IEDA 5270 Engineering Statistics and Data Analytics

Wei You

March 4, 2026

Contents

1 Review of Probability Theory

1.1 Introduction to Statistics

Probability theory provides a mathematical language for randomness. It starts from a *known* probability model (a distribution F) and studies the behavior of random quantities generated under that model. Statistics works in the opposite direction: we observe data and use it to learn about the underlying law of randomness.

Probability versus statistics A typical probability question is: if $X_1, \ldots, X_n$ are generated from a distribution F, what should we expect to see? For instance, one may model a stock return by an abstract distribution and ask about typical values (mean/median) and risk measures (variance, value-at-risk).

A typical statistics question is: given observations $X_1, \ldots, X_n$, what can we infer about F and its features? With 100 observed prices, we may assess whether a lognormal model is plausible (goodness-of-fit), estimate average level and variability (sample mean/variance), test a claim about expected return (hypothesis testing), study relationships with predictors (regression), or build prediction rules (classification).

Populations, samples, and variables

Definition 1.1 (Population)**.** The *population* is the full collection of individuals/units we want to draw conclusions about.

Definition 1.2 (Sample)**.** A *sample* is the subset of the population we actually observe. The recorded measurements from the sample are the data.

Definition 1.3 (Variable)**.** A *variable* is a measurable attribute that can take different values across units in the population.

Variables can be categorical (discrete and unordered, such as gender or department), ordinal (discrete but ordered, such as a 1–5 rating), or continuous (numerical values on a continuum, such as height, blood pressure, or stock return). In modeling we often distinguish a *response* (dependent variable) from one or more *predictors* (independent variables). For example, we may use the items in a shopping cart to predict the customer's gender.

What statistics does Statistics uses patterns in finite sample data to draw conclusions about population-level quantities, while accounting for randomness. In this course we focus on the mathematical principles that justify common statistical methods.

Main topics in this course

1. **Statistical inference**: uses data to estimate unknown features of an underlying probability model. Typical tasks include point estimation, hypothesis testing, and confidence intervals/sets.

2. **Regression analysis**: describes how a response variable changes with one or more predictors. For example, linear regression and generalized linear models.

3. **Classification**: learns a rule from labeled data and uses it to predict the label of a new observation. Typical methods include linear discriminant analysis, logistic regression, support vector machine, random forest.

4. **Unsupervised learning**: extracts structure from data without labels. Typical methods include principal component analysis, clustering, and dimensionality reduction.

1.2 Probability of Events

We now formalize events and probability measures. These basic objects support everything that follows, from conditional probability to random variables and expectations.

Sample spaces and events

Definition 1.4 (Sample space)**.** The *sample space* S is the set of all possible outcomes of an experiment.

Example 1.1 (Three-horse race)**.** If horses are labeled A, B, C, then the sample space consists of all possible orderings:

$$S = \{(A,B,C),(A,C,B),(B,A,C),(B,C,A),(C,A,B),(C,B,A)\}.$$

Definition 1.5 (Event)**.** An *event* is a subset of the sample space.

Example 1.2. In the race example, the event "horse B wins" is

$$E = \{(B,A,C),(B,C,A)\} \subset S.$$

Set relations and operations For events $E, F \subset S$, the containment relation $E \subset F$ means that every outcome in E also lies in F. Equivalently,

$$E \subset F \quad \Longleftrightarrow \quad (\forall x \in E)\ x \in F.$$

Two events are equal, $E = F$, if and only if $E \subset F$ and $F \subset E$.

Definition 1.6 (Basic set operations)**.** For events $E, F \subset S$, define

$$\begin{aligned} E^c &= \{x \in S : x \notin E\}, && \text{(complement)} \\ E \cup F &= \{x \in S : x \in E \text{ or } x \in F\}, && \text{(union)} \\ E \cap F &= \{x \in S : x \in E \text{ and } x \in F\}. && \text{(intersection)} \end{aligned}$$

We say E and F are *disjoint* (mutually exclusive) if $E \cap F = \varnothing$.

Example 1.3. In the race example, let

$$\begin{aligned} E &= \text{“}A \text{ wins the first place.”} = \{(A,B,C),(A,C,B)\}, \\ F &= \text{“}B \text{ wins the first place.”} = \{(B,A,C),(B,C,A)\}. \end{aligned}$$

Then,

$$\begin{aligned} E^c &= \{(B,A,C),(B,C,A),(C,A,B),(C,B,A)\}, \\ E \cup F &= \{(A,B,C),(A,C,B),(B,A,C),(B,C,A)\}, \\ E \cap F &= \emptyset. \end{aligned}$$

Sometimes the intersection symbol is omitted, writing EF for $E \cap F$.

For a collection $\{E_i\}$, the union $\bigcup_i E_i$ occurs when at least one E_i occurs, while the intersection $\bigcap_i E_i$ occurs when all E_i occur. These operations are well-defined for finite, countable, and even uncountable collections of sets.

Mutual exclusivity, exhaustiveness, and partitions

Definition 1.7 (Mutually exclusive)**.** Events $E_1, E_2, \ldots$ are mutually exclusive if $E_i \cap E_j = \varnothing$ for all $i \neq j$.

Definition 1.8 (Collectively exhaustive)**.** Events $E_1, E_2, \ldots$ are collectively exhaustive if $\bigcup_{i=1}^{\infty} E_i = S$.

Definition 1.9 (Partition)**.** A collection $\{E_i\}_{i \geq 1}$ is a *partition* of S if it is mutually exclusive and collectively exhaustive.

Probability measures A probability model assigns a number to each event, interpreted as its likelihood.

Definition 1.10 (Probability measure)**.** A function $\mathbb{P}$ that assigns a number $\mathbb{P}(E)$ to each event $E \subset S$ is a *probability measure* if it satisfies:

$$\mathbb{P}(E) \geq 0, \qquad \mathbb{P}(S) = 1, \qquad \mathbb{P}\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(E_i) \ \text{ for any disjoint } E_1, E_2, \ldots.$$

Two useful inequalities/identities The following results are used repeatedly when manipulating event probabilities.

Proposition 1.1 (Law of total probability)**.** *If $\{C_i\}_{i \geq 1}$ is a partition of S, then for any event A,*

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap C_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid C_i)\mathbb{P}(C_i),$$

where the conditional probability is defined whenever $\mathbb{P}(C_i) > 0$.

Proposition 1.2 (Boole's inequality (union bound))**.** *For any events $A_1, A_2, \ldots$,*

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

1.3 Conditional Probability

Conditional probability updates probabilities after observing partial information. It is the basic tool behind Bayesian calculations and many dependence/independence arguments.

Definition and a dice example

Definition 1.11 (Conditional probability)**.** For events E and F with $\mathbb{P}(F) > 0$, the conditional probability of E given F is

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}.$$

Example 1.4 (Two dice)**.** The sample space is

$$\begin{aligned} S = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\} \end{aligned}$$

Let $E = \{\text{sum of two dice equals 8}\}$ and $F = \{\text{first die equals 3}\}$. Then

$$\mathbb{P}(E) = \frac{5}{36}, \qquad \mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \frac{1/36}{6/36} = \frac{1}{6}.$$

The value of $\mathbb{P}(E \mid F)$ depends on how F overlaps with E. For instance, if $E \cap F = \varnothing$ then $\mathbb{P}(E \mid F) = 0$, while if $F \subset E$ then $\mathbb{P}(E \mid F) = 1$.

Example 1.5. What can you say about $\mathbb{P}(E \mid F)$ if

1. $E \cap F = \varnothing$: Two events cannot occur simultaneously.
2. $E \subset F$: if E occur, then F occur.
3. $F \subset E$: if F occur, then E occur.

Think about the case where $E =$ "Rain", $F =$ "Cloud".

Chain rule for probabilities The definition implies a sequential factorization that is often the cleanest way to compute joint probabilities.

Proposition 1.3 (Chain rule)**.** *For events $A_1, \ldots, A_n$ with $\mathbb{P}(A_1 \cdots A_{k-1}) > 0$ for $k \geq 2$,*

$$\mathbb{P}(A_1 A_2 \cdots A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 A_2) \cdots \mathbb{P}(A_n \mid A_1 \cdots A_{n-1}).$$

Example 1.6 (Insurance mixture)**.** An insurer classifies customers as accident-prone (B) or not (B^c). Suppose $\mathbb{P}(B) = 0.3$, $\mathbb{P}(A \mid B) = 0.4$, and $\mathbb{P}(A \mid B^c) = 0.2$, where A is the event "an accident occurs within a year." By the law of total probability,

$$\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c) = (0.4)(0.3) + (0.2)(0.7) = 0.26.$$

1.4 Bayes' Formula

Bayes' formula is a direct consequence of conditional probability and the chain rule. It is the standard way to "invert" conditioning.

Bayes' formula Starting from $\mathbb{P}(E \cap F) = \mathbb{P}(E \mid F)\mathbb{P}(F) = \mathbb{P}(F \mid E)\mathbb{P}(E)$, we obtain the following.

Proposition 1.4 (Bayes' formula)**.** *If $\mathbb{P}(F) > 0$, then*

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(F \mid E)\mathbb{P}(E)}{\mathbb{P}(F)}.$$

Example 1.7 (Insurance mixture continued)**.** With the notation above, if we observe an accident (A), then

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{(0.4)(0.3)}{0.26} = \frac{6}{13}.$$

A warning: the prosecutor's fallacy

Example 1.8 (A base-rate effect)**.** Suppose $\mathbb{P}(C) = 0.005$ that a randomly chosen person committed a crime. An evidence procedure satisfies $\mathbb{P}(E \mid C) = 0.99$ and $\mathbb{P}(E \mid C^c) = 0.02$. If evidence is found on a person, then

$$\mathbb{P}(C \mid E) = \frac{\mathbb{P}(E \mid C)\mathbb{P}(C)}{\mathbb{P}(E \mid C)\mathbb{P}(C) + \mathbb{P}(E \mid C^c)\mathbb{P}(C^c)} = \frac{(0.99)(0.005)}{(0.99)(0.005) + (0.02)(0.995)} \approx 0.199.$$

A small false-positive rate can still lead to a low posterior probability if the base rate $\mathbb{P}(C)$ is very small.

Remark 1.1. A visual explanation of Bayes' theorem is given in the 3Blue1Brown video: `https://youtu.be/HZGCoVF3YvM?si=Q-8Hfh6CQbPg2RAK`.

Bayes and total probability for finite partitions If $F_1, \dots, F_n$ form a partition of S, then for any event E,

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E \mid F_i)\mathbb{P}(F_i).$$

Moreover, for any j with $\mathbb{P}(E) > 0$,

$$\mathbb{P}(F_j \mid E) = \frac{\mathbb{P}(E \mid F_j)\mathbb{P}(F_j)}{\sum_{i=1}^{n} \mathbb{P}(E \mid F_i)\mathbb{P}(F_i)}.$$

1.5 Independent Events

Independence formalizes the idea that learning one event occurred does not change the probability of another event.

Definition 1.12 (Independence)**.** Two events E and F are independent if

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F).$$

Equivalently, if $\mathbb{P}(F) > 0$, then $\mathbb{P}(E \mid F) = \mathbb{P}(E)$.

Proposition 1.5. *If E and F are independent, then E is also independent of F^c, and E^c is independent of F and F^c.*

Proof. Using $\mathbb{P}(E) = \mathbb{P}(EF) + \mathbb{P}(EF^c)$ and $\mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F)$, we obtain

$$\mathbb{P}(EF^c) = \mathbb{P}(E) - \mathbb{P}(E)\mathbb{P}(F) = \mathbb{P}(E)\big(1 - \mathbb{P}(F)\big) = \mathbb{P}(E)\mathbb{P}(F^c),$$

which is the independence condition for E and F^c. The remaining statements follow similarly. □

Multiple events For more than two events, pairwise independence is not enough.

Definition 1.13 (Independence of multiple events)**.** Events $E_1, \dots, E_n$ are independent if for every subcollection $\{E_{i_1}, \dots, E_{i_r}\}$,

$$\mathbb{P}(E_{i_1} \cdots E_{i_r}) = \mathbb{P}(E_{i_1}) \cdots \mathbb{P}(E_{i_r}).$$

Example 1.9 (Pairwise independent but not mutually independent)**.** Toss two fair coins, and encode outcomes as 0 (tails) and 1 (heads). The sample space is

$$S = \{(0,0), (1,0), (0,1), (1,1)\}.$$

Let

$$E = \{(0,0), (1,1)\}, \qquad F = \{(1,0), (1,1)\}, \qquad G = \{(0,1), (1,1)\}.$$

Then $\mathbb{P}(E) = \mathbb{P}(F) = \mathbb{P}(G) = 1/2$ and $\mathbb{P}(EF) = \mathbb{P}(EG) = \mathbb{P}(FG) = 1/4$, so the events are pairwise independent. However, $\mathbb{P}(EFG) = 1/4 \neq 1/8$, so they are not mutually independent. The key here is that knowing E and F collectively give us information about G. To see this, try to think about $\mathbb{P}(G|EF)$.

A classic example: Monty Hall

Example 1.10 (Monty Hall problem)**.** You pick one of three doors. One door hides a car and the other two hide goats. After your choice, the host opens a different door that is known to hide a goat, and then offers you the option to switch to the remaining unopened door. The optimal strategy is to *switch*; the probability of winning by switching is 2/3.

One clean way to compute this is to condition on whether your initial choice was correct. With probability 1/3 you initially picked the car; then switching loses. With probability 2/3 you initially picked a goat; then the host's action forces the remaining unopened door to be the car, and switching wins. Hence the overall success probability of switching is 2/3.

Mathematically, let H_i, C_i, P_i be the event of host choosing door i, car in i, and player choosing i. Thus, C_i and P_i are independent.

$$\mathbb{P}(C_2|P_1) = \mathbb{P}(C_2) = \frac{1}{3}.$$

You want to compare $\mathbb{P}(C_2|H_3P_1)$ and $\mathbb{P}(C_1|H_3P_1)$.

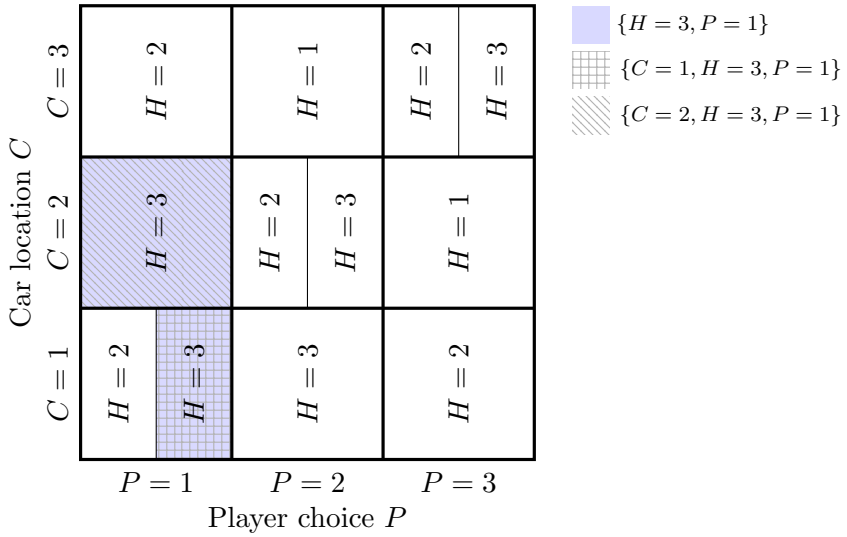
Let us calculate $\mathbb{P}(C_2|H_3P_1)$. Note taht the behavior of the host depends on the location of the car and the player's choice:

$$\mathbb{P}(H_3|C_1P_1) = \frac{1}{2}, \quad \mathbb{P}(H_3|C_2P_1) = 1, \quad \mathbb{P}(H_3|C_3P_1) = 0.$$

Apply Bayes formula

$$\begin{aligned}\mathbb{P}(C_2|H_3P_1) = \frac{\mathbb{P}(H_3C_2P_1)}{\mathbb{P}(H_3P_1)} &= \frac{\mathbb{P}(H_3|C_2P_1)\mathbb{P}(C_2P_1)}{\sum_{i=1}^{3}\mathbb{P}(H_3|C_iP_1)\mathbb{P}(C_iP_1)} \\ &= \frac{\mathbb{P}(H_3|C_2P_1)\mathbb{P}(C_2)\mathbb{P}(P_1)}{\sum_{i=1}^{3}\mathbb{P}(H_3|C_iP_1)\mathbb{P}(C_i)\mathbb{P}(P_1)} = \frac{2}{3}.\end{aligned}$$

Independence of C_i and P_i is key here. If the player has some idea of the location of the car, we may not have the same conclusion.



1.6 Random Variables

Random variables allow us to replace the abstract sample space by numerical values.

Definition 1.14 (Random variable)**.** A *random variable* X is a function from a sample space S into the real line $\mathbb{R}$. Its *range* is the set of values it can take.

