MATH 541A Introduction to Mathematical Statistics

Yizhe Zhu[*]

April 29, 2025

Abstract

This is supplementary Lecture Notes for MATH 541A Introduction to Mathematical Statistics, Spring 2025 at USC. The material is based on [6, 8].

Contents

[*] Department of Mathematics, University of Southern California, yizhezhu@usc.edu

1 Parametric families of distributions

In statistics, a **parametric family** refers to a specific class of statistical models. These models consist of families of probability distributions defined by a finite number of parameters that capture key aspects such as shape, scale, and location.

In this course, we assume that the underlying data follows a specific probability distribution characterized by such parameters. Our primary goal is to estimate these parameters, a process known as **point estimation**. This forms the foundation for analyzing and interpreting data using parametric models.

Example 1.1. The following common distributions are characterized by finitely many parameters.

- **Normal distribution (Gaussian):** Defined by two parameters: the mean μ and the variance σ^2. The probability density function (PDF) is:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Exponential distribution:** Defined by a single parameter $\lambda > 0$, which is the rate parameter. The PDF is:

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0$$

- **Binomial distribution:** Defined by two parameters: n, the number of trials, and p, the probability of success in a single trial. The probability mass function (PMF) is:

$$P(X = k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

- **Poisson distribution:** Defined by a single parameter $\lambda > 0$, which represents the mean and variance of the distribution. The PMF is:

$$P(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

- **Gamma distribution:** Defined by two parameters: the shape parameter $\alpha > 0$ and the rate parameter $\beta > 0$. The PDF is:

$$f(x|\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \quad x > 0$$

- **Multivariate Normal Distribution:** Defined by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The PDF for a d-dimensional vector $\mathbf{x}$ is:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\det \boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

A **statistical model** is a collection of probability distributions on some sample space. We assume that the collection $\mathcal{P}$, is indexed by some set Θ. The set Θ is called the **parameter space**. The model is a parametric model if $\Theta \subseteq \mathbb{R}^d$ for some positive integer d. We can specify $\mathcal{P}$ in terms of the corresponding probability density functions:

$$\mathcal{P} = \{f_\theta \mid \theta \in \Theta\}.$$

The goal of point estimation is: given a collection of samples $X_1, \dots, X_n$ from a distribution f_θ in $\mathcal{P}$, we aim to answer the following questions:

- How do we estimate θ or a function $t(\theta)$ of θ using a function of the samples $W(X_1, \dots, X_n)$?
- How can we evaluate and compare different estimators?

Remark 1.2. A statistical model is considered **nonparametric** when the parameter space is infinite-dimensional.

1.1 Exponential families

The exponential family is a broad class of probability distributions characterized by their mathematical structure, which makes them particularly useful in statistics and machine learning. These distributions share a common form that enables efficient computation of likelihoods, gradients, and sufficient statistics, making them foundational in various modeling techniques.

Exponential family A probability distribution belongs to the **exponential family** if its probability density function (or probability mass function, for discrete distributions) can be expressed in the following general form:

$$f(x \mid \theta) = h(x) \exp\left(\sum_{i=1}^{k} \eta_i(\theta) t_i(x) - A(\theta)\right), \tag{1}$$

{eq:exp_fami

where:

- θ is the parameter of the distribution, where $\theta \in \Theta \subset \mathbb{R}^d$.
- $h(x) \geq 0, t_i(x)$ are functions of x. $h(x)$ is a *base measure*, and $T(x) = (t_1(x), \dots, t_k(x))$ is the sufficient statistic.
- $\eta(\theta) = (\eta_1(\theta), \dots, \eta_k(\theta))$ is the *canonical parameter*.
- $$A(\theta) = \log \int \exp(\sum_{i=1}^{k} \eta_i(\theta) t_i(x)) h(x) dx$$

  is the *log-partition function*, which ensures that the probability distribution $f_X(x|\theta)$ integrates to 1.

Exponential family in canonical form We can re-parameterize the family by

$$f(x|\eta) = h(x) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x) - A^*(\eta)\right)$$

with

$$A^*(\eta) = \log \int \exp(\sum_{i=1}^{k} \eta_i t_i(x)) h(x) dx.$$

Note that in the new form, $h(x)$ and $t_i(x)$ are the same, but the interaction between parameters $\eta = (\eta_1, \ldots, \eta_k)$ and $T(x) = (t_1(x), \ldots, t_k(x))$ becomes linear. The set

$$\mathcal{H} = \{\eta : A^*(\eta) < \infty, \eta \in \mathbb{R}^k\}$$

is called the **natural parametric space** for the family.

{example:exp

Example 1.3. The exponential family includes many commonly used distributions such as:

- **Normal distribution with a constant variance:**

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

  where $\eta(\mu) = \mu/\sigma^2$, $T(x) = x$, $h(x) = \exp(-x^2/(2\sigma^2))$, $A(\mu) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$. This gives

$$A^*(\eta) = \frac{1}{2}\eta^2\sigma^2 + \frac{1}{2}\log(2\pi\sigma^2).$$

- **Normal distribution with two parameters:** The PDF of the normal distribution with mean μ and variance σ^2 is given by:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

  The normal distribution with unknown mean and variance can be written in the exponential family form:

$$f(x \mid \mu, \sigma^2) = h(x) \exp\left(\eta_1 T_1(x) + \eta_2 T_2(x) - A(\mu, \sigma^2)\right),$$

  where

$$T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \quad \eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = -\frac{1}{2\sigma^2},$$

$$A(\mu, \sigma^2) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2), \quad h(x) = 1.$$

  We can rewrite $A(\mu, \sigma^2)$ as

$$A^*(\eta_1, \eta_2) = -\frac{1}{4}\eta_1^2/\eta_2 + \frac{1}{2}\log(-\pi/\eta_2).$$

  The natural parameter space is $\{\eta : \eta_1 \in \mathbb{R}, \eta_2 < 0\}$.

- **Bernoulli distribution:** When $p \in (0,1)$, it can be rewritten in exponential family form for $x \in \{0,1\}$,

$$\begin{aligned} f(x \mid p) &= p^x (1-p)^{1-x} \\ &= (1-p) \exp\left(\log\left(\frac{p}{1-p}\right) x\right) \end{aligned}$$

  with $\eta(p) = \log\left(\frac{p}{1-p}\right)$, $T(x) = x$,

$$A^*(\eta) = -\log(1-p) = \log(1+e^\eta), \tag{2}$$

{eq:Bernoull

  and $h(x) = 1$ for $x \in \{0,1\}$.

- **Poisson distribution:**

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots,$$

  where $\eta(\lambda) = \log(\lambda)$ and $T(x) = x$.

In the examples above, the choices for $T(x)$ and η are not unique, but $\langle \eta, T(x) \rangle$ is uniquely determined. For the same reason, the choices for $h(x), A^*(\eta)$ are not unique.

Not all parameter families are exponential families. Here are some examples.

{example:cau

Example 1.4 (Cauchy distribution is not an exponential family)**.** The probability density function (PDF) of the Cauchy distribution with location parameter μ and scale parameter $\gamma > 0$ is given by:

$$f(x \mid \mu, \gamma) = \frac{1}{\pi\gamma} \frac{1}{1 + \left(\frac{x-\mu}{\gamma}\right)^2}, \quad x \in \mathbb{R}.$$

The Cauchy PDF can be rewritten as:

$$f(x \mid \mu, \gamma) = \frac{1}{\pi\gamma} \exp\left(-\ln\left(1 + \left(\frac{x-\mu}{\gamma}\right)^2\right)\right).$$

The term $-\ln\left(1 + \left(\frac{x-\mu}{\gamma}\right)^2\right)$ depends on both x and μ, γ in a way that cannot be separated into a linear combination of $T(x)$ and $\eta(\theta)$.

{rmk:exp_sup

Remark 1.5. From the definition 1, the support of an exponential family is independent of θ. Namely,

$$\{x : f(x|\theta) > 0\} = \{x : h(x) > 0\}.$$

Therefore any parametric family whose support depends on θ is not an exponential family.

Example 1.6. Define the indicator function of a set A as

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases}$$

The parametric family

$$f(x|\theta) = \theta^{-1} \exp(1 - x/\theta) \mathbf{1}\{x \geq \theta\}$$

is not an exponential family. In the expression above, the set $\{x : f(x|\theta) > 0\} = [\theta, \infty)$, which depends on θ.

Curved exponential family An exponential family is **curved** if the dimension of the vector θ, d, is less than k in (1). If $d = k$, the family is a *full exponential family.*

{example:cur

Example 1.7. $N(\mu, \mu^2)$ is a curved exponential family for $\mu \neq 0$. Its PDF is

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\mu^2}\right) = \frac{1}{\sqrt{2\pi\mu^2}} \exp(-1/2) \exp\left(-\frac{x^2}{2\mu^2} + \frac{x}{\mu}\right)$$

with $\eta(\mu) = (\frac{1}{\mu}, -\frac{1}{2\mu^2})$ and $T(x) = (x, x^2)$. In this example, $\eta = (\eta_1, \eta_2) \in \mathbb{R}^2$ with a constraint $\eta_2 = -\frac{1}{2}\eta_1^2$ and $\eta_1 \neq 0$. The domain of the canonical parameter η is a quadratic curve in $\mathbb{R}^2$ with one point $(0, 0)$ removed.

1.2 Differential identities

The mean and variance of probability distributions are defined as integrals with respect to the distribution. A nice property of the exponential family is that the mean and variance of the distribution can be obtained by calculating derivatives of the log-partition function $A^*(\eta)$. Below we write $A = A^*(\eta)$.

Example 1.8. We start with a simple example of Bernoulli distribution. In (2), we have $A^*(\eta) = \log(1 + e^\eta)$ with $p = \frac{1}{1+e^{-\eta}}$. Taking the first derivative with respect to η, we obtain

$$\frac{dA}{d\eta} = \frac{1}{1 + e^{-\eta}} = p,$$

which is the mean. Taking a second derivative, we have

$$\frac{d^2A}{d\eta^2} = \frac{dp}{d\eta} = \frac{e^{-\eta}}{(1 + e^{-\eta})^2} = p(1-p),$$

which is the variance of a Bernoulli random variable.

{exp:gaussia

Example 1.9. Recall for the normal distribution $N(\mu, \sigma^2)$,

$$\eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = -\frac{1}{2\sigma^2}, \quad A^*(\eta_1, \eta_2) = -\frac{1}{4}\eta_1^2/\eta_2 + \frac{1}{2}\log(-\pi/\eta_2).$$

Then

$$\begin{aligned} \frac{dA}{d\eta_1} &= -\frac{1}{2}\frac{\eta_1}{\eta_2} = \mu \\ \frac{d^2A}{d\eta_1^2} &= -\frac{1}{2\eta_2} = \sigma^2. \end{aligned}$$

The two examples above are due to a general property of the log-partition function $A^*(\eta)$: by taking derivatives of $A^*(\eta)$, we obtain information on the density function $f(x|\eta)$. For simplicity, we now write $A = A^*(\eta)$.

Theorem 1.10. *If $A(\eta)$ is differentiable*[1], *then*

$$\frac{\partial A(\eta)}{\partial \eta_j} = \mathbb{E}_\eta[t_j(X)], \tag{3}$$ {eq:first_de

where the expectation is taken with respect to the distribution $f(x|\eta)$. In vector notation, we can write

$$\nabla A(\eta) = \frac{\partial A(\eta)}{\partial \eta} = \mathbb{E}_\eta[T(X)].$$

If $A(\eta)$ is twice differentiable, then

$$\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j} = \mathrm{Cov}(t_i(X) t_j(X)).$$

Equivalently,

$$H(A(\eta)) = \frac{\partial^2 A(\eta)}{\partial \eta \partial \eta^\top} = \mathrm{Cov}(T(X)).$$

Proof. W can compute the derivatives of $e^{A(\eta)}$ as

$$\begin{aligned} e^{A(\eta)} \frac{\partial A(\eta)}{\partial \eta_j} &= \frac{\partial}{\partial \eta_j} \left(\int \exp\left(\sum_{i=1}^k \eta_i t_i(x) \right) h(x) dx \right) \\ &= \int t_j(x) \exp\left(\sum_{i=1}^k \eta_i t_i(x) \right) h(x) dx. \end{aligned}$$

In this line, we take the derivatives inside the interval. This can be justified using the Dominated Convergence Theorem; see [8]. Then

$$\frac{\partial A(\eta)}{\partial \eta_j} = \int t_j(x) h(x) \exp\left(\sum_{i=1}^k \eta_i t_i(x) - A(\eta) \right) dx = \int t_j(x) f(x|\eta) dx = \mathbb{E}_\eta[t_j(X)].$$

Taking the second derivative,

$$\begin{aligned} \frac{\partial^2 A(\eta)}{\partial \eta_j \partial \eta_l} &= \int t_j(x) h(x) (t_l(x) - \frac{\partial}{\partial \eta_l} A(\eta)) \exp\left(\sum_{i=1}^k \eta_i t_i(x) - A(\eta) \right) dx \\ &= \int t_j(x) t_l(x) h(x) \exp\left(\sum_{i=1}^k \eta_i t_i(x) - A(\eta) \right) dx \\ &\quad - \frac{\partial}{\partial \eta_l} A(\eta)) \int t_j(x) h(x) \exp\left(\sum_{i=1}^k \eta_i t_i(x) - A(\eta) \right) dx \\ &= \mathbb{E}[t_j(X) t_l(X)] - \mathbb{E}[t_l(X)] \mathbb{E}[t_j(X)] = \mathrm{Cov}(t_j(X), t_l(X)), \end{aligned}$$

where in the last identity we use (3). □

[1] $A(\eta)$ is differentiable if η is an interior point of $\mathcal{H}$, see [8, Theorem 2.4]

The **cumulant generating function** of a random vector $X \in \mathbb{R}^k$ is defined as the natural logarithm of the moment generating function (MGF) of X,

$$K_X(u) = \log M_X(u) = \log \mathbb{E}\left[e^{\langle u, X\rangle}\right], \quad u \in \mathbb{R}^k.$$

The corresponding derivatives of $K_X(u)$ at $u = 0$ are called cumulants of X denoted by

$$\kappa_{r_1,\ldots,r_k} = \frac{\partial^{r_1}}{\partial u_1^{r_1}} \cdots \frac{\partial^{r_k}}{\partial u_k^{r_k}} K_X(u) \mid_{u=0} .$$

Proposition 1.11. In the 1-dimensional case, $\kappa_1 = \mathbb{E}[X], \kappa_2 = \mathrm{Var}(X)$.

Proof. The moment generating function $M_X(u) = \mathbb{E}e^{uX}$. We have $\mathbb{E}[X^k] = M_X^{(k)}(0)$ is the k-th derivative of $M_X(u)$ at $u = 0$. Then by the chain rule,

$$\begin{aligned}\kappa_1 &= K_X'(0) = \frac{d}{du} \log M_X(u) \mid_{u=0} = \mathbb{E}X,\\ \kappa_2 &= K_X''(0) = \frac{d^2}{du^2} \log M_X(u) \mid_{u=0} = \mathrm{Var}(X).\end{aligned}$$

□

{prop:cumula

Proposition 1.12. Let X be a random variable with distribution $f(x|\eta)$ from an exponential family. Let $\kappa_{r_1,\ldots,r_k}$ be the cumulants of $T(X)$. Then

$$K_{T(X)}(u) = A(\eta + u) - A(\eta)$$

and

$$\kappa_{r_1,\ldots,r_k} = \frac{\partial^{r_1}}{\partial \eta_1^{r_1}} \cdots \frac{\partial^{r_k}}{\partial \eta_k^{r_k}} A(\eta).$$

Proof. The moment generating function of $T(X)$ is

$$\begin{aligned}\mathbb{E}_\eta e^{\langle u, T(X)\rangle} &= \int e^{\langle u, T(X)\rangle + \langle \eta, T(X)\rangle - A(\eta)} h(x) dx\\ &= e^{A(\eta+u) - A(\eta)} \int e^{\langle u+\eta, T(X)\rangle - A(\eta+u)} h(x) dx\\ &= e^{A(\eta+u) - A(\eta)} \int f(x|u+\eta) dx = e^{A(\eta+u)-A(\eta)},\end{aligned}$$

for $u + \eta \in \mathcal{H}$. Taking the natural log we have

$$K_{T(X)}(u) = A(\eta + u) - A(\eta).$$

Then

$$\kappa_{r_1,\ldots,r_k}(T(X)) = \frac{\partial^{r_1}}{\partial u_1^{r_1}} \cdots \frac{\partial^{r_k}}{\partial u_k^{r_k}} K_{T(X)}(u) \mid_{u=0} = \frac{\partial^{r_1}}{\partial \eta_1^{r_1}} \cdots \frac{\partial^{r_k}}{\partial \eta_k^{r_k}} A(\eta).$$

□

From Proposition 1.12, for an exponential family with a canonical form index by $\eta \in \mathcal{H}$, one can obtain cumulants of $T(X)$ from derivatives of $A(\eta)$. $A(\eta)$ is also called the *cumulant function* of the exponential family.

Example 1.13. In Example 1.9 for $X \sim N(\mu, \sigma^2)$, we have $T(X) = (X, X^2)$. By Proposition 1.12,

$$K_{(X,X^2)}(u_1, u_2) = A(\eta_1 + u_1, \eta_2 + u_2) - A(\eta_1, \eta_2).$$

Then

$$K_X(u) = K_{(X,X^2)}(u, 0) = A(\eta_1 + u, \eta_2) - A(\eta_1, \eta_2).$$

This gives

$$\mu = \kappa_1(X) = \frac{\partial A}{\partial \eta_1}, \quad \sigma^2 = \kappa_2(X) = \frac{\partial^2 A}{\partial \eta_1^2}.$$

1.3 Convexity

Let's recall the definition of convex sets and convex functions.

Definition 1.14. A set $S \subseteq \mathbb{R}^n$ is called *convex* if, for any two points $x_1, x_2 \in S$ and any $\lambda \in [0, 1]$, the following condition holds:

$$\lambda x_1 + (1 - \lambda) x_2 \in S.$$

In other words, the line segment connecting any two points in S lies entirely within S.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *convex* if its domain is a convex set and, for any $x_1, x_2 \in \mathrm{dom}(f)$ and $\lambda \in [0, 1]$, the following inequality holds:

$$f(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

Geometrically, this means that the line segment connecting $f(x_1)$ and $f(x_2)$ lies above or on the graph of f.

Another useful property of the exponential family is the convexity of the log partition function and the natural parameter space. We show it in the next theorem.

Theorem 1.15. *The natural parameter space $\mathcal{H}$ is convex and the cumulant function $A(\eta)$ is convex.*

Proof. We will use Holder's inequality: for $p, q \geq 1$ with $p^{-1} + q^{-1} = 1$,

$$\int |fg| d\mu \leq \left(\int |f|^p d\mu \right)^{1/p} \left(\int |f|^q d\mu \right)^{1/q}.$$

Namely,

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Choose $\eta_1 \neq \eta_2$ in $\mathcal{H}$ and let $\eta = \lambda\eta_1 + (1-\lambda)\eta_2$ with $\lambda \in (0,1)$. We have

$$\begin{aligned}
e^{A(\eta)} &= \int e^{\langle \lambda\eta_1 + (1-\lambda)\eta_2, T(x)\rangle} h(x)dx \\
&= \int e^{\lambda\langle \eta_1, T(x)\rangle} e^{(1-\lambda)\langle \eta_1, T(x)\rangle} h(x)dx \\
&\leq \left(\int e^{\langle \eta_1, T(x)\rangle} h(x)dx\right)^{\lambda} \left(\int e^{\langle \eta_2, T(x)\rangle} h(x)dx\right)^{1-\lambda} \\
&= e^{\lambda A(\eta_1) + (1-\lambda)A(\eta_2)} < \infty,
\end{aligned}$$

where we use Holder's inequality with $p = 1/\lambda, q = 1/(1-\lambda)$. Taking log on both sides, we find

$$A(\eta) \leq \lambda A(\eta_1) + (1-\lambda)A(\eta_2) < \infty.$$

This implies $\mathcal{H}$ is convex and $A(\eta)$ is convex as a function of η. □

Remark 1.16. Is there a contradiction to the curved exponential family in Example 1.7? The parameter space in Example 1.7 is

$$\mathcal{N} = \{\eta : \eta_2 = -\frac{1}{2}\eta_1^2, \eta_1 \neq 0\} = \eta(\Theta)$$

where $\Theta = \{(\mu, \mu^2), \mu \neq 0\}$. We see $\mathcal{N}$ is a subset of the natural parameter space

$$\mathcal{H} = \{\eta = (\eta_1, \eta_2) : A(\eta) < \infty, \eta \in \mathbb{R}^2\}.$$

The convexity result does not apply to $\mathcal{N} \subset \mathcal{H}$.

1.4 Location and scale families

Given a probability density f, one can form a distribution family by deforming f in certain ways. Below are two examples of shifting and rescaling f.

{def:locatic

Definition 1.17. Let $f(x)$ be a probability density function. The family

$$f(x|\mu) = f(x-\mu),$$

indexed by the parameter $\mu \in \mathbb{R}$ is called a location family.

Definition 1.18. Let $f(x)$ be a probability density function. The family

$$f(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right),$$

indexed by σ is a scale family, where $\sigma > 0$ is the scale parameter. Similarly, the family

$$f(x|\mu,\sigma) = \frac{1}{\sigma} f((x-\mu)/\sigma)$$

is a *location-scale* family.

By definition, the Cauchy distribution in Example 1.4 with parameter μ, γ is a location-scale family.

Remark 1.19. For the connection between the exponential family and the graphical model, see [12, 13].

2 Sufficient statistic

2.1 Sufficient statistic

Given i.i.d. samples $X_1, \dots, X_n$ sampled from a parametric family $f(x|\theta)$, where $\theta \in \Theta$. To estimate θ, we may construct a map $T : \mathbf{X} = (X_1, \dots, X_n) \to \mathbb{R}^k$ and estimate θ based on $T(\mathbf{X})$. Here $T(\mathbf{X})$ can be seen as a way of data reduction or summary. For example, the sample mean $T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$ is a statistic.

We are interested in a way to construct $T(\mathbf{X})$ which does not lose any information about θ in the data $X_1, \dots, X_n$. The definition of **sufficient statistic** makes this intuition rigorous.

Definition 2.1. A statistic $T(X_1, \dots, X_n)$ is sufficient for θ if the conditional distribution of $(X_1, \dots, X_n)$ given the value of $T(X_1, \dots, X_n)$ does not depend on θ. Namely, the posterior density of $(X_1, \dots, X_n)$ given $T(X_1, \dots, X_n)$ denoted by

$$f_\theta(x_1, \dots, x_n | T(X_1, \dots, X_n) = y)$$

does not depend on θ for any $\mathbf{x}, y$.

From this definition, any invertible function of a sufficient statistic is sufficient. Moreover, we have the following way to check if $T(\mathbf{X})$ is sufficient for θ. We write $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{X} = (X_1, \dots, X_n)$.

{thm:ratio_s

Theorem 2.2. *If $p(\mathbf{x}|\theta)$ is the joint pdf of $\mathbf{X}$ and $q(t, \theta)$ is the pdf of $T(\mathbf{X})$, then if for every $\mathbf{x}$, the ratio $\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}$ is a independent of θ, then $T(\mathbf{X})$ is a sufficient statistic for θ.*

For discrete random variables, we use PMFs instead of PDFs in Theorem 2.2.

Proof. We use Bayes' rule:

$$f_\theta(\mathbf{x}|T(\mathbf{X}) = y) = \frac{f_{X_1,\dots,X_n,T(\mathbf{X})}(x_1, \dots, x_n, y)}{f_{T(\mathbf{X})}(y)}$$

When $y \neq T(x_1, \dots, x_n)$, $f_{X_1,\dots,X_n,T(\mathbf{X})}(x_1, \dots, x_n, y) = 0$ and there is nothing to show. When $y = T(x_1, \dots, x_n)$, we get

$$f_\theta(\mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) = \frac{f_{X_1,\dots,X_n,T(\mathbf{X})}(x_1, \dots, x_n, T(\mathbf{x}))}{f_{T(\mathbf{X})}(T(\mathbf{x}))} = \frac{f_{X_1,\dots,X_n}(x_1, \dots, x_n)}{q(T(\mathbf{x})|\theta)} = \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)},$$

which, by our assumption, is independent of θ. □

Theorem 2.2 shows that if $p(\mathbf{x}|\theta)$ and $q(T(\mathbf{x})|\theta)$ has the same dependence on θ, then $T(\mathbf{X})$ is a sufficient.

Example 2.3 (Binomial distribution)**.** Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter $\theta \in (0, 1)$, then $T(\mathbf{X}) = X_1 + \cdots + X_n$ is a sufficient statistic for θ.

Here $T(\mathbf{X}) \sim B(n, \theta)$ is a binomial random variable. For each $\mathbf{x} \in \{0, 1\}^n$, denote $t = \sum_{i=1}^n x_i$. The ratio of $p(\mathbf{x}|\theta)$ and $q(T(\mathbf{x})|\theta)$ is

$$\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} = \frac{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}(T(\mathbf{X}) = t)} = \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{x_1 + \cdots + x_n}}.$$

Hence $T(\mathbf{X})$ is a sufficient statistic for θ. And the sample mean $\frac{1}{n}T(\mathbf{X})$ is also a sufficient statistic for θ.

Example 2.4 (Gaussian with a known variance)**.** Let $X_1, \dots, X_n$ be i.i.d. samples with distribution $N(\mu, \sigma^2)$ where σ^2 is known. Let $T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i = \overline{\mathbf{x}}$.

$$\begin{aligned}
f(\mathbf{x}|\mu) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i - \overline{\mathbf{x}} + \overline{\mathbf{x}} - \mu)^2}{2\sigma^2} \right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i - \overline{\mathbf{x}})^2 + (\overline{\mathbf{x}} - \mu)^2 + 2(x_i - \overline{\mathbf{x}})(\overline{\mathbf{x}} - \mu)}{2\sigma^2} \right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i - \overline{\mathbf{x}})^2 + (\overline{\mathbf{x}} - \mu)^2}{2\sigma^2} \right) \qquad (4) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n (x_i - \overline{\mathbf{x}})^2/(2\sigma^2) \right) \exp(-n(\overline{\mathbf{x}} - \mu)^2/(2\sigma^2),
\end{aligned}$$

{eq:density_

where in the fourth identity we use $\sum_{i=1}^n x_i = n\overline{\mathbf{x}}$. Since $T(X) \sim N(\mu, \sigma^2/n)$,

$$q(T(\mathbf{x})|\mu) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp(-n(\overline{\mathbf{x}} - \mu)^2/(2\sigma^2)).$$

We can immediately check that $\frac{f(\mathbf{x}|\mu)}{q(T(\mathbf{x})|\mu)}$ is independent of μ.

We know how to check whether $T(\mathbf{X})$ is sufficient or not. But how to construct a sufficient $T(\mathbf{X})$? The following factorization theorem gives a necessary and sufficient condition.

{thm:factori

Theorem 2.5 (Fisher–Neyman factorization theorem)**.** *Let $f(x|\theta)$ denote the joint pdf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is sufficient for θ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that for all sample points $\mathbf{x}$ and all parameter $\theta \in \Theta$,*

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}). \tag{5}$$

{eq:factor}

Proof. We prove it for discrete distributions. The case for continuous distributions is proved in [8, Section 6.4].

