Chapter 7: Estimation

Sections

- 7.1 Statistical Inference

Bayesian Methods:

- 7.2 Prior and Posterior Distributions
- 7.3 Conjugate Prior Distributions
- 7.4 Bayes Estimators

Frequentist Methods:

- 7.5 Maximum Likelihood Estimators
- 7.6 Properties of Maximum Likelihood Estimators
  - Skip: p. 434-441 (EM algorithm and Sampling Plans)
- 7.7 Sufficient Statistics
- Skip: 7.8 Jointly Sufficient Statistics
- Skip: 7.9 Improving an Estimator

Bayes Estimator

- In principle, Bayesian inference is the posterior distribution
- However, often people wish to estimate the unknown parameter θ with a single number
- A *statistic*: Any function of observable random variables $X_1, \ldots, X_n$, $T = r(X_1, X_2, \ldots, X_n)$.
  - Example: The sample mean $\overline{X}_n$ is a statistic

Def: Estimator / Estimate

Suppose our observable data $X_1, \ldots, X_n$ is i.i.d. $f(x|\theta)$, $\theta \in \Omega \subset \mathbb{R}$.

- *Estimator* of θ: A real valued function $\delta(X_1, \ldots, X_n)$
- *Estimate* of θ: $\delta(x_1, \ldots, x_n)$, i.e. estimator evaluated at the observed values

- An estimator is a statistic and a random variable

Bayes Estimator

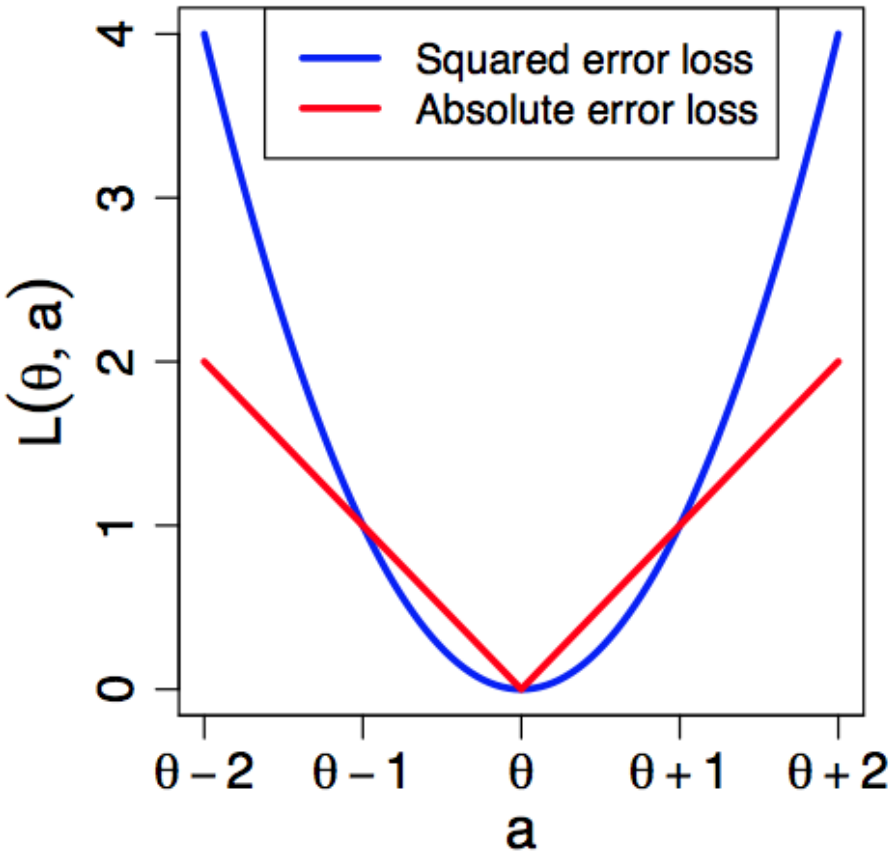
Def: Loss Function

Loss function: A real valued function $L(\theta, a)$ where $\theta \in \Omega$ and $a \in \mathbb{R}$.

- $L(\theta, a) =$ what we loose by using a as an estimate when θ is the true value of the parameter.

Examples:

- *Squared error loss function*: $L(\theta, a) = (\theta - a)^2$
- *Absolute error loss function*: $L(\theta, a) = |\theta - a|$



Bayes Estimator

- Idea: Choose an estimator $\delta(\mathbf{X})$ so that we minimize the expected loss

Def: Bayes Estimator – Minimum expected loss

An estimator is called the *Bayesian estimator* of θ if for all possible observations $\mathbf{x}$ of $\mathbf{X}$ the expected loss is minimized. For given $\mathbf{X} = \mathbf{x}$ the expected loss is

$$E\left(L(\theta, a)|\mathbf{x}\right) = \int_{\Omega} L(\theta, a)p(\theta|\mathbf{x})d\theta$$

Let $a^*(\mathbf{x})$ be the value of a where the minimum is obtained. Then $\delta^*(\mathbf{x}) = a^*(\mathbf{x})$ is the *Bayesian estimate* of θ and $\delta^*(\mathbf{X})$ is the *Bayesian estimator* of θ.

Bayes Estimator

For squared error loss: The posterior mean $\delta^*(\mathbf{X}) = E(\theta|\mathbf{X})$

- $\min_a E\left(L(\theta, a)|\mathbf{x}\right) = \min_a E\left((\theta - a)^2|\mathbf{x}\right)$. The mean of $\theta|\mathbf{x}$ minimizes this, i.e. the posterior mean.