Example 1.11 (Sum of two dice)**.** Let the outcome be $(D_1, D_2) \in \{1, \dots, 6\}^2$ and define $X = D_1 + D_2$. Then, for example, $X(3, 1) = 4$ and $X(6, 6) = 12$.

Remark 1.2. A random variable "pushes" the probability model on S forward onto $\mathbb{R}$. This lets us work directly with real numbers without repeatedly referring to the original sample space.

Distribution and CDF For any set $A \subset \mathbb{R}$, the event $\{X \in A\}$ means $\{s \in S : X(s) \in A\} \subset S$. The induced probability measure on $\mathbb{R}$ is

$$P_X(A) \triangleq \mathbb{P}(X \in A).$$

Definition 1.15 (Cumulative distribution function)**.** The *cumulative distribution function* (CDF) of X is

$$F_X(x) = \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}.$$

The CDF contains all distributional information about X. In particular, many calculations can be performed using F_X without explicit reference to S.

Proposition 1.6 (Basic properties of a CDF)**.** *A CDF F satisfies: (i) F is nondecreasing; (ii) F is right-continuous; (iii)* $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$.

Theorem 1.1 (Existence of a random variable with a given CDF)**.** *For any function F that satisfies the CDF properties above, there exists a random variable whose CDF is F.*

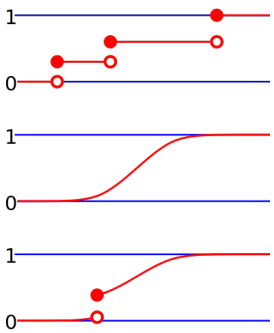


Figure 1: A typical CDF, showing right-continuity and possible jumps.

Definition 1.16 (Identically distributed)**.** Random variables X and Y are *identically distributed* if $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for every Borel set $A \subset \mathbb{R}$.

Proposition 1.7. *X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.*

Discrete random variables

Definition 1.17 (Probability mass function)**.** A random variable X is *discrete* if it takes values in a countable set $\{x_1, x_2, \dots\}$. Its probability mass function (pmf) is

$$p_X(x_i) = \mathbb{P}(X = x_i), \qquad i = 1, 2, \dots.$$

The pmf satisfies $\sum_{i=1}^{\infty} p_X(x_i) = 1$, and the CDF can be recovered from the pmf via

$$F_X(a) = \sum_{x_i \leq a} p_X(x_i).$$

Conversely, at a support point x_i, the jump size equals the probability mass:

$$p_X(x_i) = F_X(x_i) - \lim_{x \uparrow x_i} F_X(x).$$

Continuous random variables

Definition 1.18 (Probability density function)**.** A random variable X is *continuous* (more precisely, absolutely continuous) if its CDF can be written as

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\mathrm{d}s$$

for some nonnegative function f_X, called the probability density function (pdf).

If f_X exists, then $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$, and (at points of differentiability) $f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(x)$.

Remark 1.3 (Interpreting the density)*.* For a small interval $(a - \varepsilon/2, a + \varepsilon/2]$,

$$\mathbb{P}\Big(a - \frac{\varepsilon}{2} < X \le a + \frac{\varepsilon}{2}\Big) = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f_X(x)\mathrm{d}x \approx \varepsilon f_X(a).$$

In particular, for a continuous random variable, $\mathbb{P}(X = a) = 0$ for any fixed a.

1.7 Expectation

Expectation summarizes the average value of a random variable under its distribution.

Definition 1.19 (Expectation)**.** Let X be a random variable. If X is discrete with pmf p_X, define

$$\mathbb{E}[X] = \sum_i x_i p_X(x_i),$$

whenever the series is absolutely convergent. If X has pdf f_X, define

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x,$$

whenever the integral is absolutely convergent.

Expectation need not be finite for all distributions.

Theorem 1.2 (Linearity of expectation)**.** *For constants $a_1, \ldots, a_n$ and functions $g_1, \ldots, g_n$ for which the expectations exist,*

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i g_i(X)\right] = \sum_{i=1}^{n} a_i \mathbb{E}[g_i(X)].$$

Remark 1.4 (A recurring question)*.* How do we compute $\mathbb{E}[g(X)]$ efficiently? One option is to find the distribution of $Y = g(X)$ and then apply the definition of expectation to Y. A more direct option is the law of the unconscious statistician (LOTUS), derived after we discuss transformations.

1.8 Transformations

Given a random variable X and a function g, the transformed variable $Y = g(X)$ is also random. We often want the distribution of Y, or at least its moments.

Inverse images For any set $A \subset \mathbb{R}$,

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}(X \in g^{-1}(A)), \qquad g^{-1}(A) = \{x \in \mathbb{R} : g(x) \in A\}.$$

This identity is the starting point for both discrete and continuous transformation rules.

Discrete transformations If X is discrete with $\mathbb{P}(X = x_i) = p_i$, then for any value y,

$$\mathbb{P}(Y = y) = \sum_{i:g(x_i)=y} p_i.$$

Example 1.12. Let X be a fair die roll and $Y = |X - 3|$. Then $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 3) = 1/6$, while $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = 1/3$.

Continuous monotone transformations Let X have pdf f_X, and let $Y = g(X)$ with g strictly monotone and differentiable, with inverse g^{-1}. Then for y in the range of g,

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\big(X \le g^{-1}(y)\big) \quad \text{if } g \text{ is increasing,}$$

and similarly with the inequality reversed if g is decreasing.

The corresponding density transformation is the familiar "change-of-variables" rule.

Lemma 1.1 (Leibniz's integral rule)**.** *If $a(\theta)$, $b(\theta)$ and $f(x, \theta)$ are differentiable with respect to θ, then*

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)\mathrm{d}x = \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x,\theta)\mathrm{d}x + f(b(\theta),\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}b(\theta) - f(a(\theta),\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}a(\theta)$$

Lemma 1.2 (Derivatives of inverse function)**.** *If g is an invertible function, differentiable at $g^{-1}(y)$ and $g'(g^{-1}(y)) \neq 0$, then*

$$\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) = \frac{1}{g'(g^{-1}(y))}.$$

Theorem 1.3 (One-dimensional change of variables)**.** *If g is strictly monotone and differentiable with inverse g^{-1}, then*

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|$$

for y in the range of g.

Theorem 1.4 (Piecewise monotone case)**.** *Let X have pdf f_X. Suppose $\mathbb{R}$ can be partitioned into regions $A_0, \ldots, A_K$ such that g restricted to each A_k is monotone and invertible with inverse g_k^{-1}. Then, for y in the range of g,*

$$f_Y(y) = \sum_{k=0}^{K} f_X(g_k^{-1}(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}g_k^{-1}(y)\right|.$$

Think about applying the formula to $Y = |X|$ or $Y = X^2$.

Remark 1.5. Fully generalized version is proved using a measure-theoretic approach: `https://en.wikipedia.org/wiki/Pushforward_measure`

Proposition 1.8 (Law of the unconscious statistician (LOTUS))**.** *If g is measurable and the expectation exists, then*

$$\mathbb{E}[g(X)] = \sum_i g(x_i)p_X(x_i) \quad \text{(discrete)}, \quad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)\mathrm{d}x \quad \text{(continuous)}.$$

Definition 1.20 (Variance)**.** If $\mu = \mathbb{E}[X]$, the variance of X is

$$\mathrm{Var}(X) = \mathbb{E}\big[(X-\mu)^2\big],$$

whenever the expectation exists.

Expanding the square gives the commonly used identity

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

Proposition 1.9 (Scaling and shifting)**.** *For constants a, b,*

$$\mathrm{Var}(aX + b) = a^2 \,\mathrm{Var}(X).$$

Proof. Write $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$, and expand $\mathrm{Var}(aX + b) = \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2] = \mathbb{E}[(a(X - \mathbb{E}[X]))^2] = a^2 \,\mathrm{Var}(X)$. □

1.9 Random Vectors

Many statistical models are multivariate. We extend the one-dimensional notions to random vectors and joint distributions.

Definition 1.21 (Random vector)**.** An n-dimensional random vector is a function from a sample space S into $\mathbb{R}^n$.

For a bivariate random vector (X, Y), the joint CDF is

$$F_{X,Y}(x, y) = \mathbb{P}(X \le x,\ Y \le y).$$

Marginal distributions The marginal CDFs are obtained by sending the other variable to $+\infty$:

$$F_X(x) = F_{X,Y}(x, \infty), \qquad F_Y(y) = F_{X,Y}(\infty, y).$$

Different joint distributions can share the same marginals; dependence is carried by the joint law, not by the marginals alone.

Discrete and continuous cases If (X, Y) is discrete, the joint pmf is $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$, and probabilities are computed by summation:

$$\mathbb{P}\big((X, Y) \in A\big) = \sum_{(x,y)\in A} p_{X,Y}(x, y).$$

The marginal pmf is obtained by summing out the other variable, e.g.,

$$p_X(x) = \sum_y p_{X,Y}(x, y).$$

If (X, Y) is continuous, the joint pdf $f_{X,Y}$ satisfies

$$\mathbb{P}(X \le x, Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)\mathrm{d}v\mathrm{d}u,$$

and for any region $A \subset \mathbb{R}^2$,

$$\mathbb{P}\big((X, Y) \in A\big) = \iint_A f_{X,Y}(u, v)\mathrm{d}u\mathrm{d}v.$$

Consider $(a, a + \mathrm{d}a] \times (b, b + \mathrm{d}b]$,

$$\begin{aligned}\mathbb{P}\{a < X \le a + \mathrm{d}a, b < Y \le b + \mathrm{d}b\} &= \int_b^{b+\mathrm{d}b}\int_a^{a+\mathrm{d}a} f(x, y)\mathrm{d}x\mathrm{d}y \\ &\approx f(a, b)\mathrm{d}a\mathrm{d}b\end{aligned}$$

$f(a, b)$ indicates how likely (X, Y) will be near (a, b)

The marginal density is obtained by integration, e.g., $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, v)\mathrm{d}v$.

Example 1.13. Suppose the joint density is

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x}e^{-2y}, & x > 0,\ y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\mathbb{P}(X < 1) = \int_0^1 \int_0^\infty 2e^{-x}e^{-2y}\mathrm{d}y\mathrm{d}x = \int_0^1 e^{-x}\mathrm{d}x,$$

and similarly

$$\mathbb{P}(Y > 1) = \int_0^\infty \int_1^\infty 2e^{-x}e^{-2y}\mathrm{d}y\mathrm{d}x = \int_1^\infty 2e^{-2y}\mathrm{d}y.$$

1.10 Bivariate Transformations

We now extend change-of-variables ideas to two dimensions. The Jacobian determinant describes how areas distort under a transformation.

Two-dimensional LOTUS Let $g : \mathbb{R}^2 \to \mathbb{R}$. If (X, Y) is discrete,

$$\mathbb{E}[g(X,Y)] = \sum_{i,j} g(x_i, y_j) p_{X,Y}(x_i, y_j),$$

and if (X, Y) has joint density $f_{X,Y}$,

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^\infty \int_{-\infty}^\infty g(x,y) f_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x,$$

whenever the expectation exists.

Jacobian formula Let $(U, V) = g(X, Y)$ where $g : \mathbb{R}^2 \to \mathbb{R}^2$ is one-to-one with inverse $h = g^{-1}$, written as $h(u, v) = (h_1(u, v), h_2(u, v))$. The Jacobian determinant is

$$J(u,v) = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix}.$$

Theorem 1.5 (Change of variables in $\mathbb{R}^2$)**.** *If (X, Y) has joint density $f_{X,Y}$ and g is one-to-one and differentiable with inverse h, then the joint density of (U, V) is*

$$f_{U,V}(u,v) = f_{X,Y}\big(h_1(u,v), h_2(u,v)\big)|J(u,v)|$$

for (u, v) in the range of g.

Consider a simplified case where $g : \mathbb{R}^2 \to \mathbb{R}^2$ is one-to-one, thus has inverse $h = g^{-1}$. What is the PDF of (U, V)?

$$(u,v) = g(x,y) \iff (x,y) = h(u,v) = (h_1(u,v), h_2(u,v))$$

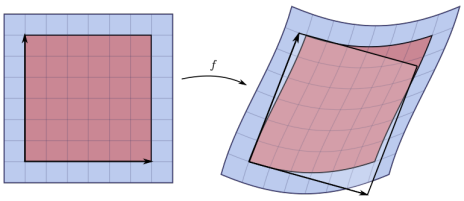
Recall that the density indicates how likely (U, V) will be near (u, v), measured using an infinitesimal area with length $\mathrm{d}u$ and width $\mathrm{d}v$:

$$f_{U,V}(u,v) \approx \frac{\mathbb{P}(u < U \le u + \mathrm{d}u, v < V \le v + \mathrm{d}v)}{\mathrm{d}u\mathrm{d}v}.$$

Let's consider an infinitesimal area with the same volume/probability

$$f_{U,V}(u,v)\mathrm{d}u\mathrm{d}v = f_{X,Y}(h_1(u,v), h_2(u,v))\mathrm{d}x\mathrm{d}y$$
$$\iff f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v))\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}u\mathrm{d}v}$$

The term $\frac{\mathrm{d}x\mathrm{d}y}{\mathrm{d}u\mathrm{d}v} = J(u, v)$ gives information about the distortion of area under the transformation.



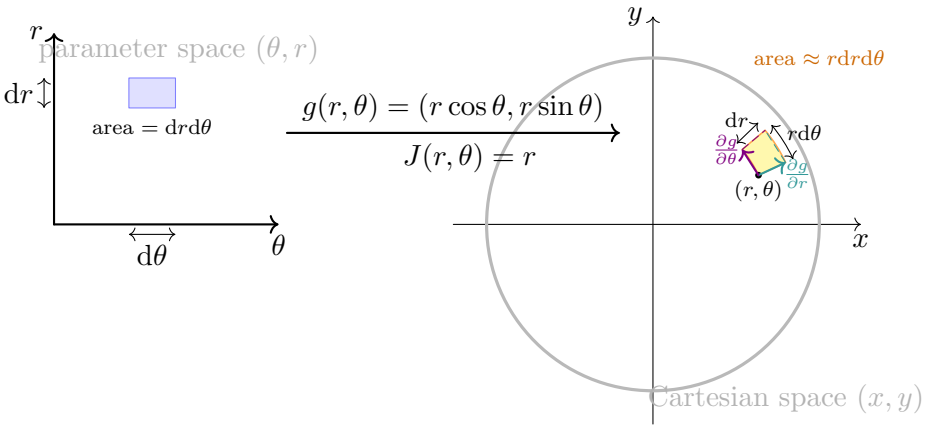
Example 1.14 (Polar coordinates)**.** Let (X, Y) be uniform on the unit disc $\{x^2 + y^2 \le 1\}$, so $f_{X,Y}(x, y) = \frac{1}{\pi}\mathbb{1}_{x^2+y^2\le 1}$. Define $(R, \Theta) = (\sqrt{X^2 + Y^2}, \arctan(Y/X))$. The inverse map is $(x, y) = (r\cos\theta, r\sin\theta)$ and

$$J(r, \theta) = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r.$$

Hence

$$f_{R,\Theta}(r, \theta) = \frac{r}{\pi}\mathbb{1}_{0\le r\le 1,\ 0\le\theta\le 2\pi}.$$

This factors into $f_R(r) = 2r\mathbb{1}_{0\le r\le 1}$ and $f_\Theta(\theta) = \frac{1}{2\pi}\mathbb{1}_{0\le\theta\le 2\pi}$, so R and Θ are independent.



1.11 Conditional Distribution and Independence

We connect conditional laws with operational criteria for independence. Conditional distributions also motivate hierarchical (mixture) models.

Independence for random variables Random variables X and Y are independent if $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for all sets $A, B \subset \mathbb{R}$. In terms of distribution functions, this is equivalent to

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \text{for all } x, y.$$

If densities exist, independence is equivalent to factorization:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y.$$

A useful consequence is that, for any functions g and h for which expectations exist,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)].$$

Conditional distributions

Definition 1.22 (Conditional pmf/pdf)**.** If (X, Y) is discrete and $p_Y(y) = \mathbb{P}(Y = y) > 0$, define

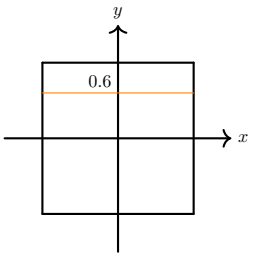
$$p_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

If (X, Y) has joint density $f_{X,Y}$ and $f_Y(y) > 0$, define

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(s, y)\mathrm{d}s.$$

Remark 1.6. The conditional distribution is a genuine distribution in x (it integrates/sums to 1). If $f_{X|Y}(x \mid y)$ does not depend on y, then X and Y are independent.

