

Lecture 13: Fisher information, Cramér-Rao Lower Bound

Lecturer: Yiqiao Zhong

Last update: 12:42:35, 2024-11-06

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

In this lecture, We will introduce an important lower-bound result for unbiased estimation.

1 Fisher Information

Suppose that X is drawn (e.g., single observation or sequence of i.i.d. data) from $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where the parameter space $\Theta \subset \mathbb{R}^k$ is an open set. Also suppose that P_θ has p.d.f. f_θ w.r.t. measure ν on $\mathbb{R}^k$ for any $\theta \in \Theta$, and f_θ is differentiable in θ .

Definition 1 (Fisher Information). *For the parametric function $\{P_\theta : \theta \in \Theta\}$, the Fisher information (matrix) is defined as*

$$I(\theta) = \mathbb{E} \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^\top \right],$$

where the expectation is taken over P_θ . We also write $I_X(\theta)$ to emphasize the dependence on X should confusion arises.

Note that by definition, $I(\theta)$ is always positive semidefinite. We call $\frac{\partial}{\partial \theta} \log f_\theta(X)$ the *score function*.

Proposition 1. (i) *If X is independent of Y , then the Fisher information about θ in (X, Y) is*

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$

If $X_1, \dots, X_n$ are i.i.d., then

$$I_{(X_1, \dots, X_n)}(\theta) = nI_{X_1}(\theta)$$

(ii) *Suppose that f_θ is twice differentiable in θ . Under a “regularity condition”,*

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta \partial \theta^\top} \log f_\theta(X) \right].$$

Proof. (i) Due the property of independence, joint p.d.f. becomes a sum under the logarithm (exercise).

(ii) A useful observation is the following identity.

$$\frac{\partial^2}{\partial \theta \partial \theta^\top} \log f_\theta(X) = \frac{\frac{\partial^2}{\partial \theta \partial \theta^\top} f_\theta(X)}{f_\theta(X)} - \frac{\partial}{\partial \theta} \log f_\theta(X) \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right]^\top.$$

When taking expectation, we get

$$\begin{aligned}\mathbb{E} \left[\frac{\frac{\partial^2}{\partial \theta \partial \theta^\top} f_\theta(X)}{f_\theta(X)} \right] &= \int \frac{\frac{\partial^2}{\partial \theta \partial \theta^\top} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx = \int \frac{\partial^2}{\partial \theta \partial \theta^\top} f_\theta(x) dx \\ &= \frac{\partial^2}{\partial \theta \partial \theta^\top} \int f_\theta(x) dx \quad (\text{regularity condition}) \\ &= 0. \quad (\text{since } \int f_\theta(x) dx \equiv 1)\end{aligned}$$

□

The regularity condition is an annoying technicality that ensures the exchangeability of integrals with differentials (again, physicists don't care). They hold under quite general scenarios such as smooth p.d.f. with a compact support or the exponential family.

2 Cramér-Rao Lower Bound

Recall that a symmetric matrix A is positive definite if $u^\top A u > 0$ for all nonzero vectors u , and it is positive semidefinite if $u^\top A u \geq 0$ for all vectors u . By convention, we write $A \geq B$ for two symmetric matrices of the same size if $A - B$ is positive semidefinite.

Let $\nu = g(\theta)$ be the quantity of interest where $g : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is differentiable. Suppose that X is drawn from a distribution in $\{P_\theta : \theta \in \Theta\}$ and $T(X)$ is an unbiased estimator of ν . Let f_θ be the density function with respect to a reference measure μ .

Theorem 1. Suppose that $I(\theta)$ is positive definite, and that for any $\theta \in \Theta$,

$$\frac{\partial}{\partial \theta} \int h(x) f_\theta(x) d\mu = \int h(x) \frac{\partial}{\partial \theta} f_\theta(x) d\mu, \quad \text{for any } \theta \in \Theta, \quad (1)$$

holds for $h(x) \equiv 1$ and $h(x) = T(x)$ (regularity conditions). Then,

$$\text{Var}(T(X)) \geq \left[\frac{\partial}{\partial \theta} g(\theta) \right]^\top [I(\theta)]^{-1} \frac{\partial}{\partial \theta} g(\theta).$$

Fisher Information provides a lower bound on variance (and thus MSE) of unbiased estimators for a parametric family. It may not be always achievable, but if we can find an estimator that achieves this lower bound, then we certify the optimality of this estimator in the class of unbiased estimators (namely certifying UMVUE). See wiki for the multivariate case.

