Lecture #3 Notes Summary

Least square estimator, Gauss-Markov theorem

Linear regression and the Gauss-Markov theorem

Let $\boldsymbol{\theta}$ be an unknown parameter of dimension m, and let K be a $m \times r-$matrix of full column rank (i.e., the columns of K are independent).

Assume that we have a vector $\boldsymbol{y} \in \mathbb{R}^N$ of measurements, satifying

$$\boldsymbol{y} = A\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \qquad \operatorname{Var}[\boldsymbol{\epsilon}] = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 \boldsymbol{I}_N,$$

where A is a $N \times m-$matrix and $\boldsymbol{I}_N$ denotes the $N \times N-$identity matrix. In words, the errors ϵ_i on the measurements are unbiased and uncorrelated.

We want to recover $\boldsymbol{\theta}$ from the vector of measurements $\boldsymbol{y}$. More generally, assume that we want to estimate the value of $\boldsymbol{\zeta} = K^T\boldsymbol{\theta}$. An estimator $\hat{\boldsymbol{\zeta}}$ is a function of $\boldsymbol{y}$ that is used to estimate $\boldsymbol{\zeta}$. We say that the estimator $\hat{\boldsymbol{\zeta}}$ is linear if $\hat{\boldsymbol{\zeta}} = H^T\boldsymbol{y}$ for some matrix H.

> **Definition 1.** An estimator $\hat{\boldsymbol{\zeta}}$ for $K^T\boldsymbol{\theta}$ is called *unbiased* if $\mathbb{E}[\hat{\zeta}] = K^T\boldsymbol{\theta}$.

> **Proposition 1.** *The linear estimator $\hat{\boldsymbol{\zeta}} = H^T\boldsymbol{y}$ is unbiased if and only if $A^T H = K$.*

> **Theorem 2** (Least square estimate)**.** *Assume that the system $A\boldsymbol{\theta} = \boldsymbol{y}$ is overdetermined (i.e.,* rank $A =$ *m). Then, the least square estimate of $\boldsymbol{\theta}$ (that is, a minimizer of $\|A\boldsymbol{\theta} - \boldsymbol{y}\|^2$) is unique and given by*
>
> $$\hat{\boldsymbol{\theta}} = (A^T A)^{-1} A^T \boldsymbol{y}.$$

Proof. We differentiate the least square criterion:

$$\nabla_\theta\Big((A\boldsymbol{\theta} - \boldsymbol{y})^T(A\boldsymbol{\theta} - \boldsymbol{y})\Big) = \nabla_\theta\Big(\boldsymbol{\theta}^T A^T A\boldsymbol{\theta} - 2\boldsymbol{\theta}^T A^T\boldsymbol{y} + \text{constant}\Big) = 2A^T A\boldsymbol{\theta} - 2A^T\boldsymbol{y}$$

The minimizer is found by equating the above expression with $\mathbf{0}$, and it is unique because $A^T A$ is invertible. □

> **Theorem 3** (Gauss Markov)**.** *Let A be a matrix of full column rank. Let $\hat{\boldsymbol{\zeta}} = H^T\boldsymbol{y}$ be an unbiased estimator for $\boldsymbol{\zeta} = K^T\boldsymbol{\theta}$, that is, $A^T H = K$. Then, we have*
>
> $$\operatorname{Var}[\hat{\zeta}] = H^T \operatorname{Var}[\boldsymbol{y}] H = \sigma^2 H^T H \succeq \sigma^2\, K^T (A^T A)^{-1} K.$$
>
> *Moreover, the lower bound is attained for $H^* = A(A^T A)^{-1}K$. In other words, the best linear unbiased estimator (BLUE) for $K^T\boldsymbol{\theta}$ is $K^T\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is the least square estimate of $\boldsymbol{\theta}$.*

Proof. The fact that the lower bound is attained for H^* is clear by substituting its expression in $H^{*T}H^*$. Hence, the only thing to prove is the matrix inequality. The matrix

$$\begin{pmatrix} A^T A & K \\ K^T & H^T H \end{pmatrix}$$

is positive semidefinite, because it can be written as

$$\begin{pmatrix} A^T \\ H^T \end{pmatrix} \begin{pmatrix} A^T \\ H^T \end{pmatrix}^T.$$

Hence, since $A^T A \succ 0$ (because A has full column rank and $\operatorname{rank} A^T A = \operatorname{rank} A$), the Schur complement lemma gives us

$$HH^T \succeq K^T (A^T A)^{-1} K.$$

□

In fact, there is a more general result for the case where A does not has full rank, but the estimability condition $\operatorname{im} K \subseteq \operatorname{im} A^T$ is satisfied (i.e., $\exists H : A^T H = K$). The proof is similar to the above, but relies on the extended version of the Schur complement lemma.

Theorem 4 (Extended Gauss Markov)**.** *Let A be a matrix such that an unbiased estimator for $\boldsymbol{\zeta} = K^T \boldsymbol{\theta}$ exists, that is, $\operatorname{im} K \subseteq \operatorname{im} A^T$ and $A^T H = K$. Then, we have*

$$\operatorname{Var}[\hat{\zeta}] = H^T \operatorname{Var}[\boldsymbol{y}] H = \sigma^2 H^T H \succeq \sigma^2 \, K^T (A^T A)^- K.$$

Moreover, the lower bound is attained for $H^* = A(A^T A)^\dagger K$.

An alternative formulation of this theorem is as follows:

Theorem 5. *Gauss Markov (extremal version) Let A be a matrix satisfying the estimability condition $\operatorname{im} K \subseteq \operatorname{im} A^T$. Then,*

$$\begin{array}{rl} \min_{\preceq} & \big(H^T H\big) = K^T (A^T A)^- K \\ s.t. & A^T H = K, \end{array}$$

where $\min_{\preceq}$ denotes a minimum with respect to the Löwner ordering.

Exercises

Recall the Hotelling's weighing problem. There are N objects, and in each trial you can put a subset L of the objects on the left pan of the balance, and another (disjoint) subset R of the objects on the right pan, so you measure $y_{LR} = \sum_{i \in L} \theta_i - \sum_{j \in R} \theta_j + \epsilon$.

1. Compute the information matrix of the approximate design which assigns weight $w_{LR} = \frac{1}{2^N}$ on each partition $(L, R) = (S, \bar{S})$ of the N objects (that is, in each weighing all objects are on the balance).

2. Now, consider a spring balance (there is only one pan, on which you can put a subset S of the N objects. For a vector $\boldsymbol{w} \in \Delta_N$, Compute the information matrix of the approximate design which assigns the weight $w_S = \frac{w_i}{\binom{N}{i}}$ to each subset S of $[N] = \{1, \dots, N\}$ of cardinality i.