Suppose $T(\mathbf{X})$ is a sufficient statistic. Choose $g(t|\theta) = \mathbb{P}(T(\mathbf{X}) = t)$ and $h(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$. Since $T(\mathbf{X})$ is sufficient, $h(\mathbf{x})$ is independent of θ. We find

$$\begin{aligned}
f(\mathbf{x}|\theta) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) &= \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x})) \\
&= \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))\mathbb{P}(T(\mathbf{X}) = T(\mathbf{x})) \\
&= h(\mathbf{x})g(T(\mathbf{x})).
\end{aligned}$$

Now suppose (5) holds. We apply Theorem 2.2. Let $q(T(\mathbf{X})|\theta)$ be the pmf of $T(\mathbf{X})$. Define

$A_{\mathbf{x}} = \{\mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})\}$.

$$\begin{aligned}
\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\mathbb{P}(T(\mathbf{X}) = T(\mathbf{x})|\theta)} \\
&= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{\mathbf{x}}} \mathbb{P}(\mathbf{X} = y|\theta)} = \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{\mathbf{x}}} g(T(\mathbf{y})|\theta)h(\mathbf{y})},
\end{aligned}$$

where in the third identity we use the definition of the pmf of T. Since T is constant on the set $A_{\mathbf{x}}$, we further have

$$\frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{\mathbf{x}}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} = \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\sum_{\mathbf{y}\in A_{\mathbf{x}}} h(\mathbf{y})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{y}\in A_{\mathbf{x}}} h(\mathbf{y})}.$$

□

Below is a non-exponential family example where we can find $T(\mathbf{X})$ using Theorem 2.5.

{example:uni

Example 2.6 (Uniform distribution)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables with the discrete uniform distribution on $\{1, \ldots, \theta\}$, where θ is a positive integer. The pdf of X_i is

$$f(x|\theta) = \begin{cases} 1/\theta & x = 1, 2, \ldots, \theta \\ 0 & \text{otherwise.} \end{cases}$$

$f(x|\theta)$ is not an exponential family since the support of $f(x|\theta)$ depends on θ, (see Remark 1.5).

The joint pdf of $X_1, \ldots, X_n$ is

$$\begin{aligned}
f(\mathbf{x}|\theta) &= \theta^{-n}\mathbf{1}\{x_1, \ldots, x_n \in \{1, \ldots, \theta\}\} \\
&= \theta^{-n}\mathbf{1}\{x_1, \ldots, x_n \in \mathbb{Z}_+, \max_i x_i \le \theta\}\} \\
&= \prod_{i=1}^n \mathbf{1}\{x_i \in \mathbb{Z}_+\}\theta^{-n}\mathbf{1}\{\max_{1\le i\le n} x_i \le \theta\}.
\end{aligned}$$

Let

$$T(\mathbf{x}) = \max_{1\le i\le n} x_i, \quad h(\mathbf{x}) = \prod_{i=1}^n \mathbf{1}\{x_i \in \mathbb{Z}_+\}, \quad g(x) = \theta^{-n}\mathbf{1}\{x \le \theta\}.$$

Then we have the factorization

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x}))h(\mathbf{x}).$$

By the factorization theorem 2.5, $T(\mathbf{X}) = \max_{1\le i\le n} X_i$ is a sufficient statistic for θ, and we denote

{example:uni

Example 2.7. Suppose $X_1, \ldots, X_n$ are random samples from uniform distribution on $[\theta, \theta + 1]$, namely:

$$f(x|\theta) = \mathbf{1}_{[\theta,\theta+1]}(x).$$

The joint density is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n \mathbf{1}_{[-\theta,\theta+1]}(x_i) = \mathbf{1}_{[\theta,\infty)}(\min_i x_i)\mathbf{1}_{(-\infty,\theta+1]}(\max_i x_i).$$

By the factorization theorem, $T(\mathbf{X}) = (\min_i X_i, \max_i X_i)$ is sufficient.

Remark 2.8. Compared to Example 2.6, why do we need both $\min_i X_i$ and $\max_i X_i$ in Example 2.7? Intuitively it is true that both $\min_i X_i$ and $\max_i X_i$ contains information about θ.

Example 2.9 (Gaussian $N(\mu, \sigma^2)$ with both parameters unknown)**.** Since we know $N(\mu, \sigma^2)$ is an exponential family, the corresponding $T(\mathbf{X})$ can be constructed directly from the density function $f(x|\theta)$. Recall from (4),

$$f(\mathbf{x}|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(- \sum_{i=1}^{n} \frac{(x_i - \overline{\mathbf{x}})^2 + (\overline{\mathbf{x}} - \mu)^2}{2\sigma^2} \right). \qquad (6)$$

{eq:joint_pd

Let $T_1(\mathbf{x}) = \overline{\mathbf{x}}, T_2(\mathbf{x}) = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2$, and

$$g(t_1, t_2|\theta) = (2\pi\sigma^2)^{-n/2} \exp(-(n(t_1 - \mu)^2 + (n-1)t_2)/(2\sigma^2)).$$

We can write

$$f(\mathbf{x}|\mu, \sigma^2) = g(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu, \sigma^2) h(\mathbf{x}).$$

Then $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\overline{\mathbf{X}}, S^2)$ is a sufficient statistic for (μ, σ^2).

With Theorem 2.5, we next show that it is easy to find a sufficient statistic for an exponential family.

{thm:suffici

Theorem 2.10 (Sufficient statistics for an exponential family)**.** *Let $X_1, \ldots, X_n$ be i.i.d. samples from an exponential family $f(x|\theta)$ where*

$$f(x|\theta) = h(x) \exp\left(\sum_{i=1}^{k} \eta_i(\theta) t_i(x) - A(\theta) \right),$$

then $T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j) \right)$ is a sufficient statistic for θ.

Proof. Let $t_i(\mathbf{X}) = \sum_{j=1}^{n} t_i(X_j)$. We have

$$\begin{aligned} f(\mathbf{x}|\theta) &= h(x_1) \cdots h(x_n) \exp\left(\sum_{j=1}^{n} \sum_{i=1}^{k} \eta_i(\theta) t_i(x_j) - nA(\theta) \right) \\ &= \prod_{i=1}^{n} h(x_i) \exp\left(\sum_{i=1}^{k} \eta_i(\theta) \sum_{j=1}^{n} t_i(X_j) - nA(\theta) \right) \\ &= \prod_{i=1}^{n} h(x_i) \exp\left(\sum_{i=1}^{k} \eta_i(\theta) t_i(\mathbf{X}) - nA(\theta) \right) \\ &= \prod_{i=1}^{n} h(x_i) \exp(\langle \eta(\theta), T(\mathbf{X}) \rangle - nA(\theta)). \end{aligned}$$

Then by Theorem 2.5, $T(\mathbf{X})$ is a sufficient statistic for θ. □

Example 2.11 ($N(\mu, \sigma^2)$ as an exponential family)**.** From Example 1.3, $N(\mu, \sigma^2)$ is an exponential family with $T(x) = (x, x^2)$. Theorem 2.10 suggests

$$T'(\mathbf{X}) = (\sum_{j=1}^{n} X_j, \sum_{j=1}^{n} X_j^2)$$

is a sufficient statistic for (μ, σ^2). Note that sufficient statistics are not unique. And $T'(\mathbf{X}) = (\sum_{j=1}^{n} X_j, \sum_{j=1}^{n} X_j^2)$ is a one-to-one map of $(\overline{\mathbf{X}}, S^2)$.

Example 2.12. Suppose $X_1, \dots, X_n$ are i.i.d. random variables with common pdf

$$f(x|\theta) = \begin{cases} (\theta+1)x^\theta, & x \in (0,1), \\ 0 & \text{otherwise.} \end{cases}$$

for $\theta > -1$. We can write $f_\theta(x) = \exp(\theta \ln(x) + \ln(\theta + 1))$ as an exponential family, where $T(x) = \ln(x)$, which is 1-dimensional. Then by Theorem 2.10,

$$T(\mathbf{X}) = \sum_{j=1}^{n} t(X_j) = \sum_{j=1}^{n} \log(X_j) = \log\left(\prod_{j=1}^{n} X_j\right)$$

is a sufficient statistic for θ. Since $\log(x)$ is a one-to-one function, we can conclude $T'(\mathbf{X}) = \prod_{j=1}^{n} X_j$ is also a sufficient statistic for θ.

Example 2.13 (Cauchy distribution)**.** Suppose $X_1, \dots, X_n$ are sampled from the Cauchy distribution $f(x|\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$, we find

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\pi(1 + (x_i - \theta)^2)} = \prod_{i=1}^{n} f(x_{s(i)})$$

for any permutation $s : [n] \mapsto [n]$. Therefore $T(\mathbf{X}) = (X_{s(1)}, \dots, X_{s(n)})$ is a sufficient statistic for any permutation s. But there are no other ways to reduce $\mathbf{X}$ further (we will show it's minimal later).

In particular, the variables $X_{(1)} \le \cdots \le X_{(n)}$ by listing $X_1, \dots, X_n$ in increasing order are called the order statistics. By the factorization theorem, $T(\mathbf{X}) = (X_{(1)}, \cdots, X_{(n)})$ is a sufficient statistic.

2.2 Minimal sufficient statistic

From the factorization theorem 2.5, there are many sufficient statistics for a parameter family $f(x|\theta)$. For example, we can just take $T(X_1, \dots, X_n) = (X_1, \dots, X_n)$. Also, any one-to-one map r applied to $T(X)$ is a sufficient statistic. We may ask what $T(\mathbf{X})$ compresses the data in the most effective way.

Definition 2.14. A sufficient statistic $T(\mathbf{X})$ is called a minimal sufficient statistic if for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$.

Here $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ means whenever $T'(\mathbf{x}) = T'(\mathbf{y})$, we must have $T(\mathbf{x}) = T(\mathbf{y})$, since a function cannot take two different values at one point.

Example 2.15. If $X_1, \ldots, X_{2n}$ are i.i.d. samples from $N(\theta, 1)$. Then $T'(\mathbf{X}) = (\sum_{i=1}^{n} X_i, \sum_{i=n+1}^{2n} X_i)$ is sufficient for θ, but not minimal since we can write $T(\mathbf{X}) = f(T'(\mathbf{X}))$ where $f(t) = t_1 + t_2$.

This definition means the minimal sufficient statistic $T(\mathbf{X})$ cannot be compressed further. But how do we check if $T(\mathbf{X})$ is minimal? From the definition itself, it seems impossible to check. The next theorem is due to Lehmann and Scheffé (1950).

{thm:ratio_m

Theorem 2.16. *Let $f(\mathbf{x}|\theta)$ be the pdf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{x})$ such that for every two sample points $\mathbf{x}, \mathbf{y}$, the ratio $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for θ.*

Proof. We first show that $T(\mathbf{X})$ is a sufficient statistic.

Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x}\}$ be the image of T. For each $t \in \mathcal{T}$, define

$$A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$$

be the preimage of t. Here $\{A_t\}_{t \in \mathcal{T}}$ gives a partition of the space $\mathcal{X}$ where $f(\mathbf{x}|\theta)$ is defined.

For each A_t, we choose one fixed element $\mathbf{x}_t \in A_t$. Then for all $\mathbf{x} \in A_t$,

$$t = T(\mathbf{x}) = T(\mathbf{x}_t) = T(\mathbf{x}_{T(\mathbf{x})}).$$

Since $T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$ for all $\mathbf{x} \in \mathcal{X}$, by our assumption in the statement of Theorem 2.16,

$$h(\mathbf{x}) = \frac{f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)}$$

is a constant function as a function of θ (we can also see this by checking that for each $\mathbf{x} \in A_t$, $T(\mathbf{x}) = T(\mathbf{x}_t)$ and by our assumption $h(\mathbf{x})$ is a constant function of θ on A_t. Since A_t is independent of θ, $h(\mathbf{x})$ is independent of θ on $\mathcal{X}$).

Let $g(t|\theta) = f(\mathbf{x}_t|\theta)$. Then

$$f(\mathbf{x}|\theta) = \frac{f(\mathbf{x}_{T(\mathbf{x})}|\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)} = g(T(\mathbf{x})|\theta) \frac{f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)} = g(T(\mathbf{x})|\theta) h(\mathbf{x}).$$

By the Factorization Theorem, $T(\mathbf{X})$ is sufficient.

Next, we show $T(\mathbf{X})$ is minimal. Let $T'(\mathbf{X})$ be another sufficient statistic for θ, then

$$f(\mathbf{x}|\theta) = g'(T'(\mathbf{x})|\theta) h'(\mathbf{x})$$

for some function g', h'. Let $\mathbf{x}, \mathbf{y}$ be two points such that $T'(\mathbf{x}) = T'(\mathbf{y})$. Then

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g'(T'(\mathbf{x})|\theta) h'(\mathbf{x})}{g'(T'(\mathbf{y})|\theta) h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}.$$

This implies $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ is a constant as a function independent of θ. By our assumption, $T(\mathbf{x}) = T(\mathbf{y})$. Therefore $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$. □

Remark 2.17 (The proof of sufficiency in Theorem 2.16)**.** Let's recap the proof strategy and work backward. To use the Factorization Theorem, we want to construct a function $g(T(\mathbf{x})|\theta)$ and write $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$. The proof starts by picking $g(t|\theta) = f(\mathbf{x}_t|\theta)$, where $\mathbf{x}_t$ is a point indexed by t (to be decided later). This gives us $g(T(\mathbf{x})|\theta) = f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ and a factorization

$$f(\mathbf{x}|\theta) = \frac{f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)} f(\mathbf{x}_{T(\mathbf{x})}|\theta).$$

Now to use the assumption in Theorem 2.16, we need

$$T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})}) \tag{7}$$ {eq:Txt}

so that $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)}$ is independent of θ. Let $T(\mathbf{x}) = t$, Equation (7) implies $t = T(\mathbf{x}_t)$. Hence we need to choose a point $\mathbf{x}_t \in T^{-t}(t) = A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. This suggests the choice $g(t|\theta) = f(\mathbf{x}_t|\theta)$ where $\mathbf{x}_t$ is any point from A_t.

Example 2.18 (minimal sufficient statistic for $N(\mu, \sigma^2)$)**.** Let $X_1, \ldots, X_n$ be i.i.d. samples from $N(\mu, \sigma^2)$. From (6), we have

$$\begin{aligned}\frac{f(\mathbf{x}|\mu,\sigma^2)}{f(\mathbf{y}|\mu,\sigma^2)} &= \frac{\exp\left(-[n(\overline{\mathbf{x}}-\mu)^2 + (n-1)s_{\mathbf{x}}^2]/(2\sigma^2)\right)}{\exp\left(-[n(\overline{\mathbf{y}}-\mu)^2 + (n-1)s_{\mathbf{y}}^2]/(2\sigma^2)\right)} \\ &= \exp\left([-n(\overline{\mathbf{x}}^2 - \overline{\mathbf{y}}^2) + 2n\mu(\overline{\mathbf{x}} - \overline{\mathbf{y}}) - (n-1)(s_{\mathbf{x}}^2 - s_{\mathbf{y}}^2)]/(2\sigma^2)\right).\end{aligned}$$

The ratio is independent of μ, σ^2, if and only if $\overline{\mathbf{x}} = \overline{\mathbf{y}}$ and $s_{\mathbf{x}}^2 = s_{\mathbf{y}}^2$. Therefore $(\overline{\mathbf{X}}, S^2)$ is a minimal sufficient statistic for (μ, σ^2).

Example 2.19 (Cauchy distribution revisited)**.** Suppose $X_1, \ldots, X_n$ are sampled from the Cauchy distribution $f(x|\theta) = \frac{1}{\pi[1+(x-\theta)^2]}$, we find

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\pi[1 + (x_i - \theta)^2]}.$$

Then the ratio

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{\prod_{i=1}^n [1 + (y_i - \theta)^2]}{\prod_{i=1}^n [1 + (x_i - \theta)^2]} = \frac{p(\theta)}{q(\theta)},$$

which is a function of the ratio of two degree-n polynomials $p(\theta), q(\theta)$. This ratio is a constant if and only if the roots of $p(\theta), q(\theta)$ are the same, which is equivalent to $T(\mathbf{x}) = T(\mathbf{y})$ where $T(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$. By Theorem 2.16, the order statistic is a minimal sufficient statistic for θ.

Example 2.20 (minimal sufficient statistic for uniform distribution)**.** Suppose $X_1, \ldots, X_n$ are random samples from uniform distribution on $[\theta, \theta + 1]$, namely:

$$f(x|\theta) = \mathbf{1}_{[\theta,\theta+1]}(x).$$

The joint density is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \mathbf{1}_{[\theta,\theta+1]}(x_i) = \mathbf{1}_{[\theta,\infty)}(\min_i x_i)\mathbf{1}_{(-\infty,\theta+1]}(\max_i x_i).$$

The ratio $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ is independent of θ if and only if $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$. This shows $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is minimal sufficient for θ.

2.3 Ancillary statistics

Definition 2.21. Let $X_1, \dots, X_n$ be i.i.d. samples from $f(x|\theta)$. A statistic $S(\mathbf{X})$ whose distribution is independent of the parameter θ is called an ancillary statistic.

Good examples of ancillary statistics are from the location and scale family.

Example 2.22 (Location family)**.** Let $X_1, \dots, X_n$ be i.i.d. observations from a location family with cdf $F(x-\theta)$. The range $R = X_{(n)} - X_{(1)}$ is an ancillary statistic. Let $Z_1 = X_1 + \theta, \dots, Z_n = X_n + \theta$. Then $Z_1, \dots, Z_n$ are i.i.d. random variables with cdf $F(x)$. Then the cdf of R is

$$F_R(r|\theta) = \mathbb{P}(R \le r) = \mathbb{P}(X_{(n)} - X_{(1)} \le r) = \mathbb{P}(Z_{(n)} - Z_{(1)} \le r),$$

where the last probability is independent of θ.

{rmk:minimal

Remark 2.23 (Ancillary statistics can be a function of minimal sufficient statistics)**.** $f(x|\theta) = \mathbf{1}_{[\theta,\theta+1]}(x)$ is a location family, and we have shown that $(X_{(1)}, X_{(n)})$ is minimal sufficient and by the example above, $X_{(n)} - X_{(1)}$ is ancillary.

Similarly, the Cauchy distribution $f(x|\theta) = \frac{1}{\pi(x-\theta)^2}$ is a scale family. We have shown $(X_{(1)}, \dots, X_{(n)})$ is minimal sufficient, and $X_{(n)} - X_{(1)}$ is ancillary.

Example 2.24 (Scale family)**.** Let $X_1, \dots, X_n$ be i.i.d. observations from a location family with cdf $\frac{1}{\sigma} f(x/\sigma)$ where $\sigma > 0$. Then

$$S(\mathbf{X}) = \left(\frac{X_1}{X_n}, \dots, \frac{X_{n-1}}{X_n} \right)$$

is an ancillary statistic. Let $Z_i = \sigma X_i$. Then Z_i are i.i.d. samples with pdf $f(x)$. The joint pdf of $X_1/X_n, \dots, X_{n-1}/X_n$ is the same as the joint pdf of $Z_1/Z_n, \dots, Z_{n-1}/Z_n$, which is independent of σ.

Let X_1, X_2 be two samples from $N(0, \sigma^2)$. X_1/X_2 has a Cauchy distribution with pdf $\frac{1}{\pi(1+x^2)}$.

2.4 Complete statistics

{def:complet

Definition 2.25 (Complete statistic)**.** Let $f(t|\theta)$ be a family of pdfs for a statistic $T(\mathbf{X})$. The family of probability distributions is called complete if $\mathbb{E}_\theta g(T) = 0$ for all θ implies $\mathbb{P}_\theta(g(T) = 0) = 1$ for all θ and all measurable function g. Equivalently, we call $T(\mathbf{X})$ a complete statistic.

Example 2.26 (Binomial distribution)**.** Suppose T has a binomial (n, p) distribution with $0 < p < 1$. Let g be a function such that $\mathbb{E}_p(g(T)) = 0$. Then

$$0 = \mathbb{E}_p g(T) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} = (1-p)^{-n} \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p} \right)^t.$$

This implies

$$\sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p} \right)^t = 0$$

for all p. The left-hand side is a polynomial with degree at most n in $p/(1-p)$, which cannot be 0 for all p unless the coefficients $g(t) = 0$ for $t = 1, \dots, n$. Since T can only take values in $\{1, \dots, n\}$, we conclude $P_p(g(T) = 0) = 1$.

{example:uni

Example 2.27. Let $X_1, \dots, X_n$ be i.i.d. samples from a uniform distribution on $(0, \theta)$ with $\theta > 0$. We know $T(\mathbf{X}) = X_{(n)}$ is a sufficient statistic by the same argument as in Example 2.6. And we have

$$\mathbb{P}_\theta(T \le t) = \mathbb{P}_\theta(X_1 \le t, \cdots, X_n \le t) = (t/\theta)^n.$$

So the pdf of $T(\mathbf{X})$ is

$$f(t|\theta) = nt^{n-1}\theta^{-n}\mathbf{1}_{(0,\theta)}(t).$$

Suppose $\mathbb{E}_\theta(g(T)) = 0$ for all θ, the following identity holds for all $\theta > 0$:

$$0 = \int_0^\theta g(t)nt^{n-1}\theta^{-n}dt. \tag{8}$$

{eq:const_in

Then we have

$$\begin{aligned}
0 = \frac{d}{d\theta}\mathbb{E}_\theta g(T) &= \frac{d}{d\theta}\int_0^\theta g(t)nt^{n-1}\theta^{-n}dt \\
&= \frac{d}{d\theta}\left(\theta^{-n}\int_0^\theta g(t)nt^{n-1}dt\right) \\
&= (-n)\theta^{-1-n}\int_0^\theta g(t)nt^{n-1}dt + \theta^{-n}g(\theta)n\theta^{n-1} \qquad (9) \\
&= (-n)\theta^{-1}\int_0^\theta g(t)nt^{n-1}\theta^{-n}dt + n\theta^{-1}g(\theta) \\
&= -n\theta^{-1}\mathbb{E}_\theta g(t) + n\theta^{-1}g(\theta),
\end{aligned}$$

{eq:fundamen

where we use the product rule for differentiation and the condition $\mathbb{E}_\theta g(T) = 0$. This implies $g(\theta) = 0$ for all $\theta > 0$.

Remark 2.28 (A more rigorous proof of Example 2.27 based on measure theory)**.** The calculation in Example 2.27 (presented in [6]) ignored the fact that in (9), to use Lebesgue differentiation theorem,

$$\frac{d}{d\theta}\int_0^\theta g(t)t^{n-1}dt = g(\theta)\theta^{n-1}$$

requires g to be a locally integrable function ($\int_K |g(t)|dt < \infty$ for all compact sets in $\mathbb{R}$). Below we sketch another proof without the locally integrable assumption for g in [8]. We will use the following property: suppose $\int_c^\theta f(x)dx = 0$ for all $\theta > c$, then $f(x) = 0$ almost surely for all $x > c$. From (8), we obtain for all $\theta > 0$, $0 = \int_0^\theta g(t)t^{n-1}dt$. This implies $g(t)t^{n-1} = 0$ a.e. for $t > 0$, hence $g(t) = 0$ for $t > 0$ almost everywhere.

We have seen in Remark 2.23 that a minimal sufficient statistic may still contain information on ancillary statistics. The next theorem shows that for a complete sufficient statistic, it is independent of every ancillary statistic.

Theorem 2.29 (Basu's theorem)**.** *If $T(\mathbf{X})$ is a complete and sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.*

Proof. We prove it for discrete distributions. Let $S(\mathbf{X})$ be an ancillary statistic, whose distribution is independent of θ. To show $S(\mathbf{X})$ and $T(\mathbf{X})$ are independent, it suffices to show that for all

$$\mathbb{P}(S(\mathbf{X}) = s | T(\mathbf{X}) = t) = \mathbb{P}(S(\mathbf{X}) = s) \tag{10}$$

{eq:complete

for all possible values s, t. Note that since $T(\mathbf{X})$ is sufficient, the law of $\mathbf{X}$ given $T(\mathbf{X})$ is independent of θ, therefore we have the conditional probability $\mathbb{P}(S(\mathbf{X}) = s | T(\mathbf{X}) = t)$ is independent of θ. From the law of total probability,

$$\mathbb{P}(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} \mathbb{P}(S(\mathbf{X}) = s | T(\mathbf{X}) = t) \mathbb{P}_\theta(T(\mathbf{X}) = t).$$

We can also write

$$\mathbb{P}(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} \mathbb{P}(S(\mathbf{X}) = s) \mathbb{P}_\theta(T(\mathbf{X}) = t).$$

The two equations above imply

$$0 = \sum_{t \in \mathcal{T}} \left[\mathbb{P}(S(\mathbf{X}) = s | T(\mathbf{X}) = t) - \mathbb{P}(S(\mathbf{X}) = s)\right] \mathbb{P}_\theta(T(\mathbf{X}) = t). \tag{11}$$

{eq:total_pr

Let

$$g_s(t) = \mathbb{P}(S(\mathbf{X}) = s | T(\mathbf{X}) = t) - \mathbb{P}(S(\mathbf{X}) = s).$$

Then (11) is equivalent to $\mathbb{E}_\theta g_s(T) = 0$. Since $T(\mathbf{X})$ is complete, this implies $g_s(t) = 0$ for all possible values $t \in \mathcal{T}$ and all possible values of s, which is equivalent to (10). □

Remark 2.30. Basu's theorem implies the order statistic in a Cauchy distribution family is not complete.

Bahadur's theorem shows that complete and sufficient statistics must be minimal if a minimal sufficient statistic exists.

Theorem 2.31 (Bahadur's Theorem)**.** *Assume a minimal sufficient statistic for $f(x|\theta)$ exists. If $T(\mathbf{X})$ is complete and sufficient, then $T(\mathbf{X})$ is a minimal sufficient statistic.*

Proof. Let $\tilde{T}(\mathbf{X})$ be a minimal sufficient statistic. Then since $T(\mathbf{X})$ is sufficient, $\tilde{T}(\mathbf{X}) = f(T(\mathbf{X}))$ for some function f.

Define $g(t) = \mathbb{E}_\theta(T(\mathbf{X}) | \tilde{T}(\mathbf{X}) = t)$. Since $\tilde{T}$ is sufficient, g is a function independent of θ, and

$$\mathbb{E}_\theta g(\tilde{T}(\mathbf{X})) = \mathbb{E}_\theta \mathbb{E}_\theta[T(\mathbf{X}) | \tilde{T}(\mathbf{X}) = \tilde{T}(\mathbf{X})] = \mathbb{E}_\theta T(\mathbf{X}).$$

Then for all θ,

$$\mathbb{E}_\theta(g(f(T(\mathbf{X})) - T(\mathbf{X})) = 0.$$

By completeness,

$$\mathbb{P}_\theta(gf(T(\mathbf{X})) = T(\mathbf{X})) = 1.$$

This shows $T(\mathbf{X}) = g(f(T(\mathbf{X})) = g(\tilde{T}(\mathbf{X}))$. Since $T(\mathbf{X})$ is a function of $\tilde{T}(\mathbf{X})$, by the minimality of $\tilde{T}(\mathbf{X})$, $T(\mathbf{X})$ is minimal. □

Under additional assumptions of an exponential family, we can show that $T(\mathbf{X})$ is complete and sufficient.

{thm:complet

Theorem 2.32 (Complete statistics in the exponential family)**.** *Let $X_1, \ldots, X_n$ be i.i.d. observations from an exponential family with pdf*

$$f(x|\theta) = h(x)\exp(\sum_{j=1}^{k} \eta_j(\theta)t_j(x) - A(\theta)),$$

where $\theta = (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k$. Then $T(\mathbf{X}) = \left(\sum_{i=1}^n t_1(X_i), \ldots, \sum_{j=1}^n t_k(X_i)\right)$ is a complete statistic as long as the parameter space Θ contains an open set in $\mathbb{R}^k$.

Remark 2.33. Note that in this theorem we need $\theta \in \mathbb{R}^k$ and an open set condition. So the curved exponential family example in Example 1.7 does not apply.

