Stochastic Processes

Week 07 (Version 2.0)

Estimation Theory - Part II

Hamid R. Rabiee

Fall 2021

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Introduction to Optimal Frequentist Estimator

- In the Frequentist's point of view, an optimal estimator is both unbiased and minimum variance.
- How can we obtain an estimator $\hat{\theta}$ that is unbiased?
  - Given any biased estimator $\hat{\theta}_b$ with bias b, then we can remove the bias to obtain an unbiased estimator $\hat{\theta}$ from $\hat{\theta}_b$, i.e. $\hat{\theta} = \hat{\theta}_b - b$.
- How can we obtain a minimum variance estimator $\hat{\theta}_{mv}$ from an unbiased estimator?
  - We need to obtain a lower bound for an unbiased estimator and make sure $\hat{\theta}$mv achieve that bound.

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Score and Fisher Information

- The **score** $s(\theta)$ is defined as the gradient of the log-likelihood function with respect to the parameter vector.

$$s(\theta) = \frac{\partial \log L(\theta|x)}{\partial \theta} = \frac{\partial \log f(x|\theta)}{\partial \theta}$$

- When evaluated at a particular value of the parameter vector, the score indicates the sensitivity of the log-likelihood function to infinitesimal changes to the parameter values.

Score and Fisher Information

- The mean of score $s(\theta)$:
- Although $s(\theta)$ is a function of θ, it also depends on the observations X, at which the likelihood function is evaluated, and the expected value of the score, evaluated at the true parameter value θ, is zero.

$$\mathrm{E}(s \mid \theta) = \int_{\mathcal{X}} f(x|\,\theta) \frac{\partial}{\partial \theta} \log \mathcal{L}(\theta\,|x)\, dx$$

$$= \int_{\mathcal{X}} f(x|\,\theta) \frac{1}{f(x|\,\theta)} \frac{\partial f(x\,|\,\theta)}{\partial \theta}\, dx = \int_{\mathcal{X}} \frac{\partial f(x\,|\,\theta)}{\partial \theta}\, dx$$

Score and Fisher Information

- We can interchange the derivative and integral by using Leibniz integral rule:

$$\frac{\partial}{\partial\theta}\int_{\mathcal{X}} f(x\,|\theta)\,dx = \frac{\partial}{\partial\theta}1 = 0$$

- If we repeatedly sample from some distribution, and repeatedly calculate its score, then the mean value of the scores would tend to zero asymptotically.

Score and Fisher Information

- The **Fisher Information** is defined as the variance of score. It is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution that models X.

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2 \middle|\, \theta\right] = \int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2 f(x|\theta)\, dx$$

- The Fisher information is not a function of a particular observation, as the random variable X has been averaged out.

Score and Fisher Information

- If $\log f(x|\theta)$ is twice differentiable with respect to θ, and under certain regularity conditions, then the Fisher information may also be written as:

$$\mathcal{I}(\theta) = -\mathrm{E}\left[\left.\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right|\theta\right]$$

- The regularity conditions are as follows:
  - The partial derivative of $f(X \mid \theta)$ with respect to θ exists.
  - The integral of $f(X \mid \theta)$ can be differentiated under the integral sign with respect to θ.
  - The support of $f(X \mid \theta)$ does not depend on θ.

Fisher Information

For i.i.d. samples $x_1, \ldots, x_n$:

Since $f(X|\theta) = \prod f(x_i|\theta)$, the Fisher Information is:

$$E_\theta \left[\left(\frac{\partial}{\partial\theta} \log\big(f(X|\theta)\big)\right)^2 \right] = n\, E_\theta \left[\frac{\partial}{\partial\theta} \log\big(f(x_i|\theta)\big) \right]^2$$

Proof:

$$E_\theta \left[\frac{\partial}{\partial\theta} \log\big(f(X|\theta)\big) \right]^2 = E_\theta \left[\frac{\partial}{\partial\theta} \log\left(\prod f(x_i|\theta) \right) \right]^2$$

Cramer-Rao Bound

$$= E_\theta\left[\sum \frac{\partial}{\partial\theta}\log\big(f(x_i|\theta)\big)\right]^2 = nE_\theta\left[\frac{\partial}{\partial\theta}\log\big(f(x|\theta)\big)\right]^2$$

If $f(X|\theta)$ satisfies $\frac{\partial}{\partial\theta}E_\theta\left[\frac{\partial}{\partial\theta}\log\big(f(X|\theta)\big)\right]$

$$= \int \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\log\big(f(X|\theta)\big)\right]f(X|\theta)dx$$

Then:

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\log\big(f(X|\theta)\big)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log\big(f(X|\theta)\big)\right]$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Cramer-Rao Bound

- The **Cramer–Rao bound (CRB)** expresses a lower bound on the variance of unbiased estimators of a deterministic (fixed, though unknown) parameter θ, stating that the variance of any such estimator is at least as high as the inverse of the Fisher information.
- An unbiased estimator which achieves this lower bound is said to be efficient.
- Suppose θ is an unknown deterministic parameter which is to be estimated from n independent observations of x, each from a distribution according to some probability density function $f(x|\theta)$.