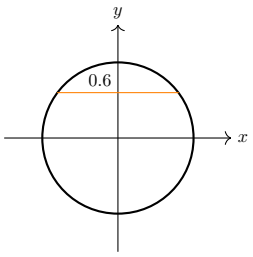
Example 1.15 (Uniform square)**.** If (X, Y) is uniform on $[-1, 1] \times [-1, 1]$, then $f_{X,Y}(x, y) = \frac{1}{4}$ on the square and 0 otherwise. Here $f_{X|Y}(x \mid y) = \frac{1}{2}\mathbb{1}_{-1 \le x \le 1}$ does not depend on y, so X and Y are independent.



Example 1.16 (Uniform disc)**.** If (X, Y) is uniform on the unit disc, then the conditional distribution $X \mid Y = y$ is uniform on

$$\left[-\sqrt{1 - y^2},\ \sqrt{1 - y^2}\right],$$

which depends on y, so X and Y are not independent.



Conditional expectation

Definition 1.23 (Conditional expectation given $Y = y$)**.** For a function g, define

$$\mathbb{E}[g(X) \mid Y = y] = \sum_x g(x) p_{X|Y}(x \mid y) \quad \text{(discrete)},$$

$$\mathbb{E}[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\mathrm{d}x \quad \text{(continuous)}.$$

The mapping $y \mapsto \mathbb{E}[g(X) \mid Y = y]$ is a function of y. When we write $\mathbb{E}[g(X) \mid Y]$ without fixing y, we mean the random variable obtained by evaluating this function at the random input Y.

Proposition 1.10 (Law of total expectation)**.** *If* $\mathbb{E}[|X|] < \infty$*, then*

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

Proposition 1.11 (Law of total variance)**.** *If* $\mathbb{E}[X^2] < \infty$*, then*

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

Proof. Use $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ and apply the law of total expectation to X and X^2:

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

The last two terms equal $\mathrm{Var}(\mathbb{E}[X \mid Y])$. □

Remark 1.7. Knowing more information typically reduces uncertainty: $\mathbb{E}[\mathrm{Var}(X \mid Y)] \le \mathrm{Var}(X)$, with equality if $\mathbb{E}[X \mid Y]$ is almost surely constant (for instance, when X is independent of Y).

A hierarchical example: Poisson thinning

Example 1.17. Let $N \sim \mathrm{Poisson}(\lambda)$ be the number of eggs produced. Given N, each egg hatches independently with probability p. Let X be the number of hatches and Y the number of non-hatches. Conditioning on $N = i + j$ gives

$$\mathbb{P}(X = i, Y = j) = \binom{i+j}{i} p^i (1-p)^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!} = \left(e^{-\lambda p} \frac{(\lambda p)^i}{i!}\right)\left(e^{-\lambda(1-p)} \frac{(\lambda(1-p))^j}{j!}\right).$$

Thus $X \sim \mathrm{Poisson}(\lambda p)$, $Y \sim \mathrm{Poisson}(\lambda(1-p))$, and X and Y are independent.

1.12 Covariance and Correlation

Covariance and correlation quantify linear dependence between random variables.

Definitions

Definition 1.24 (Covariance)**.** Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. The covariance between X and Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)],$$

whenever the expectation exists.

Expanding the product yields

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Proposition 1.12. *Covariance satisfies* $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ *and* $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$*. It is linear in each argument, for example,*

$$\mathrm{Cov}(aX + bZ, Y) = a\,\mathrm{Cov}(X, Y) + b\,\mathrm{Cov}(Z, Y).$$

More generally, for random variables $\{X_i\}_{i=1}^n$ *and* $\{Y_j\}_{j=1}^m$ *and constants* $\{a_i\}, \{b_j\}$*,*

$$\mathrm{Cov}\left(\sum_{i=1}^n a_i X_i,\ \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j\,\mathrm{Cov}(X_i, Y_j).$$

A useful special case is the variance of a sum:

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i \ne j} \mathrm{Cov}(X_i, X_j).$$

Covariance matrix For a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, the covariance matrix is

$$\Sigma = \big(\mathrm{Cov}(X_i, X_j)\big)_{1 \le i,j \le n}.$$

For any $a \in \mathbb{R}^n$, $\mathrm{Var}(a^\top \boldsymbol{X}) = a^\top \Sigma a$. In particular, Σ is positive semidefinite, since $a^\top \Sigma a \ge 0$ for all a.

Independence implies zero covariance

Proposition 1.13. *If X and Y are independent and $\mathbb{E}[|XY|] < \infty$, then $\mathrm{Cov}(X, Y) = 0$.*

Proof. Independence gives $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, so $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$. □

Remark 1.8. Zero covariance does *not* imply independence in general. It does imply independence under additional assumptions, such as joint normality.

Correlation

Definition 1.25 (Correlation)**.** If $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$, the correlation is

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Theorem 1.6 (Correlation bounds and equality cases)**.** *We always have $-1 \le \rho_{XY} \le 1$. Moreover, $\rho_{XY} = 1$ if and only if $Y = aX + b$ for some $a > 0$ almost surely, and $\rho_{XY} = -1$ if and only if $Y = -aX + b$ for some $a > 0$ almost surely.*

Remark 1.9 (Geometric view). If we regard centered random variables as vectors with inner product $\langle X, Y \rangle = \mathrm{Cov}(X, Y)$, then ρ_{XY} plays the role of the cosine of the angle between X and Y.

1.13 Marginal and Joint Distributions

Marginals do not determine the joint law. Dependence lives in the coupling between variables.

Same marginals, different dependence

Example 1.18. Let $U \sim \mathrm{Uniform}(0, 1)$. The pairs (U, U) and $(U, 1-U)$ share the same marginal distribution for each coordinate, but

$$\mathrm{Cov}(U, U) = \mathrm{Var}(U) = \frac{1}{12}, \qquad \mathrm{Cov}(U, 1-U) = -\mathrm{Var}(U) = -\frac{1}{12}.$$

Tail-integral formulas for expectation The following identities are useful when working with bounds and dependence.

Proposition 1.14 (Expectation as a tail integral)**.** *If $X \ge 0$ almost surely with CDF F, then*

$$\mathbb{E}[X] = \int_0^\infty (1 - F(x))\mathrm{d}x.$$

For a general random variable Y with CDF G,

$$\mathbb{E}[Y] = \int_0^\infty (1 - G(y))\mathrm{d}y - \int_{-\infty}^0 G(y)\mathrm{d}y,$$

whenever the integrals are finite.

Proof. For $X \geq 0$, integrate by parts with $u = x$, $\mathrm{d}v = f(x)\mathrm{d}x$:

$$\mathbb{E}[X] = \int_0^\infty x f(x)\mathrm{d}x = -x(1-F(x))|_0^\infty + \int_0^\infty (1-F(x))\mathrm{d}x = \int_0^\infty (1-F(x))\mathrm{d}x$$

Alternatively, by Fubini's theorem,

$$\int_0^\infty \mathbb{P}(X > x)\mathrm{d}x = \int_0^\infty \int_x^\infty \mathrm{d}F(t)\mathrm{d}x = \int_0^\infty \int_0^t \mathrm{d}x\mathrm{d}F(t) = \int_0^\infty t\mathrm{d}F(t) = \mathbb{E}[X]$$

For general Y, split $Y = Y^+ - Y^-$ and apply the first result:

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^\infty (1-G(y))\mathrm{d}y - \int_{-\infty}^0 G(y)\mathrm{d}y \\ &= \mathbb{E}[Y^+] - \mathbb{E}[Y^-] \quad \text{where } Y^+ = \max(Y,0), Y^- = \max(-Y,0).\end{aligned}$$ □

Proposition 1.15. *For any positive random vector $X, Y \sim H$, we have*

$$\mathbb{E}[XY] = \int_0^\infty \int_0^\infty \mathbb{P}(X > x, Y > y)\mathrm{d}y\mathrm{d}x.$$

Proof. Use Tonelli's theorem.

$$\begin{aligned}\mathbb{E}[XY] &= \mathbb{E}\Big[\Big(\int_0^\infty \mathbf{1}\{x < X\}\mathrm{d}x\Big)\Big(\int_0^\infty \mathbf{1}\{y < Y\}\mathrm{d}y\Big)\Big] = \mathbb{E}\Big[\int_0^X \int_0^Y \mathrm{d}y\mathrm{d}x\Big] \\ &= \int_0^\infty \int_0^\infty \Big[\int_0^u \int_0^v \mathrm{d}y\mathrm{d}x\Big] f(u,v)\mathrm{d}v\mathrm{d}u = \int_0^\infty \int_0^\infty \Big[\int_x^\infty \int_y^\infty f(u,v)\mathrm{d}v\mathrm{d}u\Big]\mathrm{d}y\mathrm{d}x \\ &= \int_0^\infty \int_0^\infty \mathbb{P}(X > x, Y > y)\mathrm{d}y\mathrm{d}x.\end{aligned}$$ □

Hoeffding's covariance identity and Fréchet bounds

Proposition 1.16 (Hoeffding's covariance identity)**.** *Let (X, Y) have joint CDF H and marginals F and G. Under mild integrability conditions,*

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^\infty \int_{-\infty}^\infty \big(H(x,y) - F(x)G(y)\big)\mathrm{d}y\mathrm{d}x.$$

Proof for positive random variables.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \int_0^\infty \int_0^\infty \mathbb{P}(X > x, Y > y)\mathrm{d}y\mathrm{d}x - \int_0^\infty \mathbb{P}(X > x)\mathrm{d}x \int_0^\infty \mathbb{P}(Y > y)\mathrm{d}y \\ &= \int_0^\infty \int_0^\infty \mathbb{P}(X > x, Y > y) - \mathbb{P}(X > x)\mathbb{P}(Y > y)\mathrm{d}y\mathrm{d}x \\ &= \int_0^\infty \int_0^\infty \mathbb{P}(X > x, Y > y) - (1-F(x))(1-G(y))\mathrm{d}y\mathrm{d}x\end{aligned}$$

We then use the following probability identity

$$1 + H(x,y) - F(x) - G(y) = \mathbb{P}(X > x, Y > y).$$ □

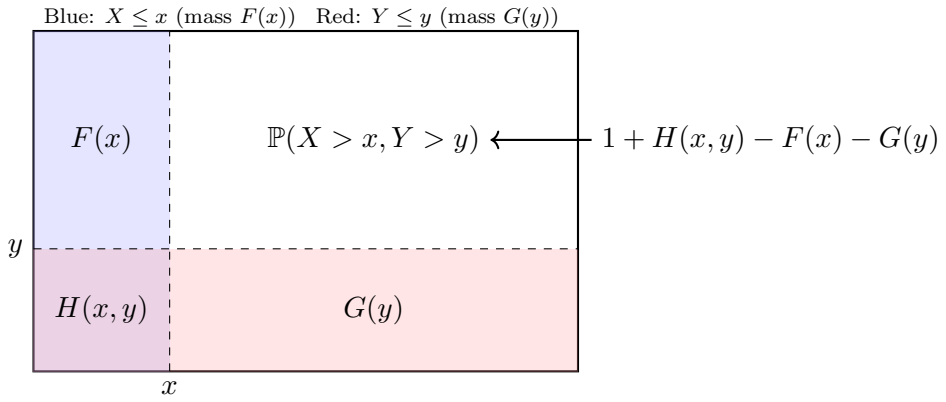


Figure 2: Inclusion-exclusion identity for joint and marginal distributions.

Proof for general random variables. We decompose a general random variable using its positive and negative parts

$$x = x^+ - x^-, \quad \text{where} \quad x^+ = \max\{x, 0\}, x^- = -\min\{x, 0\}.$$

$$\begin{aligned}
\operatorname{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[(X^+ - X^-)(Y^+ - Y^-)] - \mathbb{E}[(X^+ - X^-)]\mathbb{E}[(Y^+ - Y^-)] \\
&= (\mathbb{E}[X^+Y^+] - \mathbb{E}[X^+]\mathbb{E}[Y^+]) - (\mathbb{E}[X^-Y^+] - \mathbb{E}[X^-]\mathbb{E}[Y^+]) \\
&\quad - (\mathbb{E}[X^+Y^-] - \mathbb{E}[X^+]\mathbb{E}[Y^-]) + (\mathbb{E}[X^-Y^-] - \mathbb{E}[X^-]\mathbb{E}[Y^-]).
\end{aligned}$$

We have addressed the first term $\mathbb{E}[X^+Y^+] - \mathbb{E}[X^+]\mathbb{E}[Y^+]$. Let's look at the second term $-(\mathbb{E}[X^-Y^+] - \mathbb{E}[X^-]\mathbb{E}[Y^+])$. We want to show

$$-\operatorname{Cov}(X^-, Y^+) = \int_{-\infty}^{0}\int_{0}^{\infty} H(x,y) - F(x)G(y)\mathrm{d}y\mathrm{d}x$$

$$\begin{aligned}
\operatorname{Cov}(X^-, Y^+) &= \int_0^\infty\int_0^\infty \mathbb{P}(X^- \le x, Y^+ \le y) - \mathbb{P}(X^- \le x)\mathbb{P}(Y^+ \le y)\mathrm{d}y\mathrm{d}x \\
&= \int_0^\infty\int_0^\infty \mathbb{P}(X \ge -x, Y \le y) - \mathbb{P}(X \ge -x)\mathbb{P}(Y \le y)\mathrm{d}y\mathrm{d}x \\
&= \int_0^\infty\int_0^\infty (\mathbb{P}(Y \le y) - \mathbb{P}(X \le -x, Y \le y)) - (1 - \mathbb{P}(X \le -x))\mathbb{P}(Y \le y)\mathrm{d}y\mathrm{d}x \\
&= -\int_0^\infty\int_0^\infty H(-x,y) - F(-x)G(y)\mathrm{d}y\mathrm{d}x
\end{aligned}$$

Then we perform a change-of-variable $u = -x$, yielding

$$\operatorname{Cov}(X^-, Y^+) = -\int_{-\infty}^{0}\int_{0}^{\infty} H(x,y) - F(x)G(y)\mathrm{d}y\mathrm{d}x. \qquad \square$$

Proposition 1.17 (Fréchet–Hoeffding bounds)**.** *Given marginals F and G, any joint CDF H satisfies, for all x, y,*

$$H_*(x,y) \le H(x,y) \le H^*(x,y), \quad H^*(x,y) = \min\{F(x), G(y)\}, \quad H_*(x,y) = \big(F(x) + G(y) - 1\big)^+,$$

where $t^+ = \max\{t, 0\}$.

Proof. Note that

$$H(x,y) \le H(x,\infty) = F(x), \quad H(x,y) \le H(\infty,y) = G(y).$$

$$1 - F(x) - G(y) + H(x,y) = \mathbb{P}(X > x, Y > y) \ge 0. \qquad \square$$

Remark 1.10 (Extremal couplings via uniform rv). If $U \sim \text{Uniform}(0,1)$, then $(F^{-1}(U), G^{-1}(U))$ has joint CDF H^* and attains the largest possible correlation among all couplings with marginals F and G. Similarly, $(F^{-1}(U), G^{-1}(1-U))$ has joint CDF H_* and attains the smallest possible correlation. Intuitively, when U is large then $F^{-1}(U)$ and $G^{-1}(U)$ are large, while $1-U$ is small and hence $G^{-1}(1-U)$ is small.

Generating extremal random variables. Let $U \sim \text{Uniform}(0,1)$, then

a. $(X^*, Y^*) = (F^{-1}(U), G^{-1}(U))$ has CDF H^*.

b. $(X_*, Y_*) = (F^{-1}(U), G^{-1}(1-U))$ has CDF H_*.

Among all (X,Y) with marginal distribution F, G,

a. (X^*, Y^*) attains the largest correlation.

b. (X_*, Y_*) attains the smallest correlation.

Intuitively, when U is large then $F^{-1}(U)$ and $G^{-1}(U)$ are large, while $1-U$ is small and hence $G^{-1}(1-U)$ is small.

2 Properties of a Random Sample

2.1 Random Samples

This section fixes sampling terminology and records distributional properties of common sample statistics. The key mathematical model is an i.i.d. random sample, which is a good approximation for many large-population sampling schemes.

Population, samples, and bias A *population* is the full set of units we want to understand. A *sample* is the subset we observe. The way we select the sample matters: if selection depends on unobserved characteristics, the resulting data may not represent the population.

Remark 2.1 (Selection bias). Common selection biases include the inspection paradox (e.g., waiting-time[1] or length-of-stay bias), survivorship bias, and Simpson's paradox. These phenomena are reminders that "data" are not the same as "random draws" unless the sampling mechanism is controlled.

Random samples and the i.i.d. model

Definition 2.1 (Random sample)**.** If $X_1, \dots, X_n$ are independent and identically distributed with common distribution F, we call $(X_1, \dots, X_n)$ a *random sample of size n from F* (or from $f(\cdot \mid \theta)$ when a parametric model is specified).

If a model has density/pmf $f(\cdot \mid \theta)$, then the joint law factorizes:

$$f(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

Example 2.1 (Exponential sample)**.** If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$, then

$$f(\boldsymbol{x} \mid \lambda) = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i \Big) \prod_{i=1}^{n} \mathbb{1}_{x_i \geq 0}.$$

Remark 2.2 (Finite-population sampling). Sampling *without replacement* from a finite population is not exactly i.i.d., but the difference is often negligible when the population size is much larger than the sample size. In this course we focus on the i.i.d. framework.