Example 2.34 (Application of Basu's theorem)**.** Let $X_1, \ldots, X_n$ be i.i.d. samples from

$$f(x|\theta) = \theta e^{-\theta x}\mathbf{1}\{x \geq 0\}, \quad \theta > 0.$$

Let

$$g(\mathbf{X}) = \frac{X_n}{X_1 + \cdots + X_n} = \frac{1}{\frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1}.$$

Since $f(x|\theta)$ is a scale family, $g(\mathbf{X})$ is an ancillary statistic (this is a function of $(X_1/X_n, \ldots, X_{n-1}/X_n)$). Since $f(x|\theta)$ is also an exponential family with $\Theta = \mathbb{R}_+$, we can apply Theorem 2.32 to conclude $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a complete statistic. By Basu's theorem, $g(\mathbf{X})$ and $T(\mathbf{X})$ are independent. Now we have

$$\frac{1}{\theta} = \mathbb{E}_\theta X_n = \mathbb{E}_\theta T(\mathbf{X})g(\mathbf{X}) = \mathbb{E}_\theta T(\mathbf{X})\mathbb{E}_\theta g(\mathbf{X}) = \frac{n}{\theta}\mathbb{E}_\theta g(\mathbf{X}).$$

Then $\mathbb{E}_\theta g(\mathbf{X}) = 1/n$ for any θ.

3 Point estimation

In this section, we will study the task of estimating θ in a parametric family $f(x|\theta)$ given i.i.d. samples $X_1, \ldots, X_n$. An estimator $\hat{\theta}$ of θ is a function of $X_1, \ldots, X_n$. We will introduce several methods of finding estimators and evaluate their performance.

3.1 Method of moments (MoM)

The method of moments is an intuitive way to estimate parameters. Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with pdf $f(x|\theta_1, \ldots, \theta_k)$. Let $m_l = \frac{1}{n}\sum_{i=1}^n X_i^l$ and $\mu_l' = \mathbb{E}X^l$. Since μ_l is a function of $\theta_1, \ldots, \theta_k$, the method of moments estimator $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ is obtained by solving the system of equations

$$\begin{aligned} m_1 &= \mu_1' \\ m_2 &= \mu_2' \\ &\vdots \\ m_k &= \mu_k' \end{aligned}$$

for $(\theta_1, \ldots, \theta_k)$ in terms of $(m_1, \ldots, m_k)$. Namely, we let the first k-th true moments match the first k-th sample moments.

By the weak law of large numbers, if the first k-th moments are finite, we have $m_l \to \mu_l'$ in probability for all $1 \le l \le k$. Therefore the method of moments should yield *consistent* (we will discuss consistency later in this chapter) estimators of $\theta_1, \ldots, \theta_k$ under very mild assumptions.

Example 3.1 ($N(\theta, \sigma^2)$)**.** Suppose $X_1, \ldots, X_n$ are i.i.d. samples from $N(\theta, \sigma^2)$. The system of equations is given by

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \theta,$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \theta^2 + \sigma^2.$$

We find $\tilde{\theta} = \overline{X}$ and

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Note that $\widetilde{\sigma^2}$ is different from the sample variance S^2.

Example 3.2 (Binomial (N, p))**.** We assume both N, p are unknown. Let $X_1, \ldots, X_n$ sampled from Binomial (N, p). The system of equations is given by

$$\overline{X} = Np$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = Np(1-p) + N^2p^2.$$

Solving these two equations we obtain

$$\tilde{N} = \frac{\overline{X}^2}{\overline{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}, \quad \tilde{p} = \frac{\overline{X}}{\tilde{N}}.$$

Example 3.3 (Uniform $(0, \theta)$)**.** Let $X_1, \ldots, X_n$ be i.i.d. samples from Uniform$(0, \theta)$. The first-moment equation is

$$\frac{1}{n}X_i = \frac{\theta}{2}.$$

Therefore the MoM estimator is $\tilde{\theta} = \frac{2}{n}\sum_{i=1}^{n} X_i$.

By the Law of Large Numbers, we know $\tilde{\theta} \to \theta$ in probability as $n \to \infty$. But this estimator is different from what we would choose based on sufficiency in Example 2.6.

Remark 3.4. The method of moments has been applied for Gaussian mixture and latent variable models; see [3, 9, 2].

3.2 Maximum likelihood estimators (MLE)

The maximum likelihood estimator (MLE) is one of the most widely used approaches. Its core idea is straightforward: given samples $(X_1, \dots, X_n)$ we choose the value of θ that maximizes the probability of observing those specific samples.

Definition 3.5 (Likelihood function)**.** Let $f(\mathbf{x}|\theta)$ denote the joint density of the sample $\mathbf{X} = (X_1, \dots, X_n)$. Then given $\mathbf{X} = \mathbf{x}$ is observed,

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

is called the likelihood function.

For each sample point $\mathbf{x}$, the maximum likelihood estimator $\hat{\theta}(\mathbf{x})$ is a parameter value at which $L(\theta|\mathbf{x})$ is maximized as a function of θ with $\mathbf{x}$ fixed.

If the likelihood function is differentiable in θ_i, possible candidates for the MLE are the values of $\theta = (\theta_1, \dots, \theta_k)$ such that

$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, \quad 1 \le i \le k.$$

This is only a first-order condition, but the MLE is the global maximizer of the likelihood function.

Since $\log(x)$ is monotone, it's often convenient to find the maximum of the log-likelihood function $\log L(\theta|\mathbf{x})$.

{example:Gau

Example 3.6 (MLE for normal distribution)**.** Let $X_1, \dots, X_n$ be i.i.d. $N(\theta, 1)$. Let $L(\theta|\mathbf{x})$ be the likelihood function. Then

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right).$$

The log-likelihood is

$$\log L(\theta|\mathbf{x}) = c_n - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2,$$

which is maximized at $\theta = \overline{x}$. By checking the first order condition, we find $\sum_{i=1}^{n} x_i - \theta = 0$ and we can check the second derivative

$$\frac{d^2}{d\theta^2} \log L(\theta|\mathbf{x}) = -n < 0.$$

Therefore $\hat{\theta} = \overline{x}$ is the MLE for θ.

Note that the MoM estimator and the MLE is the same for this example.

Example 3.7 (MLE for Bernoulli)**.** Let $X_1, \dots, X_n \sim \mathrm{Ber}(p)$. The likelihood function is

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^y(1-p)^{n-y}, \quad y = \sum_{i=1}^{n} x_i.$$

The log-likelihood is

$$\log L(p|\mathbf{x}) = y \log p + (n-y)\log(1-p).$$

If $y \in (0,n)$, we can check the global maximum is $\hat{p} = y/n$.

If $y = 0$, $\log L(p|\mathbf{x}) = n\log(1-p)$ with $\hat{p} = 0$. If $y = n$, similarly, $\hat{p} = 1$. Thus we have shown that $\overline{x}$ is the MLE for p for all $p \in [0,1]$.

Remark 3.8. In this example, we assume $p \in [0,1]$. If we assume $p \in (0,1)$, then $\hat{p} = y/n$ cannot be the MLE when $y = 0, n$.

MLE for regressions

Example 3.9 (Linear regression)**.** Given data points $(\mathbf{x}_i, y_i), 1 \le i \le n$, where $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$. Suppose $y_i = \langle \mathbf{x}_i, \theta\rangle + \varepsilon_i$, where ε_i are i.i.d. $N(0,\sigma^2)$. Based on the the observations $(\mathbf{x}_i, y_i), 1 \le i \le n$, we would like to derive the MLE for $\theta \in \mathbb{R}^d$. Since $\varepsilon_i = y_i - \langle \mathbf{x}_i, \theta\rangle$, the likelihood function is given by

$$L(\theta|\mathbf{x}_1,\dots,\mathbf{x}_n, y_1,\dots,y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i - \langle\theta,\mathbf{x}_i\rangle)^2}{2\sigma^2}\right)$$

To find the MLE, it's equivalent to maximize

$$\max_\theta -\sum_{i=1}^n (y_i - \langle\theta,\mathbf{x}_i\rangle)^2,$$

or equivalently,

$$\min_\theta \sum_{i=1}^n (y_i - \langle\theta,\mathbf{x}_i\rangle)^2. \tag{12}$$

{eq:least_sq

Let $\mathbf{y} = (y_1,\dots,y_n)^\top, \mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top\end{bmatrix} \in \mathbb{R}^{n\times d}$. Then we can write the problem (12) as

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{X}\theta - \mathbf{y}\|_2^2,$$

which is a linear least-squares regression problem. Therefore we have derived that under i.i.d. Gaussian noise assumption, the MLE estimator for linear regression is the same as the least square solution in (12).

Example 3.10 (Logistic regression)**.** Given data points $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{0,1\}$. Assume y_i is a Bernoulli random variable with parameter p_i, and y_i are all independent. We have

$$f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = \exp\left(y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right).$$

Take the canonical parameter $\eta_i = \log(\frac{p_i}{1-p_i})$, we have $p_i = \frac{1}{1+\exp(-\eta_i)}$. We can write the canonical form for the exponential family as

$$f(y_i|\eta_i) = \exp(y_i\eta_i - \log(1 + \exp \eta_i)),$$

where the natural parameter space is $\eta_i \in \mathbb{R}$. If we make the assumption $\eta_i = \langle \mathbf{x}_i, \theta\rangle$. Then the likelihood function of θ is given by

$$L(\theta|x_1, \dots, x_n, y_1, \dots, y_n) = \exp\left(\sum_{i=1}^{n} y_i\langle \mathbf{x}_i, \theta\rangle - \log(1 + \exp(\langle \mathbf{x}_i, \theta\rangle)\right).$$

Therefore the MLE for θ is the minimum of the following expression:

$$\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n} (\log(1 + \exp(\langle \mathbf{x}_i, \theta\rangle) - y_i\langle \mathbf{x}_i, \theta\rangle). \tag{13}$$

{eq:Logistic

In the matrix form, the optimization problem can be written as

$$\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n} \log(1 + \exp(\langle \mathbf{x}_i, \theta\rangle)) - \mathbf{y}^\top \mathbf{X}\theta.$$

Recall $p_i = \frac{1}{1+\exp(-\langle\theta, x_i\rangle)}$. We can rewrite (13) as

$$\min_{\theta} -\sum_{i=1}^{n} y_i \log p_i + (1 - y_i) \log(1 - p_i), \tag{14}$$

{eq:cross_en

which is the sum of the cross-entropy loss

$$H(y_i, p_i) = -y_i \log p_i - (1 - y_i) \log(1 - p_i)$$

commonly used in binary classification problems. Therefore under the model assumptions that $y_i \sim \text{Ber}(p_i)$ where $p_i = \frac{1}{1+\exp(-\langle \mathbf{x}_i, \theta\rangle)}$, the MLE for θ is the solution for the minimizer of the cross-entropy loss.

Remark 3.11. In machine learning terms, the minimization problems (12), (14) are called *empirical risk minimization* (ERM), often solved by gradient-based methods numerically. Linear regression and logistic regression are examples of generalized linear models, where $\mathbb{E}[y_i|\mathbf{x}_i] = f(\langle \mathbf{x}_i, \theta\rangle)$ for some function f.

Remark 3.12. MLE can be used in other contexts beyond parameter estimation. In unsupervised learning problems, MLE can be the "optimal" algorithm statistically but might not be computationally feasible (even NP hard). We often look for other computationally efficient methods (e.g., spectral, semidefinite programming, gradient descent) to achieve the performance of the MLE asymptotically; see e.g., [1].

Example 3.13 (Restricted rangle MLE)**.** Let $X_1, \dots, X_n$ be i.i.d. $N(\theta, 1)$ where $\theta \geq 0$. In this case, we are looking for

$$\max_{\theta\geq 0} L(\theta|\mathbf{x}) = \max_{\theta\geq 0} \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

We have

$$\frac{d}{d\theta} \log L(\theta|\mathbf{x}) = \sum_{i=1}^{n} (x_i - \theta).$$

When $\overline{x} < 0$, the log-likelihood is decreasing for $\theta \geq 0$, hence $\hat{\theta}_{MLE} = 0$ if $\overline{X} < 0$. When $\overline{x} > 0$, following the computation in Example 3.6 we obtain $\hat{\theta}_{MLE} = \overline{x}$.

Example 3.14 (MLE for Binomal(N, p) with unknown N)**.** Let $X_1, \ldots, X_n$ be i.i.d. samples from Binomal(N, p) with p known and N is unknown. The likelihood function is

$$L(N|\mathbf{x}, p) = \prod_{i=1}^{n} \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}.$$

Since the $L(N|\mathbf{x}, p)$ only takes integer values, we cannot differentiate with respect to N.

The MLE should be an integer $N \geq \max_i x_i$ which satisfies $L(N|\mathbf{x}, p) \geq L(N+1|\mathbf{x}, p)$ and $L(N|\mathbf{x}, p) \geq L(N-1|\mathbf{x}, p)$. This gives some constraints but one has to solve for $\hat{N}_{MLE}$ numerically.

Invariance property of MLEs If $\tau(\theta)$ is a one-to-one map of θ. Then $\tau(\hat{\theta}_{MLE})$ is the MLE of $\tau(\theta)$. This can be seen by defining

$$L^*(\eta|\mathbf{x}) := f(\mathbf{x}|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|\mathbf{x}).$$

Then

$$\sup_{\eta} L^*(\eta|\mathbf{x}) = \sup_{\eta} L(\tau^{-1}(\eta)|\mathbf{x}) = \sup_{\theta} L(\theta|\mathbf{x}).$$

Then the maximum of $L^*(\eta|\mathbf{x})$ is obtained at $\tau^{-1}(\eta) = \hat{\theta}$, which is $\eta = \tau(\hat{\theta})$.

If τ is not one-to-one, for any given value η, there might be more than one values of θ such that $\tau(\theta) = \eta$. The following definition is needed.

Definition 3.15 (induced likelihood function)**.** Define the induced likelihood function L^* given by

$$L^*(\eta|\mathbf{x}) = \sup_{\{\theta : \tau(\theta) = \eta\}} L(\theta|x).$$

The value $\hat{\eta}$ which maximizes $L^*(\eta|\mathbf{x})$ will be called the MLE for $\eta = \tau(\theta)$.

Theorem 3.16 (Invariance property of MLEs)**.** *If $\hat{\theta}$ is the MLE for θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.*

Proof. Let $\hat{\eta} = \arg\max L^*(\eta|\mathbf{x})$. Then

$$\begin{aligned} L^*(\hat{\eta}|\mathbf{x}) = \sup_{\eta} L^*(\eta|\mathbf{x}) &= \sup_{\eta} \sup_{\{\theta : \tau(\theta) = \eta\}} L(\theta|x) \\ &= \sup_{\theta} L(\theta|\mathbf{x}) \\ &= L(\hat{\theta}|\mathbf{x}). \end{aligned}$$

Moreover, since $\hat{\theta}$ is the MLE,

$$\begin{aligned} L(\hat{\theta}|\mathbf{x}) &= \sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta})\}} L(\theta|\mathbf{x}) \\ &= L^*(\tau(\hat{\theta})|\mathbf{x}). \end{aligned}$$

We obtain

$$L^*(\hat{\eta}|\mathbf{x}) = L^*(\tau(\hat{\theta})|\mathbf{x}),$$

which means $\tau(\hat{\theta})$ maximizes $L^*(\eta|\mathbf{x})$. □

Example 3.17 (Normal MLEs, μ, σ^2 unknown)**.** Let $X_1, \ldots, X_n$ be i.i.d. samples from $N(\theta, \sigma^2)$. Then

$$\log L(\theta, \sigma^2|\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2.$$

The partial derivatives w.r.t. θ and σ^2 are

$$\begin{aligned} \frac{\partial}{\partial\theta}\log(L(\theta, \sigma^2|\mathbf{x})) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta) \\ \frac{\partial}{\partial\sigma^2}\log(L(\theta, \sigma^2|\mathbf{x})) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \theta)^2 \end{aligned}$$

Setting these partial derivatives equal to 0 and solving yields the solution $\hat{\theta} = \bar{\mathbf{x}}$ and

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2.$$

We remain to verify this solution is a global maximum. Since $\forall\theta \neq \bar{x}$,

$$\sum_{i=1}^{n}(x_i - \theta)^2 > \sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2,$$

for any value of σ^2

$$\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}{2\sigma^2}\right) \geq \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{\sum_{i=1}^{n}(x_i - \theta)^2}{2\sigma^2}\right).$$

Therefore, the problem reduces to a one-dimensional problem for σ^2 only, which achieves a global maximum at

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2.$$

Therefore, $(\bar{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2)$ are the MLEs.

Remark 3.18. Recall a matrix A is positive definite if $A = A^\top$ and all eigenvalues of A are positive. Alternatively, we can check the second order optimality condition to obtain the MLE, namely the Hessian matrix of $-\log L(\theta, \sigma^2|\mathbf{x})$ is positive definite at $(\hat{\mu}, \widehat{\sigma^2})$.

3.3 Bayes Estimators

Recall our setting for point estimation is: given i.i.d. samples $X_1, \dots, X_n$ from a parametric family $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter. In the classical approach, θ is a fixed unknown parameter (Frequentist approach).

In the Bayesian framework, we consider a parameter $\theta \in \Theta$ that we wish to estimate based on observed data $\mathbf{x}$. We start by assigning a *prior* distribution $\pi(\theta)$ to express our initial beliefs about θ. Given θ, the data are generated through the density function $f(\mathbf{x} \mid \theta)$. The core of Bayesian inference is Bayes' theorem: the *posterior* distribution of θ given the data is

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

where $m(\mathbf{x})$ is the marginal distribution

$$m(\mathbf{x}) = \int_\Theta f(\mathbf{x} \mid \theta)\, \pi(\theta)\, d\theta.$$

The posterior distribution, $\pi(\theta \mid \mathbf{x})$, represents our updated belief about θ after taking the data $\mathbf{x}$ into account.

Definition 3.19. A random variable X is said to follow a *Beta distribution* with parameters $\alpha > 0$ and $\beta > 0$, denoted by

$$X \sim \text{Beta}(\alpha, \beta),$$

if its probability density function is given by

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha - 1}(1 - x)^{\beta - 1}, \quad 0 < x < 1,$$

where $\Gamma(\cdot)$ denotes the Gamma function such that

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha - 1} e^{-t} dt.$$

Alternatively, using the Beta function,

$$B(\alpha, \beta) = \int_0^1 t^{\alpha - 1}(1 - t)^{\beta - 1}\, dt,$$

the pdf can be written as

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\, x^{\alpha - 1}(1 - x)^{\beta - 1}, \quad 0 < x < 1.$$

The mean of a Beta distribution is $\mathbb{E}X = \frac{\alpha}{\alpha + \beta}$.

{example:Ber

Example 3.20 (Binomial Bayes estimation)**.** Suppose we wish to estimate the probability of success θ in a series of Bernoulli trials. Assume we observe $x = \sum_{i=1}^n \mathbf{1}\{x_i = 1\}$ successes in n independent trials. The likelihood function for θ is given by the Binomial distribution:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n - x}.$$

A natural choice for the prior distribution is the Beta distribution:

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

where $\alpha > 0$ and $\beta > 0$ are **hyperparameters** that express our prior beliefs about θ. Note that Beta$(1, 1)$ is a uniform prior, where we assume θ is Uniform$(0, 1)$.

Using Bayes' theorem, the posterior distribution is proportional to the product of the likelihood and the prior:

$$p(\theta \mid x) = \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{x}(1-\theta)^{n-x}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} = \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.$$

Since

$$p(\theta \mid x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1},$$

the posterior distribution is also a Beta distribution with updated parameters:

$$\theta \mid x \sim \text{Beta}(\alpha + x, \beta + n - x).$$

Imagine the data $x_1, \dots, x_n$ arrive sequentially, then for each $1 \le i \le n$, after observing x_i, we obtain an updated posterior distribution of θ depending on $x_1, \dots, x_i$.

A natural estimate for θ is the **mean of the posterior distribution**, which is

$$\hat{\theta}_B = \frac{\alpha + x}{\alpha + \beta + n}.$$

The prior distribution $\pi(\theta)$ has mean $\frac{\alpha}{\alpha+\beta}$. After seeing the data $X_1, \dots, X_n \sim$ Ber(θ), we updated our estimate on θ by a convex combination of $\frac{\alpha}{\alpha+\beta}$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Namely,

$$\hat{\theta}_B = \frac{n}{\alpha+\beta+n}\overline{x} + \frac{\alpha+\beta}{\alpha+\beta+n}\cdot\frac{\alpha}{\alpha+\beta}.$$

For the uniform prior, we find

$$\hat{\theta}_B = \frac{n}{n+2}\overline{x} + \frac{1}{n+2}.$$

As $n \to \infty$, we can see $\hat{\theta}_B \approx \overline{x}$ for any choices of α, β. So the prior distribution has an effect only for finite n.

But why we assume the prior distribution $\pi(\theta) \sim$ Beta(α, β)? The Beta distribution makes the update on the posterior distribution easier. It has a closed form expression on $\pi(\theta|x)$ such that $\pi(\theta)$ and $\pi(\theta|x)$ are in the same parametric family. Such a prior distribution is called a *conjugacy family.*

Definition 3.21 (Conjugacy family)**.** Let $\mathcal{F}$ be the class of pdfs $f(x|\theta)$. A class Π of prior distributions is a conjugacy family for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all prior in Π, and all $x \in \mathcal{X}$.

Example 3.22 (Conjugacy family for Poisson)**.** Suppose $X_1, \dots, X_n$ are i.i.d. samples from Poisson(λ) with pmf

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots,$$

The conjugate family for $f(\mathbf{x}|\lambda)$ is the Gamma distribution: $\lambda \sim \text{Gamma}(\alpha, \beta)$.

A random variable X follows a Gamma distribution with shape parameter α and rate parameter β if its probability density function (PDF) is given by:

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0, \quad \alpha, \beta > 0$$

Since

$$\begin{aligned}\pi(\lambda|\mathbf{x}) &\propto f(\mathbf{x}|\lambda)\pi(\lambda) \\ &\propto \lambda^{x_1+\cdots+x_n} \exp(-n\lambda)\lambda^{\alpha-1} \exp(-\beta\lambda) \\ &\propto \lambda^{\alpha+\sum_{i=1}^n x_i - 1} \exp(-(\beta+n)\lambda),\end{aligned}$$

the posterior distribution is also Gamma:

$$\lambda|\mathbf{x} \sim \text{Gamma}(\alpha + \sum x_i, \beta + n),$$

with a posterior mean estimator

$$\hat{\lambda} = \frac{\sum_{i=1}^n x_i + \alpha}{\beta + n}.$$

{example:Gau

Example 3.23. Consider a normal distribution with known variance such that $X_1, \dots, X_n$ are i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$, A conjugate prior for μ in this setting is a normal distribution $N(\mu_0, \tau^2)$:

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\tau^2}\right),$$

Since

$$\begin{aligned}\pi(\mu|\mathbf{x}) &\propto f(\mathbf{x}|\mu)\pi(\mu) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right) \exp\left(-\frac{1}{2\tau^2}(\mu-\mu_0)^2\right),\end{aligned}$$

we find the posterior distribution remains normal:

$$\mu|\mathbf{x} \sim \mathcal{N}\left(\frac{\frac{\mu_0}{\tau^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right).$$

Using the posterior mean, we find the Bayes estimator of μ is

$$\hat{\mu}_B = \frac{\frac{\mu_0}{\tau^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}.$$

$\hat{\mu}_B$ is again a convex combination of the prior and sample means. As n increases, the dependence on the prior mean becomes weaker. Moreover, the posterior distribution has a decreasing variance that goes to zero, and $\mu|\mathbf{x}$ becomes a Gaussian variable highly concentrated around the mean.

Conjugacy family for the exponential family A probability distribution in the **exponential family** can be written in the canonical form:

$$f(x|\theta) = h(x) \exp\left(\eta(\theta)^\top T(x) - A(\theta)\right).$$

A *conjugate prior* for an exponential family likelihood takes the form:

$$\pi(\theta|\xi, \nu) = f(\xi, \nu) \exp\left(\eta(\theta)^\top \xi - \nu A(\theta)\right),$$

where:

- ξ is the prior's sufficient statistic,
- ν is a scaling parameter, often interpreted as prior sample size.
- $f(\xi, \nu)$ is a normalization constant.

Given n i.i.d. observations $x_1, x_2, \ldots, x_n$, the joint density is:

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} h(x_i) \exp\left(\eta(\theta)^\top T(x_i) - A(\theta)\right).$$

Multiplying this likelihood with the prior:

$$\pi(\theta|\xi, \nu) \propto \exp\left(\eta(\theta)\xi - \nu A(\theta)\right),$$

we find the posterior distribution is:

$$\pi(\theta|x_1, \ldots, x_n, \xi, \nu) \propto \exp\left(\eta(\theta)^\top \left(\xi + \sum_{i=1}^{n} T(x_i)\right) - (\nu + n)A(\theta)\right).$$

Thus, the updated posterior parameters are:

- Updated sufficient statistic: $\xi' = \xi + \sum_{i=1}^{n} T(x_i)$,
- Updated scale parameter: $\nu' = \nu + n$.

Example 3.24 (Multinomial distribution with Dirichlet prior)**.** The **Multinomial distribution** is a generalization of the Binomial distribution to more than two categories. Suppose we have an experiment with K possible outcomes, and each outcome has a probability $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$ where:

$$\sum_{i=1}^{K} \theta_i = 1, \quad \theta_i \geq 0.$$

If we conduct N independent trials, where each trial results in exactly one of the K possible outcomes, the probability mass function (PMF) of the Multinomial distribution is:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K|\theta) = \frac{N!}{x_1! x_2! \cdots x_K!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_K^{x_K},$$

where x_i is the number of occurrences of outcome i, and $\sum_{i=1}^K x_i = N$.

The **Dirichlet distribution** is the conjugate prior for the Multinomial distribution. If we assume a Dirichlet prior for θ with parameters $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$ with $\alpha_i > 0, 1 \leq i \leq K$, then the prior distribution is given by:

$$\pi(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1}, \quad \text{for } \sum_{i=1}^K \theta_i = 1, \theta_i \geq 0.$$

The support of $\pi(\theta|\alpha)$ is called a standard $k-1$ simplex:

$$\left\{ (x_1, \dots, x_K) : \sum_{i=1}^K x_i = 1, x_i \in [0,1], 1 \leq i \leq k \right\}.$$

Given observed data $X = (x_1, x_2, \dots, x_K)$, we update our belief about θ using Bayes' theorem:

$$\pi(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta)\pi(\theta).$$

Substituting the Multinomial likelihood and the Dirichlet prior:

$$\pi(\theta|\mathbf{x}) \propto \left[\prod_{i=1}^K \theta_i^{x_i} \right] \times \left[\prod_{i=1}^K \theta_i^{\alpha_i - 1} \right] = \prod_{i=1}^K \theta_i^{x_i + \alpha_i - 1}.$$

Thus, the posterior distribution remains a Dirichlet distribution with updated parameters:

$$\theta|X \sim \text{Dirichlet}(\alpha_1 + x_1, \alpha_2 + x_2, \dots, \alpha_K + x_K).$$

The expected value of θ_i under the posterior distribution is:

$$E[\theta_i|\mathbf{x}] = \frac{\alpha_i + x_i}{\sum_{j=1}^K (\alpha_j + x_j)}.$$

Maximum a Posteriori Estimation (MAP) Another common estimator of θ is the MAP; it relies on the same principle as the MLE – we choose the one that is the most likely. Here the 'likely' is interpreted as our posterior belief about the parameter of interest θ. Formally, MAP is defined as

$$\hat{\theta}_{MAP} = \arg\max_\theta \pi(\theta|X_1, \cdots, X_n).$$

Equivalently, $\hat{\theta}_{MAP}$ outputs the **mode** of the posterior distribution (the point θ where the density of $\pi(\theta, X_1, \dots, X_n)$ is maximized.