Only $k = 1$ and $d = 1$, similar for $k > 1$. Our goal is

$$\text{Var}(T(X)) \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] \geq (g'(\theta))^2. \quad (2)$$

It suffices to show the following two equalities.

$$\mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] = \text{Var} \left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right), \quad g'(\theta) = \text{Cov} \left(T(X), \frac{\partial}{\partial \theta} \log f_\theta(X) \right).$$

Once they are proved, Eqn. 4 follows from Cauchy-Schwarz Inequality:

$$\text{Var}(\xi_1) \text{Var}(\xi_2) \geq [\text{Cov}(\xi_1, \xi_2)]^2.$$

Note that

$$\mathbb{E} \left[\frac{\partial}{\partial \theta} \log f_{\theta}(X) \right] = \int \frac{\frac{\partial}{\partial \theta} f_{\theta}(x)}{f_{\theta}(x)} f_{\theta}(x) dx = \int \frac{\partial}{\partial \theta} f_{\theta}(x) dx = \frac{\partial}{\partial \theta} \underbrace{\int f_{\theta}(x) dx}_{\equiv 1} = 0,$$

and that

$$\begin{aligned} \mathbb{E} \left[T(X) \frac{\partial}{\partial \theta} \log f_{\theta}(X) \right] &= \int T(x) \frac{\frac{\partial}{\partial \theta} f_{\theta}(x)}{f_{\theta}(x)} f_{\theta}(x) dx \\ &= \int T(x) \frac{\partial}{\partial \theta} f_{\theta}(x) dx \\ &= \frac{\partial}{\partial \theta} \int T(x) f_{\theta}(x) dx \\ &= \frac{\partial}{\partial \theta} g(\theta) \quad (\text{by unbiasedness}) \\ &= g'(\theta). \end{aligned}$$

□

Again, the regularity conditions hold for “nice” functions, e.g., smooth functions with bounded support or fast decaying tails.

Remark 1. (Reparameterization) If $\theta = \psi(\eta)$, $\psi \in C^1$ is one-one mapping in $\mathbb{R}^k$, then

$$I_X(\eta) = \frac{\partial}{\partial \eta} \psi(\eta) I_X(\psi(\eta)) \frac{\partial}{\partial \eta} \psi(\eta)^T,$$

is different (where $\frac{\partial}{\partial \eta} \psi(\eta)$ is the Jacobian matrix of full rank), but

$$\frac{\partial}{\partial \eta} g(\theta) = \frac{\partial}{\partial \theta} g(\theta) \frac{\partial}{\partial \eta} \psi(\eta),$$

so the Cramér-Rao lower bound is unchanged (we cannot cheat the lower bound by reparametrization).

Remark 2 (Asymptotic optimality of MLE). We briefly discussed in Lecture 11 and MLE has asymptotic variance $(nI_X(\theta))^{-1}$, which matches the lower bound asymptotically. This is a key advantage of MLE.

3 Two interpretations of Fisher information

3.1 Geometric view of fisher information

$\log f_{\theta}(x)$ is also known as log-likelihood (or written as loglik) function. A natural loss that we want to minimize is the following negative loglik.

$$\ell(\theta; X) = -\log f_{\theta}(X).$$

If we have i.i.d. data $X = (X_1, \dots, X_n)$, then

$$\ell(\theta; X) = -\sum_{i=1}^n \log f_{\theta}(X_i).$$



Figure 1: Small $I(\theta)$

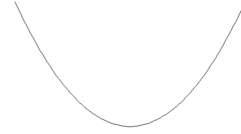


Figure 2: Large $I(\theta)$

Note that Fisher information is related to the expected curvature (in one dimension, the rate of change of tangent vector) of $\ell(\theta; X)$, since

$$I(\theta) = \mathbb{E} \left[\frac{\partial^2}{\partial \theta \partial \theta^\top} \ell(\theta; X) \right].$$

Recall that in MLE, we estimate θ by maximizing $\ell(\theta; X)$. We provide some geometric heuristics. An obvious property of MLE is that the maximum must correspond to a stationary point of $\ell(\theta; X)$, which means

$$\mathbb{E} \left[\frac{\partial}{\partial \theta} \ell(\hat{\theta}; X) \right] = 0,$$

Intuitively, consider the one-dimensional case for simplicity, a large $I(\theta)$ means larger local convexity, so noise perturbation is easier to be controlled. Figure 1 and 2 provide examples of small $I(\theta)$ and large $I(\theta)$. The following is an intuitive understanding of Fisher information.

large $I(\theta) \Rightarrow$ easy to find $\theta \Rightarrow$ less uncertainty / smaller lower bound.