Statistics: functions of the sample

Definition 2.2 (Statistic)**.** Given data $X_1, \dots, X_n$, a *statistic* is any random variable of the form $T = T(X_1, \dots, X_n)$ that depends only on the observed sample (and not on unknown parameters).

Two basic statistics summarize location and spread:

Definition 2.3 (Sample mean and sample variance)**.** The *sample mean* and *sample variance* are

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad (n \geq 2).$$

The *sample standard deviation* is $S = \sqrt{S^2}$.

[1] `https://youtu.be/wS54Gsq_4sE?si=ja4WTlpxBIV-bSwW`

Proposition 2.1 (Unbiasedness and variance of $\bar{X}$)**.** *If* $\mathbb{E}[X_i] = \mu$ *and* $\mathrm{Var}(X_i) = \sigma^2 < \infty$*, then*

$$\mathbb{E}[\bar{X}] = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Proof. By linearity, $\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \mu$. By independence and $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$,

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n}. \qquad \square$$

The variance shrinks as n grows, which captures the stabilizing effect of averaging.

Proposition 2.2 (Sample mean minimizes squared error)**.** *For observed numbers* $x_1, \ldots, x_n$*, the function* $a \mapsto \frac{1}{n}\sum_{i=1}^n (x_i - a)^2$ *is minimized at* $a = \bar{x}$*.*

Proof. Use the identity

$$\sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - a)^2,$$

which separates the within-sample variation from the deviation of a from the sample mean. $\square$

Sample variance The algebraic identity

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

is often convenient.

Proposition 2.3 (Unbiasedness of S^2)**.** *If* $\mathrm{Var}(X_i) = \sigma^2 < \infty$*, then* $\mathbb{E}[S^2] = \sigma^2$*.*

Proof. Starting from $(n-1)S^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$, take expectations:

$$(n-1)\mathbb{E}[S^2] = n\mathbb{E}[X_1^2] - n\mathbb{E}[\bar{X}^2] = n(\sigma^2 + \mu^2) - n\big(\mathrm{Var}(\bar{X}) + \mu^2\big) = n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2.$$

$\square$

Remark 2.3 (Statistics versus parameters). The statistics $\bar{X}$ and S^2 are random because they depend on random data. Parameters such as μ and σ^2 are fixed (but typically unknown). Keeping track of what is random versus fixed is essential when computing sampling distributions.

Convolution: sums of independent variables

Proposition 2.4 (Convolution formula)**.** *If* X *and* Y *are independent with densities* f_X *and* f_Y*, then* $Z = X + Y$ *has density*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w)\mathrm{d}w.$$

(Conditioning or a Jacobian on $(Z, W) = (X + Y, X)$ gives the same result.) *Video:* `https://youtu.be/IaSGqQa5O-M?si=L0sXvsZUABCURCQu`

Remark 2.4. The same formula holds for discrete random variables with summation in place of integration.

2.2 Moment Generating Function

Moment generating functions (MGFs) provide a unified tool for moments, distribution identification, and convergence arguments.

Definition and basic properties

Definition 2.4 (Moment generating function)**.** The MGF of a random variable X is

$$\phi_X(t) = \mathbb{E}[e^{tX}],$$

whenever the expectation exists (it may be infinite for some t).

Proposition 2.5 (MGF of a sum)**.** *If X and Y are independent and both MGFs are finite at t, then*

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$

Proposition 2.6 (MGF generates moments)**.** *If $\phi_X(t)$ is finite in a neighborhood of 0, then for each $k \geq 1$,*

$$\phi_X^{(k)}(0) = \mathbb{E}[X^k].$$

Proof. Differentiate $\phi_X(t) = \mathbb{E}[e^{tX}]$ under the expectation:

$$\phi'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[e^{tX}] = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}e^{tX}\right] = \mathbb{E}[Xe^{tX}],$$
$$\phi''(t) = \frac{\mathrm{d}}{\mathrm{d}t}\phi'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[Xe^{tX}] = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}Xe^{tX}\right] = \mathbb{E}[X^2e^{tX}].$$

Evaluating at $t = 0$ yields $\phi_X^{(k)}(0) = \mathbb{E}[X^k]$. The interchange of differentiation and expectation is justified, for example, by dominated convergence when ϕ_X is finite near 0. □

Interchanging limits and expectations The expectation operator is an integral, while differentiation is a limit. A standard tool is the dominated convergence theorem.

Theorem 2.1 (Dominated convergence theorem (DCT))**.** *If $X_n \to X$ pointwise and $|X_n| \leq Y$ for all n, where $\mathbb{E}[|Y|] < \infty$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.*

DCT justifies many common steps such as exchanging limits with integrals and passing derivatives through expectations when an integrable dominating function exists.

Corollary 2.1 (Differentiation Under the Integral)**.** *If f is differentiable in t and $\left|\frac{\partial f}{\partial t}(x,t)\right| \leq g(x)$ with $\int g < \infty$ near t_0, then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int f(x,t)\mathrm{d}x = \int \frac{\partial f}{\partial t}(x,t)\mathrm{d}x.$$

Identifiability via MGFs

Theorem 2.2 (Uniqueness of the MGF)**.** *If two random variables have the same MGF in an open interval containing 0, then they have the same distribution.*

This theorem implies that, for many standard families, matching MGFs is a convenient way to prove equality in distribution.

However, consider two random variables X, Y with CDF F_X, F_Y and MGF $\phi_X(\cdot), \phi_Y(\cdot)$. If $\phi_X(\cdot) \approx \phi_Y(\cdot)$, can we assert that $F_X \approx F_Y$?

Binomial and Poisson

Definition 2.5 (Binomial distribution)**.** A random variable X has a binomial distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1)$, written $X \sim \text{Binomial}(n,p)$, if

$$\mathbb{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad i = 0, 1, \dots, n.$$

Proposition 2.7 (Binomial mean, variance, and MGF)**.** *If $X \sim \text{Binomial}(n,p)$ and $q = 1-p$, then*

$$\mathbb{E}[X] = np, \qquad \text{Var}(X) = npq, \qquad \phi_X(t) = (pe^t + q)^n.$$

Proof. Write $X = \sum_{i=1}^{n} Z_i$ with $Z_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. Then $\mathbb{E}[Z_i] = p$, $\text{Var}(Z_i) = pq$, and $\phi_{Z_i}(t) = pe^t + q$. Independence gives

$$\phi_X(t) = \prod_{i=1}^{n} \phi_{Z_i}(t) = (pe^t + q)^n,$$

and the moment formulas follow. □

Proposition 2.8 (Additivity of binomials)**.** *If $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$ are independent, then $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$.*

Proof. Use $\phi_{X_1+X_2}(t) = \phi_{X_1}(t)\phi_{X_2}(t) = (pe^t + q)^{n_1+n_2}$ and uniqueness of MGFs. □

Definition 2.6 (Poisson distribution)**.** A random variable X has a Poisson distribution with parameter $\lambda > 0$, written $X \sim \text{Poisson}(\lambda)$, if

$$\mathbb{P}(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = 0, 1, 2, \dots.$$

Proposition 2.9 (Poisson mean, variance, and MGF)**.** *If $X \sim \text{Poisson}(\lambda)$, then*

$$\mathbb{E}[X] = \lambda, \qquad \text{Var}(X) = \lambda, \qquad \phi_X(t) = \exp\big(\lambda(e^t - 1)\big).$$

Proof. Compute

$$\phi_X(t) = \sum_{i=0}^{\infty} e^{ti} e^{-\lambda} \frac{\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda e^t)^i}{i!} = \exp\big(\lambda(e^t - 1)\big),$$

and differentiate at $t = 0$ to obtain the mean and second moment. □

Proposition 2.10 (Poisson as a binomial limit)**.** *If $X_n \sim \text{Binomial}(n, \dfrac{\lambda}{n})$, then $X_n \xrightarrow{d} \text{Poisson}(\lambda)$ as $n \to \infty$.*

Proof. The binomial MGF is

$$\phi_{X_n}(t) = \left(1 + \frac{\lambda}{n}(e^t - 1)\right)^n \to \exp\big(\lambda(e^t - 1)\big),$$

which is the Poisson MGF. □

Remark 2.5 (Why Poisson models?). Poisson laws often model counts of "rare and roughly independent" events in a fixed window (e.g., arrivals per hour, defect counts per batch, misprints per book).

Convergence in distribution via MGFs

Theorem 2.3 (MGF convergence implies distributional convergence)**.** *If there exists $\delta > 0$ such that $\phi_{X_n}(t) \to \phi_X(t)$ for all $|t| < \delta$, and ϕ_X is the MGF of X, then $X_n \xrightarrow{d} X$.*

Proposition 2.11 (MGF of $\bar{X}$)**.** *If $X_1, \ldots, X_n$ are i.i.d. with MGF ϕ, then*

$$\phi_{\bar{X}}(t) = \big[\phi(t/n)\big]^n.$$

Proof. Since $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$,

$$\phi_{\bar{X}}(t) = \mathbb{E}\left[\exp\left(\frac{t}{n}\sum_{i=1}^n X_i\right)\right] = \prod_{i=1}^n \mathbb{E}\left[e^{(t/n)X_i}\right] = \big[\phi(t/n)\big]^n.$$

□

2.3 Sample from a Normal Distribution

Normal-distribution facts appear repeatedly in classical inference, so we collect them here.

Standard normal and location–scale

Definition 2.7 (Standard normal)**.** A standard normal random variable $Z \sim \mathcal{N}(0,1)$ has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

and CDF $\Phi(z) = \int_{-\infty}^z f_Z(u)\mathrm{d}u$.

Proposition 2.12 (Location–scale transformation)**.** *If $Z \sim \mathcal{N}(0,1)$ and $X = \mu + \sigma Z$ with $\sigma > 0$, then $X \sim \mathcal{N}(\mu, \sigma^2)$ with*

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \qquad f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right).$$

More generally, if f is a pdf, then $g(x \mid \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ is a pdf; μ is a location parameter and σ is a scale parameter.

Normal MGF

Proposition 2.13 (MGF of $\mathcal{N}(\mu, \sigma^2)$)**.** *If $X \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$\phi_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

In particular, $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Proof. Completing the square in the exponent yields

$$\begin{aligned}\mathbb{E}[e^{tX}] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(tx - \frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\ &= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}\right) \mathrm{d}x,\end{aligned}$$

and the remaining integral equals 1 because it is a normal density. □

Sums and the multivariate normal

Proposition 2.14 (Sum of independent normals)**.** *If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ are independent, then*

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i,\ \sum_{i=1}^{n} \sigma_i^2\right).$$

Proof. Multiply MGFs and match the normal MGF form. □

Definition 2.8 (Multivariate normal)**.** A random vector $\boldsymbol{X} \in \mathbb{R}^k$ is multivariate normal, written $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its density is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

where $\boldsymbol{\Sigma}$ is symmetric positive definite.

If $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and A is an $m \times k$ matrix, then $A\boldsymbol{X}$ is normal with mean $A\boldsymbol{\mu}$ and covariance $A\boldsymbol{\Sigma}A^\top$. For a jointly normal vector, zero covariance implies independence, but this implication does not hold for arbitrary (non-normal) joint distributions.

2.4 Properties of $\bar{X}$ and S^2 in the Normal Case

The normal model is special: the sample mean and sample variance are independent, and their distributions have closed forms.

Theorem 2.4 (Normal sample)**.** *If $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, then*

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \qquad \bar{X} \text{ is independent of } S^2, \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

The first statement follows from the sum-of-normals property. The remaining two are consequences of orthogonal decompositions of a multivariate normal vector.

Chi-squared and a projection lemma

Definition 2.9 (Chi-squared distribution)**.** If $Z_1, \ldots, Z_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then

$$Q = \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k.$$

Lemma 2.1 (Projection matrix trick)**.** *Let A be symmetric idempotent ($A^2 = A$) with $\operatorname{tr}(A) = r$. If $X \sim \mathcal{N}(0, \sigma^2 I)$, then*

$$\frac{X^\top A X}{\sigma^2} \sim \chi^2_r.$$

Proof. A symmetric idempotent matrix has eigenvalues 0 or 1, and $\operatorname{tr}(A) = r$ implies rank r. Hence $A = Q^\top \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Q$ for some orthogonal Q. With $Y = QX \sim \mathcal{N}(0, \sigma^2 I)$,

$$X^\top A X = Y^\top \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} Y = \sum_{i=1}^{r} Y_i^2,$$

so dividing by σ^2 gives a sum of r squared standard normals. □

Applying the lemma to the centered vector $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ yields $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Independence of $\bar{X}$ and S^2 follows because the mean and the centered residuals are uncorrelated linear transforms of a jointly normal vector.

2.5 Student's t and Snedecor's F

Two pivotal distributions arise from ratios involving normal and chi-squared variables.

Student's t distribution

Definition 2.10 (t distribution)**.** If $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_\nu$ are independent, then

$$T = \frac{U}{\sqrt{V/\nu}}$$

has a Student t distribution with ν degrees of freedom, written $T \sim t_\nu$.

For a normal sample, the standardized mean

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

because $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0,1)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ are independent.

Snedecor's F distribution

Definition 2.11 (F distribution)**.** If $V_1 \sim \chi^2_p$ and $V_2 \sim \chi^2_q$ are independent, then

$$F = \frac{(V_1/p)}{(V_2/q)}$$

has an F distribution with (p, q) degrees of freedom, written $F \sim F_{p,q}$.

If $X_1, \ldots, X_n \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent samples, then

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{n-1,m-1}.$$

Also, if $T \sim t_q$, then $T^2 \sim F_{1,q}$.

2.6 Order Statistics

Order statistics capture the distribution of sorted sample values and are used in quantiles, medians, and nonparametric methods.

Discrete and continuous laws Let $X_1, \ldots, X_n$ be i.i.d. with continuous CDF F and pdf f. Write the sorted values as

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}.$$

Proposition 2.15 (CDF of the minimum and maximum)**.** *For a continuous distribution,*

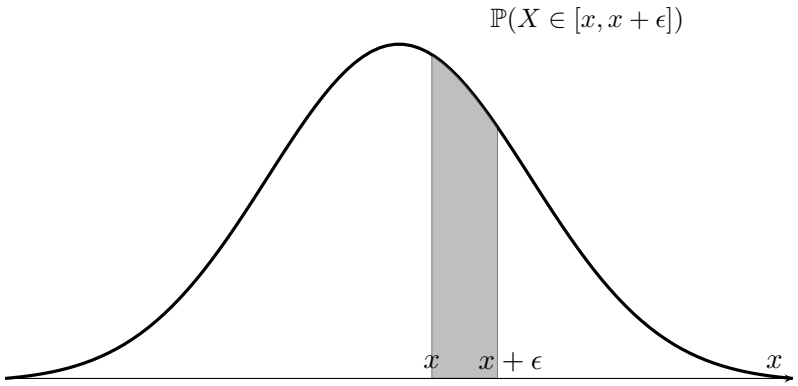
$$\mathbb{P}(X_{(1)} > x) = \mathbb{P}(X_1 > x, \ldots, X_n > x) = (1 - F(x))^n,$$

so $F_{X_{(1)}}(x) = 1 - (1 - F(x))^n$*. Similarly,* $F_{X_{(n)}}(x) = F(x)^n$*.*

Proposition 2.16 (Density of the kth order statistic)**.** *For* $1 \le k \le n$*,*

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} \big(1 - F(x)\big)^{n-k} f(x).$$

Remark 2.6 (Heuristic). The factor $F(x)^{k-1}(1-F(x))^{n-k}$ represents the probability that $k-1$ sample points lie below x and $n-k$ lie above x, while $f(x)\mathrm{d}x$ captures the chance that one observation falls in an infinitesimal neighborhood of x. The multinomial coefficient counts which observation plays which role.



Theorem 2.5 (Joint Density of Two Order Statistics)**.** *If $i<j$ and $u<v$,*

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f(u)f(v)[F(u)]^{i-1}[F(v)-F(u)]^{j-i-1}[1-F(v)]^{n-j}.$$

- Informal proof: $f_{i,j}(u,v) = \mathbb{P}(i-1$ less than u, $n-j$ greater than v, one at u, one at $v)$.
- Be careful about the domain!
- $f_{X_{(1)},\ldots,X_{(n)}}(x_1,\ldots,x_n) = n!f(x_1)\cdots f(x_n)$, $x_1 < x_2 < \cdots < x_n$.

Example 2.2 (Range and Midrange for Uniform$(0,a)$.)**.** For $0 < x_1 < x_n < a$,

$$f_{X_{(1)},X_{(n)}}(x_1,x_n) = \frac{n(n-1)}{a^n}(x_n - x_1)^{n-2}.$$

Let $R = X_{(n)} - X_{(1)}$, $V = \frac{X_{(n)}+X_{(1)}}{2}$; then with Jacobian 1,

$$f_{R,V}(r,v) = \frac{n(n-1)}{a^n} r^{n-2}, \qquad 0 < r < a,\ \frac{r}{2} < v < a - \frac{r}{2}.$$

Range and midrange are independent.