Recall the posterior distribution satisfies

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta) = L(\theta|\mathbf{x})\pi(\theta).$$

We can write

$$\hat{\theta}_{MAP} = \arg\max_\theta L(\theta|\mathbf{x})\pi(\theta).$$

The MAP estimator differs from MLE, where the function we are trying to maximize is reweighted by the prior distribution $\pi(\theta)$. MAP estimation can therefore be seen as a regularization of the MLE.

Example 3.25 (MAP for Binomial distribution)**.** Assume that we have an observation $Y \sim \text{Bin}(N, \theta)$ where N is known and the parameter of interest is θ:

$$P(Y = y|\theta) = \binom{N}{y} \theta^y (1-\theta)^{N-y}.$$

We use a Beta distribution with parameters (α, β) as our prior distribution for θ. Namely,

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

The posterior distribution is proportional to:

$$\pi(\theta|Y) \propto \theta^{y+\alpha-1}(1-\theta)^{N-y+\beta-1}.$$

Thus, the posterior follows a Beta distribution:

$$\theta|Y \sim \text{Beta}(\alpha + y, \beta + N - y).$$

So it is a Beta distribution with parameters $(Y + \alpha, N - Y + \beta)$. Then the posterior mean and MAP are:

$$\hat{\theta}_\pi = \frac{Y+\alpha}{N+\alpha+\beta}, \quad \hat{\theta}_{MAP} = \frac{Y+\alpha-1}{N+\alpha+\beta-2}.$$

These are the mean and the mode of a Beta distribution. We have computed

$$\hat{\theta}_{MLE} = \frac{Y}{N}.$$

The three methods output different estimators.

Example 3.26 (MAP for Gaussian)**.** Consider a normal distribution with known variance such that $X_1, \ldots, X_n$ are i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$, A conjugate prior for μ in this setting is a normal distribution $N(\mu_0, \tau^2)$:

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\tau^2}\right),$$

Since the posterior distribution remains normal:

$$\mu|\mathbf{x} \sim \mathcal{N}\left(\frac{\frac{\mu_0}{\tau^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right).$$

and the mean and mode of a Gaussian is the same, we immediately get

$$\hat{\mu}_{MAP} = \frac{\frac{\mu_0}{\tau^2} + \frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}.$$

4 Evaluating estimators

In this section, we will discuss some basic criteria for evaluating estimators.

4.1 Mean squared error

In statistics, the *mean squared error* (MSE) is a central metric for assessing the quality of an estimator. MSE encapsulates both the *accuracy* and the *precision* of an estimator $\hat{\theta}$ in a single measure.

Definition 4.1. The MSE of an estimator $\hat{\theta}$ for a parameter θ is

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big].$$

This expectation is taken over the distribution of the random sample (i.e., the distribution of $X_1, \ldots, X_n$) from which $\hat{\theta}$ is constructed.

Decomposition of MSE An important result in estimation theory is that the MSE can be decomposed into the variance and the squared bias of the estimator:

$$\mathrm{MSE}(\hat{\theta}) = \underbrace{\mathrm{Var}(\hat{\theta})}_{\text{variance term}} + \underbrace{\big[\mathbb{E}(\hat{\theta}) - \theta\big]^2}_{\text{bias term}}.$$

Bias The bias of an estimator $\hat{\theta}$ is defined as

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

An estimator is said to be **unbiased** if $\mathbb{E}(\hat{\theta}) = \theta$, in which case $\mathrm{Bias}(\hat{\theta}) = 0$.

For example, the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator for the population mean μ, i.e., $\mathbb{E}(\overline{X}) = \mu$.

Variance The variance of $\hat{\theta}$ measures how spread out the estimator $\hat{\theta}$ is around its mean:

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\big].$$

Lower variance typically implies that the estimator is more *consistent* (less spread out) around its expected value. However, a low-variance estimator might be biased, illustrating a fundamental tradeoff.

{example:MSE

Example 4.2 (MSE for normal distribution)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution $N(\mu, \sigma^2)$. Then $\overline{\mathbf{X}}$ and S^2 are unbiased estimator since

$$\mathbb{E}\overline{\mathbf{X}} = \mu, \quad \mathbb{E}S^2 = \frac{1}{n-1}\mathbb{E}\sum_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2 = \sigma^2.$$

The mean squared error of the estimators are given by

$$\mathbb{E}(\overline{\mathbf{X}} - \mu)^2 = \mathrm{Var}(\overline{\mathbf{X}}) = \frac{\sigma^2}{n}$$
$$\mathbb{E}(S^2 - \sigma^2)^2 = \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

First, recall that for $X_1, \ldots, X_n$ i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, the sample variance S^2 satisfies

$$(n-1)\,\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

See Theorem 5.3.1 in [6].

The variance of a chi-square random variable with k degrees of freedom is $2k$. Hence,

$$\mathrm{Var}\Big((n-1)\,\tfrac{S^2}{\sigma^2}\Big) = 2\,(n-1).$$

We have

$$\mathrm{Var}(S^2) = \frac{2\,\sigma^4}{n-1}.$$

Bias–Variance Tradeoff The decomposition

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \big[\mathrm{Bias}(\hat{\theta})\big]^2$$

shows that for a small MSE, two factors must be considered simultaneously. Balancing these two quantities is known as the *bias–variance tradeoff*.

Example 4.3 (A biased estimator with lower variance)**.** Consider estimating the population mean μ by $\hat{\theta} = (1-c)\,\overline{X}$, where c is a constant and $\overline{X}$ is the sample mean. For simplicity, assume the X_i are i.i.d. with variance σ^2. Then:

$$\mathbb{E}(\hat{\theta}) = (1-c)\mu,$$

which implies

$$\mathrm{Bias}(\hat{\theta}) = -c\mu.$$

The variance of $\hat{\theta}$ is

$$\mathrm{Var}(\hat{\theta}) = (1-c)^2\,\mathrm{Var}(\overline{X}) = (1-c)^2\frac{\sigma^2}{n}.$$

As we adjust c, we see:

- $\mathrm{Bias}(\hat{\theta})$ can be made large (in absolute value) if c is far from 0.
- $\mathrm{Var}(\hat{\theta})$ decreases as $|1-c|$ gets smaller.

Hence, you could choose c to reduce variance at the cost of introducing a bias. The MSE of this estimator becomes:

$$\mathrm{MSE}(\hat{\theta}) = (1-c)^2\frac{\sigma^2}{n} + \big(c\mu\big)^2.$$

Minimizing this MSE involves balancing the squared bias term $\big(c\mu\big)^2$ and the variance term $(1-c)^2\frac{\sigma^2}{n}$.

Remark 4.4. The bias-variance trade-off is an idea beyond the setting of point estimation. See more modern analysis of bias-variance trade-off in machine learning [4, 7].

Example 4.5 (MLE for $N(\mu, \sigma^2)$)**.** Recall for $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, the MLE for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2 = \frac{n-1}{n}S^2.$$

The MLE is biased, since $\mathbb{E}\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2$. The variance can also be calculated

$$\mathrm{Var}\hat{\sigma}^2 = \frac{2(n-1)\sigma^4}{n^2}.$$

We thus have the MSE of $\hat{\sigma}^2$ is

$$\mathbb{E}(\hat{\sigma}^2 - \sigma^2)^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4 < \frac{2}{n-1}\sigma^4.$$

This shows $\hat{\sigma}^2$ has a smaller MSE than S^2.

Example 4.6 (MSE for Binomial Bayes estimator)**.** Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{Ber}(p)$. The MSE of the MLE $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is

$$\mathbb{E}_p(\hat{p} - p)^2 = \mathrm{Var}_p\overline{X} = \frac{p(1-p)}{n}.$$

In Example 3.20, the posterior mean estimator is $\hat{p}_B = \frac{Y+\alpha}{\alpha+\beta+n}$. The MSE for $\hat{p}_B$ is

$$\mathbb{E}_p(\hat{p}_B - p)^2 = \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2.$$

This suggests in terms of MSE, the comparison of the two estimators depends on the range of p and α, β. Choosing $\alpha = \beta = \sqrt{n/4}$, we obtain

$$\mathbb{E}_p(\hat{p}_B - p)^2 = \frac{n}{4(n+\sqrt{n})^2},$$

which is a constant independent of p. When p is close to $1/2$ and n is sufficiently large, $\overline{X}$ has a larger MSE.

4.2 Best unbiased estimators

Definition 4.7. An estimator W^* is a *best unbiased estimator* of $\tau(\theta)$ if $\mathbb{E}_\theta W^* = \tau(\theta)$ for all $\theta \in \Theta$ and for any other estimator W with $\mathbb{E}_\theta W = \tau(\theta)$, we have

$$\mathrm{Var}_\theta(W^*) \leq \mathrm{Var}_\theta(W)$$

for all θ. Then W^* is called a uniform minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$.

Example 4.8 (Unbiased estimator for Poisson)**.** Let $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$ be i.i.d. samples. We have $\mathbb{E}X_i = \text{Var}(X_i) = \lambda$. Therefore

$$\mathbb{E}_\lambda \overline{X} = \mathbb{E}_\lambda S^2 = \lambda$$

for all $\lambda > 0$. Then both $\overline{X}$ and S^2 are unbiased estimators for λ. For any constant $a > 0$,

$$a\overline{X} + (1-a)S^2$$

is also unbiased. We have $\text{Var}_\lambda(\overline{X}) = \frac{\lambda}{n}$. The variance of S^2 is not easy to obtain but one can check $\text{Var}_\lambda S^2 \geq \text{Var}_\lambda \overline{X}$.

Example 4.9 (Unique unbiased estimator for Gaussian)**.** Let $X \sim N(\theta, 1)$. $W(X) = X$ is an unbiased estimator for θ. For any other unbiased estimator $\delta(X)$, since $\mathbb{E}\delta(X) = \theta$, we have

$$\mathbb{E}[\delta(X) - X] = 0.$$

Since $N(\theta, 1)$ is an exponential family, by the open set condition in Theorem 2.32, $W(X) = X$ is a complete sufficient statistic. Let $g(X) = \delta(X) - X$. From the definition of complete statistics (Definition 2.25), $\mathbb{E}g(X) = 0$ implies

$$\mathbb{P}_\theta(\delta(X) - X = 0) = 1.$$

This implies X is the unique unbiased estimator of θ.

The Cramér-Rao Inequality provides a lower bound for a class of estimators with certain regularity conditions.

{thm:CR}

Theorem 4.10 (Cramér-Rao Inequality)**.** *Let $X_1, \dots, X_n$ be a sample with joint pdf $f(\mathbf{x}|\theta)$ and let $W(\mathbf{X}) = W(X_1, \dots, X_n)$ be any estimator satisfying*

$$\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X}) = \int_{\mathcal{X}} W(\mathbf{x}) \frac{\partial}{\partial\theta}[f(\mathbf{x}|\theta)]d\mathbf{x} \tag{15}$$

{eq:fubini}

and $\text{Var}_\theta W(\mathbf{X}) < \infty$. Then

$$\text{Var}_\theta W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X})\right)^2}{\mathbb{E}_\theta\left((\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta))^2\right)}.$$

Proof. We will use the following covariance inequality: for any two random variables X, Y,

$$(\text{Cov}(X,Y))^2 \leq (\text{Var}X)(\text{Var}Y).$$

This implies a lower bound on $\text{Var}(X)$:

$$\text{Var}(X) \geq \frac{(\text{Cov}(X,Y))^2}{\text{Var}(Y)}.$$

Define the *score function* by

$$S(\mathbf{X};\theta) = \frac{\partial}{\partial\theta}\log f(\mathbf{X} \mid \theta).$$

Note that

$$\mathbb{E}_\theta\big[S(\mathbf{X};\theta)\big] = \int S(\mathbf{x};\theta)\, f(\mathbf{x}\mid\theta)\, dx = \int \frac{\partial}{\partial\theta} f(\mathbf{x}\mid\theta)\, dx = 0, \tag{16}$$ {eq:score}

where the last identity is due to (15) by taking $W(\mathbf{x}) = 1$. Hence,

$$\mathrm{Var}_\theta\big[S(\mathbf{X};\theta)\big] = \mathbb{E}_\theta\big[S(\mathbf{X};\theta)^2\big].$$

By interchanging differentiation and expectation from (15),

$$\begin{aligned}\frac{d}{d\theta}\,\mathbb{E}_\theta\big[W(\mathbf{X})\big] &= \int_{\mathcal{X}} W(\mathbf{x})\frac{\frac{\partial}{\partial\theta}[f(\mathbf{x}|\theta)]}{f(\mathbf{x}|\theta)} f(\mathbf{x}|\theta) d\mathbf{x}\\ &= \mathbb{E}_\theta\left[W(\mathbf{X})\frac{\frac{\partial}{\partial\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right]\\ &= \mathbb{E}_\theta\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right]\\ &= \mathbb{E}_\theta[W(\mathbf{X})S(\mathbf{X};\theta)].\end{aligned}$$

Since $\mathrm{Cov}(W(\mathbf{X}), S(\mathbf{X};\theta)) = \mathbb{E}_\theta\big[W(\mathbf{X})\, S(\mathbf{X};\theta)\big]$ and $|\mathrm{Var}(S(\mathbf{X};\theta)) = \mathbb{E}[S(\mathbf{X};\theta)^2]$, we now apply the Cauchy–Schwarz inequality to $W(\mathbf{X})$ and $S(\mathbf{X};\theta)$:

$$\Big(\mathbb{E}_\theta\big[W(\mathbf{X})\, S(\mathbf{X};\theta)\big]\Big)^2 \le \mathrm{Var}_\theta\big[W(\mathbf{X})\big]\mathbb{E}_\theta\big[S(\mathbf{X};\theta)^2\big].$$

Hence,

$$\Big(\frac{d}{d\theta}\,\mathbb{E}_\theta\big[W(\mathbf{X})\big]\Big)^2 = \Big(\mathbb{E}_\theta\big[W(\mathbf{X})\, S(\mathbf{X};\theta)\big]\Big)^2 \le \mathrm{Var}_\theta\big[W(\mathbf{X})\big]\mathbb{E}_\theta\big[S(\mathbf{X};\theta)^2\big].$$

Consequently,

$$\mathrm{Var}_\theta\big[W(\mathbf{X})\big] \ge \frac{\Big(\frac{d}{d\theta}\,\mathbb{E}_\theta\big[W(\mathbf{X})\big]\Big)^2}{\mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\Big)^2\Big]},$$

which completes the proof. □

If $W(\mathbf{X})$ is an unbiased estimator of θ, the Cramér-Rao lower bound is given by

$$\mathrm{Var}_\theta\big[W(\mathbf{X})\big] \ge \frac{1}{\mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\Big)^2\Big]}.$$

If $\mathbb{E}W(\mathbf{X}) \ne \theta$, the lower bound still holds. Denote

$$I_{\mathbf{X}}(\theta) = \mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\Big)^2\Big].$$

This is also called the Fisher information of the sample $\mathbf{X}$. the Fisher information is a way of measuring the amount of information that $\mathbf{X}$ carries about an unknown parameter θ of a distribution that models X. Formally, it is the variance of the score.

Remark 4.11. Score function is a useful concept in diffusion model [10].

{cor:iid}

Corollary 4.12 (Cramér-Rao Inequality for i.i.d. data)**.** Under the same assumptions as in Theorem 4.10, assume $X_1, \dots, X_n$ are i.i.d. with pdf $f(x|\theta)$, then

$$\mathrm{Var}_\theta W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X})\right)^2}{n\mathbb{E}_\theta\left((\frac{\partial}{\partial\theta}\log f(X|\theta))^2\right)}.$$

Proof. It suffices to show

$$\mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\Big)^2\Big] = n\mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\Big)^2\Big].$$

By independence, for $i \neq j$

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right] = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right],$$

And from (16), by the i.i.d. assumption,

$$0 = \mathbb{E}_\theta[\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)] = n\mathbb{E}_\theta[\frac{\partial}{\partial\theta}\log f(X\mid\theta)].$$

Hence for $i \neq j$

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right] = 0.$$

Therefore

$$\begin{aligned}\mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}\log f(\mathbf{X}\mid\theta)\Big)^2\Big] &= \mathbb{E}_\theta\left[\left(\sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right] \\ &= \mathbb{E}_\theta\left[\sum_{i=1}^n\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right] = n\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2.\end{aligned}$$

□

Example 4.13 (CR lower bound for Poisson)**.** Let $X_1, \dots, X_n \sim \mathrm{Poisson}(\lambda)$ be i.i.d. samples. We have

$$I_X(\lambda) = \mathbb{E}\left[\frac{\partial}{\partial\lambda}\log\left(\frac{e^{-\lambda}\lambda^X}{X!}\right)\right]^2 = \frac{1}{\lambda}.$$

For any unbiased estimator W satisfying the regularity condition in Corollary 4.12,

$$\mathrm{Var}_\lambda(W) \geq \frac{1}{nI_{X_1}(\lambda)} = \frac{\lambda}{n}.$$

Remark 4.14. If $f(x|\theta)$ is an exponential family, then the regularity condition in Theorem 4.10 hold; see [8, Theorem 2.4]. So in the case of Poisson, $\overline{X}$ is a best unbiased estimator of λ.

The CR lower bound might not hold if the regularity condition in Theorem 4.10 fails. Here is one example.

{example:sca

Example 4.15 (Unbiased estimator for the scale uniform)**.** Let $X_1, \ldots, X_n$ be i.i.d. with pdf $f(x|\theta) = 1/\theta, x \in (0, \theta)$. We have

$$I_X(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial\theta} \log f(X \mid \theta) \right)^2 \right] = \frac{1}{\theta^2}.$$

The CR lower bound is $\frac{\theta^2}{n}$.

Recall $Y = \max\{X_1, \ldots, X_n\}$ is a sufficient statistic for θ. The pdf of Y is given by

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

Therefore we get $\mathbb{E}Y = \frac{n}{n+1}\theta$, and $\frac{n+1}{n}Y$ is an unbiased estimator of θ. We can compute

$$\mathrm{Var}_\theta(\frac{n+1}{n}Y) = \frac{\theta^2}{n(n+2)},$$

which is below the CR lower bound. This is because the regularity condition (15) does not hold for $f(x|\theta)$.

The next lemma provides a convenient way to compute $I_X(\theta)$:

{lem:Fisher_

Lemma 4.16. *if $f(x|\theta)$ satisfies*

$$\frac{d}{d\theta}\mathbb{E}_\theta \left(\frac{\partial}{\partial\theta} \log f(X|\theta) \right) = \int \frac{\partial}{\partial\theta} \left[\left(\frac{\partial}{\partial\theta} \log f(x|\theta) \right) f(x|\theta) \right] dx,$$

(this condition holds for an exponential family), then

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

{example:CR_

Example 4.17 (CR lower bound for Gaussian)**.** Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be i.i.d. samples where μ is unknown. With Lemma 4.16,

$$I_X(\sigma^2) = -\mathbb{E} \left(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6} \right) = \frac{1}{2\sigma^4}.$$

Then any unbiased estimator W of σ^2 satisfies

$$\mathrm{Var}(W) \geq \frac{2\sigma^4}{n}.$$

From Example 4.2, The variance of sample variance S^2 is $\frac{2\sigma^4}{n-1}$.

This raises the question of whether the CR lower bound is attainable. Since the proof of CR lower bound is an application of Cauchy-Schwartz inequality, we can examine the condition where the Cauchy-Schwartz inequality is tight.

Corollary 4.18 (Attainment)**.** Let $X_1, \dots, X_n$ be i.i.d. random variables with pdf $f(x \mid \theta)$ satisfying the conditions of the Cramér–Rao Theorem. Let

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

denote the likelihood function. If $W(\mathbf{X})$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramér–Rao Lower Bound if and only if

$$a(\theta)\big[\, W(\mathbf{x}) - \tau(\theta)\big] = \frac{\partial}{\partial\theta} \log L(\theta \mid \mathbf{x}) \tag{17}$$

{eq:12.34}

for some function $a(\theta)$.

Proof. The Cramér–Rao Inequality can be written as

$$\Big[\mathrm{Cov}_\theta\big(W(\mathbf{X}), \tfrac{\partial}{\partial\theta} \log \textstyle\prod_{i=1}^n f(X_i \mid \theta)\big)\Big]^2 \le \mathrm{Var}_\theta\big(W(\mathbf{X})\big)\mathrm{Var}_\theta\big(\tfrac{\partial}{\partial\theta} \log \textstyle\prod_{i=1}^n f(X_i \mid \theta)\big).$$

{eq:12.35}

Recalling that $E_\theta W = \tau(\theta)$ and $E_\theta\Big[\frac{\partial}{\partial\theta} \log \prod_{i=1}^n f(X_i \mid \theta)\Big] = 0$. Thus, by the necessary and sufficient condition for equality in Cauchy–Schwarz, we must have $W(\mathbf{x}) - \tau(\theta)$ proportional to $\frac{\partial}{\partial\theta} \log \prod_{i=1}^n f(X_i \mid \theta)$, which is exactly (17). □

Example 4.19 (Normal Variance Bound Continued)**.** Consider the normal likelihood

$$L(\mu, \sigma^2 \mid x) = (2\pi\,\sigma^2)^{-\frac{1}{2}} \exp\Big(-\tfrac{(x-\mu)^2}{2\,\sigma^2}\Big),$$

{eq:12.37}

and hence

$$\tfrac{\partial}{\partial\sigma^2} \log L(\mu, \sigma^2 \mid x) = \tfrac{n}{2\,\sigma^4}\Big(\sum_{i=1}^{n} \tfrac{(x_i-\mu)^2}{n} - \sigma^2\Big).$$

{eq:12.38}

Thus, choosing

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Taking $a(\sigma^2) = \frac{n}{2\sigma^4}$ shows that the best unbiased estimator of σ^2 is

$$\tfrac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

When μ is unknown, the CR lower bound is not attainable. Note that if μ is unknown, this particular estimator is no longer unbiased for σ^2.

4.3 Sufficiency and Unbiasedness

Theorem 4.20 (Law of Total Variance))**.** *Let X and Y be random variables on the same probability space. Then*

$$\mathrm{Var}(X) \;=\; \mathbb{E}[\mathrm{Var}(X \mid Y)] \;+\; \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big).$$

Proof. Recall that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2$. By the law of total expectation, we have:

$$\mathbb{E}[X^2] = \mathbb{E}\big[\mathbb{E}[X^2 \mid Y]\big].$$

Next, observe that

$$\mathbb{E}[X^2 \mid Y] = \mathrm{Var}(X \mid Y) \; + \; \big(\mathbb{E}[X \mid Y]\big)^2.$$

Putting this into the expression for $\mathrm{Var}(X)$, we get

$$\mathrm{Var}(X) = \mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] + \mathbb{E}\Big[\big(\mathbb{E}[X \mid Y]\big)^2\Big] - \big(\mathbb{E}[X]\big)^2.$$

But again by the law of total expectation,

$$\big(\mathbb{E}[X]\big)^2 = \Big(\mathbb{E}\big[\mathbb{E}[X \mid Y]\big]\Big)^2.$$

We can rewrite

$$\mathbb{E}\Big[\big(\mathbb{E}[X \mid Y]\big)^2\Big] - \Big(\mathbb{E}\big[\mathbb{E}[X \mid Y]\big]\Big)^2 = \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big).$$

Therefore,

$$\mathrm{Var}(X) = \mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] + \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big),$$

which completes the proof. □

{thm:rao_bla

Theorem 4.21 (Rao–Blackwell)**.** *Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define*

$$\phi(T) = \mathbb{E}_\theta[\, W \mid T\,].$$

Then

$$\mathbb{E}_\theta[\phi(T)] = \tau(\theta) \quad \textit{and} \quad \mathrm{Var}_\theta[\phi(T)] \le \mathrm{Var}_\theta[W] \quad \textit{for all } \theta.$$

In other words, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

Proof. From the law of total expectation we have

$$\tau(\theta) = \mathbb{E}_\theta[W] = \mathbb{E}_\theta\big[\mathbb{E}_\theta[\, W \mid T\,]\big] = \mathbb{E}_\theta\big[\phi(T)\big],$$

{eq:13.1}

so $\phi(T)$ is unbiased for $\tau(\theta)$. Next, using the law of total variance:

$$\mathrm{Var}_\theta[W] = \mathrm{Var}_\theta\big(\mathbb{E}_\theta[\, W \mid T\,]\big) + \mathbb{E}_\theta\big(\mathrm{Var}_\theta[\, W \mid T\,]\big) = \mathrm{Var}_\theta[\phi(T)] + \mathbb{E}_\theta\big(\mathrm{Var}_\theta[\, W \mid T\,]\big) \ge \mathrm{Var}_\theta[\phi(T)].$$

{eq:13.2}

Hence, $\phi(T)$ has variance no larger than W, making $\phi(T)$ a uniformly better estimator (in the sense of smaller or equal variance while remaining unbiased).

It remains to show that $\phi(T)$ is indeed a valid estimator—that is, it depends only on the observed sample (through T) and not on θ. This follows from:

- the definition of sufficiency of T, and
- the fact that W itself is a function only of the sample, so the conditional distribution of W given T does not depend on θ.

Therefore, $\phi(T)$ is a (uniformly) better unbiased estimator of $\tau(\theta)$. □

If we condition on the insufficient statistics, the resulting quantity might not be an estimator: it will depend on the parameter θ we want to estimate.

Example 4.22. Let $X_1, X_2 \sim N(\theta, 1)$ be i.i.d. random variables. The statistic $\overline{X} = \frac{1}{2}(X_1 + X_2)$ has $\mathbb{E}_\theta \overline{X} = \theta$, $\mathrm{Var}(\overline{X}) = \frac{1}{2}$.

Let $\phi(X) = \mathbb{E}_\theta(\overline{X}|X_1)$. By the same argument as in Theorem 4.21, $\mathbb{E}\phi(X) = \theta$, $\mathrm{Var}(\phi(X)) \leq \mathrm{Var}(\overline{X})$. However,

$$\phi(X_1) = \mathbb{E}_\theta(\overline{X}|X_1) = \frac{1}{2}X_1 + \frac{1}{2}\theta$$

is not an estimator.

Theorem 4.23. *If W is a best unbiased estimator of $\tau(\theta)$, then W is unique.*

Proof. Suppose W_1 and W_2 are two *best* unbiased estimators of $\tau(\theta)$. We will show that W_1 and W_2 must coincide almost surely.

Because both W_1 and W_2 have minimum variance, we have $\mathrm{Var}_\theta(W_1) = \mathrm{Var}_\theta(W_2)$ for all θ. Consider their average $A := \frac{1}{2}(W_1 + W_2)$. A is also an unbiased estimator of $\tau(\theta)$. We can compute its variance as follows:

$$\mathrm{Var}_\theta(A) \;=\; \mathrm{Var}_\theta\Big(\frac{W_1 + W_2}{2}\Big) = \frac{1}{4}\Big(\mathrm{Var}_\theta(W_1) + \mathrm{Var}_\theta(W_2) + 2\,\mathrm{Cov}_\theta(W_1, W_2)\Big)\,.$$

Using $\mathrm{Var}_\theta(W_1) = \mathrm{Var}_\theta(W_2) = V_{\min}$) we get:

$$\mathrm{Var}_\theta(A) \;=\; \frac{1}{4}\Big(2V_{\min} + 2\,\mathrm{Cov}_\theta(W_1, W_2)\Big) = \frac{1}{2}\Big(V_{\min} + \mathrm{Cov}_\theta(W_1, W_2)\Big)\,.$$

Now, by the Cauchy–Schwarz inequality, $\mathrm{Cov}_\theta(W_1, W_2) \leq \sqrt{\mathrm{Var}_\theta(W_1)\mathrm{Var}_\theta(W_2)} = V_{\min}$. In fact, $\mathrm{Cov}_\theta(W_1, W_2) = V_{\min}$ if and only if W_1 and W_2 are perfectly linearly correlated (i.e. one is an affine function of the other with probability 1).