Cramer-Rao Bound

- The variance of any *unbiased* estimator $\hat{\theta}$ of θ is then bounded by the reciprocal of the Fisher information $I(\theta)$:

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

- The efficiency of an unbiased estimator $\hat{\theta}$ measures how close this estimator's variance comes to this lower bound; estimator efficiency is defined as:

$$e(\hat{\theta}) = \frac{I(\theta)^{-1}}{\mathrm{var}(\hat{\theta})}$$

- The Cramer–Rao lower bound gives: $e(\hat{\theta}) \leq 1$

Cramer-Rao Bound

Let $x_1, \ldots, x_n$ have joint pdf $f(X|\theta)$:

Let $\hat{\theta} = w(X) = w(x_1, \ldots, x_n)$ be any estimator where $E_\theta[w(X)]$ is differentiable by θ, and for any function h with $E_\theta[x(X)] < \infty$, Suppose:

$$\frac{\partial}{\partial\theta} \int \ldots \int h(x)\, f(x|\theta) dx_1 \ldots dx_n = \int \ldots \int h(x) \frac{\partial}{\partial\theta} f(x|\theta) dx_1 \ldots dx_n$$

Then: $var_\theta\,(\hat{\theta}) = var_\theta[w(X)] \ \geq \ \dfrac{(\frac{\partial}{\partial\theta} E_\theta[w(X)])^2}{E_\theta\left[(\frac{\partial}{\partial\theta} \log\left(f(X|\theta)\right))^{\,2}\right]} \rightarrow Fisher\ Information$

If $w(X)$ is unbiased then: $E_\theta[w(X)] = \theta$ and $\frac{\partial}{\partial\theta} E_\theta[w(X)] = 1$, and:

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

Cramer-Rao Bound

Example: $x_1, \dots, x_n \; iid \; p(\lambda)$

$$E_\lambda w(x) = \lambda \rightarrow \frac{\partial}{\partial\lambda} E[w(x)] = 1$$

$$logf(x|\lambda) = log\frac{e^{-\lambda}\lambda^x}{x!} = -\lambda + xlog\lambda - log\lambda!$$

$$\frac{\partial}{\partial\lambda} logf = -1 + \frac{x}{\lambda} \rightarrow \frac{\partial^2}{\partial\lambda^2} logf = -\frac{x}{\lambda^2}$$

$$\text{Fisher Information} = -nE_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log\big(f(x|\theta)\big)\right] = -n\left(-\frac{E[X]}{\lambda^2}\right) = \frac{n}{\lambda}$$

$$Var[w] \geq \frac{\lambda}{n} \quad Var_\lambda \bar{X} = \frac{\lambda}{n}$$

$\rightarrow \quad (\bar{X} \; is \; unbiased \; and \; achieves \; the \; CRB \; (i.e. it \; is \; is \; an \; UMVUE)$

Cramer-Rao Bound

Example: $x_1, \ldots, x_n$ iid $\quad f(X|\theta)$ $uniform;\quad 0 < X < \theta$

$$\frac{\partial}{\partial\theta} logf = -\frac{1}{\theta} \rightarrow E_\theta\left[\frac{\partial}{\partial\theta}\log(f)\right]^2 = \frac{1}{\theta^2} \quad Fisher\ Information$$

$$\text{Fisher Information} = -nE_\lambda\left[\frac{\partial^2}{\partial\lambda^2}\log\big(f(x|\theta)\big)\right] = -n\left(-\frac{E[X]}{\lambda^2}\right) = \frac{n}{\lambda}$$

$$CR\ boumd\!:\quad if\ w\ is\ unbiased\ for\ \theta,\qquad Var_\theta w \geq \frac{\theta^2}{n}$$

How to find estimator:

$Y = \max(X_i) \quad \leftarrow Sufficient\ Statistic$

$$f_Y(y|\theta) = \frac{ny^{n-1}}{\theta^n} \quad 0 < y < \theta$$

Cramer-Rao Bound

Problem with CR approach:

- gives you a lower bound
- can it be attained?