For absolute error loss: The posterior median

- $\min_a E\left(L(\theta, a)|\mathbf{x}\right) = \min_a E\left(|\theta - a| \mid \mathbf{x}\right)$. The median of $\theta|\mathbf{x}$ minimizes this, i.e. the posterior median.

The Posterior mean is a more common estimator because it is often difficult to obtain a closed expression of the posterior median.

Examples

Normal Bayes Estimator, with respect to squared error loss:

- If $X_1, \ldots, X_n$ are $N(\theta, \sigma^2)$ and $\theta \sim N(\mu_0, \nu_0^2)$ then the Bayesian estimator of θ is

$$\delta^*(\mathbf{X}) = \frac{\sigma^2 \mu_0 + n\nu_0^2 \overline{\mathbf{X}}_n}{\sigma^2 + n\nu_0^2}$$

Binomial Bayes Estimator, with respect to squared error loss:

- If $X \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(\alpha, \beta)$ then the Bayesian estimator of θ is

$$\delta^*(\mathbf{X}) = \frac{\alpha + X}{\alpha + \beta + n}$$

Consistency

Def: Consistent estimators

An estimator $\delta_n(\mathbf{X}) = \delta(X_1, \ldots, X_n)$ is *consistent* if

$$\delta(\mathbf{X}) \xrightarrow{P} \theta \quad \text{as } n \to \infty$$

- Under fairly general conditions and for a wide range of loss functions, the Bayes estimator is consistent

Bayesian Inference – Pros and cons

Pros:

- Gives a coherent theory for statistical inference such as estimation.
- Allows for incorporation of prior scientific knowledge about parameters

Cons:

- Selecting a scientifically meaningful prior distributions (and loss functions) is often difficult, especially in high dimensions

Frequentist Inference

Likelihood

When the joint pdf/pf $f(\mathbf{x}|\theta)$ is regarded as a function of θ for given observations $x_1, \ldots, x_n$ it is called the *likelihood function*.

Maximum Likelihood Estimator

Maximum likelihood estimator (MLE): For any given observations $\mathbf{x}$ we pick the $\theta \in \Omega$ that maximizes $f(\mathbf{x}|\theta)$.

Frequentist Inference

Likelihood

When the joint pdf/pf $f(\mathbf{x}|\theta)$ is regarded as a function of θ for given observations $x_1, \ldots, x_n$ it is called the *likelihood function*.

Maximum Likelihood Estimator

Maximum likelihood estimator (MLE): For any given observations $\mathbf{x}$ we pick the $\theta \in \Omega$ that maximizes $f(\mathbf{x}|\theta)$.

- Given $\mathbf{X} = \mathbf{x}$, the *maximum likelihood estimate (MLE)* will be a function of $\mathbf{x}$. Notation: $\hat{\theta} = \delta(\mathbf{X})$
- Potentially confusing notation: Sometimes $\hat{\theta}$ is used for both the estimator and the estimate.

Frequentist Inference

Likelihood

When the joint pdf/pf $f(\mathbf{x}|\theta)$ is regarded as a function of θ for given observations $x_1, \ldots, x_n$ it is called the *likelihood function*.

Maximum Likelihood Estimator

Maximum likelihood estimator (MLE): For any given observations $\mathbf{x}$ we pick the $\theta \in \Omega$ that maximizes $f(\mathbf{x}|\theta)$.

- Given $\mathbf{X} = \mathbf{x}$, the *maximum likelihood estimate (MLE)* will be a function of $\mathbf{x}$. Notation: $\hat{\theta} = \delta(\mathbf{X})$
- Potentially confusing notation: Sometimes $\hat{\theta}$ is used for both the estimator and the estimate.
- Note: The MLE is required to be in the parameter space Ω.

Frequentist Inference

Likelihood

When the joint pdf/pf $f(\mathbf{x}|\theta)$ is regarded as a function of θ for given observations $x_1, \ldots, x_n$ it is called the *likelihood function*.

Maximum Likelihood Estimator

Maximum likelihood estimator (MLE): For any given observations $\mathbf{x}$ we pick the $\theta \in \Omega$ that maximizes $f(\mathbf{x}|\theta)$.

- Given $\mathbf{X} = \mathbf{x}$, the *maximum likelihood estimate (MLE)* will be a function of $\mathbf{x}$. Notation: $\hat{\theta} = \delta(\mathbf{X})$
- Potentially confusing notation: Sometimes $\hat{\theta}$ is used for both the estimator and the estimate.
- Note: The MLE is required to be in the parameter space Ω.
- Often it is easier to maximize the *log-likelihood* $L(\theta) = \log(f(\mathbf{x}|\theta)$

Examples

- Let $X \sim \text{Binomial}(\theta)$. Find the maximum likelihood estimator of θ. Say we observe $X = 3$, what is the maximum likelihood estimate of θ?
- Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$.
  - Find the MLE of μ when σ^2 is known
  - Find the MLE of μ and σ^2 (both unknown)
- Let $X_1, \ldots, X_n$ be i.i.d. $\text{Uniform}[0, \theta]$, where $\theta > 0$. Find $\hat{\theta}$
- Let $X_1, \ldots, X_n$ be i.i.d. $\text{Uniform}[\theta, \theta + 1]$. Find $\hat{\theta}$

MLE

Intuition:

- We pick the parameter that makes the observed data most likely
- But: The likelihood is not a pdf/pf: If the likelihood of θ_1 is larger than the likelihood of θ_1, i.e. $f(\mathbf{x}|\theta_2) > f(\mathbf{x}|\theta_1)$ it does NOT mean that θ_2 is more likely
  - Remember: θ is not random here

Limitations:

- Does not always exist
- Not always appropriate - we cannot incorporate "external" (prior) knowledge
- May not be unique