There are two cases to consider:

(1) If $\mathrm{Cov}_\theta(W_1, W_2) < V_{\min}$ for some θ, then $\mathrm{Var}_\theta(A) = \frac{1}{2}(V_{\min} + \mathrm{Cov}_\theta(W_1, W_2)) < \frac{1}{2}(V_{\min} + V_{\min}) = V_{\min}$. This means A has smaller variance than $V_{\min}$ for that θ, contradicting the assumption that W_1 and W_2 had the minimum possible variance.

(2) For all θ, $\mathrm{Cov}_\theta(W_1, W_2) = V_{\min}$. In this case, the equality condition in Cauchy–Schwarz implies W_2 is almost surely an affine function of W_1 : there exists $a(\theta), b(\theta)$ such that $W_2 = a(\theta)\, W_1 + b(\theta)$ with probability 1.

From the identity $\mathrm{Cov}(W_1, W_2) = \mathrm{Var}(W_1)$, we find $a(\theta) = 1$. Since $\mathbb{E}_\theta[W_1] = \mathbb{E}_\theta[W_2] = \tau(\theta)$ for all θ, we obtain $b(\theta) = 0$ for all θ. Therefore $W_2 = W_1$ almost surely. □

The following theorem gives a necessary and sufficient condition for the best unbiased estimator.

{thm:zero_un

Theorem 4.24. *Let W be an unbiased estimator of $\tau(\theta)$, i.e. $E_\theta[W] = \tau(\theta)$. Then W is the* best *unbiased estimator of $\tau(\theta)$ if and only if*

$$\mathrm{Cov}_\theta(W, U) \;=\; 0 \quad \textit{for all } \theta \textit{ and for every unbiased estimator } U \textit{ of } 0.$$

Proof. Suppose W is the best unbiased estimator of $\tau(\theta)$. We claim that for any unbiased estimator U of 0, we must have $\mathrm{Cov}_\theta(W, U) = 0$ for all θ. To see why, consider a new estimator

$$\phi_a \;=\; W \;+\; a\,U,$$

where a is any real constant. Since $E_\theta[U] = 0$ and $E_\theta[W] = \tau(\theta)$, it follows that ϕ_a is also unbiased for $\tau(\theta)$. Next, we look at the variance of ϕ_a:

$$\mathrm{Var}_\theta(\phi_a) \;=\; \mathrm{Var}_\theta\bigl(W + aU\bigr) \;=\; \mathrm{Var}_\theta(W) \;+\; 2\,a\,\mathrm{Cov}_\theta(W,U) \;+\; a^2\,\mathrm{Var}_\theta(U).$$

If there existed some θ_0 for which $\mathrm{Cov}_{\theta_0}(W, U) < 0$, then by choosing a suitably (for instance $a \in \Bigl(0,\; -2\,\mathrm{Cov}_{\theta_0}(W,U)\,/\,\mathrm{Var}_{\theta_0}(U)\Bigr)$), we can make

$$2\,a\,\mathrm{Cov}_{\theta_0}(W,U) \;+\; a^2\,\mathrm{Var}_{\theta_0}(U) \;<\; 0,$$

so that $\mathrm{Var}_{\theta_0}(\phi_a) < \mathrm{Var}_{\theta_0}(W)$. Then ϕ_a would be a strictly better unbiased estimator of $\tau(\theta_0)$ than W, contradicting the assumption that W is the best. A parallel argument shows that if $\mathrm{Cov}_\theta(W, U) > 0$ for some other θ_0, one can pick a negative a in the same way to reduce the variance at that θ_0. Hence, the only way W can be the best estimator is if

$$\mathrm{Cov}_\theta(W,U) \;=\; 0 \quad \text{for all } \theta \text{ and all unbiased } U \text{ of } 0.$$

Conversely, suppose we have an unbiased estimator W of $\tau(\theta)$ such that W is uncorrelated with *every* unbiased estimator of 0. Take any other unbiased estimator W' of $\tau(\theta)$, and write

$$W' \;=\; W \;+\; \bigl(W' - W\bigr).$$

Note that $W' - W$ is itself an unbiased estimator of 0, by hypothesis, $\mathrm{Cov}_\theta\bigl(W, W' - W\bigr) = 0$. Hence,

$$\mathrm{Var}_\theta(W') \;=\; \mathrm{Var}_\theta\bigl(W + (W' - W)\bigr) \;=\; \mathrm{Var}_\theta(W) \;+\; \mathrm{Var}_\theta\bigl(W' - W\bigr) \;+\; 2\,\mathrm{Cov}_\theta\bigl(W, W' - W\bigr).$$

The last term vanishes since $W - W'$ and W are uncorrelated, so

$$\mathrm{Var}_\theta(W') \;=\; \mathrm{Var}_\theta(W) \;+\; \mathrm{Var}_\theta\bigl(W' - W\bigr) \;\geq\; \mathrm{Var}_\theta(W).$$

Thus, for every unbiased estimator W', we have $\mathrm{Var}_\theta(W') \geq \mathrm{Var}_\theta(W)$ at all θ, so W is indeed the best unbiased estimator of $\tau(\theta)$. □

It's hard to check if an unbiased estimator is uncorrelated with every unbiased estimator of 0. But Theorem 4.24 provides a way to determine an estimator is not the best.

Example 4.25. Let X be an observation from $\mathrm{Unif}(\theta, \theta + 1)$. Then

$$E_\theta[X] \;=\; \int_\theta^{\theta+1} x\,dx \;=\; \theta \;+\; \frac{1}{2}. \qquad \texttt{\{eq:13.9\}}$$

So $X - \frac{1}{2}$ is an unbiased estimator of θ, and $\mathrm{Var}_\theta[X] \;=\; \frac{1}{12}$. Now we proceed to find an unbiased estimator of 0. If a function $h(x)$ satisfies

$$\int_\theta^{\theta+1} h(x)\,dx \;=\; 0, \quad \forall\theta, \qquad \texttt{\{eq:13.10\}}$$

then

$$0 \;=\; \frac{d}{d\theta}\int_{\theta}^{\theta+1} h(x)\,dx \;=\; h(\theta+1) \;-\; h(\theta), \quad \forall\theta.$$

{eq:13.11}

Hence $h(x)$ must be periodic with period 1. One such function is $h(x) \;=\; \sin(2\pi x)$.

Next, observe that

$$\mathrm{Cov}_{\theta}\Big(X-\tfrac{1}{2},\ \sin(2\pi X)\Big) = \mathrm{Cov}_{\theta}\big(X,\ \sin(2\pi X)\big) \;=\; \int_{\theta}^{\theta+1} x\,\sin(2\pi x)\,dx.$$

We can integrate by parts:

$$\int_{\theta}^{\theta+1} x\,\sin(2\pi x)\,dx \;=\; \frac{x\cos(2\pi x)}{2\pi}\Big|_{\theta}^{\theta+1} \;+\; \int_{\theta}^{\theta+1}\frac{\cos(2\pi x)}{2\pi}\,dx = -\frac{-\cos(2\pi\theta)}{2\pi}.$$

So $X-\frac{1}{2}$ is correlated with an unbiased estimator of 0 (namely $\sin(2\pi X)$). Therefore it cannot be the best unbiased estimator of θ.

Example 4.26 (Application of Theorem 4.24)**.** For $X_1,\ldots,X_n$ sampled i.i.d. from Uniform$(0,\theta)$, we have shown in Example 4.15 that $\frac{n+1}{n}Y$ is an unbiased estimator of θ, where

$$Y = \max\{X_1,\ldots,X_n\}.$$

In Example 2.27, we have shown that Y is a complete sufficient statistic. Then for any unbiased estimator W of 0, $T(Y) = \mathbb{E}(W|Y)$ is an unbiased estimator of 0 and $\mathbb{E}_{\theta}[T(Y)] = 0$. Since Y is complete, this implies $\mathbb{E}(W|Y) = 0$ with probability 1 and

$$\mathrm{Cov}(Y,W) = \mathbb{E}[YW] = \mathbb{E}[\mathbb{E}[YW|Y]] = \mathbb{E}[Y\mathbb{E}[W|Y]] = 0.$$

Since Y is uncorrelated with any unbiased estimator of 0, Y is the UMVUE of θ.

{thm:LS}

Theorem 4.27 (Lehmann–Scheffé Theorem)**.** *Let T be a complete sufficient statistic for a parameter θ and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expectation.*

Proof. Assume $\mathbb{E}\phi(T) = \tau(\theta)$. Let W be any unbiased estimator of $\tau(\theta)$. Let $\phi'(T) = \mathbb{E}[W|T]$ Then $\phi'(T)$ is an unbiased estimator of $\tau(\theta)$ and

$$\mathrm{Var}_{\theta}(\phi'(T)) \le \mathrm{Var}_{\theta}(W).$$

Note that $\mathbb{E}_{\theta}[\phi(T)-\phi'(T)] = 0$. Since T is a complete sufficient statistic, $\mathbb{P}_{\theta}(\phi(T)-\phi'(T) = 0) = 1$ for any θ. Therefore $\mathrm{Var}_{\theta}(\phi'(T)) = \mathrm{Var}_{\theta}(\phi(T))$ and

$$\mathrm{Var}_{\theta}(\phi(T)) \le \mathrm{Var}_{\theta}(W).$$

□

The theorem above suggests a way to find the best estimator. If W is any unbiased estimator of $\tau(\theta)$, and T is a complete sufficient statistic for θ, then $\phi(T) = \mathbb{E}[W|T]$ is the best unbiased estimator of $\tau(\theta)$.

Corollary 4.28. If T is a complete sufficient statistic for θ and W is any unbiased estimator of $\tau(\theta)$, then $\phi(T) = \mathbb{E}[W|T]$ is the best unbiased estimator of $\tau(\theta)$.

Proof. This follows immediately from Theorem 4.27 since $\phi(T) = \phi(T) = \mathbb{E}[W|T]$ is an estimator based only on T. □

Example 4.29. Suppose $X_1, \dots, X_n$ are i.i.d. samples from $N(\mu, 1)$. Find the best unbiased estimator of $\tau(\mu) = \mu^2$.

We know $\overline{X} = N(\mu, \frac{1}{n})$. Therefore

$$\mathbb{E}[\overline{X}^2] = \mu^2 + \frac{1}{n}$$

and $W = \overline{X}^2 - \frac{1}{n}$ is an unbiased estimator for μ^2.

From the general results for exponential families, $T = \overline{X}$ is a complete sufficient statistic for μ. So

$$\phi(T) = \mathbb{E}[W|T] = \mathbb{E}[\overline{X}^2 - \frac{1}{n}|T] = \overline{X}^2 - \frac{1}{n}$$

is the UMVUE. And $\mathrm{Var}(\phi(T)) = \mathrm{Var}(\overline{X}^2) = \frac{4\mu^2}{n} + \frac{2}{n^2}$. The CR lower bound gives

$$\mathrm{Var}(W) \geq \frac{(\tau'(\mu))^2}{nI_X(\mu)} = \frac{4\mu^2}{n},$$

where we use $I_X(\mu) = \frac{1}{2}$ due to Example 4.17. This suggests a gap between the variance of UMVUE and the CR lower bound.

Example 4.30 (Binomial Best Unbiased Estimation)**.** Let $X_1, X_2, \dots, X_n$ be i.i.d. $\mathrm{Bin}(k, \theta)$ with k known. We wish to estimate the probability of exactly one success in $\mathrm{Bin}(k, \theta)$:

$$\tau(\theta) = \mathbb{P}_\theta(X = 1) \ = \ k\,\theta\,(1-\theta)^{k-1}.$$

Because the complete sufficient statistic is $T = \sum_{i=1}^n X_i \sim \mathrm{Bin}(kn, \theta)$, but no unbiased estimator based solely on T is immediately evident, we try a simpler approach first.

A simple unbiased estimator. Define

$$h(X_1) \ = \ \begin{cases} 1, & X_1 = 1, \\ 0, & \text{otherwise.} \end{cases}$$

We compute

$$\mathbb{E}_\theta\big[h(X_1)\big] \ = \ \sum_{x_1=0}^{k} h(x_1) \binom{k}{x_1} \theta^{x_1}(1-\theta)^{k-x_1} \ = \ k\,\theta\,(1-\theta)^{k-1},$$

showing that $h(X_1)$ is indeed an unbiased estimator of $k\,\theta\,(1-\theta)^{k-1}$.

Conditioning on the sufficient statistic. Define

$$\phi\Big(\sum_{i=1}^n X_i\Big) = \mathbb{E}\Big[h(X_1)\,\Big|\,\sum_{i=1}^n X_i\Big].$$

Then $\phi\big(\sum X_i\big)$ is the best unbiased estimator of $k\,\theta\,(1-\theta)^{k-1}$.
Explicit form of ϕ. Suppose we observe $T = \sum_{i=1}^n X_i = t$. Then

$$\phi(t) \;=\; E\big[h(X_1)\,\big|\,T=t\big] \;=\; P\big(X_1 = 1\,\big|\,T=t\big) = \frac{P\big(X_1=1,\;\sum_{i=2}^n X_i = t-1\big)}{P\big(\sum_{i=1}^n X_i = t\big)}.$$

Since $X_1 \sim \mathrm{Bin}(k,\theta)$ and $\sum_{i=2}^n X_i \sim \mathrm{Bin}(k(n-1),\theta)$ and $T \sim \mathrm{Bin}(kn,\theta)$, the dependence on θ cancels out (because T is sufficient). Hence the ratio of the probabilities reduces to

$$\phi(t) \;=\; k\,\frac{\dbinom{k(n-1)}{t-1}}{\dbinom{kn}{t}}.$$

Therefore, our final best unbiased estimator is

$$\phi\Big(\sum_{i=1}^n X_i\Big) \;=\; k\,\frac{\dbinom{k(n-1)}{\sum_{i=1}^n X_i - 1}}{\dbinom{kn}{\sum_{i=1}^n X_i}}.$$

Checking directly that $\phi\Big(\sum_{i=1}^n X_i\Big)$ with the formula above is an unbiased estimator is not easy but we know by definition $\mathbb{E}\phi(t) = \mathbb{E}h(X_1) = \tau(\theta)$.

4.4 Loss function optimality

A *loss function* $L(\theta,a)$ is a way to measure the cost associated with deciding a when the true parameter (or state of nature) is θ. In this context:

- θ: the unknown (but fixed or random) parameter we want to estimate;
- a: an action or decision, such as an estimator $\hat{\theta}$;
- $L(\theta,a)$: a non-negative cost (penalty) that increases with the discrepancy between θ and a.

Commonly used loss function

- Squared error loss: $L(\theta,a) = (\theta-a)^2$.
- Absolute error loss: $L(\theta,a) = |\theta-a|$.
- Zero-one Loss: $L(\theta,a) = \begin{cases} 0, & \text{if } a=\theta, \\ 1, & \text{otherwise.} \end{cases}$

Risk Function Given a function $L(\theta,a)$, we can define a risk function $R(\theta,W)$ for an estimator W of θ:

$$R(\theta,W) = \mathbb{E}_\theta L(\theta,W).$$

For a given θ, this is the average loss that will be incurred if W is used. The expectation is taken over the randomness of data.

Taking $L(\theta,W) = (W-\theta)^2$, then $R(\theta,W)$ is the mean-squared error of an estimator W.

Example 4.31. Assume $X_1, \ldots, X_n$ are i.i.d. samples from a distribution with variance σ^2. Now we use the loss function

$$L(\sigma^2, a) = \frac{a}{\sigma^2} - 1 - \log\left(\frac{a}{\sigma^2}\right).$$

If $a = \sigma^2$, the loss is 0 and when $a \to 0$ or $a \to \infty$, $L(\sigma^2, a) \to \infty$. Consider estimators of the form $\delta_b = bS^2$, where S^2 is the sample variance. Since $\mathbb{E}S^2 = \sigma^2$, we have

$$R(\sigma^2, \delta_b) = \mathbb{E}\left(\frac{bS^2}{\sigma^2} - 1 - \log\frac{bS^2}{\sigma^2}\right) = b - 1 - \log b - \log\frac{S^2}{\sigma^2}.$$

$R(\sigma^2, \delta_b)$ is minimized if and only if $b - \log b$ is minimized. Therefore the optimal choice is $b = 1$.

Example 4.32 (Risk for squared error loss, Gaussian case)**.** Let $X_1, \ldots, X_n$ be i.i.d. samples from $N(\mu, \sigma^2)$. We consider the risk under the square loss for estimating σ^2. We consider $\delta_b(\mathbf{X}) = bS^2$ for $b \geq 0$. Then

$$\begin{aligned} R(\sigma^2, \delta_b) &= \mathrm{Var}(bS^2) + (\mathbb{E}bS^2 - \sigma^2)^2 \\ &= b^2\mathrm{Var}(S^2) + (b\sigma^2 - \sigma^2)^2 \\ &= \frac{b^2 2\sigma^4}{n-1} + (b-1)^2\sigma^4. \end{aligned}$$

Therefore we find the estimate with the minimal risk, it suffices to minimize

$$\frac{2b^2}{n-1} + (b-1)^2,$$

which implies $b = \frac{n-1}{n+1}$. Therefore

$$\frac{n-1}{n+1}S^2 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

has the smallest risk among all estimators of the form bS^2.

Bayes risk If θ has a prior $\pi(\theta)$, the *Bayes risk* for an estimator $\delta(\mathbf{X})$ is:

$$r(\pi, \delta) \;=\; \int_\Theta R(\theta, \delta)\pi(\theta)d\theta \;=\; \mathbb{E}_\pi\big[L(\theta, \delta(\mathbf{X}))\big].$$

Now $r(\pi, \delta)$ is independent of θ, since $R(\theta, \delta)$ is averaged over the distribution $\pi(\theta)$.

An estimator δ that achieves the minimal Bayes risk is called the *Bayes rule with respect to a prior* π denoted by δ^π.

For $\mathbf{X} \sim f(\mathbf{x}|\theta)$, $\theta \sim \pi$, the Bayes risk of an estimator δ can be written as

$$\begin{aligned} \int_\Theta R(\theta, \delta)\pi(\theta)d\theta &= \int_\Theta \left(\int_\mathcal{X} L(\theta, \delta(\mathbf{x}))f(\mathbf{x}|\theta)d\mathbf{x}\right)\pi(\theta)d\theta \\ &= \int_\Theta\int_\mathcal{X} L(\theta, \delta(\mathbf{x}))f(\mathbf{x}|\theta)\pi(\theta)d\mathbf{x}d\theta. \end{aligned}$$

We can write $f(\mathbf{x}|\theta)\pi(\theta) = \pi(\theta|\mathbf{x})m(\mathbf{x})$. Then we can rewrite

$$\int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta \int_\mathcal{X} L(\theta,\delta(\mathbf{x}))\pi(\theta|\mathbf{x})m(\mathbf{x})d\mathbf{x}d\theta = \int_\mathcal{X}\left(\int_\Theta L(\theta,\delta(\mathbf{x}))\pi(\theta|\mathbf{x})d\theta\right)m(\mathbf{x})d\mathbf{x}.$$

Note that

$$\int_\Theta L(\theta,\delta(\mathbf{x}))\pi(\theta|\mathbf{x})d\theta$$

is the expected loss under the posterior distribution $\pi(\theta|\mathbf{x})$ (posterior expected loss). Therefore to find the Bayes rule, it suffices to minimize the posterior expected loss.

Example 4.33. For squared error loss $L(\theta,a) = (\theta-a)^2$, we find

$$\int_\Theta (\theta-a)^2\pi(\theta|\mathbf{x})d\theta = \mathbb{E}[(\theta-a)^2|\mathbf{X}=\mathbf{x}].$$

Note that for any random variable Y and a constant b, we have

$$\mathbb{E}(Y-b)^2 = \mathbb{E}(Y-\mathbb{E}Y)^2 + (\mathbb{E}Y-b)^2 \geq \mathbb{E}(Y-\mathbb{E}Y)^2.$$

Then the posterior expected loss is minimized at $\delta^\pi = \mathbb{E}[\theta|\mathbf{x}]$, which is the posterior mean. So the Bayes rule for squared error loss is the posterior mean estimator.

{example:med

Example 4.34. For absolute error loss, we have

$$\int_\Theta |\theta-a|\;\pi(\theta|\mathbf{x})d\theta = \mathbb{E}\left[|\theta-a|\mid\mathbf{X}=\mathbf{x}\right].$$

This is minimized by choosing a to be the median of $\pi(\theta|\mathbf{x})$.

Therefore for the absolute error loss, the Bayes rule is the median of the posterior distribution.

Example 4.35. Let $X_1,\ldots,X_n$ be i.i.d. samples from $N(\mu,\sigma^2)$. and let $\pi(\mu)\sim N(\mu_0,\tau^2)$, where σ^2,μ,τ^2 are known. In Example 3.23, we get

$$\mu|\mathbf{x}\sim\mathcal{N}\left(\frac{\frac{\mu_0}{\tau^2}+\frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}},\frac{1}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right).$$

Its mean and median are equal to

$$\mathbb{E}[\theta|\mathbf{x}] = \frac{\frac{\mu_0}{\tau^2}+\frac{\sum x_i}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}.$$

Example 4.36. Let $X_1,\ldots,X_n$ be i.i.d. $\mathrm{Ber}(p)$ and $Y=\sum_{i=1}^n X_i$. Suppose the prior distribution on p is $\mathrm{Beta}(\alpha,\beta)$. Its probability density function is given by

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},\quad 0<x<1.$$

The posterior distribution $\pi(p|\mathbf{x})$ only depends on $Y=y$ and from Example 3.20,

$$p|y\sim\mathrm{Beta}(\alpha+y,\beta+n-y).$$

The Bayes estimator for the square loss is $\delta^\pi = \frac{y+\alpha}{\alpha+\beta+n}$.

If we consider the absolute error loss, from (4.34), we need to find the median m of the posterior distribution $\mathrm{Beta}(\alpha+y,\beta+n-y)$, which is given by solving the integral equation

$$\int_0^m \frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+a)\Gamma(n-y+\beta}p^{y+\alpha-1}(1-p)^{n-y+\beta-1}dp = \frac{1}{2}.$$

This equation can only be solved numerically.

5 Algorithms for point estimation

5.1 The EM algorithm

The Expectation-Maximization (EM) algorithm is an iterative approach used to find maximum likelihood estimates (MLE) for models with latent variables. It is widely applied in statistics and machine learning, especially for problems involving missing data or mixture models.

The EM algorithm consists of two main steps:

- *Expectation Step (E-step):* Compute the expected value of the complete-data log-likelihood, given the observed data and current parameter estimates.
- *Maximization Step (M-step):* Maximize this expectation with respect to the parameters to obtain updated estimates.

These steps are iterated until convergence, usually when changes in parameter estimates are below a certain threshold.

Gaussian mixture (GMM) The generative process for GMM is:

1. Choose a cluster assignment $z_i, 1 \le i \le K$

$$z_i \sim \text{Categorical}(w), \quad P(z_i = k|w) = w_k,$$

where w_k are the mixture weights, satisfying:

$$\sum_{k=1}^{K} w_k = 1, \quad 0 \le w_k \le 1.$$

2. Generate data from the assigned cluster:

$$x_i|z_i = k \sim \mathcal{N}(\mu_k, \Sigma_k).$$

Here

- $x_i \in \mathbb{R}^p$: Data point.
- z_i: Cluster assignment.
- μ_k: Mean of cluster k.
- Σ_k: Covariance matrix of cluster k.
- w_k: Mixture weight for cluster k.

The probability density function of a multivariate normal distribution is:

$$f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Let θ be the collection of all parameters w_k, μ_k, Σ_k. Our goal is to find the MLE of θ and estimate the probability $\mathbb{P}(z_i = k|x_i, \theta)$, the probability that data point i belongs to the k-th cluster.

The Expectation-Maximization (EM) algorithm iteratively updates cluster assignments and model parameters:

- **E-step:** Compute the posterior probabilities (responsibilities):
$$\gamma_{ik} = P(z_i = k|x_i, \theta^{(t)}) = \frac{w_k^{(t)} \mathcal{N}(x_i; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k'=1}^{K} w_{k'}^{(t)} \mathcal{N}(x_i; \mu_{k'}^{(t)}, \Sigma_{k'}^{(t)})}.$$

- **M-step:** Update the parameters by maximizing the auxiliary function:
$$\begin{aligned}
\mu_k^{(t+1)} &= \frac{\sum_i \gamma_{ik} x_i}{\sum_i \gamma_{ik}}, \\
\Sigma_k^{(t+1)} &= \frac{\sum_i \gamma_{ik} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^\top}{\sum_i \gamma_{ik}}, \\
w_k^{(t+1)} &= \frac{\sum_i \gamma_{ik}}{n}.
\end{aligned}$$

Each iteration increases the log-likelihood and converges to a local maximum.
The likelihood function is:

$$\mathcal{L}(\theta) = f(\mathbf{x}|w, \mu, \Sigma) = \prod_i \sum_{k=1}^{K} f(x_i|z_i = k, \theta)\mathbb{P}(z_i = k|\theta),$$

where the last identity is due to the law of total probability. Taking the log:

$$\log \mathcal{L}(\theta) = \sum_i \log \sum_k f(x_i|z_i = k, \theta)\mathbb{P}(z_i = k|\theta). \tag{31}$$

{eq:GMM_like

This sum inside the logarithm makes direct maximization difficult due to the fact that all z_i are not observed.

We now just focus on (31) in the abstract setting where

- $x_1, \ldots, x_n$ are observed data
- $z_1, \ldots, z_n$ are unobserved (hidden variables) taking values in $k = 1, \ldots, K$
- θ represents all the model parameters.

The idea of Expectation Maximization (EM) is to maximize a lower bound of $\log \mathcal{L}(\theta)$ using the following Jensen's inequality:

{lem:Jensen}

Lemma 5.1 (Jensen's inequality)**.** *Let f be a convex function on $\mathbb{R}$ and let X be a random variable. Then*

$$\mathbb{E}f(X) \geq f(\mathbb{E}X).$$

Let us continue with (31), we can rewrite it as

$$\begin{aligned}
\log \mathcal{L}(\theta) &= \sum_i \log \sum_k f(x_i|z_i = k, \theta)\mathbb{P}(z_i = k|\theta) \\
&= \sum_i \log \sum_k p(X_i = x_i, z_i = k|\theta) \\
&= \sum_i \log \sum_k \mathbb{P}(z_i = k|x_i, \theta_t) \frac{p(X_i = x_i, z_i = k|\theta)}{\mathbb{P}(z_i = k|x_i, \theta_t)},
\end{aligned}$$

here θ_t is a parameter we will decide later. Now for each z_i,

$$\sum_k \mathbb{P}(z_i = k|x_i, \theta_t)\frac{p(X_i = x_i, z_i = k|\theta)}{\mathbb{P}(z_i = k|x_i, \theta_t)} = \mathbb{E}_{z_i}\left(\frac{p(X_i = x_i, z_i = k|\theta)}{\mathbb{P}(z_i = k|x_i, \theta_t)}\right)$$

is a weighted average over the distribution $\mathbb{P}(z_i|x_i, \theta_t)$. By Jensen's inequality, since $-\log(x)$ is convex,

$$\begin{aligned}\log \mathcal{L}(\theta) &= \sum_i \log \mathbb{E}_{z_i}\left(\frac{p(X_i = x_i, z_i = k|\theta)}{\mathbb{P}(z_i = k|x_i, \theta_t)}\right)\\ &\geq \sum_i \mathbb{E}_{z_i} \log\left(\frac{p(X_i = x_i, z_i = k|\theta)}{\mathbb{P}(z_i = k|x_i, \theta_t)}\right)\\ &= \sum_i \sum_k \mathbb{P}(z_i = k|x_i, \theta_t) \log\left(\frac{p(X_i = x_i, z_i = k|\theta)}{\mathbb{P}(z_i = k|x_i, \theta_t)}\right)\end{aligned}$$

Denote

$$A(\theta, \theta_t) = \sum_i \sum_k \mathbb{P}(z_i = k|x_i, \theta_t) \log\left(\frac{p(X_i = x_i, z_i = k|\theta)}{\mathbb{P}(z_i = k|x_i, \theta_t)}\right).$$

It's called an evidence lower bound (ELBO). By the definition of conditional probability, we have

$$p(X_i = x_i, z_i = k|\theta_t) = \mathbb{P}(z_i = k|x_i, \theta_t)f(x_i|\theta_t)$$

And we can check that

$$\begin{aligned}A(\theta_t, \theta_t) &= \sum_i \sum_k \mathbb{P}(z_i = k|x_i, \theta_t) \log f(x_i|\theta_t)\\ &= \sum_i \log f(x_i|\theta_t) = \log \mathcal{L}(\theta_t).\end{aligned}$$

Therefore $L(\theta)$ and $A(\theta, \theta_t)$ as functions of θ, coincides at $\theta = \theta_t$.