Yes, if $f(X|\theta)$ is a regular one-parameter exponential family and an unbiased estimator exists.

Example: $x_1, \ldots, x_n \quad iid \quad N(\mu, \ \delta^2)$ interested in δ^2

$$logf(X|\mu, \delta^2) = -\frac{1}{2} log 2\pi\delta^2 \ - \frac{1}{2}\frac{(x-\mu)^2}{\delta^2}$$

$$\frac{\partial}{\partial \delta^2} logf \ = -\frac{1}{2\delta^2} + \frac{1}{2}\frac{(x-\mu)^2}{\delta^4}$$

$$\frac{\partial^2}{\partial (\delta^2)^2} logf = \frac{1}{2\delta^4} - \frac{(x-\mu)^2}{\delta^6} \qquad -E\left[\frac{\partial^2}{\partial (\delta^2)^2} logf\right] = \frac{1}{2\delta^4}$$

Cramer-Rao Bound

$\Rightarrow$ any unbiased estimator w for δ^2 satisfies:

$$Var(w) \geq \frac{2\delta^4}{n}$$

$$Var(\delta^2) = \frac{2\delta^4}{n-1}$$

- When is bound attainable?

$$\left(cov(w,y)\right)^2 \leq (var\ w)(var\ y)$$

$$y = \frac{\partial}{\partial\theta}\log f(X|\theta)$$

Cramer-Rao Bound

When do we have equality in Cauchy-Schwartz?

$a(w - Ew) = y - Ey$

$y = \frac{\partial}{\partial\theta}\log f(X|\theta) \quad E[y] = 0$

Bound is attained when: $a(\theta)[w - \theta] = \frac{\partial}{\partial\theta}\log f(X|\theta)$.

Corollary: $x_1, \dots, x_n \quad iid \quad f(X|\theta)$ satisfies CRB,

Let likelihood function $L(\theta|X) = \pi f(X_i|\theta)$

If w is any unbiased estimator of θ, then it attains the CRB lower bound,

iff:

$\frac{\partial}{\partial\theta}\log L(X|\theta) = a(\theta)[w(X) - \theta]$ for some function $a(\theta)$.

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Rao-Blackwell Theorem

- The Rao-Blackwell theorem uses sufficiency to characterizes the transformation of an arbitrarily estimator into an estimator that is optimal by the mean-squared-error (MSE) criterion.
- Recall: x and y are random variables:

$$E[X] = E\big[E[X|Y]\big]$$

$$var(X) = var(E[X|Y]) + E[var(X|Y)]$$

Rao-Blackwell Theorem:

Let w be unbiased for θ, and let T be a sufficient statistic for θ:

Define $\phi(T) = E[w|T]$, then:

$E[\phi(T)] = \theta$

and $\quad var\big(\phi(T)\big) \leq var_{\theta}(T).$

Rao-Blackwell Theorem

Proof:

(1) $\phi(T) = \mathrm{E}_{\theta}(\mathrm{w}|\mathrm{T})$ is an estimator because T is sufficient

$\Longrightarrow$ conditioned dist. of $\underline{\mathrm{X}}$ given T does not depend on θ

and w is a function of $\underline{\mathrm{X}}$ only:

$$E_{\theta}\big(\phi(T)\big) = E_{\theta}\big(E(w|T)\big) = E_{\theta}(w) = \theta$$

(2) $Var_{\theta}(w) = Var_{\theta}[E(w|T)] + E_{\theta}[Var(w|T)]$

$$= Var_{\theta}\big(\phi(T)\big) + E_{\theta}\underbrace{\big(Var(w|T)\big)}_{positive} \geq Var_{\theta}\big(\phi(T)\big)$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

UMVUE

- The minimum-variance unbiased estimator (MVUE) or uniformly minimum-variance unbiased estimator (UMVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter.
- How to find an UMVUE?

2 strategies for finding UMVUE′s:

(1) Let T be a complete suff. Stat. for θ, find a function $\phi(T)$ of T, such that $E_\theta[\phi(T)] = \theta$.

(2) Let w be any unbiased estimator and T be a suff. Stat. for θ, compute $\phi(T) = \mathrm{E}(\mathrm{w}|\mathrm{T})$.

UMVUE

Example:

$x_1, \dots, x_n \quad iid \quad N(\mu, 1)$,

Median $(x_1, \dots, x_n)$ is unbiased.

However, it can't be UMVUE since it is not a

sufficient statistics for θ,

(i.e. sufficient statistics is $\bar{X}$).

Is UMVUE unique?