Example 1. Let $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known.

$$\frac{\partial}{\partial \mu} \log f_\mu(X) = \sum_{i=1}^n (X_i - \mu) / \sigma^2 \implies I(\mu) = \frac{n}{\sigma^2}.$$

We know $\bar{X}$ is a UMVUE. It is easy to check

$$\text{Var}(\bar{X}) = [I(\mu)]^{-1},$$

attaining the Cramér-Rao lower bound.

Remark 3. $\sigma/\sqrt{n}$ is the “effective noise level”. A large sample size n reduces uncertainty.

3.2 Information theoretic interpretation

Let us recall the information notations.

- Mutual Information: Defined generally for random variables, unchanged under one-one mapping.
- Fisher Information: Defined for specific parametric models, changed under reparameterization.

Both information quantities measure the amount of information about θ given observation X . Here we give a concrete connection: Under certain “regularity conditions”,

$$D(P_{\theta_0} || P_{\theta_0 + \xi}) = \frac{1}{2} \xi^\top I(\theta_0) \xi + o(\|\xi\|^2); \quad (3)$$

this is because by Taylor expansion,

$$\log f_{\theta_0+\xi}(X) = \log f_{\theta_0}(X) + \xi^\top \nabla_\theta \log f_\theta(X) + \frac{1}{2} \xi^\top \nabla_\theta^2 (\log f_\theta(X)) \xi + o(\|\xi\|^2),$$

and so by the definition of KL divergence and the fact $\mathbb{E}[\nabla_\theta \log f_\theta(X)] = 0$, we obtain Eqn. 3.

Thus, Fisher information characterizes how hard to distinguish two distributions in a parametric family under the KL divergence. To draw an analogy, Fisher information $I(\theta_0)$ in Eqn. 3 plays a similar role as local convexity plays in optimization.

4 Cramér-Rao lower bound may not be attained

Let $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ known, and we are interested in estimating $\nu = g(\mu) = \mu^2$. Check (exercise)

- (1) $h(\bar{X}) = \bar{X}^2 - \frac{\sigma^2}{n}$ is UMVUE.
- (2) $\text{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}$.
- (3) Cramér-Rao lower bound for ν is $\frac{4\mu^2\sigma^2}{n}$.
- (4) If, however, we are interested in estimating μ , then as in Example 1, both $\text{Var}(\bar{X})$ and the Cramér-Rao lower bound are σ^2/n .

Note that $h(\bar{X})$ matches the leading order term but not the exact lower bound. Why?

- We know that $\bar{X}$ is the MLE for estimating μ with optimal asymptotic variance $V = \sigma^2/n$. By the Delta method, the asymptotic variance of $g(\bar{X})$ is $[g'(\mu)]^2 V$ —this form is exactly the Fisher information under reparametrization, so the asymptotic variance matches the lower bound. However, the asymptotic variance does not need to equal the actual variance of $h(\bar{X})$; nevertheless, it matches the leading-order term in this case.
- The reason for the gap between the variance and the lower bound is that we have a strict inequality in Cauchy-Schwarz Inequality in Eqn. 4. Recall that the equality condition for Cauchy-Schwarz Inequality is $T(X) = c \frac{\partial}{\partial \theta} \log f_\theta(X)$ a.s. for some constant c . This does not hold for $T(X) = h(\bar{X})$.

5 Examples of Cramér-Rao Lower Bound

5.1 Exponential Family

Let $\Theta \subset \mathbb{R}^k$ be an open set. Let $\{f_\theta : \theta \in \Theta\}$ be the exponential family with p.d.f.

$$f_\theta(x) = \exp\{\eta(\theta)^\top T(x) - \xi(\theta)\} c(x).$$

Suppose that X is drawn from a distribution P_θ with corresponding p.d.f. f_θ in this exponential family.