The EM algorithm maximizes the lower bound of $\log \mathcal{L}(\theta)$ by updating θ_t. Let

$$\gamma_{ik} = \mathbb{P}(Z_i = k|x_i, \theta_t).$$

We can write

$$\begin{aligned}A(\theta, \theta_t) &= \sum_i \sum_k \mathbb{P}(z_i = k|x_i, \theta_t) \log\left(\frac{p(X_i = x_i, z_i = k|\theta)}{\mathbb{P}(z_i = k|x_i, \theta_t)}\right)\\ &= \sum_i \sum_k \gamma_{ik} \log\left(\frac{p(X_i = x_i, z_i = k|\theta)}{\gamma_{ik}}\right).\end{aligned} \tag{32}$$

{eq:ELBO}

The EM algorithm can be summarized as follows:

- E-step: Compute the responsibilities γ_{ik} for each i, k based on current θ_t.
- M-step: Maximize the ELBO: $\max_\theta A(\theta, \theta_t)$. Since only the numerator in (32) depends on θ, it's equivalent to consider

$$\max_\theta \sum_i \sum_k \gamma_{ik} \log p(X_i = x_i, z_i = k|\theta). \tag{33}$$

{eq:AELBO}

EM method for Gaussian Mixture Models Assuming $x_1, \dots, x_n$ are sampled from a Gaussian mixture model. After choosing a desired cluster number K, the EM algorithm provides a clustering method to classify $x_1, \dots, x_n$ into K clusters and output estimators for μ, Σ, w and responsibility γ_{ik} for each $i \in [n], k \in [K]$.

Let us compute each step in detail.

1. For the E-step, by Bayes Rule, we have

$$\begin{aligned}\gamma_{ik} = \mathbb{P}(Z_i = k | x_i, \theta_t) &= \frac{p(X_i = x_i | z_i = k, \theta_t)\mathbb{P}(Z_i = k|\theta_t)}{f(x_i|\theta_t)} \\ &= \frac{w_k^{(t)} \mathcal{N}(x_i; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k'=1}^K w_{k'}^{(t)} \mathcal{N}(x_i; \mu_{k'}^{(t)}, \Sigma_{k'}^{(t)})}.\end{aligned}$$

2. For the M-step. From (33), it suffices to maximize over μ, Σ, w of the following quantity:

$$\begin{aligned}&\sum_i \sum_k \gamma_{ik} \log p(X_i = x_i, z_i = k | w, \mu, \Sigma) \\ =&\sum_i \sum_k \gamma_{ik} \log[p(X_i = x_i|\mu, \Sigma, z_i = k)\mathbb{P}(z_i = k|w)] \\ =&\sum_i \sum_k \gamma_{ik} \log[p(X_i = x_i|\mu, \Sigma, z_i = k) \cdot w_k] \\ =&\sum_i \sum_k \gamma_{ik} \log p(X_i = x_i|\mu, \Sigma, z_i = k) + \sum_{i,k} \gamma_{ik} \log(w_k). \qquad (34)\end{aligned}$$

{eq:ELBO_exp

Let us first consider the update rule for μ. The only term related to μ is

$$\begin{aligned}&\sum_i \sum_k \gamma_{ik} \log p(X_i = x_i | w, \mu, \Sigma, z_i = k) \\ =&\sum_i \sum_k \gamma_{ik} \log \left(\frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right) \right)\end{aligned}$$

Taking the derivative with respect to μ_l and setting it to zero we find

$$-\frac{1}{2} \sum_i \gamma_{il} \nabla_{\mu_l} \left[(x_i - \mu_l)^\top \Sigma_l^{-1} (x_i - \mu_l) \right] = 0. \qquad (35)$$

{eq:mu_l}

We can compute that

$$\nabla_{\mu_l} \left[(x_i - \mu_l)^\top \Sigma_l^{-1} (x_i - \mu_l) \right] = 2\Sigma_l^{-1}\mu_l - 2\Sigma_l^{-1} x_i.$$

Then (35) implies

$$\sum_i \gamma_{il}(x_i - \mu_l) = 0.$$

Hence

$$\mu_k^{(t+1)} = \frac{\sum_i \gamma_{ik} x_i}{\sum_i \gamma_{ik}}.$$

Applying the same analysis we can obtain the updates for Σ_k as

$$\Sigma_k^{(t+1)} = \frac{\sum_i \gamma_{ik}(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^\top}{\sum_i \gamma_{ik}}.$$

Next, we consider the update for w_k. From (34), we need to maximize

$$\sum_{i,k} \gamma_{ik} \log(w_k)$$

subject to the constraint that $\sum_{k=1}^K w_i = 1$. Consider the Lagrangian

$$L(w) = \sum_{i,k} \gamma_{ik} \log(w_k) + \beta(\sum_{k=1}^K w_k - 1).$$

Taking derivative w.r.t. w_k we find

$$w_k = -\frac{\sum_k \gamma_{ik}}{\beta}.$$

Since $\sum_k w_k = 1$, we find $-\beta = \sum_i(\sum_k \gamma_{ik}) = n$. Hence

$$w_k^{(t+1)} = \frac{\sum_k \gamma_{ik}}{n}.$$

Therefore we have obtained the closed-form expression for the update rules.

General EM algorithms Suppose we have n independent random variables $X_1, \ldots, X_n$ and latent variables $Z_1, \ldots, Z_n$. Here X_i depends only on Z_i and $Z_i, 1 \le i \le n$ are not observed. Now we assume Z_i are discrete random variables for simplicity. The PDF of x is given by

$$f(x|\theta) = \sum_{z_i} p(x, z_i|\theta),$$

where $p(x, z_i|\theta)$ is the joint density. The log-likelihood function $l(\theta)$ given data $X_1, \ldots, X_n$ is given by

$$\begin{aligned} l(\theta) = \log(f(\mathbf{x}|\theta)) &= \sum_{i=1}^n \log f(x_i, \theta) \\ &= \sum_{i=1}^n \log \sum_{z_i} p(x_i, z_i|\theta) \end{aligned}$$

Now we introduce n distributions $q_1, \ldots, q_n$ of the latent variables $Z_1, \ldots, Z_n$ (namely, $\sum_{z_i} q_i(z_i) = 1$, and write

$$\begin{aligned} l(\theta) &= \sum_{i=1}^n \log \sum_{z_i} q_i(z_i) \frac{p(x_i, z_i|\theta)}{q_i(z_i)} \\ &\ge \sum_{i=1}^n \sum_{z_i} q_i(z_i) \log \left(\frac{p(x_i, z_i|\theta)}{q_i(z_i)} \right), \qquad (36) \end{aligned}$$

{eq:jensen_E

where we use Jensen's inequality. Denote

$$\mathrm{ELBO}(x, q, \theta) = \sum_z q(z) \log \left(\frac{p(x, z|\theta)}{q(z)} \right).$$

We can rewrite (36) as

$$l(\theta) \geq \sum_{i=1}^{n} \mathrm{ELBO}(x_i, q_i, \theta). \tag{37}$$

{eq:anyq}

(36) holds for all distributions $q_1, \ldots, q_n$. We know the *equality* case holds for Jensen's inequality (Lemma 5.1) when the random variable X is a constant. In our case, this means

$$\frac{p(x_i, z_i|\theta)}{q_i(z_i)} = c$$

for some constant c independent of z_i. Hence $q_i(z_i) \propto p(x_i, z_i|\theta)$. Since $\sum_{z_i} q_i(z_i) = 1$, we can find

$$q_i(z_i) = \frac{p(x_i, z_i|\theta)}{\sum_{z_i} p(x_i, z_i|\theta)} = \frac{p(x_i, z_i|\theta)}{f(x_i|\theta)} = p(z_i|x_i, \theta).$$

This implies the following identity for any θ:

$$l(\theta) = \sum_{i=1}^{n} \mathrm{ELBO}(x_i, p(z_i|x_i, \theta), \theta). \tag{38}$$

{eq:MLE_ELBO

We now have the general EM algorithm for each iterate. At step $t+1$:

- Let $q_i^{(t)}(z_i) = p(z_i|x_i, \theta^{(t)})$. Compute ELBO: $\sum_{i=1}^{n} \mathrm{ELBO}(x_i, q_i^{(t)}, \theta)$.
- Set

$$\begin{aligned} \theta^{(t+1)} &= \arg\max_{\theta} \sum_{i=1}^{n} \mathrm{ELBO}(x_i, q_i^{(t)}, \theta) \qquad (39) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \sum_{z_i} q_i^{(t)}(z_i) \log \left(\frac{p(x_i, z_i|\theta)}{q_i^{(t)}(z_i)} \right). \end{aligned}$$

{eq:deft+1}

See Figure 5.1 for an illustration.

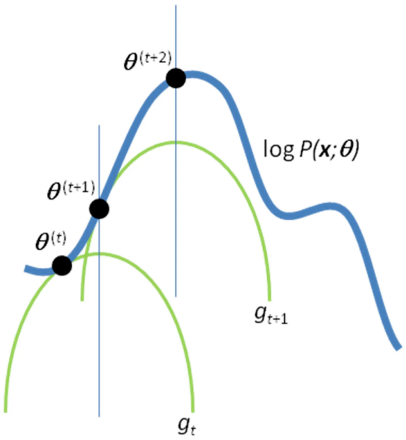
Since $q_i^{(t)}(z_i)$ does not depend on θ, simplifying (39) yields the following:

Proposition 5.2 (The general EM algorithm)**.** The EM algorithm seeks to find the MLE of θ by iteratively applying these two steps:

- Expectation step: compute

  $$Q(\theta|\theta^{(t)}) := \sum_{i=1}^{n} \mathbb{E}_{Z_i \sim p(\cdot|x_i, \theta^{(t)})} \left[\log(p(x_i, Z_i|\theta) \right]$$

  where the expectation of Z_i is taken over the distribution $p(\cdot|x_i, \theta^{(t)})$. This is the expected value of the log-likelihood function of θ, with respect to the current conditional distribution of Z_i given x_i and the current estimate $\theta^{(t+1)}$.



{fig:ELBO}

Figure 1: An illustration of the EM algorithm, where g_t is the ELBO at step t.

- Maximization step: find

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}). \quad \{eq:Mstep\}$$

This holds for both discrete and continuous random variables.

Next, we show the log-likelihood function is monotonically increasing along the iteration.

Lemma 5.3. *We have* $l(\theta^{(t+1)}) \geq l(\theta^{(t)})$.

Proof. We have

$$\begin{aligned} l(\theta^{(t+1)}) &\geq \sum_{i=1}^{n} \mathrm{ELBO}(x_i, q_i^{(t)}, \theta^{(t+1)}) \\ &\geq \sum_{i=1}^{n} \mathrm{ELBO}(x_i, q_i^{(t)}, \theta^{(t)}) = l(\theta^{(t)}). \end{aligned}$$

We explain each step here. In the first inequality, we use the fact that the lower bound (37) holds for any choice of q_i. In the second inequality, we use the definition of $\theta^{(t+1)}$ in (39) being a maximizer. In the last identity, we use (38). □

Remark 5.4. A generalization of the EM algorithm is called variational inference [5].

5.2 Jackknife

Jackknife estimator The jackknife estimator of a parameter is found by systematically leaving out each observation from a dataset and calculating the parameter estimate over the remaining observations and then aggregating these calculations.

Let $X_1, \ldots, X_n$ be a random sample, and let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be some estimator of a parameter θ. Define the *leave-one-out estimator*

$$\hat{\theta}_{(i)} = \text{Estimator computed without } X_i$$

The Jackknife estimator of θ denoted by $\mathrm{JK}(\theta)$ is given by

$$\mathrm{JK}(\theta) = n\hat{\theta} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}.$$

Example 5.5. Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with mean θ. The Jackknife estimator of the sample mean is the sample mean. Let $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. Then

$$\mathrm{JK}(\theta) = \overline{X}.$$

Example 5.6. Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{Ber}(\theta)$. We aim to estimate θ^2. The maximum likelihood estimator of θ^2 is $Y_n = (\overline{X})^2$.

This estimator is biased since

$$\mathbb{E}Y_n = \theta^2 + \frac{1}{n}(\theta - \theta^2).$$

The Jackknife estimator is

$$Z_n = n\left(\frac{1}{n}\sum_i X_i\right)^2 - \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j\neq i} X_j\right)^2$$

and we can check $\mathbb{E}Z_n = 0$.

Jackknife resampling Estimating the bias and variance of an estimator is not always analytically feasible. Resampling techniques allow us to empirically approximate these quantities. The Jackknife resampling is a leave-one-out resampling procedure to approximate the bias and variance of the estimators.

Example 5.7 (Jackknife estimate of bias and variance)**.** Define:

$$\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$$

Then the Jackknife estimate of the bias of an estimator $\hat{\theta}$ is

$$\mathrm{Bias}_{\mathrm{jack}} = (n-1)(\bar{\theta} - \hat{\theta})$$

Jackknife estimate of variance:

$$\mathrm{Var}_{\mathrm{jack}}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \bar{\theta})^2.$$

5.3 Bootstrapping

Bootstrap is a powerful and flexible non-parametric resampling method that allows us to approximate the sampling distribution of an estimator without relying on strong parametric assumptions.

Given a dataset $X = \{X_1, X_2, \ldots, X_n\}$, the idea is to simulate the sampling distribution of an estimator $\hat{\theta}$ by resampling the data with replacement (allowing the same data to be chosen multiple times).

- Each **bootstrap sample** is given by:

  $$X_b^* = \{X_1^*, X_2^*, \ldots, X_n^*\} \quad \text{where } X_i^* \text{ drawn with replacement from } X.$$

  We can think of X_i^* is sampled from the empirical distribution $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

- Then we compute an estimator based on X_b^*:

  $$\hat{\theta}_b^* = \hat{\theta}(X_b^*)$$

- Repeat B times to obtain:

  $$\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$$

Example 5.8 (Bootstrap estimation)**.** Variance estimation:

$$v_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \bar{\theta}^*\right)^2$$

where

$$\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*.$$

We use v_{boot} as an approximation of the variance of $\hat{\theta}$.

Bias estimation:

$$\text{Bias}_{\text{boot}} = \bar{\theta}^* - \hat{\theta}.$$

6 Asymptotic evaluations

In this section, we discuss the asymptotic properties of estimators when the sample size $n \to \infty$.

6.1 Law of large numbers

Definition 6.1 (Convergence in probability)**.** A sequence of random variables $X_1, X_2, \cdots$ *converges in probability* to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0 \quad \text{or equivalently} \quad \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

{thm:WLLN}

Theorem 6.2 (Weak law of large numbers)**.** *Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\bar{X}_n$ converges in probability to μ, i.e.,*

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| < \epsilon) = 1, \quad \forall \epsilon > 0$$

Proof. By Chebychev inequality:

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) = \mathbb{P}((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(\bar{X}_n - \mu)^2}{\epsilon^2} = \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Hence,

$$\mathbb{P}(|\bar{X}_n - \mu| < \epsilon) = 1 - \mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2} \to 1 \quad \text{as } n \to \infty.$$

□

An estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is consistent if $\hat{\theta}$ converges in probability to θ as $n \to \infty$. Theorem 6.2

{thm:continc

Theorem 6.3 (Continous mapping theorem)**.** *Suppose $X_1, X_2, \cdots, X_n$ converges in probability to a random variable X and h is a continuous function. Then $h(X_1), h(X_2), \cdots$ converges in probability to $h(X)$.*

Proof. Since h is a continuous function, for any $\epsilon > 0$, there exists some $\delta > 0$ such that as long as $|x_n - x| < \delta$, $|h(x_n) - h(x)| < \epsilon$. Since $X_1, X_2, \cdots, X_n$ converges in probability to X, thus, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$1 \geq \mathbb{P}(|h(X_n) - h(x)| < \epsilon) \geq \mathbb{P}(|X_n - x| < \delta) \to 1 \text{ as } n \to \infty.$$

Thus, $\mathbb{P}(|h(X_n) - h(x)| < \epsilon) \to 1$ as desired. □

Example 6.4 (Consistency of S^2)**.** Suppose $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$.

The variance of the sample variance is given by

$$\text{Var}(S_n^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right).$$

When $\mu_4 = \mathbb{E}(X_i - \mu)^4 < \infty$, $\text{Var}(S_n^2) \to 0$. If $\text{Var}(S_n^2) \to 0$, then

$$\mathbb{P}\left(|S_n^2 - \sigma^2| > \varepsilon\right) \leq \frac{\mathbb{E}(S_n^2 - \sigma^2)^2}{\varepsilon^2} = \frac{\text{Var}(S_n^2)}{\varepsilon^2} \to 0.$$

Example 6.5 (Consistency of S)**.** The standard deviation S_n is a consistent estimator of σ from Theorem 6.3.

Note that $\mathbb{E}S_n \leq \sqrt{\mathbb{E}S_n^2} = \sigma$. The equal case holds if and only if S_n is a constant. So when $\sigma > 0$, we have $\mathbb{E}S_n < \sigma$. As an estimator of σ, S_n is biased, but the bias disappears asymptotically.

Definition 6.6 (Almost surely convergence)**.** A sequence of random variables $X_1, X_2, \cdots$ *converges almost surely* to a r.v. X if for every $\epsilon > 0$,

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1.$$

Theorem 6.7 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables with $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\bar{X}_n$ converges to μ almost surely, i.e.*

$$\mathbb{P}\left(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon\right) = 1, \quad \forall \epsilon > 0.$$

6.2 Central limit theorem

Definition 6.8 (Convergence in distribution)**.** A sequence of random variables $X_1, X_2, \cdots$ *converges in distribution* to a r.v. X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous.

{example:max

Example 6.9 (Maximum of uniforms)**.** If $X_1, \dots, X_n$ are i.i.d. Uniform$(0,1)$ and $X_{(n)} = \max_{1\le i\le n} X_i$. Then for any $t > 0$,

$$\mathbb{P}(X_{(n)} \le 1-t) = \mathbb{P}(X_i \le 1-t, 1 \le i \le n) = (1-t)^n.$$

Since

$$\mathbb{P}(|X_{(n)} - 1| \ge \varepsilon) = \mathbb{P}(X_{(n)} \le 1-\varepsilon) \to 0$$

as $n \to \infty$, this shows $X_{(n)}$ converges to 1 in probability.

Similarly, we have

$$\mathbb{P}(X_{(n)} \le 1 - t/n) = (1-t/n)^n \to e^{-t},$$

which is equivalent to

$$\mathbb{P}(n(1 - X_{(n)}) \le t) \to 1 - e^{-t}.$$

Therefore $(n(1-X_{(n)})$ converges in distribution to Exponential (1).

Theorem 6.10. *If a sequence of random variables X_n converges in probability to a random variable X, then X_n also converges in distribution to X.*

{thm:converg

Theorem 6.11. *The sequence of random variables $X_1, X_2, \cdots$ converges in probability to a constant μ if and only if the sequence also converges in distribution to μ. That is*

$$P(|X_n - \mu| > \epsilon) \to 0 \quad \forall \epsilon > 0 \iff P(X_n \le x) \to \begin{cases} 0 & \textit{if } x < \mu \\ 1 & \textit{if } x > \mu \end{cases}$$

Lemma 6.12 (portmanteau lemma)**.** *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and X a random variable in $\mathbb{R}$. Then X_n converges to X in distribution if and only if*

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X)$$

for any bounded continuous functions f.

{thm:CLT}

Theorem 6.13 (Central limit theorem)**.** *Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with $EX_i = \mu$ and $Var(X_i) = \sigma^2 > 0$. Let $G_n(x)$ denote the cdf of $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$, then for any $-\infty < x < \infty$,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

That is, $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ converge in distribution to standard normal random variable.

Theorem 6.14 (Slutsky's theorem)**.** *If $X_n \xrightarrow{d} X$ in distribution and $Y_n \xrightarrow{p} a$, where a is a constant, then*

(a) $Y_n X_n \xrightarrow{d} aX$;

(b) $X_n + Y_n \xrightarrow{d} X + a$

(c) $X_n / Y_n \xrightarrow{d} X/a$ *when* $a \ne 0$.

Remark 6.15. The requirement that Y_n converges to a constant is important. For example, let $X_n \sim \text{Uniform}(0,1)$ and $Y_n = -X_n$. The sum $X_n + Y_n = 0$ for all values of n. Moreover, $Y_n \xrightarrow{p} \text{Uniform}(-1,0)$, but $X_n + Y_n$ does not converge in distribution to $X+Y$, where $X \sim \text{Uniform}(0,1)$, $Y \sim \text{Uniform}(-1,0)$, and X and Y are independent.

For convergence in probability, the following statement holds for random variables X, Y:

Lemma 6.16. *If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ then $X_n Y_n \xrightarrow{p} XY$ and $X_n + Y_n \xrightarrow{p} X + Y$.*

With Slutsky's theorem, we can replace the variance with the sample variance in the CLT statement.

Example 6.17 (CLT with sample variance)**.** From the CLT, we have

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

Since $S_n^2 \xrightarrow{p} \sigma^2$, we have $\frac{\sigma}{S_n} \xrightarrow{p} 1$. Then by Slutsky's theorem

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n} \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

We introduce two theorems that relax the i.i.d. assumption in Theorem 6.13

Theorem 6.18 (Lyapunov CLT)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of independent random variables with means $\mu_i = \mathbb{E}[X_i]$ and variances $\sigma_i^2 = \text{Var}(X_i)$, and define the partial sums*

$$S_n = \sum_{i=1}^n X_i, \quad \text{and} \quad s_n^2 = \sum_{i=1}^n \sigma_i^2.$$

Suppose that $s_n^2 \to \infty$ as $n \to \infty$, and that there exists $\delta > 0$ such that Lyapunov's condition is satisfied:

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}\left[|X_i - \mu_i|^{2+\delta}\right] = 0.$$

Then

$$\frac{S_n - \sum_{i=1}^n \mu_i}{s_n} \xrightarrow{d} \mathcal{N}(0,1).$$

The Lindeberg CLT relaxes the i.i.d. assumption and replaces it with a Lindeberg condition. One can check that Lyapunov's condition implies Lindeberg's condition (hint: apply Holder's inequality).

Theorem 6.19 (Lindeberg CLT)**.** *Let $X_1, \dots, X_n$ be independent with finite variance. Assume that $s_n > 0$ for all $n \geq 1$. If for any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^n \mathbb{E}(|X_k|^2 1_{|X_k| > \epsilon s_n}) = 0,$$

then the random variable

$$\frac{S_n - \sum_{i=1}^n \mu_i}{\sigma_n}$$

converge in distribution to a standard Gaussian random variable.

If $X_1, \ldots, X_n$ are i.i.d. vectors in $\mathbb{R}^d$ (where d is fixed), we have the following multivariate CLT.

Theorem 6.20 (Multivariate CLT)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$ with mean vector $\mu = \mathbb{E}[X_i] \in \mathbb{R}^d$ and covariance matrix $\Sigma = Cov(X_i) \in \mathbb{R}^{d \times d}$. Define the sample mean:*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then, as $n \to \infty$,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

To prove the Multivariate CLT, we can use the following useful theorem.

Theorem 6.21 (Cramér–Wold Theorem)**.** *A sequence of random vectors $Y_n \in \mathbb{R}^d$ converges in distribution to a random vector Y if and only if for every $\lambda = (\lambda_1, \ldots, \lambda_d)^\top \in \mathbb{R}^d$, the scalar sequence $\lambda^\top Y_n$ converges in distribution to $\lambda^\top Y$.*

Proof of Multivariate CLT. Fix $\lambda \in \mathbb{R}^d$, and consider:

$$\lambda^\top \sqrt{n}(\bar{X}_n - \mu) = \sqrt{n} \left(\lambda^\top \bar{X}_n - \lambda^\top \mu \right).$$

Define the scalar random variable:

$$Y_i := \lambda^\top X_i.$$

Then $\{Y_i\}$ are i.i.d. real-valued random variables with

$$\mathbb{E}[Y_i] = \lambda^\top \mu, \quad \mathrm{Var}(Y_i) = \lambda^\top \Sigma \lambda.$$

By the classical (univariate) Central Limit Theorem:

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} Y_i - \lambda^\top \mu \right) \xrightarrow{d} \mathcal{N}(0, \lambda^\top \Sigma \lambda).$$

That is,

$$\lambda^\top \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \lambda^\top \Sigma \lambda).$$

Since the above holds for every $\lambda \in \mathbb{R}^d$, the Cramér–Wold theorem implies that:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}_d(0, \Sigma).$$

□

Berry–Esseen inequality, gives a more quantitative CLT because it also specifies the rate at which this convergence takes place by giving a bound on the maximal error of approximation between the normal distribution and the true distribution of the scaled sample mean. The approximation is measured by the Kolmogorov–Smirnov distance.

{thm:BerryEs

Theorem 6.22 (Berry–Esseen Theorem)**.** *Let $X_1, X_2, \dots, X_n$ be independent and identically distributed (i.i.d.) random variables with mean $\mu = \mathbb{E}[X_i]$, variance $\sigma^2 = \mathrm{Var}(X_i) > 0$, and third absolute central moment $\rho = \mathbb{E}[|X_i - \mu|^3] < \infty$. Define the standardized sum:*

$$S_n = \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} (X_i - \mu).$$

Let F_n be the cumulative distribution function (CDF) of S_n, and let Φ be the CDF of the standard normal distribution $\mathcal{N}(0,1)$. Then there exists a constant $C > 0$ such that for all n,

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \le \frac{C\rho}{\sigma^3 \sqrt{n}}.$$

Non-asymptotic results can also be derived by using concentration inequalities. We state a simple inequality about the bounds on the sum of independent bounded variables:

Lemma 6.23 (Hoeffding's inequality)**.** *Let $X_1, X_2, \dots, X_n$ be independent random variables such that for each i, the variable X_i is almost surely bounded:*

$$a_i \le X_i \le b_i.$$

Define the sum of centered variables:

$$S_n = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]).$$

Then, for any $t > 0$, Hoeffding's inequality states:

$$\mathbb{P}\left(|S_n| \ge t\right) \le 2 \exp\left(- \frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

This gives a probability bound on how close S_n to 0 at every n, and the tail probability behaves like a Gaussian CDF of the form $\exp(-ct^2)$.

Beyond the case for bounded random variables, there are many inequalities available to control the concentration of a sum of independent random variables, including Bernstein's inequality, Bennett's inequality, Chernoff's inequality, etc; see [11].

6.3 The Delta method

We now consider the distribution of some function of a random variable. The Delta method is a method of showing CLTs for a function of a random variable, based on Taylor series approximation.

We will need the following continuous mapping theorem for convergence in distribution.

{thm:contin

Theorem 6.24 (Continuous mapping theorem)**.** *Let $X_n \xrightarrow{d} X$, and let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then,*

$$g(X_n) \xrightarrow{d} g(X).$$

{thm:Delta}

Theorem 6.25 (Delta method)**.** *Let Y_n be a sequence of random variables such that $\sqrt{n}(Y_n - \theta) \to N(0, \sigma^2)$ in distribution. For a given function g and θ, suppose $g'(\theta)$ exists and is not 0, then*

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} N(0, \sigma^2[g'(\theta)]^2).$$

Proof. From Slutsky's theorem,

$$\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

implies that $Y_n - \theta \xrightarrow{d} 0$. Then from Theorem 6.11, it implies that $Y_n \xrightarrow{p} \theta$.