2.7 LLN and CLT

We now record the two fundamental large-sample results: the law of large numbers (LLN) and the central limit theorem (CLT).

Modes of convergence Let X_n be random variables and X a random variable.

Definition 2.12 (Almost sure convergence)**.** $X_n \to X$ almost surely, written $X_n \xrightarrow{a.s.} X$, if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.

Definition 2.13 (Convergence in probability)**.** $X_n \to X$ in probability, written $X_n \Rightarrow X$, if for every $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0.$$

Definition 2.14 (Convergence in distribution)**.** $X_n \xrightarrow{d} X$ if $F_{X_n}(x) \to F_X(x)$ at all continuity points x of F_X.

Almost sure convergence implies convergence in probability, which implies convergence in distribution:

$$\xrightarrow{a.s.} \quad \text{implies} \quad \Rightarrow \quad \text{implies} \quad \xrightarrow{d}.$$

The reverse implications do not hold in general.

Example 2.3 (Convergence in probability does not imply a.s.)**.** Work on $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}, \lambda)$. For $k \geq 1$ and $n = 2^k, \ldots, 2^{k+1} - 1$, write $j = n - 2^k + 1$ and set

$$A_n = \left(\tfrac{j-1}{2^k}, \tfrac{j}{2^k}\right], \qquad X_n(\omega) = \mathbb{1}_{A_n}(\omega).$$

Then for $n \in [2^k, 2^{k+1} - 1]$, $\mathbb{P}(X_n = 1) = \lambda(A_n) = 2^{-k} \to 0$, so $X_n \to 0$ in probability. However, for any fixed $\omega \in (0, 1]$ and each k there is exactly one $n \in [2^k, 2^{k+1} - 1]$ with $\omega \in A_n$, hence $X_n(\omega) = 1$ infinitely often. Thus X_n does not converge to 0 almost surely.

Example 2.4 (Convergence in distribution does not imply convergence in probability)**.** Let X be a standard normal random variable and let $X_n = -X$. X_n converges in distribution to X, but not in probability.

Theorem 2.6 (Continuous Mapping Theorem)**.** *If $X_n \to X$ (in any of the three modes) and g is continuous, then $g(X_n) \to g(X)$ in the same mode.*

Theorem 2.7. *If the limit is deterministic, $\Rightarrow$ and $\xrightarrow{d}$ are equivalent.*

Example 2.5. Let $X_{(n)}$ be the max of n independent Uniform $(0, 1)$. The CDF of $X_{(n)}$ is

$$F_{X_{(n)}} = F^n(x) = x^n \to \mathbb{1}_{x \geq 1}.$$

The CDF converges to that of a constant 1, so $X_{(n)} \xrightarrow{d} 1$. This implies that $X_{(n)} \Rightarrow 1$. In fact, it can be directly verified:

$$\mathbb{P}(|X_{(n)} - 1| \geq \epsilon) = \mathbb{P}(X_{(n)} \leq 1 - \epsilon) = (1 - \epsilon)^n \to 0$$

So we have convergence in probability. Furthermore,

$$\mathbb{P}(|X_{(n)} - 1| \geq \epsilon) = (1 - \epsilon)^n.$$

Let $\epsilon = t/n$

$$\mathbb{P}(n(1 - X_{(n)}) \leq t) = \mathbb{P}(X_{(n)} \leq 1 - t/n) = (1 - t/n)^n \to e^{-t}.$$

So $n(1 - X_{(n)}) \xrightarrow{d}$ Exponential(1).

Laws of large numbers

Theorem 2.8 (Weak law of large numbers)**.** *If $X_1, \ldots$ are i.i.d. with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}(X_1) < \infty$, then*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \Rightarrow \mu.$$

Proof. By the variance calculation, $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$. Chebyshev's inequality gives

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0. \qquad \square$$

Definition 2.15 (Consistency)**.** An estimator T_n is **consistent** for θ if $T_n \Rightarrow \theta$.

Example 2.6. WLLN $\Rightarrow \bar{X}_n$ is consistent for μ. Consider the sample variance

$$S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}, \quad \mathbb{P}(|S_n^2 - \sigma^2| > \epsilon) \leq \frac{\mathbb{E}(S_n^2 - \sigma^2)^2}{\epsilon^2}.$$

So if $\mathrm{Var}(S_n^2) \to 0$ as $n \to \infty$ (true for normal), then the estimator is consistent. By continuous mapping ($g(x) = \sqrt{x}$), S_n is also a consistent estimator for σ (but biased by Jensen's inequality).

Theorem 2.9 (Strong law of large numbers)**.** *If $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}[|X_1|] < \infty$, then $\bar{X}_n \xrightarrow{a.s.} \mu$.*

Remark 2.7. The strong law is deeper than the weak law. We will typically use the weak law for intuition and the strong law when we need almost sure statements.

Central limit theorem

Theorem 2.10 (Central limit theorem)**.** *If $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}(X_1) = \sigma^2 \in (0, \infty)$, then*

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

Equivalently, for large n,

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

MGF sketch. Write $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$. Using the MGF of $\bar{X}_n$,

$$\phi_{Z_n}(t) = \left[\phi_{(X_1-\mu)/\sigma}\left(\frac{t}{\sqrt{n}}\right)\right]^n.$$

A Taylor expansion of ϕ at 0 yields $\phi_{Z_n}(t) \to e^{t^2/2}$, the MGF of $\mathcal{N}(0,1)$. □

Example 2.7 (Normal Approximation to Binomial)**.** Let $X \sim \mathrm{Binomial}(n,p)$, fixed p, n large:

$$\mathbb{P}(a \leq X \leq b) \approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right).$$

Continuity correction: for $\mathbb{P}(X = a)$, use $[a - \frac{1}{2}, a + \frac{1}{2}]$ before standardizing.

Theorem 2.11 (Slutsky)**.** *If $X_n \xrightarrow{d} X$ and $Y_n \Rightarrow a$, then $X_n Y_n \xrightarrow{d} aX$ and $X_n + Y_n \xrightarrow{d} X + a$.*

Slutsky's theorem lets us replace unknown constants by consistent estimators inside limiting distributions and still keep the same limit.

Example 2.8 (Student's t statistic converges to standard normal)**.** Student's t statistic converges to standard normal:

$$T_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n} \cdot \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

Example 2.9 (Counterexample)**.** $X_n \sim \mathrm{Uniform}(0,1)$ and $Y_n = -X_n$. The sum $X_n + Y_n = 0$ for all values of n. Moreover, $Y_n \xrightarrow{d} \mathrm{Uniform}(-1,0)$, but $X_n + Y_n$ does not converge in distribution to $X + Y$.

2.8 Delta Method

The delta method translates a CLT for $\bar{X}_n$ into a CLT for smooth functions of $\bar{X}_n$.

First- and second-order delta methods

Theorem 2.12 (First-order delta method)**.** *If $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and g is differentiable at μ with $g'(\mu) \neq 0$, then*

$$\sqrt{n}\big(g(\bar{X}_n) - g(\mu)\big) \xrightarrow{d} \mathcal{N}\big(0, (g'(\mu))^2\sigma^2\big).$$

Proof. By Taylor's theorem, for some θ_n between Y_n and θ,

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta_n)}{2}(Y_n - \theta)^2.$$

Then use Slutsky's theorem to the sum. □

Theorem 2.13 (Second-order delta method)**.** *If $g'(\mu) = 0$ but $g''(\mu) \neq 0$, then*

$$n\big(g(\bar{X}_n) - g(\mu)\big) \xrightarrow{d} \frac{1}{2} g''(\mu)\sigma^2\chi_1^2,$$

under suitable regularity conditions.

Proof. By Taylor's theorem, for some θ_n between Y_n and θ,

$$g(Y_n) - g(\theta) = g'(\theta)(Y_n - \theta) + \tfrac{1}{2}\, g''(\theta_n)\,(Y_n - \theta)^2.$$

Since $g'(\theta) = 0$, we have

$$n\big(g(Y_n) - g(\theta)\big) = \tfrac{1}{2}\, g''(\theta_n)\,\big[\sqrt{n}(Y_n - \theta)\big]^2.$$

Then use Slutsky's theorem to the sum. □

Application: estimating odds

Example 2.10. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ with $\mu = p$ and $\sigma^2 = p(1-p)$. The *odds* are $\theta = p/(1-p)$, and the plug-in estimator is $\hat{\theta} = \bar{X}/(1-\bar{X})$.

Take $g(x) = x/(1-x)$, so $g'(p) = 1/(1-p)^2$. By the delta method,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{p(1-p)}{(1-p)^4}\right) = \mathcal{N}\left(0,\ \frac{p}{(1-p)^3}\right).$$

3 Principles of Data Reduction

In statistics, a central task is to turn a large sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ into *inferences* about the data-generating process—for example, about an unknown parameter θ indexing a family of distributions. A basic question is whether we must keep *all* of the data to make good inferences.

Example 3.1 (A motivating example)**.** Suppose $X_1, \ldots, X_n \sim \text{Uniform}([0, \theta])$ and we observe

$$[2.16,\ 0.72,\ 9.75,\ 0.89,\ 2.21].$$

What can we say about θ?

Not all features of the sample are relevant to a given statistical question. A *data reduction* procedure discards information that is irrelevant for the parameter of interest, leading to simpler inference.

Definition 3.1 (Statistic)**.** A *statistic* $T(\boldsymbol{X})$ is any function of the sample; it does not depend on unknown parameters.

Example 3.2. In the Uniform($[0, \theta]$) example, $X_{(n)} = \max\{X_1, \ldots, X_n\}$ is a statistic.

If T is not one-to-one, it effects data reduction: distinct samples may yield the same value of T and thus become indistinguishable through the lens of T. A "good" statistic should preserve information about the unknown parameter θ.

Question. Is there a statistic that contains all the information about θ in the sample?

If such a statistic exists, we can compress the original data to a simpler object without losing information about θ.

3.1 Sufficient statistics

Definition 3.2 (Sufficient statistic)**.** A statistic $T(\boldsymbol{X})$ is *sufficient*[2] for a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if, for every t, the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X}) = t$ does not depend on θ.

Sufficiency is model-dependent: it refers to $\mathcal{P}$[3] and hence to the parameter θ. Intuitively, once a sufficient statistic T is known, the remaining features of the data carry no additional information about θ; one can estimate θ just as well from T as from the full sample.

One useful way to see this is to regard the full data as "dummy" randomness generated around T in two stages. First draw the statistic,

$$T(\boldsymbol{X}) \sim \mathbb{P}_\theta\big(T(\boldsymbol{X}) = t\big),$$

and then draw a conditional sample $\boldsymbol{X} \mid T(\boldsymbol{X}) = t$. This reproduces the joint law via

$$\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}_\theta\big(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x})\big)\, \mathbb{P}_\theta\big(T(\boldsymbol{X}) = T(\boldsymbol{x})\big).$$

By sufficiency, $\boldsymbol{X} \mid T(\boldsymbol{X}) = t$ is free of θ, so all information about θ is contained in $T(\boldsymbol{X})$:

$$\mathbb{P}_\theta\big(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x})\big) = \mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x})\big).$$

Example 3.3 (Bernoulli model)**.** Let $X_1, \ldots, X_n$ be a random sample from Bernoulli(θ). Consider the count of successes $T = \sum_{i=1}^n X_i$. The joint pmf is

$$\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}.$$

[2] Introduced by R. A. Fisher (1922).

[3] Think of $\mathcal{P}$ as a family of distributions parametrized by θ.

For $t \in \{0, 1, \dots, n\}$,

$$\mathbb{P}_\theta\big(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = t\big) = \frac{\mathbb{P}_\theta\big(\boldsymbol{X} = \boldsymbol{x},\, T(\boldsymbol{X}) = t\big)}{\mathbb{P}_\theta\big(T(\boldsymbol{X}) = t\big)} = \frac{\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}\,\mathbb{1}_{\{t=\sum_i x_i\}}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$
$$= \frac{\mathbb{1}_{\{t=\sum_i x_i\}}}{\binom{n}{t}}, \qquad x_i \in \{0,1\}.$$

The conditional law does not depend on θ, hence $\sum_i X_i$ is sufficient for θ.
Question. How should we interpret $\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = t)$ here? Given $T = t$ we are "uniformly choosing" the t positions of the ones among the n coordinates (the coordinates are no longer independent under this conditioning).

Example 3.4 (Uniform($[0, \theta]$))**.** Conditioning on $X_n = T(\boldsymbol{X}) = X_{(n)} = t$, the remaining $n-1$ observations behave like a random sample from Uniform($[0, t]$), independent of θ:

$$\begin{aligned}
&\mathbb{P}_\theta\big(X_1 \le x_1, \dots, X_{n-1} \le x_{n-1} \mid X_n = X_{(n)} = t\big) \\
&\quad= \frac{\mathbb{P}_\theta\big(X_1 \le x_1, \dots, X_{n-1} \le x_{n-1},\, X_n = X_{(n)} = t\big)}{\mathbb{P}_\theta\big(X_n = X_{(n)} = t\big)} \\
&\quad= \frac{\mathbb{P}_\theta\big(X_1 \le x_1 \wedge t, \dots, X_{n-1} \le x_{n-1} \wedge t,\, X_n = t\big)}{\mathbb{P}_\theta\big(X_1 \le t, \dots, X_{n-1} \le t,\, X_n = t\big)} \\
&\quad= \frac{\mathbb{P}_\theta(X_1 \le x_1 \wedge t) \cdots \mathbb{P}_\theta(X_{n-1} \le x_{n-1} \wedge t)\,\mathbb{P}_\theta(X_n = t)}{\mathbb{P}_\theta(X_1 \le t) \cdots \mathbb{P}_\theta(X_{n-1} \le t)\,\mathbb{P}_\theta(X_n = t)} \\
&\quad= \prod_{i=1}^{n-1} \frac{x_i \wedge t}{t}\,\mathbb{1}_{\{x_i \ge 0\}} \overset{i.i.d.}{\sim} \text{Uniform}([0, t]).
\end{aligned}$$

Here $a \wedge b = \min\{a, b\}$. Recall that for indicator functions, $\mathbb{1}_A \mathbb{1}_B = \mathbb{1}_{A \cap B}$.

Consequently,

$$\begin{aligned}
&\mathbb{P}_\theta\big(X_1 \le x_1, \dots, X_n \le x_n \mid X_{(n)} = t\big) \\
&\quad= \sum_{i=1}^{n} \mathbb{P}_\theta\big(X_1 \le x_1, \dots, X_n \le x_n,\, X_i = X_{(n)} \mid X_{(n)} = t\big) \\
&\quad= \sum_{i=1}^{n} \mathbb{P}_\theta\big(X_1 \le x_1, \dots, X_n \le x_n \mid X_i = X_{(n)} = t\big)\,\mathbb{P}_\theta\big(X_i = X_{(n)} \mid X_{(n)} = t\big) \\
&\quad= \sum_{i=1}^{n} \prod_{j \ne i} \frac{x_j \wedge t}{t}\,\mathbb{1}_{\{x_j \ge 0\}} \times \frac{1}{n}.
\end{aligned}$$

This does not depend on θ, so by definition $X_{(n)}$ is sufficient for θ.

Example 3.5 (Normal mean, known variance)**.** Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with known σ. Consider $T(\boldsymbol{X}) = \bar{X} = \frac{1}{n}\sum_i X_i$. Writing

$$\mathbb{P}_\theta\big(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x})\big) = \frac{f(\boldsymbol{x} \mid \mu)}{q(T(\boldsymbol{x}) \mid \mu)},$$

we have

$$\begin{aligned}
f(\boldsymbol{x} \mid \mu) &= \frac{1}{(\sqrt{2\pi}\,\sigma)^n} \exp\left(-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{(\sqrt{2\pi}\,\sigma)^n} \exp\left(-\frac{\sum_i (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right),
\end{aligned}$$

and

$$q\big(T(\boldsymbol{x}) \mid \mu\big) = q(\bar{x} \mid \mu) = \frac{1}{\sqrt{2\pi}\,\frac{\sigma}{\sqrt{n}}} \exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right).$$

Thus $\bar{X}$ is sufficient for μ, but not for σ.

Example 3.6 (Order statistics)**.** Let $X_1, \dots, X_n$ be a random sample from a distribution with density f, and set $T(\boldsymbol{X}) = (X_{(1)}, \dots, X_{(n)})$. For any $\boldsymbol{x}$ that is a permutation $\pi(\boldsymbol{X})$ of the sample,

$$\mathbb{P}_\theta\big(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x})\big) = \frac{f(\boldsymbol{x})}{q\big(T(\boldsymbol{x})\big)} = \frac{\prod_{i=1}^n f(x_i)}{n! \prod_{i=1}^n f(x_{(i)})} = \frac{1}{n!}.$$

Thus, conditional on the order statistics, all $n!$ permutations are equally likely.