Proposition 2. (i) Regularity conditions in Eqn. 1 are satisfied for any measurable h with

$$\mathbb{E}|h(X)| < \infty.$$

(ii) Consider the natural parameter η . Then

$$\text{Var}(T(X)) = I(\eta).$$

(iii) Suppose $\nu = \mathbb{E}[T(X)]$ is parameters of interest, and $\bar{I}(\nu)$ denoted the Fisher Information. Then

$$\text{Var}(T) = [I(\nu)]^{-1}.$$

Proof. Under natural parameters η , we write the exponential family as

$$f_\eta(X) = \exp\{\eta^T T(X) - \xi(\eta)\} c(X).$$

(i) + (ii) proved in Homework 4. In addition, we proved that $f_\eta(x)$ is C^∞ . Also, from Homework 4,

$$\nu = \frac{\partial}{\partial \eta} \xi(\eta).$$

By reparameterization,

$$I(\eta) = \frac{\partial \nu}{\partial \eta} \bar{I}(\nu) \frac{\partial \nu}{\partial \eta}^T.$$

Note that

$$\frac{\partial \nu}{\partial \eta} = \frac{\partial^2}{\partial \eta \partial \eta^T} \xi(\eta) = \text{Var}(T) = I(\eta).$$

So,

$$\bar{I}(\nu) = [I(\eta)]^{-1} = [\text{Var}(T)]^{-1}.$$

□

5.2 Linear model

Basics. Data $(X_1, Z_1), \dots, (X_n, Z_n)$, where $X_i \in \mathbb{R}$, $Z_i \in \mathbb{R}^p$. Often independent across $i = 1, 2, \dots, n$, and they follow

$$X_i = \beta^T Z_i + \epsilon_i, \tag{4}$$

where

- X_i is the response, observation, or outcome.
- β is the coefficients.
- Z_i is the explanatory variables (or covariates, features, attributes, regressors, predictors).
- ϵ_i is the random error (or noise).

Here Z_i can be random or nonrandom. For simplicity, we will treat Z_i as fixed (If Z_i is random and independent with other random variables, we condition on Z_i and conclusions remain the same.) We can also write the matrix form.

$$\underset{n \times 1}{X} = \underset{n \times p}{Z} \cdot \underset{p \times 1}{\beta} + \underset{n \times 1}{\epsilon}.$$

Note: Eqn. 4 subsumes the case with an intercept term: $[\beta, \beta_0]^T [Z_i, 1]$. In the matrix form, we view intercept as a column of Z .

Least square estimator (LSE). LSE $\widehat{\beta} \in B$ satisfies

$$\|X - Z\widehat{\beta}\|^2 = \min_{b \in B} \|X - Zb\|^2. \quad (5)$$

For any $\ell \in \mathbb{R}^p$, $\ell^\top \widehat{\beta}$ is called an LSE of $\ell^\top \beta$. As we already know, LSE admits an explicit formula.

- If Z has full column rank, then

$$\widehat{\beta} = (Z^\top Z)^{-1} Z^\top X.$$

- If Z does not have full column rank, then there are infinitely many solutions to Eqn. 5, and the minimum ℓ_2 -norm solution is unique and given by

$$\widehat{\beta} = (Z^\top Z)^+ Z^\top X,$$

where $(\cdot)^+$ denotes the (Moore–Penrose) pseudo inverse.

Optimality of LSE. We make the following parametric assumption: $\epsilon \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_n)$, where $\sigma^2 > 0$ is unknown.

Theorem 2. Under the above assumption, if Z has full column rank, then

- (i) $\ell^\top \widehat{\beta}$ is UMVUE of $\ell^\top \beta$.
- (ii) $\widehat{\sigma}^2 = \frac{1}{n-p} \|X - Z\widehat{\beta}\|^2$ is UMVUE of σ^2 .

