^p : Xv = 0\}$.
- Then for any $v \in \ker(X)$, we have $X^\top X v = 0$.
- The set of **all** solutions of the normal equations is :
$$\widehat{\boldsymbol{\theta}} = (X^\top X)^+ X^\top \mathbf{y} + v, \quad \forall v \in \ker(X)$$

Properties of the OLS solution

Model I : The fixed design model

$$
\begin{aligned}
y_i &= \theta_0^\star + \sum_{k=1}^{p} \theta_k^\star x_{i,k} + \varepsilon_i \\
x_i^\top &= (1, x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^{p+1} \\
\varepsilon_i &\overset{i.i.d}{\sim} \varepsilon, \text{ for } i = 1, \ldots, n \\
\mathbb{E}(\boldsymbol{\varepsilon}) &= 0, \mathrm{Var}(\epsilon) = \sigma^2
\end{aligned}
$$

- x_i is deterministic
- σ^2 is called the noise level

Example :

- Physical experiment when the analyst is choosing the design *e.g.*,temperature of the experiment
- Some features are not random *e.g.*,time, location.

Model I with Gaussian noise : The fixed design Gaussian model

$$
\begin{aligned}
y_i &= \theta_0^\star + \sum_{k=1}^{p} \theta_k^\star x_{i,k} + \varepsilon_i \\
x_i^\top &= (1, x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^{p+1} \\
\varepsilon_i &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \text{ for } i = 1, \ldots, n
\end{aligned}
$$

- Parametric model : specified by the two parameters $(\boldsymbol{\theta}, \sigma)$
- Strong assumption

Model II : The random design model

$$
\begin{aligned}
y_i &= \theta_0^\star + \sum_{k=1}^{p} \theta_k^\star x_{i,k} + \varepsilon_i \\
x_i^\top &= (1, x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^{p+1} \\
(\varepsilon_i, x_i) &\overset{i.i.d}{\sim} (\varepsilon, x), \text{ for } i = 1, \ldots, n \\
\mathbb{E}(\varepsilon|x) &= 0, \ \mathrm{Var}(\varepsilon|x) = \sigma^2
\end{aligned}
$$

Rem: here, the features are modelled as random (they might also suffer from some noise)

The ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{k=1}^{p} \theta_k x_{i,k} \right)^2$$

How to deal with these two models ?

- The estimator is the same for both models
- The mathematics involved are different for each case
- The study of the fixed design case is easier as many closed formulas are available
- The two models lead to the same estimators of the variance σ^2

Prediction

$$\textbf{Prediction vector :} \qquad \widehat{\mathbf{y}} = X\hat{\boldsymbol{\theta}}$$

Rem: $\hat{\mathbf{y}}$ depends linearly on the observation vector $\mathbf{y}$

Rem: an **orthogonal projector** is a matrix H such that

1. H is symmetric : $H^\top = H$
2. H is idempotent : $H^2 = H$

Proposition Writing H the orthogonal projector onto the space span by the columns of X, one gets $\hat{\mathbf{y}} = H\mathbf{y}$

If X is full (column) rank, then $H = X(X^\top X)^{-1}X^\top$ is called the **hat matrix**

See that $\widehat{\mathbf{y}} = H\mathbf{y} = X(X^\top X)^{-1}X^\top \mathbf{y}$

Exercise: Show that H is an orthogonal projector

Prediction (continued)

If a new observation $x_{n+1} = (x_{n+1,1}, \dots, x_{n+1,p})$ is provided, the associated prediction is :

$$\hat{y}_{n+1} = \langle \hat{\boldsymbol{\theta}}, (1, x_{n+1,1}, \dots, x_{n+1,p})^\top \rangle$$
$$\hat{y}_{n+1} = \hat{\theta}_0 + \sum_{j=1}^{p} \hat{\theta}_j x_{n+1,j}$$