This is a nonparametric example: the "parameter" here is the entire density f. There is little data reduction because T is n-dimensional. Outside the exponential family, it is rare to have sufficient statistics of lower dimension than the sample size.

3.1.1 Factorization theorem

Using the definition of sufficiency directly can be awkward—both for finding a candidate sufficient statistic and for checking sufficiency. The factorization theorem streamlines both tasks.

Theorem 3.1 (Factorization theorem)**.** *Let $f(\boldsymbol{x} \mid \theta)$ denote the joint pdf/pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is sufficient for θ if and only if there exist functions $h(\boldsymbol{x})$ and $g(t \mid \theta)$ such that, for all sample points $\boldsymbol{x}$ and all θ,*

$$f(\boldsymbol{x} \mid \theta) = h(\boldsymbol{x})\, g\big(T(\boldsymbol{x}) \mid \theta\big).$$

Remark 3.1. The function h may depend on the full sample $\boldsymbol{x}$ but not on the parameter θ. The function g may depend on θ, but it can depend on $\boldsymbol{x}$ only through $t = T(\boldsymbol{x})$.

Proposition 3.1 (One-to-one transformations preserve sufficiency)**.** *If $s(\cdot)$ is one-to-one and T is sufficient for θ, then $S(\boldsymbol{X}) \triangleq s\big(T(\boldsymbol{X})\big)$ is also sufficient for θ.*

Example 3.7 (Order statistic)**.** In the order-statistic example, the empirical cdf

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \le x\}}$$

is an equivalent sufficient statistic.

Proof of the factorization theorem. "⇒" If T is sufficient, let

$$h(\boldsymbol{x}) = \mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x})\big).$$

Then

$$f(\boldsymbol{x} \mid \theta) = \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}\big(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x})\big)\, \mathbb{P}_\theta\big(T(\boldsymbol{X}) = T(\boldsymbol{x})\big),$$

which has the desired form with $g(t \mid \theta) = \mathbb{P}_\theta(T(\boldsymbol{X}) = t)$.

"⇐" (discrete case) If $f(\boldsymbol{x} \mid \theta) = h(\boldsymbol{x})\, g(T(\boldsymbol{x}) \mid \theta)$, let $q(t \mid \theta)$ be the pmf of $T(\boldsymbol{X})$ and define

$$A_t = \{\boldsymbol{y} : T(\boldsymbol{y}) = t\}.$$

Then

$$\begin{aligned}\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid T = t) &= \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x},\, T = t)}{\mathbb{P}(T = t)} = \frac{f(\boldsymbol{x} \mid \theta)\, \mathbb{1}_{\{T(\boldsymbol{x}) = t\}}}{\sum_{\boldsymbol{y} \in A_t} f(\boldsymbol{y} \mid \theta)} \\ &= \frac{h(\boldsymbol{x})\, \mathbb{1}_{\{T(\boldsymbol{x}) = t\}}}{\sum_{\boldsymbol{y} \in A_t} h(\boldsymbol{y})},\end{aligned}$$

which is free of θ and proves sufficiency. For a full treatment of the continuous case, see Lehmann and Romano, *Testing Statistical Hypotheses* (2015), Section 2.6. □

Example 3.8 (Uniform($[0,\theta]$) revisited)**.** The joint density can be written as

$$f(\boldsymbol{x} \mid \theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{\{0 \le x_i \le \theta\}} = \left[\prod_{i=1}^{n} \mathbb{1}_{\{x_i \ge 0\}} \right] \frac{1}{\theta^n} \mathbb{1}_{\{\max_i x_i \le \theta\}}.$$

By the factorization theorem, $T(\boldsymbol{X}) = \max_i X_i = X_{(n)}$ is sufficient for θ. In contrast, the sample mean is not a sufficient statistic for $\mathbb{E}[X] = \theta/2$.

Example 3.9 (Normal mean revisited)**.** Consider a normal random sample $\mathcal{N}(\mu, \sigma^2)$ with known σ.

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

Then $T(\boldsymbol{X}) = \bar{X}$ is sufficient, with $g(t \mid \mu) = \exp\left(-\frac{n(t-\mu)^2}{2\sigma^2} \right)$.

3.1.2 Multi-dimensional sufficient statistics

Definition 3.3 (Multi-dimensional case)**.** Statistics $\boldsymbol{T} = (T_1, \ldots, T_k)$ are jointly sufficient if, for each $\boldsymbol{t} = (t_1, \ldots, t_k)$, the conditional distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ given $\boldsymbol{T} = \boldsymbol{t}$ does not depend on $\boldsymbol{\theta}$.

The factorization theorem applies verbatim to multi-dimensional parameters and statistics.

Example 3.10 (Normal mean and variance.)**.** If $\boldsymbol{\theta} = (\mu, \sigma)$ is unknown, let $T_1(\boldsymbol{X}) = \bar{X}$ and $T_2(\boldsymbol{X}) = S^2$. Then

$$g(\boldsymbol{t} \mid \boldsymbol{\theta}) = g(t_1, t_2 \mid \mu, \sigma) = \frac{1}{(\sqrt{2\pi}\,\sigma)^n} \exp\left(-\frac{(n-1)t_2}{2\sigma^2} \right) \exp\left(-\frac{n(t_1 - \mu)^2}{2\sigma^2} \right),$$

with $h(\boldsymbol{x}) = 1$. Hence $(\bar{X}, S^2)$ is sufficient for (μ, σ).

Remark 3.2 (Mappings of sufficient statistics). Suppose $\boldsymbol{\theta} \in \mathbb{R}^k$. If $\boldsymbol{f}$ is a one-to-one function on $\mathbb{R}^k$ and $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$, then $\boldsymbol{f}(\boldsymbol{T})$ is also sufficient. More generally, if T is sufficient and $T = \psi(S)$ for some (measurable, not necessarily one-to-one) function ψ and some statistic S, then S is sufficient.

Example 3.11. For normal population, the statistics $T_1 = \bar{X}$, $T_2 = \left(X_1, \sum_{i=2}^{n} X_i\right)$, and $T_3 = \boldsymbol{X}$ are all sufficient for μ.

3.1.3 Partitions induced by statistics

Any statistic T induces a partition of the sample space according to its value,

$$\mathcal{A}(T) = \{A_t\}, \qquad A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}.$$

Example 3.12 (Uniform$(0,\theta)$ with $n = 2$.)**.** Figure 3 illustrates three partitions: one induced by $T(\boldsymbol{X}) = X_{(2)}$, one by the full data $T(\boldsymbol{X}) = (X_1, X_2)$, and one by $T(\boldsymbol{X}) = X_1$.

Remark 3.3. For a statistic to be sufficient, its partition should be fine enough to distinguish information about different θ. Equivalently, if T is sufficient, we should draw identical statistical conclusions about θ at every point inside a cell A_t.

It is this partition, rather than the particular formula defining T, that is the fundamental object. In measure-theoretic terms, the statistic generates a σ-algebra, and sufficiency is a statement about that generated σ-algebra.

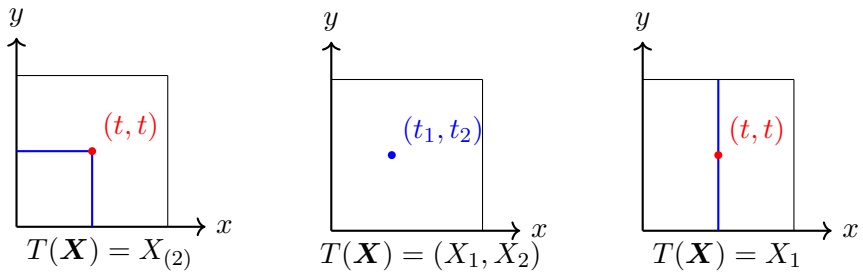


Figure 3: Partitions of the sample space induced by different statistics for a sample of size 2 from Uniform$(0, \theta)$.

3.1.4 Exponential families

Definition 3.4 (Exponential family)**.** A family of pdfs/pmfs is called a k-parameter exponential family if it can be written as

$$f(x \mid \boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \, \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})\, t_i(x) \right).$$

This special form is chosen for mathematical convenience.

Example 3.13 (Binomial(n, p).)**.**

$$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (1-p)^n \exp\left(x \log \frac{p}{1-p} \right).$$

Example 3.14 (Normal $\mathcal{N}(\mu, \sigma^2)$.)**.**

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

Example 3.15 (Counterexample.)**.** The shifted exponential distributions do not form an exponential family:

$$f(x \mid \theta) = \frac{1}{\theta} \exp\left(\frac{\theta - x}{\theta} \right) \mathbb{1}_{\{x \geq \theta\}}.$$

Natural parameters and canonical form An exponential family is sometimes reparameterized by letting $w_i(\boldsymbol{\theta}) = \eta_i$:

$$f(x \mid \boldsymbol{\eta}) = h(x)\, c^*(\boldsymbol{\eta}) \, \exp\left(\sum_{i=1}^{k} \eta_i\, t_i(x) \right),$$

where $\boldsymbol{\eta} = (w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta}))$ is called the *natural parameter*. This representation is the *canonical form*.

Definition 3.5 (Natural parameter space)**.** The natural parameter space is

$$\mathcal{H} = \left\{ \boldsymbol{\eta} = (\eta_1, \ldots, \eta_k) : \int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x) \right) \mathrm{d}x < \infty \text{ (or } \sum_{x} \cdots < \infty) \right\}.$$

Example 3.16 (Exponential distribution)**.** For the exponential distribution, $\mathcal{H} = \{\eta > 0\}$. Using Hölder's inequality, one can prove that $\mathcal{H}$ is convex.

Natural sufficient statistics Suppose $X_1, \ldots, X_n$ is a random sample from the canonical exponential family

$$f(x \mid \boldsymbol{\eta}) = h(x)\, c^*(\boldsymbol{\eta}) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right).$$

Define $\boldsymbol{T} = (T_1(\boldsymbol{X}), \ldots, T_k(\boldsymbol{X}))$ by

$$T_i(\boldsymbol{X}) = \sum_{j=1}^{n} t_i(X_j), \qquad i = 1, \ldots, k.$$

In matrix form,

$$f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\eta}) = \left(\prod_{j=1}^{n} h(x_j)\right) [c^*(\boldsymbol{\eta})]^n \exp\bigl(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x})\bigr).$$

Proposition 3.2 (Natural sufficient statistics)**.** *By the factorization theorem, the natural statistic*

$$\boldsymbol{T}(\boldsymbol{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is sufficient for $\boldsymbol{\eta}$. Moreover, $\boldsymbol{T}$ itself belongs to an exponential family:

$$f_T(\boldsymbol{u} \mid \boldsymbol{\eta}) = \tilde{h}(\boldsymbol{u})[c^*(\boldsymbol{\eta})]^n \exp\bigl(\boldsymbol{\eta}^\top \boldsymbol{u}\bigr).$$

Example 3.17 (Bernoulli(p))**.** We can write

$$f(x \mid p) = p^x (1-p)^{1-x} = (1-p) \exp\left(x \log \frac{p}{1-p}\right).$$

So $k = 1$, $t_1(x) = x$, and $\eta = \log \frac{p}{1-p}$. The natural sufficient statistic is

$$T = T_1(X_1, \ldots, X_n) = X_1 + \cdots + X_n.$$

Since $\eta = \log \frac{p}{1-p}$ is a one-to-one mapping of p, T is also sufficient for p.

3.2 Minimal sufficient statistics

For a given parameter, there are many sufficient statistics. Sufficiency implies no loss of information for θ, but by itself it does not guarantee data reduction (for example, the full data $\boldsymbol{X}$ is always sufficient). This motivates the search for a sufficient statistic that gives the "maximal" reduction.

Definition 3.6 (Minimal sufficient statistic)**.** A sufficient statistic $T(\boldsymbol{X})$ is *minimal sufficient* if, for any other sufficient statistic $S(\boldsymbol{X})$, there exists a (measurable) function ψ such that

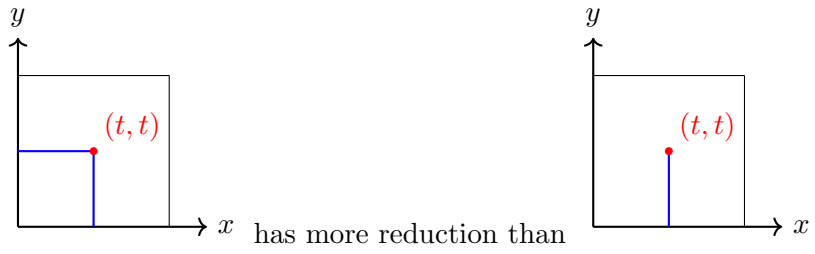
$$T = \psi(S) \qquad \text{a.s. under every } \mathbb{P}_\theta.$$

Recall the partition induced by a statistic,

$$\mathcal{A}(T) = \{A_t\}, \qquad A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}.$$

Minimality of T implies the following: for any sufficient statistic S, if $S(\boldsymbol{x}) = S(\boldsymbol{y})$, then $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Equivalently, the partition $\mathcal{A}(T)$ is coarser than $\mathcal{A}(S)$. The simpler the partition is, the more data reduction we have.

While retaining all information about θ, minimal sufficiency identifies the maximal reduction of the data, the coarsest (simplest) partition of the sample space, and (in measure-theoretic language) the coarsest σ-algebra that remains sufficient.



Proposition 3.3 (One-to-one mapping)**.** *Any one-to-one function of a minimal sufficient statistic is minimal sufficient.*

Remark 3.4 (Uniqueness). Minimal sufficient statistics are unique up to one-to-one measurable transformations: two statistics that are one-to-one functions of each other induce the same partition, and it is this partition that is the fundamental object.

Example 3.18 (Normal sample: a non-minimal sufficient statistic)**.** For a normal sample with known σ, compare

$$T(\boldsymbol{X}) = \bar{X}, \qquad T'(\boldsymbol{X}) = (\bar{X}, S^2).$$

Recall that $\bar{X}$ and S^2 are independent. Therefore $T'(\boldsymbol{X})$ cannot be written as a function of $T(\boldsymbol{X})$, so T' is not minimal.

How can we check whether $\bar{X}$ is minimal? The following "checking rule" gives a practical criterion.

Theorem 3.2 (Checking rule)**.** *Let $f(\boldsymbol{x} \mid \theta)$ be the pmf/pdf of a sample $\boldsymbol{X}$. Suppose there exists a statistic $T(\cdot)$ such that, for every two sample realizations $\boldsymbol{x}$ and $\boldsymbol{y}$,*

$$\frac{f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{y} \mid \theta)} \text{ does not depend on } \theta \quad \Longleftrightarrow \quad T(\boldsymbol{x}) = T(\boldsymbol{y}).$$

Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

Example 3.19 (Normal minimal sufficient statistics)**.** Normal minimal sufficient statistics for $\theta = \mu$ and for $\theta = (\mu, \sigma^2)$. One can examine

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

Proof of the checking rule. **Sufficiency.** Consider the partition $\mathcal{A}(T) = \{A_t\}$ induced by T, where $A_t \triangleq \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}$. Select (arbitrarily) a representative point $\boldsymbol{x}_t \in A_t$ for each cell. Since $T(\boldsymbol{x}) = T(\boldsymbol{x}_{T(\boldsymbol{x})})$, the ratio

$$h(\boldsymbol{x}) \triangleq \frac{f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{x}_{T(\boldsymbol{x})} \mid \theta)}$$

does not depend on θ. Let $g(t \mid \theta) = f(\boldsymbol{x}_t \mid \theta)$; by construction, g depends on $\boldsymbol{x}$ only through $t = T(\boldsymbol{x})$. The factorization theorem then yields

$$f(\boldsymbol{x} \mid \theta) = h(\boldsymbol{x})\, g\big(T(\boldsymbol{x}) \mid \theta\big),$$

so T is sufficient.

Minimality. Let $T'(\boldsymbol{X})$ be any sufficient statistic, so $f(\boldsymbol{x} \mid \theta) = g'\big(T'(\boldsymbol{x}) \mid \theta\big)h'(\boldsymbol{x})$. If $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$, then

$$\frac{f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{y} \mid \theta)} = \frac{g'\big(T'(\boldsymbol{x}) \mid \theta\big)h'(\boldsymbol{x})}{g'\big(T'(\boldsymbol{y}) \mid \theta\big)h'(\boldsymbol{y})} = \frac{h'(\boldsymbol{x})}{h'(\boldsymbol{y})}.$$

The right-hand side is free of θ, so by the assumed equivalence for T we must have $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Hence $T(\boldsymbol{X})$ is a function of $T'(\boldsymbol{X})$, which is exactly minimality. □

Example 3.20 (Uniform($[\theta, \theta+1]$): a two-dimensional minimal sufficient statistic)**.** Let $X_1, \dots, X_n$ be i.i.d. Uniform($[\theta, \theta+1]$). Using the indicator-function form of the joint density,

$$\frac{f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{y} \mid \theta)} = \frac{\mathbb{1}_{\{\theta \le x_{(1)} \le x_{(n)} \le \theta+1\}}}{\mathbb{1}_{\{\theta \le y_{(1)} \le y_{(n)} \le \theta+1\}}}.$$

By the checking rule, $T(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ is minimal sufficient. Notice that the minimal sufficient statistic has dimension 2, even though the parameter is one-dimensional and the sample is n-dimensional.