Since g is differentiable at θ, by the mean value theorem, there exists a point ξ_n between Y_n and θ such that

$$g(Y_n) = g(\theta) + g'(\xi_n)(Y_n - \theta).$$

By continuity of g' at θ, $g'(\xi_n) \xrightarrow{p} g(\theta)$. Then with Slutsky's theorem,

$$\sqrt{n}(g(Y_n) - g(\theta)) = g'(\xi_n)\sqrt{n}(Y_n - \theta) \xrightarrow{d} g'(\theta)N(0, \sigma^2) = N(0, \sigma^2[g'(\theta)]^2).$$

□

What if $g'(\theta) = 0$? We have the following theorem using a second-order Taylor expansion:

{thm:second_

Theorem 6.26 (Second-order Delta method)**.** *Let Y_n be a sequence of random variables that satisfies*

$$\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

For a given function g and a specific value of θ, suppose that

$$g'(\theta) = 0, \quad g''(\theta) \text{ exists and is not } 0.$$

Then

$$n[g(Y_n) - g(\theta)] \xrightarrow{d} \frac{\sigma^2 g''(\theta)}{2}\chi_1^2.$$

{lem:AN_cons

Lemma 6.27 (Asymptotic normality implies consistency)**.** *Assume $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$, then $Y_n \xrightarrow{p} \theta$.*

Proof of Lemma 6.27. Let $Z \sim N(0, \sigma^2)$. For any $\delta > 0$, there exists a constant M depending on δ such that $\mathbb{P}(|Z| \leq M) \leq \delta$. For any $\epsilon > 0$, we have

$$\begin{aligned}\limsup_n \mathbb{P}(|Y_n - \theta| > \varepsilon) &= \limsup_n \mathbb{P}(|\sqrt{n}(Y_n - \theta)| > \sqrt{n}\varepsilon) \\ &\leq \limsup_n \mathbb{P}(|\sqrt{n}(Y_n - \theta)| > M) \\ &= \mathbb{P}(|Z| \leq M) \leq \delta.\end{aligned}$$

Since δ is arbitrary, we conclude $\lim_n \mathbb{P}(|Y_n - \theta| > \varepsilon) = 0$. Therefore $Y_n \xrightarrow{p} \theta$.

Another easy way to show this is to use Slutsky's theorem. Since $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ and $1/\sqrt{n} \xrightarrow{p} 0$. Multiplying the two sequences yields:

$$Y_n - \theta = \frac{1}{\sqrt{n}}(\sqrt{n}(Y_n - \theta)) \xrightarrow{d} 0.$$

Since convergence in distribution to a constant is equivalent to convergence in distribution to a constant (Theorem 6.11), we obtain $Y_n \xrightarrow{d} \theta$. □

Proof of Theorem 6.26. Because g is twice differentiable at θ, we can apply a second-order Taylor expansion of $g(Y_n)$ around θ:

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{1}{2}g''(\xi_n)(Y_n - \theta)^2$$

Given $g'(\theta) = 0$, this simplifies to:

$$g(Y_n) - g(\theta) = \frac{1}{2}g''(\xi_n)(Y_n - \theta)^2.$$

Multiplying both sides by n, we get:

$$n[g(Y_n) - g(\theta)] = \frac{1}{2}g''(\xi_n)n(Y_n - \theta)^2.$$

Since $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$, then by Continuous Mapping Theorem 6.24,

$$n(Y_n - \theta)^2 = (\sqrt{n}(Y_n - \theta))^2 \xrightarrow{d} \sigma^2\chi_1^2,$$

because the square of a standard normal random variable has the chi-squared distribution with one degree of freedom. Then by Slutskly's theorem,

$$n[g(Y_n) - g(\theta)] \xrightarrow{d} \frac{\sigma^2 g''(\theta)}{2}\chi_1^2.$$

□

We can generalize the first-order Delta method to the multivariate setting:

Theorem 6.28 (Multivariate Delta method)**.** *Let $\mathbf{X}_n \in \mathbb{R}^k$ be a sequence of random vectors such that*

$$\sqrt{n}(\mathbf{X}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma),$$

for some $\boldsymbol{\theta} \in \mathbb{R}^k$ and positive semi-definite matrix Σ. Let $g : \mathbb{R}^k \to \mathbb{R}$ be differentiable at $\boldsymbol{\theta}$, and let

$$\nabla g(\boldsymbol{\theta}) = \left(\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_k}\right)\Bigg|_{\mathbf{x}=\boldsymbol{\theta}} \in \mathbb{R}^k$$

denote the gradient row vector of g at $\boldsymbol{\theta}$. Then, if $\nabla g(\boldsymbol{\theta})\,\Sigma\,\nabla g(\boldsymbol{\theta})^\top > 0$,

$$\sqrt{n}\,(g(\mathbf{X}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(\boldsymbol{\theta})\,\Sigma\,\nabla g(\boldsymbol{\theta})^\top\right).$$

The proof is based on the multivariate Taylor expansion.

6.4 Consistency

Definition 6.29. 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 every $\varepsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathbb{P}_\theta(|W_n - \theta| < \varepsilon) = 1.$$

Theorem 6.30. *If W_n is a sequence of estimators of a parameter θ satisfying*

- $\lim_{n\to\infty} \mathrm{Var}_\theta W_n = 0$,
- $\lim_{n\to\infty} \mathrm{Bias}_\theta W_n = 0$,

for every $\theta \in \Theta$, then W_n is a consistent sequence of estimators of θ.

Proof. From Chebyshev's inequality,

$$P_\theta(|W_n - \theta| \geq \epsilon) \leq \frac{\mathbb{E}_\theta[(W_n-\theta)^2]}{\epsilon^2},$$

so if, for every $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathbb{E}_\theta[(W_n-\theta)^2] = 0,$$

then the sequence of estimators is consistent. Furthermore, by the bias-variance decomposition,

$$\mathbb{E}_\theta[(W_n-\theta)^2] = \mathrm{Var}_\theta W_n + [\mathrm{Bias}_\theta W_n]^2.$$

Based on our assumption, W_n is consistent. □

If we have a consistent sequence of estimators, we can create many consistent sequences from the following theorem:

Theorem 6.31. *Let W_n be a sequence of estimators for θ. Let $\{a_n\}, \{b_n\}$ be sequences such that $a_n \to 1, b_n \to 0$. Then $U_n = a_n W_n + b_n$ is a consistent sequence of estimators of θ.*

Proof. By checking the definition of consistency, this theorem holds. □

Consistency of the Maximum Likelihood Estimator Let $X, X_1, X_2, \ldots$ be i.i.d. with common density f_{θ_0}, $\theta_0 \in \Theta$, and let l_n be the log-likelihood function for the first n observations:

$$l_n(\theta) = \log \prod_{i=1}^n f_\theta(X_i) = \sum_{i=1}^n \log f_\theta(X_i).$$

Then the maximum likelihood estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ from the first n observations will maximize l_n. For regularity, assume $f_\theta(x)$ is continuous in θ.

Definition 6.32. *The Kullback–Leibler information is defined as*

$$I(\theta_0, \theta) = \mathbb{E}_{\theta_0} \log \left[\frac{f_{\theta_0}(X)}{f_\theta(X)}\right].$$

It can be viewed as a measure of the information discriminating between θ_0 and θ when θ_0 is the true value of the unknown parameter.

{lem:KL}

Lemma 6.33. *If $f_{\theta_0} \neq f_\theta$, then $I(\theta_0, \theta) > 0$.*

Proof. By Jensen's inequality,

$$\begin{aligned}
-I(\theta_0,\theta) = \mathbb{E}_{\theta_0} \log\left[\frac{f_\theta(X)}{f_{\theta_0}(X)}\right] &\le \log \mathbb{E}_{\theta_0}\left[\frac{f_\theta(X)}{f_{\theta_0}(X)}\right] \\
&= \log \int \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x)\,dx \\
&= \log \int f_\theta(x)\,dx = \log 1 = 0.
\end{aligned}$$

The equality holds if and only $f_\theta(X)/f_{\theta_0}(X)$ is a constant. Since $f_{\theta_0} \neq f_\theta$, we have $I(\theta_0,\theta) > 0$. □

The next result gives consistency for the maximum likelihood estimator when Θ is compact. Define

$$W(\theta) = \log\left[\frac{f_\theta(X)}{f_{\theta_0}(X)}\right].$$

We need the following technical lemma to finish the proof, see [8, Theorem 9.2, Theorem 9.3(2)].

{lem:random_

Lemma 6.34 (Weak Law for Random Functions). *The following two properties of continuous random functions hold:*

1. *Let $W_1,\dots,W_n$ be i.i.d. random continuous functions on a compact set K with mean μ and $\mathbb{E}\|W\|_\infty = \mathbb{E}\sup_{x\in K}|W(x)| < \infty$. Let $\overline{W}_n = \frac{1}{n}(W_1+\cdots+W_n)$, then*

   $$\|\overline{W}_n - \mu\|_\infty = \sup_{x\in K}|W_n(x) - \mu| \to 0$$

   in probability.

2. *Let W_n be random continuous functions defined on a compact set K and suppose $\|W_n - g\| \to 0$ in probability with g a nonrandom continuous function on K. If g achieves its maximum at a unique value t^* and t_n are random variables maximizing W_n such that*

   $$W_n(t_n) = \sup_{t\in K} W_n(t),$$

   then $t_n \to t^*$ in probability.

Theorem 6.35. *If Θ is compact, $\mathbb{E}_{\theta_0}\|W\|_\infty < \infty$, $f_\theta(x)$ is a continuous function of θ for all x, and $f_{\theta_0} \neq f_\theta$ for all $\theta \neq \theta_0$, then under $\hat{\theta}_n$ is a consistent estimator of θ.*

Proof. Let $W_i(\theta) = \log\left(f_\theta(X_i)/f_{\theta_0}(X_i)\right)$, then under P_{θ_0}, $W_1, W_2, \dots$ are i.i.d. random continuous function in θ with mean $\mu(\theta) = -I(\theta_0,\theta)$. Note that $\mu(\theta_0) = 0$ and $\mu(\theta) < 0$ for $\theta \neq \theta_0$ by Lemma 6.33, so μ has a unique maximum at θ_0. Since

$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{j=1}^{n} W_i(\theta) = \frac{l_n(\theta) - l_n(\theta_0)}{n},$$

$\hat{\theta}_n$ maximizes $\overline{W}_n$. By Lemma 6.34 (1),

$$\|\overline{W}_n - \mu(\theta)\| = \|\overline{W}_n - (-I(\theta,\theta_0))\|_\infty \to 0$$

in probability. Since $\hat{\theta}_n$ maximizes $\overline{W}_n$ and θ_0 maximizes $\mu(\theta)$, the from the second claim of Lemma 6.34, $\hat{\theta}_n \to \theta_0$ in probability. □

By the Dominated Convergence Theorem, the theorem above can be extended to unbounded domain Θ. See [8, Theorem 9.11].

Theorem 6.36. *Suppose $\Theta = \mathbb{R}^p$, $f_\theta(x)$ is a continuous function of θ for all x, $f_\theta \neq f_{\theta_0}$ for all $\theta \neq \theta_0$, and $f_\theta(x) \to 0$ as $\theta \to \infty$. If $\mathbb{E}_{\theta_0}\|\mathbf{1}_K W\|_\infty < \infty$ for any compact set $K \subset \mathbb{R}^p$, and if $\mathbb{E}_{\theta_0}\sup_{\|\theta\|>a} W(\theta) < \infty$ for some $a > 0$, then under P_θ, $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

Bias versus consistency A sequence of estimators $\hat{\theta}_n$ is *asymptotically unbiased* if $\mathbb{E}_\theta \hat{\theta}_n \to \theta$.

Example 6.37. The two examples show that the two concepts of consistency and asymptotically unbiased are different:

1. Asymtotically unbiased but not consistent: Let $X_1, \ldots, X_n$ be i.i.d. with mean μ. Then $W_n = X_n$ is an unbiased estimator of μ but not consistent.

2. Consistent but not asymptotically biased: Let T_n be a sequence of estimators for θ such that $T_n = \begin{cases} \hat{\theta}_n & \text{with probability } 1 - 1/n \\ n\delta + \hat{\theta}_n & \text{with probability } 1/n. \end{cases}$ where $\hat{\theta}_n \xrightarrow{p} \theta$ and $\mathbb{E}\hat{\theta}_n = \theta$. Then $\mathbb{E}T_n = \theta + \delta$ and the bias does not converge to zero. And we can check $T_n \xrightarrow{p} \theta$.

6.5 Efficiency

Recall the best unbiased estimators (UMVUE) achieve the minimal mean-squared error among all unbiased estimators. Asymptotic efficiency is the asymptotic analog of UMVUE.

Definition 6.38 (Asymptotic efficiency)**.** A sequence of estimators W_n is *asymptotically efficient* for a parameter $\tau(\theta)$ if

$$\sqrt{n}[W_n - \tau(\theta)] \xrightarrow{d} \mathcal{N}[0, v(\theta)]$$

in distribution and

$$v(\theta) = \frac{[\tau'(\theta)]^2}{\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)};$$

that is, the asymptotic variance of W_n achieves the Cramér–Rao Lower Bound.

When $\tau'(\theta) = 1, \nu(\theta) = \frac{1}{I_X(\theta)}$, where $I_X(\theta)$ is the Fisher information of X. This definition suggests the asymptotic variance of $\mathrm{Var}(W_n) \approx \frac{1}{nI_X(\theta)}$. Namely W_n asymptotically achieves the Cramér–Rao Lower Bound.

Asymptotic normality of MLE We first show that under regularity conditions, MLE is asymptotically efficient.

{thm:AN_MLE}

Theorem 6.39 (Asymptotic normality of MLE)**.** *Assume:*

1. *Variables $X, X_1, X_2, \ldots$ are i.i.d. with common density f_θ, $\theta \in \Omega \subset \mathbb{R}$.*

2. *The support $A = \{x : f_\theta(x) > 0\}$ is independent of θ.*

3. *For every $x \in A$, $\partial^2 f_\theta(x)/\partial\theta^2$ exists and is continuous in θ.*

4. *Let $W(\theta) = \log f_\theta(X)$. The Fisher information $I(\theta)$ from a single observation exists, is finite, and can be found using either*

$$I(\theta) = \mathbb{E}_\theta[W'(\theta)^2] \quad \textit{or} \quad I(\theta) = -\mathbb{E}_\theta[W''(\theta)].$$

   Also, $\mathbb{E}_\theta[W'(\theta)] = 0$.

5. *For every θ in the interior of Ω there exists $\epsilon > 0$ such that $\mathbb{E}_\theta \left\|\mathbb{1}_{[\theta-\epsilon,\theta+\epsilon]} W''\right\|_\infty < \infty$.*

6. *The maximum likelihood estimator $\hat{\theta}_n$ is consistent.*

Then for any θ in the interior of Ω,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)$$

under $\mathbb{P}_\theta$ as $n \to \infty$.

Remark 6.40. Asymptotic normality implies consistency, see Lemma 6.27.

The following technical lemma shows that, when proving convergence in distribution, we only need to consider what happens under a sequence of events with probabilities converging to one.

{lem:good_ev

Lemma 6.41. *Suppose $Y_n \xrightarrow{d} Y$ and $\mathbb{P}(B_n) \to 1$ as $n \to \infty$. Then for arbitrary random variables Z_n,*

$$Y_n \mathbf{1}_{B_n} + Z_n \mathbf{1}_{B_n^c} \xrightarrow{d} Y$$

as $n \to \infty$.

Proof. Since $\mathbf{1}_{B_n} \xrightarrow{p} 1$, by Slutsky's theorem, $Y_n \mathbf{1}_{B_n} \xrightarrow{d} Y$. Since

$$\mathbb{P}(|Z_n \mathbf{1}_{B_n^c}| > \varepsilon) \leq \mathbb{P}(\mathbf{1}_{B_n^c} = 1) = \mathbb{P}(B_n^c) = 1 - \mathbb{P}(B_n) \to 0,$$

we have $Z_n \mathbf{1}_{B_n^c} \xrightarrow{p} 0$. By Slutsky's theorem again,

$$Y_n \mathbf{1}_{B_n} + Z_n \mathbf{1}_{B_n^c} \xrightarrow{d} Y$$

as $n \to \infty$. □

Now we are ready to prove Theorem 6.39.

Proof of Theorem 6.39. From Assumption 5, there exists a closed interval in the interior of Ω: $[\theta - \epsilon, \theta + \epsilon] \subset \Omega^0$ such that

$$\mathbb{E}_\theta \left\|\mathbb{1}_{[\theta-\epsilon,\theta+\epsilon]} W''\right\|_\infty < \infty.$$

Let B_n be the event that $\hat{\theta}_n \in [\theta - \epsilon, \theta + \epsilon]$. Because $\hat{\theta}_n$ is consistent (Assumption 6), $\mathbb{P}_\theta(B_n) \to 1$, and since $\hat{\theta}_n$ maximizes

$$n\overline{W_n}(w) = l_n(w) = \sum_{i=1}^{n} \log f(X_i|w),$$

on the event B_n, we have $\overline{W}'_n(\hat{\theta}_n) = 0$ and the Taylor expansion of $\overline{W}'_n(\hat{\theta}_n)$ about θ gives

$$\overline{W}'_n(\hat{\theta}_n) = \overline{W}'_n(\theta) + \overline{W}''_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta),$$

where $\tilde{\theta}_n$ is an intermediate value between $\hat{\theta}_n$ and θ. Setting the left-hand side of this equation to zero, we obtain that under event B_n, the following identity holds:

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\sqrt{n}\overline{W}'_n(\theta)}{-\overline{W}''_n(\tilde{\theta}_n)}. \tag{52}$$ {eq:MLE_CLT2

Step 1: We first show that the right-hand side of (6.5) converges to a limit.

By Assumption 4, $\overline{W}'_n(\theta) = \frac{1}{n}\sum_{i=1}^n (\log W_i(\theta))'$ is the average of n i.i.d., mean zero random variables with variance $I(\theta) < \infty$. By the central limit theorem,

$$\sqrt{n}\overline{W}'_n(\theta) \xrightarrow{d} N(0, I(\theta)).$$

Turning to the denominator of (6.5), since $|\tilde{\theta}_n - \theta| \le |\hat{\theta}_n - \theta|$ on the event B_n, and $\hat{\theta}_n \xrightarrow{p} \theta$, we have

$$\tilde{\theta}_n \xrightarrow{p} \theta.$$

By Part (1) of Lemma 6.34, since $W''_n(w)\mathbb{1}_{[\theta-\varepsilon,\theta+\varepsilon]}(w)$ is a continuous function (Assumption 3) on a compact set $[\theta - \epsilon, \theta + \epsilon]$, we have

$$\left\| \mathbb{1}_{[\theta-\varepsilon,\theta+\varepsilon]} \left(\overline{W}''_n - \mu \right) \right\|_\infty \xrightarrow{p} 0, \tag{53}$$ {eq:Uniform_

where $\mu(\omega) = \mathbb{E}_\theta W''(\omega)$. From the uniform convergence in (53), since $\mathbb{P}(\tilde{\theta}_n \in [\theta - \epsilon, \theta + \epsilon]) \to 1$, we have

$$W''_n(\tilde{\theta}_n) - \mu(\tilde{\theta}_n) \xrightarrow{p} 0.$$

From Assumption 3, $\mu(w)$ is continuous, and by continuous mapping theorem, $\mu(\tilde{\theta}_n) \xrightarrow{p} \mu(\theta)$. Therefore

$$W''_n(\tilde{\theta}_n) \xrightarrow{p} \mu(\theta) = \mathbb{E}_\theta W''(\theta) = I(\theta).$$

By Slutsky's theorem,

$$\frac{\sqrt{n}\overline{W}'_n(\theta)}{-\overline{W}''_n(\tilde{\theta}_n)} \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right).$$

Step 2: Now we show the left-hand side of (6.5) converges to the same limit. Since the behavior of $\hat{\theta}_n$ on B_n^c cannot affect convergence in distribution, we have from Lemma 6.41,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right),$$

as $n \to \infty$. □

Some of the regularity conditions are necessary. For example, when the support of $f_\theta(x)$ depends on θ, the asymptotic normality of MLE might not hold.

Example 6.42 (Non-normal limiting distribution for MLE)**.** Let $X_1, \dots, X_n$ be i.i.d. sampled from Uniform$(0, \theta)$. The MLE is $\hat{\theta}_n = X_{(n)} = \max_{1 \le i \le n} X_i$. With the same calculation in Example 6.9,

$$\begin{aligned}
\mathbb{P}(n(\theta - \hat{\theta}_n) \le t) &= \mathbb{P}(X_{(n)} \ge \frac{-t}{n} + \theta) \\
&= 1 - \mathbb{P}(X_{(n)} \le \frac{-t}{n} + \theta) \\
&= 1 - (1 - \frac{t}{n\theta})^n \to 1 - e^{-t/\theta}.
\end{aligned}$$

Hence $n(\theta - \hat{\theta}) \to \exp(\theta)$. The scaling factor is n instead of $\sqrt{n}$ and the limiting distribution is exponential, not Gaussian.

Superefficiency Suppose $X_1, X_2, \dots$ are i.i.d. with common density f_θ, $\theta \in \Omega$. By the Cramér–Rao lower bound (Theorem 4.12), if $\delta_n = \delta_n(X_1, \dots, X_n)$ is an unbiased estimator of $g(\theta)$, then

$$\mathrm{Var}_\theta(\delta_n) \ge \frac{[g'(\theta)]^2}{nI_X(\theta)}.$$

The definition of asymptotic efficiency seems to suggest if

$$\sqrt{n}(\delta_n - g(\theta)) \Rightarrow \mathcal{N}(0, \sigma^2(\theta)),$$

then

$$\mathrm{Var}_\theta(\sqrt{n}\delta_n) \to \sigma^2(\theta) \ge \frac{[g'(\theta)]^2}{I(\theta)}.$$

However, we will see that the asymptotic variance is not always bounded by the Cramér-Rao lower bound.

Example 6.43 (The Hodge estimator)**.** Let $X_1, X_2, \dots$ be i.i.d. from $N(\theta, 1)$ and take $\overline{X}_n = (X_1 + \cdots + X_n)/n$. Define δ_n by

$$\delta_n = \begin{cases} \overline{X}_n, & |\overline{X}_n| \ge 1/n^{1/4}, \\ a\overline{X}_n, & |\overline{X}_n| < 1/n^{1/4}, \end{cases}$$

where a is some constant in $(0, 1)$. Let us compute the limiting distribution of $\sqrt{n}(\delta_n - \theta)$.

Suppose $\theta < 0$. Fix x and consider

$$\mathbb{P}_\theta(\sqrt{n}(\delta_n - \theta) \le x) = \mathbb{P}(\delta_n \le \theta + x/\sqrt{n}).$$

Since $\theta + x/\sqrt{n} \to \theta < 0$ and $-1/n^{1/4} \to 0$, for n sufficiently large, $\theta + x/\sqrt{n} < -1/n^{1/4}$, and then

$$\mathbb{P}_\theta(\sqrt{n}(\delta_n - \theta) \le x) = \mathbb{P}_\theta(\overline{X}_n \le \theta + x/\sqrt{n}) = \Phi(x).$$

To show the identity above holds, note that $|\delta_n| \in (0, an^{-1/4}) \cup [n^{1/4}, \infty)$, so for sufficiently large n, the two events are the same $\{\delta_n \le \theta + x/\sqrt{n}\} = \{\overline{X}_n \le \theta + x/\sqrt{n}\}$.

In this case, $\sqrt{n}(\delta_n - \theta) \Rightarrow N(0, 1)$. A similar calculation shows that $\sqrt{n}(\delta_n - \theta) \Rightarrow N(0, 1)$ when $\theta > 0$.

Suppose now that $\theta = 0$. Fix x and consider

$$\mathbb{P}_0(\sqrt{n}\delta_n \leq x) = \mathbb{P}_0(\delta_n \leq x/\sqrt{n}).$$

For n sufficiently large, $a|x| < n^{1/4}$, then

$$\mathbb{P}_0(\sqrt{n}\delta_n \leq x) = \mathbb{P}_0(a\overline{X}_n \leq x/\sqrt{n}) = \Phi(x/a).$$

This is the cumulative distribution function for $N(0, a^2)$. So when $\theta = 0$,

$$\sqrt{n}(\delta_n - \theta) \Rightarrow N(0, a^2).$$

Altogether, we have

$$\sqrt{n}(\delta_n - \theta) \Rightarrow N(0, \sigma^2(\theta)),$$

where

$$\sigma^2(\theta) = \begin{cases} 1, & \theta \neq 0; \\ a^2, & \theta = 0. \end{cases}$$

This estimator is called "superefficient" since the variance of the limiting distribution when $\theta = 0$ is smaller than $1/I(\theta) = 1$.

Because $\sqrt{n}(\overline{X}_n - \theta) \sim N(0, 1)$, This seems to suggest that δ_n may be a better estimator than $\overline{X}_n$ when n is large.

Consider the mean squared error. Since $R(\theta, \overline{X}_n) = \mathbb{E}_\theta(\overline{X}_n - \theta)^2 = 1/n$, $nR(\theta, \overline{X}_n) = 1$. It can be shown that

$$nR(\theta, \delta_n) \to \begin{cases} 1, & \theta \neq 0; \\ a^2, & \theta = 0, \end{cases}$$

Note that δ_n never takes values in the interval

$$\left(\frac{a}{n^{1/4}}, \frac{1}{n^{1/4}}\right).$$

If we define

$$\theta_n = \frac{1 + a}{2n^{1/4}}$$

then

$$(\delta_n - \theta_n)^2 \geq \left(\frac{1 - a}{2n^{1/4}}\right)^2 = \frac{(1 - a)^2}{4\sqrt{n}}.$$

From this,

$$nR(\theta_n, \delta_n) \geq n \cdot \frac{(1 - a)^2}{4\sqrt{n}} = \frac{(1 - a)^2}{4}\sqrt{n} \to \infty,$$

as $n \to \infty$. This shows that for large n the risk of δ_n at θ_n will be much worse than the risk of $\overline{X}_n$ at θ_n.

6.6 Asymptotic relative efficiency

Definition 6.44. If two estimators W_n and V_n satisfy

$$\sqrt{n}[W_n - \tau(\theta)] \to \mathcal{N}[0, \sigma_W^2]$$

$$\sqrt{n}[V_n - \tau(\theta)] \to \mathcal{N}[0, \sigma_V^2]$$

in distribution, the *asymptotic relative efficiency* (ARE) of V_n with respect to W_n is

$$\mathrm{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}.$$

Example 6.45 (AREs of Poisson estimators)**.** Suppose that $X_1, X_2, \ldots$ are iid Poisson(λ), and we are interested in estimating $P(X = 0) = e^{-\lambda}$, and a natural estimator comes from defining $Y_i = I(X_i = 0)$ and using

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

The Y_is are Bernoulli($e^{-\lambda}$), and hence it follows that

$$\mathbb{E}(\hat{\tau}) = e^{-\lambda} \quad \text{and} \quad \mathrm{Var}(\hat{\tau}) = \frac{e^{-\lambda}(1 - e^{-\lambda})}{n}.$$

$$\sqrt{n}(\hat{\tau} - e^{-\lambda}) \to \mathcal{N}[0, e^{-\lambda}(1 - e^{-\lambda})].$$

Alternatively, the MLE of $e^{-\lambda}$ is $e^{-\hat{\lambda}}$, where $\hat{\lambda} = \sum_i X_i/n$ is the MLE of λ. We have

$$\sqrt{n}(\hat{\lambda} - \lambda) \to \mathcal{N}(0, \lambda).$$

Using Delta Method (Theorem 6.25), we have that

$$\sqrt{n}(e^{-\hat{\lambda}} - e^{-\lambda}) \to \mathcal{N}(0, \lambda e^{-2\lambda})$$

in distribution. The ARE of $\hat{\tau}$ with respect to the MLE $e^{-\hat{\lambda}}$ is

$$\mathrm{ARE}(\hat{\tau}, e^{-\hat{\lambda}}) = \frac{\lambda e^{-2\lambda}}{e^{-\lambda}(1 - e^{-\lambda})} = \frac{\lambda}{e^{\lambda} - 1}.$$

This ARE is maximized at $\lambda = 0$ with the value 1 and decreases as λ increases.