UMVUE

Claim: UMVUE is unique.

Proof:

Suppose w$^{'}$ is also UMVUE:

Then, $w^* = \frac{w+w^{'}}{2}$ is also unbiased; $E_\theta\,[w^*] = \theta$,

$Var_\theta[w^*] = \frac{1}{4}Var_\theta[w] + \frac{1}{4}Var_\theta\left[w^{'}\right] + \frac{1}{2}conv\left(w, w^{'}\right)$,

use $Cauchy - Shwartz$:

$\leq\ \frac{1}{4}Var_\theta[w] + \frac{1}{4}Var_\theta\left[w^{'}\right] + \frac{1}{2}\left[Var_\theta[w]Var_\theta\left[w^{'}\right]\right]^{\frac{1}{2}}$

$= Var_\theta[w]$ (because $\left[w^{'}\right] = Var_\theta[w]$)

UMVUE

When the equality in Cauchy-Schwartz holds?

$(w' - Ew') = a(\theta)(w - E(w))$

$cov_\theta(w, w') = E(w - Ew)(w' - Ew')$

$= a(\theta)E_\theta(w - Ew)^2 = a(\theta)var(w)$

In the above $a(\theta) = 1 \Rightarrow w = w'$.

UMVUE

Example: $x_1, \dots, x_n \quad iid \quad Bern(\theta)$

We know $\bar{X}$ is the $UMVUE$ (CRB attained)

Showed $T = \sum X_i$ is a complete suff. Stat. for θ.

$E(T) = n\theta \implies \phi(T) = \frac{T}{n}$

Example: $x_1, \dots, x_n \quad iid \quad N(\mu,\ \delta^2)$

Showed $T = (T_1,\ T_2) = (\sum X_i\ ,\ \sum X_i^2)$ is a complete suff. stat. for $N(\mu,\ \delta^2)$

Consider $(\bar{X},\ S^2) = \left(\frac{T_1}{n},\ \frac{1}{n-1}\left(T_2 - \frac{T_1^2}{n}\right)\right)$

UMVUE

Example: $x_1, \dots, x_n \quad iid \;\; U(0,\ \theta)$

We showed that $\frac{n+1}{n} y$ is an unbiased estimator for θ, $y = \max(X_i)$

Can show y is a complete suff. stat. for θ.

$\Longrightarrow \ \frac{n+1}{n} y$ is the $UMVUE.$

Example: $x_1, \dots, x_n \quad iid \;\; p(\lambda)$

Interested in estimating $\theta = e^{-\lambda} = P_\lambda(X = 0)$

$\sum x_i \sim p(n, \lambda)$ is a complete sufficient statistic and:

$\frac{\sum xi}{n}\ is\ the\ UMUVE\ for \lambda$

UMVUE

$Guess\ e^{-x}\ \leftarrow not\ unbiased.$

$$W(X) = \begin{cases} 1 & X = 0 \\ 0 & X > 0 \end{cases}$$

$E_\lambda(w) = e^{-\lambda}\ \rightarrow unbiased$

$ComputeE_\lambda(w|T)$:

$$\phi(t) = E(w|T = t) = P_\lambda\left(X_1 = 0 \middle| \sum_{i=2}^{n} X_i = t\right)$$

$$= \frac{P_\lambda(X_1 = 0,\ \sum_{i=2}^{n} X_i = t)}{P_\lambda\left(\sum_{i=1}^{n} X_i = t\right)} = \frac{P_\lambda(X_1 = 0)P_\lambda(\sum_{i=2}^{n} X_i = t)}{P_\lambda\left(\sum_{i=1}^{n} X_i = t\right)}$$

$X_i \sim P(\lambda)$ $\qquad \sum_{i=2}^{n} X_i \sim P\big((n-1)\lambda\big)$ $\qquad \sum_{i=1}^{n} X_i \sim P(n\lambda)$

UMVUE

$$\Longrightarrow \phi(t) = \frac{[e^{-\lambda}]\left[e^{-(n-1)\lambda} \times \frac{[(n-1)\lambda]^t}{t!}\right]}{e^{-n\lambda} \times \frac{[n\lambda]^t}{t!}}$$

$$\therefore \phi(\mathrm{t}) = \left(\frac{\mathrm{n}-1}{\mathrm{n}}\right)^{\mathrm{t}} = \left(1 - \frac{1}{\mathrm{n}}\right)^{\mathrm{t}} \text{ is UMUVE of } \mathrm{e}^{-\lambda}$$

We can write: $\phi(t) = \left(\frac{n-1}{n}\right)^t = \left(\left(1 - \frac{1}{n}\right)^n\right)^{\frac{1}{n}\sum x_i}$

$$as\ n \to \infty, \phi(t) \to e^{-\bar{X}}$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Bayes Estimation

Bayes estimation:

- Frequentists or classical regards the parameter θ as an unknown but fixed.