3.3 Complete statistics

This subsection introduces completeness and its interaction with ancillary statistics.

3.3.1 Ancillary statistics

Definition 3.7 (Ancillary statistic)**.** A statistic $V(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called *ancillary*.

Example 3.21 (Location family: an ancillary range)**.** Suppose F is a cdf and $X_1, \dots, X_n$ is a sample from $F(x-\theta)$. Then the range $R = X_{(n)} - X_{(1)}$ is ancillary:

$$P_\theta(R \le r) = P_\theta\big(\max_i X_i - \min_i X_i \le r\big) = P_\theta\big(\max_i (X_i - \theta) - \min_i (X_i - \theta) \le r\big).$$

Example 3.22 (Scale family: ancillary ratios)**.** Let $X_1, \dots, X_n$ be a sample from $F(x/\sigma)$. Then $X_1/X_n, \dots, X_{n-1}/X_n$ is ancillary: writing $X_i = \sigma Z_i$ with $Z_i \sim F$ shows the ratios are free of σ.

Remark 3.5. The simplest ancillary statistic is the constant statistic $V(\boldsymbol{X}) \equiv c$. More generally, a non-trivial ancillary statistic $V(\boldsymbol{X})$ induces a partition $\mathcal{A}(V) = \{\{\boldsymbol{x} : V(\boldsymbol{x}) = v\} : v\}$ that carries no information about θ.

If $T(\boldsymbol{X})$ is a statistic and $V\big(T(\boldsymbol{X})\big)$ is a non-trivial ancillary statistic, then the partition $\mathcal{A}(T)$ contains a coarser partition with no information about θ. This suggests that further data reduction may be possible.

In this sense, a sufficient statistic looks most "successful" when *no* nonconstant function of it is ancillary.

Question. Minimal sufficient statistics represent maximal data reduction while keeping information about θ. Are they "successful" in the above sense?

Ancillary statistics may in fact appear as components of a minimal sufficient statistic.

Example 3.23 (Uniform($\theta, \theta+1$): ancillarity inside a minimal sufficient statistic)**.** Let $X_1, \dots, X_n$ be a sample from Uniform($\theta, \theta+1$). The range $R = X_{(n)} - X_{(1)}$ is ancillary. We also know that $(X_{(1)}, X_{(n)})$ is minimal sufficient, hence

$$\big(X_{(n)} - X_{(1)},\, X_{(n)} + X_{(1)}\big) \quad \text{is a minimal sufficient statistic.}$$

Therefore, there can exist a nonconstant function of a minimal sufficient statistic that is ancillary, so minimal sufficiency is not necessarily "successful" in the above sense. Moreover, ancillary statistics are not always independent of minimal sufficient statistics.

This motivates the definition of *completeness*.

Definition 3.8 (Complete statistic)**.** Let $\boldsymbol{X}$ be i.i.d. with pdf/pmf $f(\cdot \mid \theta)$. A statistic $T(\boldsymbol{X})$ is *complete* for θ if every (measurable) function g *not depending on* θ that satisfies

$$\mathbb{E}_\theta[g(T(\boldsymbol{X}))] = 0 \text{ for all } \theta$$

must also satisfy

$$\mathbb{P}_\theta\big(g(T(\boldsymbol{X})) = 0\big) = 1 \text{ for all } \theta.$$

Remark 3.6. Completeness means there is no non-trivial unbiased estimator of 0 based on $T(\boldsymbol{X})$. It implies that an unbiased estimator of a target parameter, when it exists and is a function of T, is unique.

A minimal sufficient statistic is not necessarily complete (for example, Uniform($[\theta, \theta+1]$)). Conversely, a complete statistic is not necessarily sufficient (see the examples below).

3.3.2 Examples of complete and non-complete statistics

Example 3.24 (Uniform$(\theta, \theta+1)$ revisited)**.** Uniform$(\theta, \theta+1)$ revisited. $T(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ is minimal sufficient. However,

$$X_{(n)} - X_{(1)} - \mathbb{E}\big[X_{(n)} - X_{(1)}\big]$$

has mean 0 but is not 0 a.s., so neither $T(\boldsymbol{X})$ nor $X_{(n)} - X_{(1)}$ is complete. It is important here that $g(\cdot)$ does not depend on θ.

The range $R = X_{(n)} - X_{(1)}$ itself does not contain any information about θ, but combined with a sufficient statistic, it does (see Casella–Berger, Example 6.2.20).

Example 3.25 (Normal$(0, \sigma^2)$ with $\theta = \sigma$)**.** Normal$(0, \sigma^2)$ with $\theta = \sigma$ and $T = \bar{X}$. Let $g(x) = x$. Then $\mathbb{E}[g(\bar{X})] = 0$ but $\mathbb{P}(g(\bar{X}) = 0) \neq 1$, so $\bar{X}$ is not complete for σ.

Example 3.26 (Normal$(\mu, 1)$ with $\theta = \mu$)**.** Normal$(\mu, 1)$ with $\theta = \mu$ and $T = \bar{X}$. If $\mathbb{E}_\theta[g(\bar{X})] = 0$ for all θ, then $g \equiv 0$ a.s.; thus $\bar{X}$ is complete for μ.

Example 3.27 (Normal(μ, σ^2) and $T = \bar{X}$)**.** Normal(μ, σ^2) and $T = \bar{X}$. The statistic T is complete for $\theta = \mu$ and for $\theta = (\mu, \sigma^2)$, but not for $\theta = \sigma^2$. *Completeness does not imply sufficiency.* (Here T is not sufficient for σ^2, hence also not sufficient for $\theta = (\mu, \sigma^2)$.)

Example 3.28 (Bernoulli: completeness of the binomial count)**.** For a Bernoulli sample with success probability $\theta = p \in (0, 1)$, a minimal sufficient statistic is

$$T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p).$$

Suppose $\mathbb{E}_p[g(T)] = 0$ for all p. Then

$$0 = \mathbb{E}_p[g(T)] = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} = (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \phi^t,$$

where $\phi = \frac{p}{1-p} \in (0, \infty)$ and $(1-p)^n > 0$. The right-hand side is a polynomial in ϕ that is identically zero, so every coefficient must be 0, hence $g(t) = 0$ for all t. Therefore T is complete.

3.3.3 Completeness implies minimality under mild conditions

Theorem 3.3 (Bahadur's theorem)**.** *If a minimal sufficient statistic exists, then any complete sufficient statistic is also minimal sufficient. In particular, a finite-dimensional complete sufficient statistic is minimal sufficient.*

Under mild conditions, a complete sufficient statistic is "all you need": it implies minimal sufficiency.

Remark 3.7. The converse is not true. In the Uniform$(\theta, \theta+1)$ example, $(X_{(n)} - X_{(1)},\ X_{(n)} + X_{(1)})$ is minimal sufficient for θ, but not complete.

Remark 3.8. If a minimal sufficient statistic is not complete, then there does not exist any complete statistic.

Completeness formalizes an ideal notion of optimal data reduction, whereas minimal sufficiency is an achievable notion of maximal reduction.

3.3.4 Order statistics revisited: completeness

We have argued that the order statistics $\boldsymbol{T} = (X_{(1)}, \dots, X_{(n)})$ are sufficient for the nonparametric model $\theta = f \in \Theta = \{\text{all distributions with a density}\}$. We now show that $\boldsymbol{T}$ is also complete for $\theta \in \Theta$.

Proposition 3.4 (Completeness of order statistics for the nonparametric model)**.** *Let $X_1, \dots, X_n$ be i.i.d. from a distribution with density f. Then $\boldsymbol{T} = (X_{(1)}, \dots, X_{(n)})$ is complete for the model $\Theta = \{$all densities$\}$.*

Proof. First, note that a statistic δ is a function of $\boldsymbol{T}$ if and only if it is symmetric in its arguments, i.e. $\delta(\boldsymbol{x}) = \delta(\pi\boldsymbol{x})$ for every permutation π.

Consider first the case where δ is nonnegative. Fix densities $f_1, \dots, f_n$ and consider mixtures

$$f = \sum_{i=1}^{n} \alpha_i f_i \in \Theta, \qquad \alpha_i > 0,\ \sum_i \alpha_i = 1.$$

If $\mathbb{E}_f[\delta(\boldsymbol{X})] = 0$ for all such f, then

$$0 = \int \cdots \int \delta(\boldsymbol{x}) \prod_{j=1}^{n} f(x_j)\, \mathrm{d}\boldsymbol{x} = \int \cdots \int \delta(\boldsymbol{x}) \prod_{j=1}^{n} \left(\sum_{i=1}^{n} \alpha_i f_i(x_j) \right) \mathrm{d}\boldsymbol{x}.$$

The right-hand side is a polynomial in $\boldsymbol{\alpha}$, hence all coefficients must be zero.

Consider in particular the coefficient of $\prod_i \alpha_i$:

$$\begin{aligned} 0 &= \sum_{\pi} \int \cdots \int \delta(\boldsymbol{x}) \prod_{i=1}^{n} f_i\big(x_{\pi(i)}\big)\, \mathrm{d}\boldsymbol{x} \\ &= \sum_{\pi} \int \cdots \int \delta(\pi^{-1}\boldsymbol{x}) \prod_{i=1}^{n} f_i(x_i)\, \mathrm{d}\boldsymbol{x} = \sum_{\pi} \int \cdots \int \delta(\boldsymbol{x}) \prod_{i=1}^{n} f_i(x_i)\, \mathrm{d}\boldsymbol{x} \\ &= n! \int \cdots \int \delta(\boldsymbol{x}) \prod_{i=1}^{n} f_i(x_i)\, \mathrm{d}\boldsymbol{x}. \end{aligned}$$

Now let f_i be uniform on an interval $[a_i, b_i]$. Then

$$\int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \delta(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 0 \quad \Longrightarrow \quad \delta(\boldsymbol{x}) = 0 \text{ a.s.},$$

because δ is nonnegative.

For a general symmetric δ, apply the same argument to its positive and negative parts. Therefore $\boldsymbol{T}$ is complete. □

Example 3.29 (Order statistics and ranks)**.** Now that the order statistic $\boldsymbol{T}$ is complete and sufficient, consider the ranks of the observations,

$$\boldsymbol{R} = (R_1, \dots, R_n), \qquad R_i \triangleq \#\{j : X_j \le X_i\}.$$

Then $\mathbb{P}\big(\boldsymbol{R} = \pi(1, \dots, n)\big) = 1/n!$, so $\boldsymbol{R}$ is ancillary.

In fact, $\boldsymbol{T}$ and $\boldsymbol{R}$ are independent:

$$\mathbb{P}(\boldsymbol{T} = \boldsymbol{t},\ \boldsymbol{R} = \boldsymbol{r}) = \frac{1}{n!} \times n! \prod_i f(t_i).$$

3.3.5 Basu's theorem

Theorem 3.4 (Basu's theorem)**.** *If $T(\boldsymbol{X})$ is complete and sufficient for $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$ and $V(\boldsymbol{X})$ is ancillary, then $T(\boldsymbol{X})$ and $V(\boldsymbol{X})$ are independent under $\mathbb{P}_\theta$ for every θ.*

Proof. Fix a measurable set B and define

$$q_B(t) = \mathbb{P}_\theta\big(V \in B \mid T(\boldsymbol{X}) = t\big), \qquad p_B = \mathbb{P}_\theta(V \in B).$$

Let $g(t) = q_B(t) - p_B$. By sufficiency and ancillarity, g does not depend on θ. Moreover,

$$\mathbb{E}_\theta[g(T(\boldsymbol{X}))] = \mathbb{E}_\theta\big[\mathbb{P}_\theta(V \in B \mid T(\boldsymbol{X}))\big] - p_B = \mathbb{P}_\theta(V \in B) - p_B = 0.$$

By completeness, $g\big(T(\boldsymbol{X})\big) = 0$ a.s., i.e. $q_B(T) = p_B$ a.s.

Now let A be a measurable set for T. Then

$$\begin{aligned}
\mathbb{P}_\theta(T \in A,\, V \in B) &= \mathbb{E}_\theta\big[\mathbb{1}_{\{T\in A\}}\mathbb{1}_{\{V\in B\}}\big] = \mathbb{E}_\theta\Big[\mathbb{E}_\theta\big[\mathbb{1}_{\{T\in A\}}\mathbb{1}_{\{V\in B\}} \mid T\big]\Big] \\
&= \mathbb{E}_\theta\Big[\mathbb{1}_{\{T\in A\}}\mathbb{E}_\theta\big[\mathbb{1}_{\{V\in B\}} \mid T\big]\Big] = \mathbb{E}_\theta\big[\mathbb{1}_{\{T\in A\}}\, q_B(T)\big] \\
&= \mathbb{E}_\theta\big[\mathbb{1}_{\{T\in A\}}\, p_B\big] = \mathbb{P}_\theta(T \in A)\,\mathbb{P}_\theta(V \in B).
\end{aligned}$$

□

3.3.6 Completeness for exponential families

Basu's theorem is a powerful tool for deducing independence, but it requires completeness. For exponential families, we have general criteria that guarantee completeness.

Let $X_1, \ldots, X_n$ be a sample from a k-parameter *canonical* exponential family

$$f(x \mid \boldsymbol{\eta}) = h(x)\, c^*(\boldsymbol{\eta})\, \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right),$$

where $\boldsymbol{\eta} \in \Xi \subset \mathcal{H}$ is the parameter set.

Definition 3.9 (Minimal exponential family)**.** An exponential family parameterized by its natural parameters $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{H}\}$ is *minimal* if (i) there is no nonzero $\boldsymbol{\lambda} \in \mathbb{R}^{k+1}$ such that $\sum_i \lambda_i \eta_i = \lambda_0$, and (ii) there is no nonzero $\boldsymbol{\lambda} \in \mathbb{R}^{k+1}$ such that $\sum_i \lambda_i T_i(\boldsymbol{x}) = \lambda_0$.

The first condition rules out the possibility of linearly transforming the k-dimensional exponential family into an exponential family of smaller dimension. The second condition rules out unidentifiable cases (i.e., when there exist $\boldsymbol{\eta}_1 \neq \boldsymbol{\eta}_2$ such that $\mathbb{P}_{\boldsymbol{\eta}_1} = \mathbb{P}_{\boldsymbol{\eta}_2}$).

Example 3.30. Consider $X \sim \mathrm{Exp}(\eta_1, \eta_2)$, where $p(x; \eta_1, \eta_2) = \exp\big(-\eta_1 x - \eta_2 x + \log(\eta_1 + \eta_2)\big)\mathbb{1}_{\{x \geq 0\}}$.

Non-minimal families can always be reduced to minimal families via a suitable transformation and reparameterization.

Definition 3.10 (Curved exponential family)**.** Suppose $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \Xi\}$ is a k-parameter minimal canonical exponential family. If Ξ contains a k-dimensional open set, then $\mathcal{P}$ is called *full-rank*. Otherwise, $\mathcal{P}$ is *curved*.

In a curved exponential family, the η_i's are related in a nonlinear way.

Examples (normal family) Consider $\mathcal{N}(\mu, \sigma^2)$,

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

It is minimal and full-rank with $\eta_1 = \mu/\sigma^2$ and $\eta_2 = -1/(2\sigma^2)$.

If we restrict $\mu = \sigma^2 = \theta$, then $\eta_1 = 1$ and $\eta_2 = -1/(2\theta)$, which is non-minimal (for example, take $\lambda_1 = 1$, $\lambda_2 = 0$, $\lambda_0 = 1$ so that $\sum_i \lambda_i \eta_i = \lambda_0$).

If we restrict $\mu = \sigma = \theta$, then $\eta_1 = 1/\theta$ and $\eta_2 = -1/(2\theta^2)$; this yields a minimal but curved family. The statistic $T = (\bar{X}, S^2)$ is sufficient for θ, but it is not complete: to see this, one can find a nonzero function of T with mean 0 for all θ, for instance

$$g(\bar{X}, S^2) = \frac{n\bar{X}^2}{n+1} - S^2.$$

Theorem 3.5 (Completeness for full-rank exponential families)**.** *Suppose $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \Xi\}$ is a k-parameter minimal canonical exponential family of full-rank. Then*

$$T(\boldsymbol{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is complete.

Applications of Basu's theorem Basu's theorem allows us to deduce the independence of two statistics once we identify one statistic that is complete and sufficient and another that is ancillary.