Since the MLE is typically asymptotically efficient, another estimator cannot hope to beat its asymptotic variance. However, other estimators may have other desirable properties (ease of calculation, robustness to underlying assumptions) that make them desirable.

Median and quantiles Let $X_1, \dots, X_n$ be random variables. These variables, arranged in increasing order, $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$, are called *order statistics*. The first order statistic $X_{(1)}$ is the smallest value,

$$X_{(1)} = \min\{X_1, \dots, X_n\},$$

and the last order statistic $X_{(n)}$ is the largest value,

$$X_{(n)} = \max\{X_1, \dots, X_n\}.$$

The *median* is the middle order statistic when n is odd, or the average of the two middle order statistics when n is even:

$$\tilde{X} = \begin{cases} X_{(m)}, & n = 2m-1; \\ \frac{1}{2}(X_{(m)} + X_{(m+1)}), & n = 2m. \end{cases}$$

The median $\tilde{X}$ and mean $\overline{X}$ are commonly used to describe the center or overall location of the variables $X_1, \dots, X_n$. One possible advantage for the median is that it will not be influenced by a few extreme values.

For i.i.d. samples, the asymptotic distribution of $\overline{X}$ can be derived using the central limit theorem. In what follows, we derive an analogous result for $\tilde{X}$.

{thm:AN_medi

Theorem 6.46 (Asymptotic normality of the median)**.** *Assume now that $X_1, X_2, \dots$ are i.i.d. with common cumulative distribution function F, and let $\tilde{X}_n$ be the median of the first n observations. Assume that F has a unique median θ such that $F(\theta) = 1/2$, and that $F'(\theta) = f(\theta)$ exists and is finite and positive. Then*

$$\sqrt{n}(\tilde{X}_n - \theta) \Rightarrow N\left(0, \frac{1}{4f(\theta)^2}\right).$$

Proof. For any $a \in \mathbb{R}$, let us try to approximate

$$\mathbb{P}(\sqrt{n}(\tilde{X}_n - \theta) \le a) = \mathbb{P}(\tilde{X}_n \le \theta + a/\sqrt{n}).$$

Define

$$S_n = \#\{i \le n : X_i \le \theta + a/\sqrt{n}\} = \sum_{i=1}^{n} \mathbf{1}\{X_i \le \theta + a/\sqrt{n}\}.$$

Assume n is odd and $n = 2m-1$. The key to this derivation is the observation that $\tilde{X}_n \le \theta + a/\sqrt{n}$ if and only if $S_n \ge m$. Also,

$$S_n \sim \text{Binomial}\left(n, F\left(\theta + \frac{a}{\sqrt{n}}\right)\right).$$

which is a sum of i.i.d. Bernoulli random variables. The Berry-Essen bound (Theorem 6.22) gives

$$\begin{aligned} \mathbb{P}\left(\sqrt{n}(\tilde{X}_n - \theta) \le a\right) &= \mathbb{P}(S_n > m-1) \\ &= \mathbb{P}\left(\frac{S_n - nF(\theta + a/\sqrt{n})}{\sqrt{n}} > \frac{m - 1 - nF(\theta + a/\sqrt{n})}{\sqrt{n}}\right) \\ &= \Phi\left(\frac{[nF(\theta + a/\sqrt{n}) - m + 1]/\sqrt{n}}{\sqrt{F(\theta + a/\sqrt{n})(1 - F(\theta + a/\sqrt{n}))}}\right) + o(1), \qquad (56) \end{aligned}$$

{eq:BE}

where $o(1)$ means the error term vanishes as $n \to \infty$. This is because $\frac{[nF(\theta+a/\sqrt{n})-m+1]/\sqrt{n}}{\sqrt{F(\theta+a/\sqrt{n})(1-F(\theta+a/\sqrt{n}))}}$ is in an ε-neighborhood of $2aF'(\theta)$ for sufficiently large n and we can bound the error of approximation uniformly on ε-neighborhood by the Berry-Essen bound.

Since F is continuous at θ, the denominator satisfies

$$\sqrt{F(\theta + a/\sqrt{n})(1 - F(\theta + a/\sqrt{n}))} \to 1/2,$$

as $n \to \infty$. And because F is differentiable at θ, $F(\theta) = 1/2$, and $n = 2m - 1$,

$$\begin{aligned}\frac{nF(\theta + a/\sqrt{n}) - m + 1}{\sqrt{n}} &= a \cdot \frac{F(\theta + a/\sqrt{n}) - F(\theta)}{a/\sqrt{n}} + \frac{nF(\theta) - m + 1}{\sqrt{n}} \\ &= a \cdot \frac{F(\theta + a/\sqrt{n}) - F(\theta)}{a/\sqrt{n}} + \frac{1}{2\sqrt{n}} \to aF'(\theta).\end{aligned}$$

Since the numerator and denominator of the argument of Φ both converge, from (6.6),

$$\mathbb{P}(\sqrt{n}(\tilde{X}_n - \theta) \leq a) \to \Phi(2aF'(\theta)).$$

The limit here is the cumulative distribution function for the normal distribution with mean zero and variance $1/(4[F'(\theta)]^2)$ evaluated at a, therefore

$$\sqrt{n}(\tilde{X}_n - \theta) \Rightarrow N\left(0, \frac{1}{4[F'(\theta)]^2}\right) = N\left(0, \frac{1}{4f(\theta)^2}\right).$$

If $n = 2m$ is even, then we can check that the following inequality holds:

$$\mathbb{P}(S_n > m + 1) \leq \mathbb{P}\left(\sqrt{n}(\tilde{X}_n - \theta) \leq a\right) \leq \mathbb{P}(S_n > m - 1)$$

Since both $\mathbb{P}(S_n > m + 1)$ and $\mathbb{P}(S_n > m - 1)$ converge to $\Phi(2aF'(\theta))$. The conclusion also holds for even n. □

A similar derivation leads to the following central limit theorem for other quantiles.

Theorem 6.47. *Let $X_1, X_2, \ldots$ be i.i.d. with common cumulative distribution function F, let $\gamma \in (0, 1)$, and let $\tilde{\theta}_n$ be the $\lfloor \gamma n \rfloor$th order statistic for $X_1, \ldots, X_n$ (or a weighted average of the $\lfloor \gamma n \rfloor$th and $\lceil \gamma n \rceil$th order statistics). If $F(\theta) = \gamma$, and if $F'(\theta) = f(\theta)$ exists and is finite and positive, then*

$$\sqrt{n}(\tilde{\theta}_n - \theta) \Rightarrow N\left(0, \frac{\gamma(1 - \gamma)}{f(\theta)^2}\right),$$

as $n \to \infty$.

ARE of median and mean In this section, we compute the ARE of the median and the mean. We assume $X_1, \ldots, X_n$ are i.i.d. samples from a location family (see Definition 1.17) $f(x - \theta)$ and f is symmetric about zero ($f(x) = f(-x)$. Assuming X_i has a finite mean and variance σ^2, then $\mathbb{E}_\theta X_i = \theta$. By the CLT,

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

where $\sigma^2 = \int x^2 f(x)dx$. And from Theorem 6.46,

$$\sqrt{n}(\tilde{X}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{4f^2(0)}\right).$$

Therefore we have

$$\text{ARE}(\overline{X}_n, \tilde{X}_n) = \frac{1}{4f(0)^2\sigma^2}$$

Example 6.48 (ARE($\overline{X}_n, \tilde{X}_n$) for different densities)**.** Depending on f, ARE($\overline{X}_n, \tilde{X}_n$) can be different:

1. Suppose f is the standard Gaussian density, $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, then

$$\text{ARE}(\overline{X}_n, \tilde{X}_n) = \frac{\pi}{2}.$$

2. Suppose f is the standard Laplacian distribution $f(x) = \frac{1}{2}e^{-|x|}$. Then

$$\sigma^2 = \int x^2 \frac{1}{2} e^{-|x|}dx = \int_0^\infty x^2 e^{-x}dx = \Gamma(3) = 2.$$

So ARE($\overline{X}_n, \tilde{X}_n$) $= \frac{1}{2}$. In this case, the median is more efficient.

The performance of the median improves for distributions with heavy tails.

Example 6.49. Suppose $X_1, \ldots, X_n$ is a random sample from $N(\theta, 1)$, and we are interested in estimating

$$p = \mathbb{P}_\theta(X_i \leq a) = \Phi(a - \theta).$$

One natural estimator is

$$\hat{p} = \Phi(a - \bar{X}),$$

where $\bar{X} = (X_1 + \cdots + X_n)/n$. (This is the maximum likelihood estimator.) Another natural estimator is the proportion of the observations that are at most a,

$$\tilde{p} = \frac{1}{n}\#\{i \leq n : X_i \leq a\} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq a\}.$$

By the central limit theorem,

$$\sqrt{n}(\tilde{p} - p) \Rightarrow N(0, \tilde{\sigma}^2), \quad \text{as } n \to \infty,$$

where

$$\tilde{\sigma}^2 = \text{Var}_\theta(I\{X_i \leq a\}) = \Phi(a - \theta)(1 - \Phi(a - \theta)).$$

Because the first estimator is a function of the average $\bar{X}$, by the delta method,

$$\sqrt{n}(\hat{p} - p) \Rightarrow N(0, \hat{\sigma}^2), \quad \text{as } n \to \infty,$$

where

$$\hat{\sigma}^2 = \left[\frac{d}{dx} \Phi(a-x) \bigg|_{x=\theta} \right]^2 = \phi^2(a-\theta).$$

The asymptotic relative efficiency of $\hat{p}$ with respect to $\tilde{p}$ is

$$\text{ARE} = \frac{\Phi(a-\theta)(1-\Phi(a-\theta))}{\phi^2(a-\theta)}.$$

In this example, the asymptotic relative efficiency depends on the unknown parameter θ. When $\theta = a$, ARE $= \pi/2$, and the ARE increases without bound as $|\theta - a|$ increases.

Note, however, that $\tilde{p}$ is a sensible estimator even if the stated model is wrong, provided the data are indeed i.i.d. In contrast, $\hat{p}$ is only reasonable if the model is correct. Gains in efficiency using $\hat{p}$ should be balanced against the robustness of $\tilde{p}$ to departures from the model.

6.7 Robustness

Thus far, we have evaluated the performance of estimators assuming that the underlying model is the correct one. Under this assumption, we have derived estimators that are optimal in some sense. However, if the underlying model is not correct, then we cannot be guaranteed of the optimality of our estimator.

We may be concerned about small or medium-sized deviations from our assumed model. This may lead us to the consideration of *robust estimators*. Such estimators will give up optimality at the assumed model in exchange for reasonable performance if the assumed model is not the true model. Thus, we have a trade-off between optimality and robustness.

The term "robustness" can have many interpretations, but perhaps it is best summarized by Huber (1981, Section 1.2), who noted: Any statistical procedure should possess the following desirable features:

1. It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed model.

2. It should be robust in the sense that small deviations from the model assumptions should impair the performance only slightly...

3. Somewhat larger deviations from the model should not cause a catastrophe.

Is the sample mean a robust estimator? It may depend on exactly how we formalize measures of robustness.

Example 6.50 (Robustness of the sample mean))**.** Let $X_1, X_2, \ldots, X_n$ be iid $n(\mu, \sigma^2)$. We know that $\bar{X}$ has variance $\text{Var}(\bar{X}) = \sigma^2/n$, which is the Cramér-Rao Lower Bound.

Hence, $\bar{X}$ satisfies (1) in that it attains the best variance at the assumed model.

To investigate (2), the performance of $\bar{X}$ under small deviations from the model, we first need to decide on what this means. A common interpretation is to use a δ-*contamination model*; that is, for small δ, assume that we observe

$$X_i \sim \begin{cases} N(\mu, \sigma^2) & \text{with probability } 1-\delta \\ f(x) & \text{with probability } \delta, \end{cases}$$

where $f(x)$ is some other distribution. Suppose that we take $f(x)$ to be any density with mean θ and variance τ^2. Write this mixture model as a hierarchical model. Let $Y = 0$ with probability $1-\delta$ and $Y = 1$ with probability δ. We can write $X = YW + (1-Y)Z$, where $W \sim f(x)$, $Z \sim N(\mu, \sigma^2)$.

Then using the conditional variance formula

$$\text{Var}(X_i) = \mathbb{E}[\text{Var}(X_i \mid Y)] + \text{Var}(\mathbb{E}[X_i \mid Y]),$$

we obtain

$$\begin{aligned}\text{Var}(X_i) =& \mathbb{E}[Y^2\text{Var}(W) + (1-Y)^2\text{Var}(Z)] + \text{Var}((\theta Y + \mu(1-Y)) \\ =& \delta\tau^2 + (1-\delta)\sigma^2 + (\theta-\mu)^2\delta(1-\delta).\end{aligned}$$

Therefore

$$\text{Var}(\bar{X}) = \frac{1}{n}\text{Var}(X_i) = (1-\delta)\frac{\sigma^2}{n} + \delta\frac{\tau^2}{n} + \frac{\delta(1-\delta)(\theta-\mu)^2}{n}.$$

This actually looks pretty good for $\bar{X}$, since if $\theta \approx \mu$ and $\sigma \approx \tau$, $\bar{X}$ will be near optimal. However, if $f(x)$ is a Cauchy pdf. Then it immediately follows that $\text{Var}(\bar{X}) = \infty$.

Turning to item (3), we ask what happens if there is an outlier observation. Envision a particular set of sample values and then consider the effect of increasing the largest observation. For example, suppose that $X_{(n)} = x$, where $x \to \infty$. The effect of such an observation could be considered "catastrophic." Although none of the distributional properties of $\bar{X}$ are affected, the observed value would be "meaningless." This illustrates the *breakdown value*, an idea attributable to Hampel (1974).

Definition 6.51. Let $X_{(1)} < \cdots < X_{(n)}$ be an ordered sample of size n, and let T_n be a statistic based on this sample. T_n has *breakdown value* b, $0 \leq b \leq 1$, if, for every $\epsilon > 0$,

$$\lim_{X_{\{(1-b)n\}} \to \infty} T_n < \infty \quad \text{and} \quad \lim_{X_{(\{(1-(b+\epsilon))n\}})\to\infty} T_n = \infty.$$

where $\{x\}$ is defined to be the number x rounded to the nearest integer.

It is easy to see that the breakdown value of $\bar{X}$ is 0; that is, if any fraction of the sample is driven to infinity, so is the value of $\bar{X}$. In stark contrast, the sample median is unchanged by this change of the sample values. This insensitivity to extreme observations is sometimes considered an asset of the sample median, which has a breakdown value of 1/2.

Example 6.52. An estimator that is between the mean and the median in terms of sensitivity is the *α-trimmed mean*, $0 < \alpha < \frac{1}{2}$, defined as follows. $\bar{X}_n^\alpha$, the α-trimmed mean, is computed by deleting the αn smallest observations and the αn largest observations, and taking the arithmetic mean of the remaining observations. The α-trimmed mean of the sample, $0 < \alpha < \frac{1}{2}$, has breakdown value α.

Example 6.53. For a sample $X_1, \ldots, X_n$, the breakdown value of the sample variance $S^2 = \sum(X_i - \bar{X})^2/(n-1)$ is 0.

A robust alternative is the *median absolute deviation*, or *MAD*, the median of $|X_1 - M|, |X_2 - M|, \ldots, |X_n - M|$, where M is the sample median. This estimator has a breakdown value of 50%.

6.8 M-estimators

Many of the estimators that we use are the result of minimizing a particular criterion. For example, if $X_1, X_2, \ldots, X_n$ are iid from $f(x|\theta)$, possible estimators are the mean, the minimizer of $\sum(x_i - a)^2$; the median, the minimizer of $\sum|x_i - a|$; and the MLE, the maximizer of $\prod_{i=1}^n f(x_i|\theta)$ (or the minimizer of the negative likelihood).

Huber estimators Huber (1964) considered a compromise between the mean and the median. The mean criterion is a square, which gives it sensitivity, but in the "tails" the square gives too much weight to big observations. In contrast, the absolute value criterion of the median does not overweight big or small observations. The compromise is to minimize a criterion function

$$\sum_{i=1}^{n} \rho(x_i - a), \tag{57}$$

{huberloss}

where ρ is given by

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq k \\ k|x| - \frac{1}{2}k^2 & \text{if } |x| \geq k. \end{cases}$$

The function $\rho(x)$ (called the Huber loss function) acts like x^2 for $|x| \leq k$ and like $|x|$ for $|x| > k$. Moreover, since $\frac{1}{2}k^2 = k|k| - \frac{1}{2}k^2$, the function is continuous. In fact, ρ is differentiable. The constant k, which can also be called a *tuning parameter*, controls the mix, with small values of k yielding a more "median-like" estimator.

M-estimators The estimator minimizing (57) is a special case of the estimators studied by Huber. For a general function ρ, we call the estimator minimizing

$$\sum_i \rho(x_i - \theta)$$

an *M-estimator*, a name that is to remind us that these are *maximum-likelihood-type* estimators.

Remark 6.54. Note that if we choose ρ to be the negative log likelihood $-l(\theta|x)$ for a location family $f(x - \theta)$, then the M-estimator is the usual MLE.

Assume ρ is differentiable. Defining $\psi = \rho'$, we see that an M-estimator is the solution to

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0. \tag{59}$$

{eq:first_or

Theorem 6.55 (Asymptotic normality of the M-estimator)**.** *Let $X_1, \ldots, X_n$ be i.i.d. samples with a location family pdf $f(x|\theta_0) = f(x - \theta_0)$. Assume that*

1. *f and $\rho(x)$ are symmetric.*

2. *$\psi'(x)$ exists and is continuous.*

3. *Assume $\mathbb{E}_{\theta_0}[\psi(X_i - \theta_0)^2] < \infty$, and there exists $\varepsilon > 0$ such that $\mathbb{E}_{\theta_0}\left\|\mathbb{1}_{\theta\in[\theta_0-\epsilon,\theta_0+\epsilon]}\psi'(X - \theta)\right\|_\infty < \infty$. Assume $\mathbb{E}[\psi'(X - \theta_0)] \neq 0$.*

4. $\hat{\theta}_M \xrightarrow{p} \theta_0$.

Then

$$\sqrt{n}(\hat{\theta}_M - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}_{\theta_0}[\psi(X-\theta_0)^2]}{[\mathbb{E}_{\theta_0}\psi'(X-\theta_0)]^2}\right).$$

Proof. **Step 1: Taylor expansion.**

As in the proof of Theorem 6.39, by the mean-value theorem, we write a Taylor expansion for ψ as

$$\sum_{i=1}^n \psi(X_i - \theta) = \sum_{i=1}^n \psi(X_i - \theta_0) + (\theta - \theta_0)\sum_{i=1}^n \psi'(X_i - \tilde{\theta}_n),$$

where θ_0 is the true value, and $\tilde{\theta}_n$ is between θ and θ_0.

Let $\hat{\theta}_M$ be the solution to (59) and substitute this for θ to obtain

$$0 = \sum_{i=1}^n \psi(X_i - \theta_0) + (\hat{\theta}_M - \theta_0)\sum_{i=1}^n \psi'(X_i - \tilde{\theta}_n),$$

where the left-hand side is 0 because $\hat{\theta}_M$ is the solution.

Now, again analogous to the proof of Theorem 6.39, we rearrange terms, divide through by $\sqrt{n}$ to get

$$\sqrt{n}(\hat{\theta}_M - \theta_0) = \frac{\frac{-1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i - \theta_0)}{\frac{1}{n}\sum_{i=1}^n \psi'(X_i - \tilde{\theta}_n)}.$$

Step 2: Show that $\mathbb{E}_{\theta_0}\psi(X-\theta_0) = 0$.

Since $\rho(x) = \rho(-x)$, we have $\psi(x) = -\psi(-x)$. Then

$$\mathbb{E}_{\theta_0}\psi(X-\theta_0) = \int_{-\infty}^{\infty} \psi(x-\theta_0)f(x-\theta_0)dx = \int_{-\infty}^{\infty} \psi(u)f(u)du = 0,$$

where in the last identity we use the fact that $f(x)$ is an even function and $\psi(x)$ is an odd function.

Step 3: CLT for the numerator

From the assumption that $\mathbb{E}_{\theta_0}[\psi(X_i-\theta_0)^2] < \infty$, it follows the CLT that

$$\frac{-1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i - \theta_0) = \sqrt{n}\left[\frac{-1}{n}\sum_{i=1}^n \psi(X_i-\theta_0)\right] \xrightarrow{d} \mathcal{N}(0, \mathbb{E}_{\theta_0}[\psi(X-\theta_0)^2]).$$

Step 4: Convergence of the denominator

Since $\hat{\theta}_M \xrightarrow{p} \theta_0$, we have $\tilde{\theta}_n \xrightarrow{p} \theta_0$. From the assumption $\mathbb{E}_{\theta_0}\left\|\mathbb{1}_{\theta\in[\theta_0-\epsilon,\theta_0+\epsilon]}\psi'(X-\theta)\right\|_\infty < \infty$ and Lemma 6.34 (Uniform Law of Large Numbers), we obtain

$$\frac{1}{n}\sum_{i=1}^n \psi'(X_i - \tilde{\theta}_n) \xrightarrow{p} \mathbb{E}_{\theta_0}\psi'(X-\theta_0).$$

Step 5: Conclusion

Putting this all together, by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\theta}_M - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}_{\theta_0}\psi(X-\theta_0)^2}{[\mathbb{E}_{\theta_0}\psi'(X-\theta_0)]^2}\right). \tag{60}$$

{eq:variance

□

Example 6.56 (Limit distribution of the Huber estimator)**.** If $X_1, \ldots, X_n$ are iid from a pdf $f(x-\theta)$, where f is symmetric around 0, then for the Huber loss ρ, we have

$$\psi(x) = \begin{cases} x & \text{if } |x| \le k \\ k & \text{if } x > k \\ -k & \text{if } x < -k \end{cases}$$

and thus

$$\begin{aligned} \mathbb{E}_\theta \psi(X-\theta) &= \int_{\theta-k}^{\theta+k} (x-\theta) f(x-\theta)\,dx \\ &\quad - k\int_{-\infty}^{\theta-k} f(x-\theta)\,dx + k\int_{\theta+k}^{\infty} f(x-\theta)\,dx \\ &= \int_{-k}^{k} y f(y)\,dy - k\int_{-\infty}^{-k} f(y)\,dy + k\int_{k}^{\infty} f(y)\,dy = 0, \end{aligned}$$

where we substitute $y = x-\theta$. The integrals add to 0 by the symmetry of f.

To calculate the variance, we need the expected value of ψ'. While ψ is not differentiable beyond the points of nondifferentiability ($x = \pm k$), ψ' will be 0 except at those points. Thus, we only need to deal with the expectation for $|x| \le k$, and we have

$$\mathbb{E}_\theta \psi'(X-\theta) = \int_{\theta-k}^{\theta+k} f(x-\theta)\,dx = \mathbb{P}_0(|X| \le k),$$

$$\begin{aligned} \mathbb{E}_\theta \psi(X-\theta)^2 &= \int_{\theta-k}^{\theta+k} (x-\theta)^2 f(x-\theta)\,dx + k^2\int_{\theta+k}^{\infty} f(x-\theta)\,dx + k^2\int_{-\infty}^{\theta-k} f(x-\theta)\,dx \\ &= \int_{-k}^{k} x^2 f(x)\,dx + 2k^2\int_{k}^{\infty} f(x)\,dx. \end{aligned}$$

Thus, we can conclude that the Huber estimator is asymptotically normal with mean θ and asymptotic variance

$$\frac{\int_{-k}^{k} x^2 f(x)\,dx + 2k^2 \mathbb{P}_0(|X| > k)}{[\mathbb{P}_0(|X| \le k)]^2}.$$

Example 6.57 (ARE between the M-estimator and the MLE)**.** Let us look more closely at the asymptotic variance in (60). The denominator of the variance contains the term $\mathbb{E}_{\theta_0}\psi'(X-\theta_0)$, which we can write as

$$\mathbb{E}_{\theta_0}\psi'(X-\theta) = \int \psi'(x-\theta) f(x-\theta)\,dx = -\int \left[\frac{\partial}{\partial\theta}\psi(x-\theta)\right] f(x-\theta)\,dx. \tag{63}$$ {eq:M_1}

Now we use the differentiation product rule to get

$$\frac{d}{d\theta}\int \psi(x-\theta) f(x-\theta)\,dx = \int \left[\frac{d}{d\theta}\psi(x-\theta)\right] f(x-\theta)\,dx + \int \psi(x-\theta)\left[\frac{d}{d\theta} f(x-\theta)\right] dx.$$

The left-hand side is 0 because $\mathbb{E}_\theta \psi(X-\theta) = 0$, so we have

$$-\int \left[\frac{d}{d\theta}\psi(x-\theta)\right] f(x-\theta)\,dx = \int \psi(x-\theta)\left[\frac{d}{d\theta}f(x-\theta)\right]\,dx$$
$$= \int \psi(x-\theta)\left[\frac{d}{d\theta}\log f(x-\theta)\right] f(x-\theta)\,dx, \qquad (64) \quad \texttt{\{eq:M\_2\}}$$

where we use the fact that $\frac{d}{dy}g(y)/g(y) = \frac{d}{dy}\log g(y)$. This last expression on the right hand side can be written

$$\mathbb{E}_\theta[\psi(X-\theta)l'(\theta|X)],$$

where $l(\theta|X)$ is the log-likelihood. From (63) and (64),

$$\mathbb{E}_\theta \psi'(X-\theta) = -\mathbb{E}_\theta\left[\frac{d}{d\theta}\psi(X-\theta)\right] = \mathbb{E}_\theta[\psi(X-\theta)l'(\theta|X)].$$

Therefore, the asymptotic variance of the M-estimator is

$$\frac{\mathbb{E}_\theta \psi(X-\theta)^2}{[\mathbb{E}_\theta \psi(X-\theta)l'(\theta|X)]^2}.$$

Recall that the asymptotic variance of the MLE, $\hat{\theta}$, is given by

$$\frac{1}{\mathbb{E}_{\theta_0}[l'(\theta|X)]^2} = \frac{1}{I(\theta)}$$

from Theorem 6.39, so we have

$$\text{ARE}(\hat{\theta}_M, \hat{\theta}) = \frac{[\mathbb{E}_\theta \psi(X-\theta)l'(\theta|X)]^2}{\mathbb{E}_{\theta_0}[\psi(X-\theta)^2]\mathbb{E}_\theta[l'(\theta|X)]^2} \leq 1$$

by the Cauchy-Schwartz Inequality. Thus, an M-estimator is always less efficient than the MLE, and matches its efficiency only if ψ is proportional to l'.

References

[1] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487, 2015.

[2] Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014.

[3] Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. In *Conference on learning theory*, pages 33–1. JMLR Workshop and Conference Proceedings, 2012.

[4] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[5] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[6] George Casella and Roger Berger. *Statistical inference*. CRC press, 2024.

[7] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[8] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media, 2010.

[9] João M Pereira, Joe Kileel, and Tamara G Kolda. Tensor moments of gaussian mixture models: Theory and applications. *arXiv preprint arXiv:2202.06930*, 2022.

[10] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[11] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[12] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[13] Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847, 2015.