- Bayes: regards θ as random variable, with prior distribution $\pi(\theta)$.

- Observe data $x_1, \dots, x_n$
- Update the prior into a posterior distribution; $\pi(\theta|X)$.
- $\pi(\theta|X) = \frac{f(X,\theta)}{m(X)} = \frac{f(X|\theta)\pi(\theta)}{m(X)}$

  $m(x) = \int f(X|\theta)\pi(\theta)d\theta = marginal\ dist.of\ X$

Bayes estimation

Example: $x_1, \dots, x_n \quad iid \quad Bernoulli(\theta), \qquad \theta \sim \beta eta(\alpha,\ \beta)$

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$f(x)\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$m(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{\sum x_i+\alpha-1}(1-\theta)^{n-\sum x_i+\beta_{-1}} d\theta$$

$$\beta\,\mathrm{eta}\left(\sum x_{i+}\alpha, n - \sum x_i + \beta\right)$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\sum x_i + \alpha)\Gamma(n - \sum x_i + \beta)}{\Gamma(n+\alpha+\beta)}$$

$$\Gamma(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \sum^{\sum x_i+\alpha-1} (1-\theta)^{n-\sum x_i+\beta-1} \times \frac{1}{m(\alpha)}$$

$\pi(\theta|X) \sim \beta eta(\sum X_i + \alpha\,, n - \sum X_i + \beta)$

Bayes Estimation

Finding the posterior:

(a) Calculate $\pi(\theta) f(X|\theta)$

(b) Factor into piece depending on θ and piece not depending on θ.

(c) Drop piece not depending on θ, multiply and divide by constants.

(d) $\pi(\theta|X)$ is $k(X)$ times what is left.

choose $k(X)$ s.t. $\int \pi(\theta|X)\, d\theta = 1$

Bayes Estimation

Example: $x_1, \ldots, x_n$ i.i.d. $N(\mu,\ \delta^2)$, δ^2 known

$$f(\,x \mid \mu\,) = (2\Pi\delta^2)^{-\frac{n}{2}} e^{-\frac{1}{2\delta^2}\Sigma(x_i-\mu)^2}$$

$$\Pi(\mu) = N\left(\mu_0, \delta_0^2\right)$$

$$\pi(\mu) f(\,x \mid \mu\,) = \left(\frac{1}{\sqrt{2\pi\delta^2}}\right)^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2s^2}\sum(x_i-\mu)^2} e^{-\frac{1}{2\delta_0^2}(\mu-\mu_0)^2}$$

$$\alpha \exp\left[-\frac{1}{2\delta_0^2}(\mu-\mu_0)^2 - \frac{1}{2\delta^2}\sum(x_i-\bar{x})^2 - \frac{1}{2\delta^2} n(\bar{x}-\mu)^2\right]$$

$$= \exp\left[-\frac{1}{2}\left(\frac{(\mu-\mu_0)^2}{\delta_0^2} + \frac{n(\bar{x}-\mu)^2}{\delta^2}\right)\right]$$

Bayes Estimation

$$= \exp\left[-\frac{1}{2}\left(\frac{(\mu-\mu_0)^2}{\delta_0^2}+\frac{n(\bar{x}-\mu)^2}{\delta^2}\right)\right]$$

$$= \exp\left[-\frac{1}{2}\left(\left(\frac{1}{\delta_0^2}+\frac{n}{\delta^2}\right)\mu^2-2\mu\left(\frac{\mu_0}{\delta_0^2}+\frac{n\bar{x}}{\delta^2}\right)+\frac{\mu\delta^2}{\delta_0^2}+\frac{n\bar{x}^2}{\delta^2}\right)\right]$$

$$= \frac{-1}{2}a\mu^2-2b\mu=\frac{-1}{2}a\left(\mu-\frac{b}{a}\right)^2$$

$$a=\frac{1}{\delta_0^2}+\frac{n}{\delta^2} \qquad \pi(\mu)f(x\mid\mu)\propto\exp\left[-\frac{1}{2}a\left(\mu-\frac{b}{a}\right)^2\right]$$

$$b=\frac{\mu_c}{\delta_0^2}+\frac{n\bar{x}}{\delta^2} \qquad =N\left(\frac{b}{a},\frac{1}{a}\right)\sim\pi(\mu\mid\underline{x})$$

Bayes Estimation

Bayes estimator:

(1) Maximum A Posteriori (MAP) Estimator:

In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.