Rem: the normal equation ensures **equi-correlation** between observations and features :

$$(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y} \Leftrightarrow X^\top \hat{\mathbf{y}} = X^\top \mathbf{y}$$
$$\Leftrightarrow \begin{pmatrix} \langle \mathbf{x}_0, \hat{\mathbf{y}} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \hat{\mathbf{y}} \rangle \end{pmatrix} = \begin{pmatrix} \langle \mathbf{x}_0, \mathbf{y} \rangle \\ \vdots \\ \langle \mathbf{x}_p, \mathbf{y} \rangle \end{pmatrix}$$

Properties of the OLS estimator, $\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top \mathbf{y}$

Assuming full-rank X and the fixed design model with Gaussian noise,

- P1 : Equivalent expression : $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}$
- P2 : Unbiasedness : $\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ because $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$
- P3 : Covariance : $\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 (X^\top X)^{-1}$
- P4 : Distribution : $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2 (X^\top X)^{-1})$
- P5 : BLUE : $\hat{\boldsymbol{\theta}}$ is the Best Linear Unbiased Estimator
- P6 : Invariance : $\hat{\mathbf{y}}$ is invariant under linear transformations of the design matrix

---

Exercise: Prove the above statements

The trace of a matrix

Let $A \in \mathbb{R}^{n\times n}$ denote a matrix. The **trace** of A is the sum of the diagonal elements of A and is denoted by $\mathrm{tr}(A)$:

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Several properties :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- For any $A, B \in \mathbb{R}^{n\times n}$, and $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linearity)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- For any $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(A)$, hence if A is diagonalisable, the trace is the sum of the eigenvalues
- If H is an orthogonal projector $\mathrm{tr}(H) = \mathrm{rank}(H)$

Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Under model I, whenever the matrix X has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

<u>Proof</u> :

$$\begin{aligned}
R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \varepsilon) - \boldsymbol{\theta}^\star)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \varepsilon) - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top \varepsilon)^\top ((X^\top X)^{-1} X^\top \varepsilon)\right] = \mathbb{E}(\varepsilon^\top X (X^\top X)^{-2} X^\top \varepsilon) \\
&= \operatorname{tr}[\mathbb{E}(\varepsilon^\top X (X^\top X)^{-1} (X^\top X)^{-1} X^\top \varepsilon)] \text{ (thx to } \operatorname{tr}(u^\top u) = u^\top u) \\
&= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1} X^\top \varepsilon\varepsilon^\top X (X^\top X)^{-1}\right]\right) \\
&= \operatorname{tr}[(X^\top X)^{-1} X^\top \mathbb{E}(\varepsilon\varepsilon^\top) X (X^\top X)^{-1}] \\
&= \sigma^2 \operatorname{tr}((X^\top X)^{-1})
\end{aligned}$$

Prediction risk (normalized) $R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Under model I, whenever the matrix X has full rank, we have

$$R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \frac{\mathrm{rank}(X)}{n}$$

Because X has full rank, $\mathrm{rank}(X) = p + 1$.

<u>Proof</u> : As before

$$\begin{aligned}
n \cdot R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1}(X^\top X)(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}) \\
&= \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}) \\
&= \mathrm{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H \boldsymbol{\varepsilon})] = \mathrm{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H^\top H \boldsymbol{\varepsilon})] \\
&= \mathrm{tr}[\mathbb{E}(H\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top H^\top)] = \mathrm{tr}\left(H\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)H^\top\right) \\
&= \sigma^2 \,\mathrm{tr}(H) = \sigma^2 \,\mathrm{rank}(H) = \sigma^2 \,\mathrm{rank}(X)
\end{aligned}$$

More Exercises

- Compute the variance and covariance of the OLS estimator for the one-dimensional model.
- Show that the predicted value $\widehat{\mathbf{y}}$ is invariant under a full-rank linear transformation of the predictors X.
- Show that the hat matrix defined with the Moore–Penrose generalized inverse is an orthogonal projection matrix.