Proof.

$$\|X - Z\beta\|^2 = \|X - Z\widehat{\beta}\|^2 - 2\beta^\top Z^\top X + \|Z\beta\|^2 + \|Z\widehat{\beta}\|^2.$$

The joint p.d.f. of X is

$$p_\beta(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ \frac{\beta^\top Z^\top x}{\sigma^2} - \frac{\|x - Z\widehat{\beta}\|^2 + \|Z\widehat{\beta}\|^2}{2\sigma^2} - \frac{\|Z\beta\|^2}{2\sigma^2} \right\}.$$

We can use the property about exponential family (Proposition 1 from Lecture 12) to check that $(Z\widehat{\beta}, \|X - Z\widehat{\beta}\|^2)$ is complete and sufficient for $\theta = (\beta, \sigma^2)$. Note that $\ell^\top \widehat{\beta}$, $\widehat{\sigma}^2$ are functions of $Z\widehat{\beta}$, $\|X - Z\widehat{\beta}\|^2$ respectively. Further, it is easy to verify unbiasedness ([exercise](#))

$$\mathbb{E}[\ell^\top \widehat{\beta}] = \ell^\top \beta, \quad \mathbb{E}[\widehat{\sigma}^2] = \sigma^2.$$

Therefore, Lehmann-Scheffé theorem implies UMVUE. □

In fact, we can characterize the distribution of the LSE and residual explicitly, which is useful for deriving statistical tests for LSE.

Theorem 3. Under the same assumption as in Theorem 2, $\ell^\top \widehat{\beta}$ is independent of $\widehat{\sigma}^2$. Specifically,

$$\ell^\top \widehat{\beta} \sim \mathcal{N}(\ell^\top \beta, \sigma^2 \ell^\top (Z^\top Z)^+ \ell), \quad \widehat{\sigma}^2 \frac{n-p}{\sigma^2} \sim \chi_{n-p}^2.$$

Proof sketch. Denote $(P_Z, P_Z^\perp)$ to be the projection matrices to the range and null space of Z respectively. Then,

$$Z\hat{\beta} = P_Z(X), \quad X - Z\hat{\beta} = P_Z^\perp(X).$$

Orthogonal transformation of an isotropic¹ Gaussian random vector is still an isotropic Gaussian. Thus,

$$P_Z(X) \text{ is independent of } P_Z^\perp(X), \quad \frac{1}{\sigma^2} \|P_Z^\perp(X)\|^2 \sim \chi_{n-r}^2.$$

□

There is a geometric intuition behind this proof: the best linear fit is the orthogonal projection of observation vector X onto the subspace spanned by columns of Z .

Fisher Information. We consider the parametric assumption $\epsilon \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_n)$ and the unknown parameters are $\theta = (\beta, \sigma^2)$.

$$I(\theta) = \frac{1}{\sigma^2} \begin{pmatrix} Z^T Z & 0 \\ 0 & \frac{n}{2\sigma^2} \end{pmatrix}.$$

If Z is not full rank, $I(\theta)$ may not be invertible. Indeed, in the null space $\text{Ker}(Z)$, two distributions $P_\theta, P_{\theta'}$ are not distinguishable.

It is clear that LSE is the MLE for estimating β under our assumption, and it attains the Cramér-Rao lower bound; **exercise**: if we are interested in estimating σ^2 , does the UMVUE attain the lower bound?

6 Further reading

The information perspective of Fisher information can be found in many sources, e.g., Section 2.6.2 of [1]. Fisher information is a central quantity in classical mathematical statistics, and in STAT 710 you will revisit this concept, especially when studying MLE.

Linear square estimator. The discovery of least squares has an interesting history. A nice reference is Wikipedia page.

Least squares estimators can be derived using both MLE or method-of-moments, both of which will be covered in STAT 709 and STAT 710. The practical side of least squares is covered in STAT 849 and 850. Some key related notions:

- collinearity,
- basis expansion,
- least absolute deviations.

For solving nonlinear least squares, a classical approach is *iterative reweighted least squares*.

In the modern context where dimension grows proportionally with sample size, least squares estimator is inconsistent; in over-parametrized settings, least squares estimator is not uniquely defined, but min-norm solution is usually expected and enjoys good statistical properties.

¹This means that that covariance is an identity matrix.

Optimality of LSE under weaker conditions We make the following assumptions: $\mathbb{E}[\epsilon] = 0$, $\text{Var}(\epsilon) = \sigma^2 I_n$ where $\sigma^2 > 0$ is unknown.

Theorem 4 (Gauss-Markov Theorem). *Under the above assumptions, if Z has full column rank, then $\ell^\top \hat{\beta}$ is the best linear (linear in X) unbiased estimator (BLUE) of $\ell^\top \beta$.*

We make weaker assumptions, and accordingly the optimality is stated in a smaller space of candidate estimators.

References

- [1] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.