Given $x_1, \ldots, x_n$ i.i.d. with $f(x_i|\theta)$; and $\pi(\theta)$

$f(X|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$:

$$\hat{\theta}_{\text{MAP}}(X) = \arg\max_{\theta} \pi(\theta|X) = \arg\max_{\theta} [f(X|\theta)\pi(\theta)]$$

Bayes Estimation

How to compute Maximum A Posteriori (MAP):

- Analytically: when the mode(s) of the posterior distribution can be given in closed form. This is the case when conjugate priors are used.
- Numerical optimization: such as the conjugate gradient method or Newton's method. This usually requires first or second derivatives, which have to be evaluated analytically or numerically.
- Modification of an expectation-maximization (EM) algorithm. This does not require derivatives of the posterior density.
- Monte Carlo method using simulated annealing.

Bayes Estimation

(2) Bayes Minimum Loss (Risk) Estimator:

- Define a loss function $L(\theta, \hat{\theta})$

$L(\theta, \hat{\theta}) = loss\ of\ estimation\ \theta\ by\ \hat{\theta}$

- Minimize the expected loss:

$$\min \int_{\Theta} L(\theta, \hat{\theta}) \pi(\theta|X)\, d\theta$$

- Then $\hat{\theta}$ is the Bayes minimum loss estimator.

Bayes Estimation

(1) $L(\theta - \hat{\theta}) = (\theta - \hat{\theta})^2$ squared error loss

$\Rightarrow \hat{\theta} = E(\theta|X)$

(2) $L(\theta - \hat{\theta}) = |\theta - \hat{\theta}|$ absolute error loss

$\Rightarrow \hat{\theta} = Median\ of\ \pi(\theta|X)$

Example: $x_1, \dots, x_n$ iid $N(\mu,\ \delta^2)$

Posterior is normed with mean: $\left(\frac{\mu_0}{\delta_0^2} + \frac{n\bar{x}}{\delta^2}\right) / \left(\frac{1}{\delta_0^2} + \frac{n}{\delta^2}\right)$

And variance: $1/\left(\frac{1}{\delta_0^2} + \frac{n}{\delta^2}\right)$ using squared loss criterion.

$$\hat{\mu} = E(\mu \mid x) = \alpha\bar{x} + (1-\alpha)\mu_0$$

$$\alpha = n/\delta^2/\left(\frac{n}{\delta^2} + \frac{1}{\delta_0^2}\right) = \frac{n}{n + \frac{\delta^2}{\delta_0^2}}$$

Bayes Estimation

Note:

(1) As $n \longrightarrow \infty, \alpha \rightarrow 1$

$$\Rightarrow E(\mu \mid x) \longrightarrow \bar{x}$$

(2) Vague prior information:

Let $\delta_0^2 \rightarrow \infty$

$$\mu \sim N(\mu_0, \infty) \ \Rightarrow E(\mu \mid x) \longrightarrow \bar{x}$$

(3) Good prior info:

Let $\delta_0^2 \rightarrow 0 \Rightarrow E(\mu \mid x) \rightarrow \mu_0$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Conjugate Prior

In Bayesian probability theory, if the posterior distribution $\pi(\theta \mid x)$ is in the same probability distribution family as the prior probability distribution $\pi(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $f(x|\theta)$.

Examples:

Conjugate Prior	Likelihood	Posterior
Beta	Bernoulli	Beta
Gamma	Poisson	Gamma
Normal	Normal	Normal

Bayesian Estimation with Conjugate Priors

Example: $x_1, \ldots, x_n$ i.i.d. $Bern(\theta)$

Prior: $\beta eta(\alpha, \beta)$

Posterior βeta: $(\alpha + \sum X_i, n - \sum X_i + \beta)$

Use squared error loss: $E(\theta|X) = \frac{\alpha + \sum X_i}{\alpha + \beta + n}$

$E(\theta|X) = wX + (1 - w)\frac{\alpha}{\alpha + \beta}$; where $w = \frac{n}{\alpha + \beta + n}$

Problems with Bayes Estimator

Choice of prior:

- Subjective – Conjugate Priors
- What can we do when we do not have the prior?
- Use: non-informative priors:

  Prior: $\pi(\theta) = 1, \forall\theta$
- Can we do better?
- Use Jeffreys Prior

Jeffreys Prior

Jeffreys Prior: is a non-informative (objective) prior distribution for a parameter space; its density function is proportional to the square root of the determinant of the Fisher information matrix: $\pi(\theta) \propto [detI(\theta)]^{\frac{1}{2}}$.

Example: $x_1, \dots, x_n \ iid \ Bern(\theta)$

$$\log\big(f(X|\theta)\big) = xlog\theta + (1-x)\log(1-\theta)$$

$$\frac{\partial}{\partial\theta}\log\big(f(X|\theta)\big) = \frac{x}{\theta} - \frac{1-x}{1-\theta} \rightarrow \frac{\partial^2}{\partial\theta^2}\log\big(f(X|\theta)\big) = \frac{-x}{\theta^2} + \frac{1-x}{(1-\theta)^2}$$

$$E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log\big(f(X|\theta)\big)\right] = -\frac{1}{\theta} - \frac{1}{1-\theta} = -\frac{1}{\theta(1-\theta)}$$

$$\pi(\theta) \propto (\frac{1}{\theta(1-\theta)})^{\frac{1}{2}} \ i.e. \ \beta eta(\frac{1}{2}, \frac{1}{2})$$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Consistency

Def: a sequence of estimators:

$w_n = w_n(x_1, \ldots, x_n)$ is a consistent sequence of estimators of the parameter θ if for any $\epsilon > 0$, $\theta \in \Theta$:

$$\lim_{n \to \infty} P_\theta(|w_n - \theta| < \epsilon) = 1$$

or: $$lim_{n \to \infty} P_\theta(|w_n - \theta| \geq \epsilon) = 0$$

(it means w_n converges to θ in probability)

Consistency

Theorem:

If w_n is a sequence of estimators of a parameter θ with:

(a) $\lim_{n\to\infty} Var_\theta(w_n) = 0$ and

(b) w_n unbiased estimator of θ

Then w_n is a consistent sequence of estimators of θ.

Proof:

$$Chebychev \implies P_\theta(|w_n - \theta| \geq \varepsilon) \leq \frac{E_\theta(w_n - \theta)^2}{\varepsilon^2}$$

$$E_\theta(w_n - \theta)^2 = E_\theta(w_n + Ew_n - Ew_n - \theta)^2$$

$$= Var_\theta w_n + (Bias_\theta w_n)^2$$

Consistency

Why do frequentists use MLE's?

- MLE's are consistent
- MLE's are asymptotically unbiased

Theorem:

Let $x_1, \dots, x_n \quad iid \;\; f(X|\theta)$.

Let $L(\theta|X) = \prod f(X_i|\theta)$

$\hat{\theta}$ = MLE of θ

Then we have:

$\hat{\theta}_n$ is a consistent estimator of θ.

Condition: support of pdf does not depend on parameters and rules out $U(0, \theta)$

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Efficiency

Let $I(\theta) = E_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2$.

Def:

Let w be an unbiased estimator of θ. The efficiency of w is:

$$eff(w) = \frac{[1/n\, I(\theta)]}{var(w)}$$

$\longrightarrow$ CR lower bound

Efficiency

Definition:

A sequence of estimators w is said to be asymptotically efficient if:

$$\lim_{n\to\infty} eff(w_n) \to 1$$

As $n \to \infty$, $var\ w_n$ attains CR lower bound.

- MLE's are asymptotically efficient.
- MLE's are asymptotically normal.

i.e. $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to N\left(0, \frac{1}{I(\theta)}\right)$

- MLE's are (with some fairly general conditions):

(1) Consistent, (2) asymptotically unbiased, (3) asymptotically efficient, (4) asymptotically normal.

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Estimator Comparison

Example: $x_1, \ldots, x_n \quad iid \quad N(\mu, \delta^2)$, want to estimate δ^2**:**

MLE $\widehat{\delta_1}^2 = \frac{s}{n}\ when\ s = \sum(x_i - \bar{x})^2$

Bayes (Jeffery's prior) $\pi(\delta^2) \propto \frac{1}{s^2} \quad \widehat{\delta_2^2} = \frac{s}{n-2}$

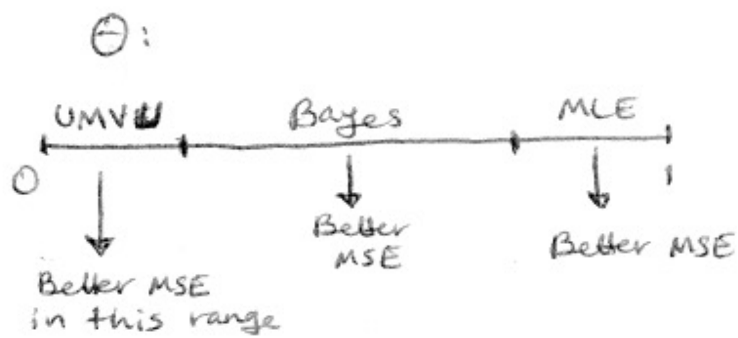
UMVUE $\widehat{\delta_3^2} = \frac{s}{n-1}$

Estimator Comparison

	$\hat{\delta}_1^2$	$\hat{\delta}_2^2$	$\hat{\delta}_3^2$
Estimator	$\frac{S}{n}$	$\frac{S}{n-2}$	$\frac{S}{n-1}$
MSE	$\delta^4\left(\frac{2n-1}{n^2}\right)$	$\delta^4\left(\frac{2n-1}{(n-2)^2}\right)$	$\delta^4\left(\frac{2}{n-1}\right)$

theta	k1	MLE	Bayes	UMVUE
0.10	2	0.0258	0.0250	0.0148
0.20	4	0.0171	0.0169	0.0125
0.30	6	0.0159	0.0151	0.0134
0.40	8	0.0154	0.0140	0.0141
0.50	10	0.0142	0.0126	0.0138
0.60	12	0.0127	0.0110	0.0128
0.70	14	0.0105	0.0090	0.0109
0.80	16	0.0077	0.0067	0.0082
0.90	18	0.0042	0.0038	0.0045
0.95	19	0.0021	0.0022	0.0023

* Mean squared error



Estimator Comparison

Example: let R= #of tosses needed to reach k heads, $\theta = p(head)$

$$P[R = r] = {}^{r-1}C_{k-1}\theta^{k}(1-\theta)^{r-k} \qquad r = k, k+1, \dots$$

R has negative binomial distribution.

(1) MLE $\widehat{\theta_1} = \frac{k}{r}$

(2) Bayes $\pi(\theta) \propto [\theta(1-\theta)]^{-\frac{1}{2}}$

$\Longrightarrow \pi(\theta|R) \propto \theta^{k-\frac{1}{2}}(1-\theta)^{r-k-\frac{1}{2}}$

$\Longrightarrow \widehat{\theta_2} = E(\theta|R) = \dfrac{k+\frac{1}{2}}{r+1}$

Estimator Comparison

(3) UMVUE: r is complete and sufficient for θ:

$$E\left[\frac{1}{r-1}\right] = \frac{\theta}{k-1}$$

$$\Longrightarrow \widehat{\theta_3} = \frac{k-1}{r-1} \quad \textit{which is the UBMUE of } \theta.$$

(4) Can't calculate MSE exactly:

Instead: simulation study:

Fix k and θ

Generate R

Calculate $\hat{\theta}_i \left(\hat{\theta}_i - \theta \right)^2$.

Outline of Week 7 Lectures

- Introduction to Optimal Frequentist Estimator
- Score and Fisher Information
- Cramer-Rao Bound (CRB)
- Rao-Blackwell Theorem
- UMVUE
- Bayesian Estimation
- Conjugate Prior
- Consistency
- Efficiency
- Estimator Comparison
- Summary

Summary

(1) **Likelihood**:

Estimate θ by the value $\hat{\theta}$ which maximizes the likelihood

(2) **Bayes**:

Let $\pi(\theta)$ be a prior distribution for θ leading to a posterior

$\pi(\theta|\underline{X})$

Let $L(\theta, \hat{\theta})$ be a loss function. Choose $\hat{\theta}$ to minimize:

$$\int_{\Theta} L(\theta, \hat{\theta})\, \pi(\theta|X) d\theta$$

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2 \implies \hat{\theta} = E[\theta|X]$$

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}| \implies \hat{\theta} = \text{median of } \pi(\theta|X)$$

Summary

(3) Frequentist:

(a) If possible, find the UMVUE of θ.

(b) If (a) impossible, use the MLE $\hat{\theta}$ which is asymptotically unbiased and whose efficiency $\longrightarrow 1$ as $n \longrightarrow \infty$.

---

(1), (2) and (3) may not exist!

Example:

MLE: $X \sim N(\mu, \delta^2)$

$$X = \mu, \quad \delta^2 \longrightarrow 0$$

UMVUE: $\text{Bern}(p)$. Then $\theta = \frac{p}{1-p} \Longrightarrow$ UMVUE ofθ does not exist.

Summary

- MLE and Bayes may not be unique, but UMVUE is unique.
- MLE has invariance property, UMVUE and Bayes do not.
- Bayes: incorporate prior information, but MLE and UMVUE don't.

Next Week:

Hypothesis Testing

Have a good day!