Example 3.31 (Exponential sample and Basu's theorem)**.** Let $X_1, \ldots, X_n$ be a sample from Exponential(θ). Then

$$g(\boldsymbol{X}) = \frac{X_n}{X_1 + \cdots + X_n}, \qquad T(\boldsymbol{X}) = X_1 + \cdots + X_n$$

are independent. Here Exponential(θ) is a scale family, so $g(\boldsymbol{X})$ is ancillary. It is also a minimal one-parameter exponential family of full-rank, so $T(\boldsymbol{X})$ is complete and sufficient. Therefore $\mathbb{E}_\theta[g(\boldsymbol{X})] = 1/n$.

One can also verify minimality using the checking rule, but that is unnecessary once the exponential-family structure is recognized.

Example 3.32. If we consider $\mathcal{N}(\mu, \sigma^2)$ with known σ, then $\bar{X}$ and S^2 are independent.

Example 3.33 (Normal with known σ: ancillary median deviation)**.** For $\mathcal{N}(\mu, \sigma^2)$ with known σ, $\bar{X}$ is sufficient and complete, and $\mathrm{med}(\boldsymbol{X}) - \bar{X}$ is ancillary, where $\mathrm{med}(\boldsymbol{X})$ is the sample median. Hence

$$\mathrm{Cov}\big(\bar{X}, \mathrm{med}(\boldsymbol{X})\big) = \frac{\sigma^2}{n}.$$

3.4 Summary

Consider two experiments: observe $X \sim \mathbb{P}_{X|\theta}$, or observe $T \sim \mathbb{P}_{T|\theta}$ and then generate $X \mid T = t \sim \mathbb{P}_{X|t}$. The variable X has the same distribution in both experiments, so inference about θ should be the same.

If T is sufficient, only the experiment of observing T is informative about θ. A sufficient statistic induces a partition on which identical statistical conclusions are drawn, and this partition (equivalently, the generated σ-algebra) is the fundamental object.

If no coarser partition of the sample space can retain sufficiency, then T is minimal sufficient. We will return to completeness in the next topic.

Reading materials

Same level

- Robert W. Keener, *Theoretical Statistics*, Chapters 2–3.
- (Not recommended) Casella and Berger, *Statistical Inference*, Section 6.2.

Measure theoretic

- Jun Shao, *Mathematical Statistics*, Section 2.2.
- Lehmann and Romano, *Testing Statistical Hypotheses*, Sections 1.9 and 2.6.

4 Point Estimation

4.1 Introduction

The statistical model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ is unknown up to a **parameter** θ. After observing a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, our goal is to estimate θ.

A **point estimator** is a statistic $T(\boldsymbol{X})$ designed to estimate the unknown parameter θ of the model $\mathcal{P}$. Keep in mind that an estimator is a random variable: it depends on the random sample.

Remark 4.1 (Notation). A generic estimator of θ is denoted by $\widehat{\theta}$. Regular letters (e.g., x, Y, θ) denote scalars, while boldface letters (e.g., $\boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{\theta}$) denote vectors. Lowercase letters (e.g., $\boldsymbol{x}, y, \beta$) denote deterministic values, whereas uppercase letters (e.g., $X, \boldsymbol{Y}$) denote random variables or random vectors.

Parameter	Estimator
$\theta = \mathbb{E}[X]$	$\bar{X}$
$\theta = \mathrm{Var}(X)$	S^2
$\theta = \mathrm{Cov}(Y_1, Y_2)$	$\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(Y_{i,1} - \bar{Y}_1)(Y_{i,2} - \bar{Y}_2)$
$\theta = \rho(Y_1, Y_2)$	$\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(\dfrac{Y_{i,1} - \bar{Y}_1}{S_{Y_1}}\right)\left(\dfrac{Y_{i,2} - \bar{Y}_2}{S_{Y_2}}\right)$

Error A natural first idea is to measure performance via the estimation error.

Definition 4.1 (Error)**.** If T is used to estimate θ, then $T - \theta$ is the **error** of T.

The raw error is rarely a stable criterion by itself. For continuous distributions,

$$\mathbb{P}_\theta\big(T(\boldsymbol{X}) = \theta\big) = 0.$$

For a Bernoulli sample of size n, if $\theta \notin \{0, 1/n, \ldots, 1\}$ then $\bar{X} \neq \theta$ almost surely. More generally, the error is random; it can be large with small probability, making single-value comparisons unreliable.

Bias To remove randomness from the error, we use its expectation.

Definition 4.2 (Bias)**.** If T is used to estimate θ, the **bias** of T is $\mathbb{E}_\theta[T - \theta]$.

Bias captures systematic error. We call T **unbiased** if $\mathbb{E}_\theta[T] = \theta$ for all $\theta \in \Theta$, **negatively biased** if $\mathbb{E}_\theta[T] \leq \theta$ for all θ, and **positively biased** if $\mathbb{E}_\theta[T] \geq \theta$ for all θ. A biased estimator need not be uniformly negative or positive; the sign can depend on θ.

Bias and Mean Squared Error Bias alone ignores dispersion. A more balanced criterion is MSE.

Definition 4.3 (Mean squared error (MSE))**.** If T estimates θ, then $\mathbb{E}_\theta\big[(T - \theta)^2\big]$ is the **mean squared error** of T.

Proposition 4.1 (Bias–variance decomposition)**.**

$$\mathrm{MSE}_\theta(T) = \mathrm{Var}_\theta(T) + \mathrm{Bias}_\theta(T)^2.$$

Proof. $\mathbb{E}[(T - \theta)^2] = \mathrm{Var}(T - \theta) + \mathbb{E}[T - \theta]^2 = \mathrm{Var}(T) + \mathrm{Bias}(T)^2$. □

Often, lowering bias increases variance and vice versa: the **bias–variance trade-off**.

Other Evaluations Beyond error, bias, and MSE, other criteria are useful:

Remark 4.2 (Other evaluation criteria). Mean absolute error: $\mathbb{E}_\theta[|T-\theta|]$. Tail probability: $\mathbb{P}\big(|T-\theta|>\varepsilon\big)$ for fixed ε.

MSE is popular because it is smooth and tractable. Mean absolute error can yield more robust procedures, and the tail criterion $\mathbb{P}(|T-\theta|>\varepsilon)$ avoids moment assumptions.

Efficiency For unbiased estimators, smaller variance means smaller MSE.

Definition 4.4 (Efficiency)**.** Suppose U and V are unbiased for θ. We say that U is more **efficient** than V if $\mathrm{Var}(U)\le\mathrm{Var}(V)$. The **relative efficiency** of U w.r.t. V is $\dfrac{\mathrm{Var}(V)}{\mathrm{Var}(U)}$.

If the relative efficiency exceeds 1, then U is more efficient than V.

Asymptotic Properties In many problems the sample size grows, so it is useful to study a *sequence* of estimators $T_n(\boldsymbol{X}_n)$ as $n\to\infty$. We say that T_n is **asymptotically unbiased** if $\lim_{n\to\infty}\mathbb{E}[T_n]=\theta$, and **consistent** if $T_n\Rightarrow\theta$, i.e., $\mathbb{P}(|T_n-\theta|>\varepsilon)\to 0$ for every $\varepsilon>0$.

Example 4.1. The estimator

$$\widehat{\sigma}^2\triangleq\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

is asymptotically unbiased for σ^2.

If U_n and V_n are asymptotically unbiased for θ, the **asymptotic relative efficiency** of U to V is

$$\lim_{n\to\infty}\frac{\mathrm{Var}(V_n)}{\mathrm{Var}(U_n)},$$

when the limit exists.

Proposition 4.2 (Consistency via MSE)**.** *If* $\mathrm{MSE}(T_n)\to 0$*, then* T_n *is consistent.*

Proof. For any $\varepsilon>0$, by Markov's inequality applied to $|T_n-\theta|^2$,

$$\mathbb{P}(|T_n-\theta|>\varepsilon)=\mathbb{P}\big(|T_n-\theta|^2>\varepsilon^2\big)\le\frac{\mathbb{E}[(T_n-\theta)^2]}{\varepsilon^2}\to 0.\qquad\square$$

Example 4.2 (Common estimators)**.** The sample mean $\bar{X}$ is unbiased and consistent for $\theta=\mathbb{E}[X]$ (by the law of large numbers). Two useful special cases are $X=\mathbb{1}_A$, where $\bar{X}$ estimates $\mathbb{P}(A)$, and $X=\mathbb{1}_{\{Y\le y\}}$, where $\bar{X}$ estimates the CDF value $F_Y(y)$. The sample variance S^2 is unbiased and consistent for $\sigma^2=\mathrm{Var}(X)$, and the sample covariance is unbiased and consistent for $\mathrm{Cov}(Y_1,Y_2)$.

Example 4.3 (Poisson(λ))**.** Two natural unbiased (and consistent) estimators of λ are the sample mean $\bar{X}$ and the sample variance S_n^2. To compare them, look at their variances (hence MSEs, since they are unbiased):

$$\mathrm{Var}(\bar{X})=\frac{\lambda}{n},\qquad \mathrm{Var}(S_n^2)=\frac{1}{n}\left(\mathbb{E}[(X-\lambda)^4]-\mathbb{E}[(X-\lambda)^2]^2\frac{n-3}{n-1}\right)=\frac{\lambda}{n}\left(1+2\lambda\frac{n}{n-1}\right).$$

This one needs tedious calculation so the steps are omitted. The asymptotic relative efficiency of $\bar{X}$ to S_n^2 is therefore $1+2\lambda>1$, so $\bar{X}$ is strictly better.

4.2 Constructing Estimators

We first present constructive principles for deriving estimators from model structure.

4.2.1 Method of Moments

Definition 4.5 (Method of moments)**.** Let $X_1, \dots, X_n \sim f(x \mid \theta_1, \dots, \theta_k)$. The classical *method of moments* (MoM) estimates $(\theta_1, \dots, \theta_k)$ by matching the first k population moments to the corresponding sample moments:

$$m_1(\theta_1, \dots, \theta_k) \triangleq \mathbb{E}[X] \approx \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \dots, \qquad m_k(\theta_1, \dots, \theta_k) \triangleq \mathbb{E}[X^k] \approx \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

This yields k equations, which are then solved for the k unknown parameters.

Example 4.4 (Normal $\mathcal{N}(\mu, \sigma^2)$)**.**

$$\mathbb{E}[X] = \mu = \bar{X}, \qquad \mathbb{E}[X^2] = \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Thus

$$\widehat{\mu} = \bar{X}, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Example 4.5 (Binomial(k, p))**.**

$$\mathbb{E}[X] = kp = \bar{X}, \qquad \mathbb{E}[X^2] = kp(1-p) + k^2p^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Solving gives

$$\widehat{k} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \ (\in \mathbb{Z}_+?), \qquad \widehat{p} = \frac{\bar{X}}{\widehat{k}}.$$

Remark 4.3 (Remarks on the method of moments). More generally, one can match moments of functions: choose g_i and use $\mathbb{E}[g_i(X)] = m_i(\boldsymbol{\theta})$, then match these to $\frac{1}{n}\sum_{j=1}^{n} g_i(X_j)$ for $1 \le i \le k$.

MoM is simple to construct and compute and requires limited distributional detail. On the other hand, it may yield estimates outside the parameter space (for example, the $\widehat{k}$ above need not be an integer), and the choice of moments is somewhat arbitrary. In many regular parametric families, MoM is typically less efficient than MLE.

Example 4.6 (Method of moments for the lognormal model)**.** Let $X = e^{\mu + \sigma Z}$ with $Z \sim \mathcal{N}(0,1)$ (i.e., $X \sim \text{Lognormal}(\mu, \sigma^2)$).

A convenient approach is to match moments of the transformed variable $\ln X$. Take

$$g_1(x) = \ln x, \qquad g_2(x) = (\ln x)^2.$$

Since $\mathbb{E}[\ln X] = \mu$ and $\mathbb{E}[(\ln X)^2] = \mu^2 + \sigma^2$, the MoM equations

$$\frac{1}{n}\sum_{i=1}^{n} \ln X_i = \mu, \qquad \frac{1}{n}\sum_{i=1}^{n} (\ln X_i)^2 = \mu^2 + \sigma^2$$

yield

$$\widehat{\mu}_{\text{MoM}} = \frac{1}{n}\sum_{i=1}^{n} \ln X_i, \qquad \widehat{\sigma}^2_{\text{MoM}} = \frac{1}{n}\sum_{i=1}^{n} (\ln X_i - \widehat{\mu}_{\text{MoM}})^2.$$

Alternatively, one can match raw moments. Here

$$m_1(\mu, \sigma^2) = \mathbb{E}[X] = e^{\mu + \sigma^2/2}, \qquad m_2(\mu, \sigma^2) = \mathbb{E}[X^2] = e^{2\mu + 2\sigma^2}.$$

Thus the MoM equations

$$\bar{X} = e^{\mu+\sigma^2/2}, \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = e^{2\mu+2\sigma^2}$$

give

$$\widehat{\sigma}^2 = \log\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\bar{X}^2}\right), \qquad \widehat{\mu} = \log\bar{X} - \frac{1}{2}\widehat{\sigma}^2.$$

This approach may not be numerically stable.

4.2.2 Maximum Likelihood

The *maximum likelihood estimator* (MLE) is foundational. Historically, it traces back to Gauss for estimating parameters in Normal models, and its general form is due to R. A. Fisher. The guiding intuition is simple: a reasonable estimator chooses the parameter value under which the observed sample is most likely.

Definition 4.6 (Likelihood)**.** The **likelihood** $L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i \mid \theta)$ is a function of the unknown $\theta \in \Theta$, evaluated at the observed $\boldsymbol{x}$.

It equals the joint density/pmf of $\boldsymbol{X}$ at $\boldsymbol{x}$, but is viewed as a function of θ.

Definition 4.7 (Maximum likelihood estimator)**.** The MLE $\widehat{\theta}(\boldsymbol{X}) = \arg\max_{\theta\in\Theta} L(\theta \mid \boldsymbol{X})$ maximizes the likelihood for the random sample $\boldsymbol{X}$ (assume a unique maximizer exists).

In practice, we first compute

$$\widehat{\theta}(\boldsymbol{x}) = \arg\max_{\theta\in\Theta} L(\theta \mid \boldsymbol{x})$$

for observed (deterministic) data $\boldsymbol{x}$, and then regard $\widehat{\theta}(\boldsymbol{X})$ as the resulting statistic.

Remark 4.4 (Sufficiency and the MLE). If $T(\boldsymbol{X})$ is sufficient for θ and an MLE exists, then there is an MLE that is a function of T, since $\log L(\theta \mid \boldsymbol{x}) = \log g(T(\boldsymbol{x}) \mid \theta) + \log h(\boldsymbol{x})$.

Computing the MLE In simple models the MLE can be obtained in closed form by solving the likelihood equations $\partial L/\partial\theta_i = 0$ (equivalently, $\partial l/\partial\theta_i = 0$) and checking that the Hessian of l is negative semidefinite at the maximizer. Working with $\log L$ is usually easier because it converts products into sums and often yields a concave objective (many exponential families are log-concave).

Definition 4.8 (Log-likelihood)**.** $l(\theta \mid \boldsymbol{x}) \triangleq \log L(\theta \mid \boldsymbol{x})$ is the log-likelihood.

If no closed form exists, one typically uses numerical optimization (Newton's method, gradient-based algorithms, etc.). In practice, a MoM estimator often provides a useful initial value.

Example 4.7 (Normal $\mathcal{N}(\mu, \sigma^2)$)**.**

$$L(\mu, \sigma \mid \boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right], \quad \log L = -n\log(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}.$$

Partial derivatives:

$$\frac{\partial \log L}{\partial \mu} = -\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}, \qquad \frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3}.$$

Setting these to zero gives

$$\widehat{\mu} = \bar{X}, \qquad \widehat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2\right)^{1/2} \neq S.$$

Example 4.8 (Bernoulli(p))**.**

$$L(p \mid \boldsymbol{x}) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, \qquad \log L = \Big(\sum_{i=1}^n x_i\Big)\log p + \Big(n - \sum_{i=1}^n x_i\Big)\log(1-p).$$

Differentiate:

$$\frac{\mathrm{d}}{\mathrm{d}p}\log L = \frac{1}{p}\sum_{i=1}^n x_i - \frac{1}{1-p}\Big(n - \sum_{i=1}^n x_i\Big).$$

The MLE is $\widehat{p} = \bar{X}$.

Example 4.9 (Poisson(λ))**.**

$$L(\lambda \mid \boldsymbol{x}) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1! \cdots x_n!}, \qquad \log L = -n\lambda + \Big(\sum_{i=1}^n x_i\Big)\log\lambda - \log(x_1! \cdots x_n!).$$

Differentiate:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log L = -n + \frac{1}{\lambda}\sum_{i=1}^n x_i.$$

The MLE is $\widehat{\lambda} = \bar{X}$.

Example 4.10 (MLE for Uniform$(0,\theta)$)**.** The log-likelihood is

$$l(\theta \mid \boldsymbol{x}) = -n\log\theta + \log \mathbb{1}\big(0 \le X_{(1)} \le X_{(n)} \le \theta\big).$$

The indicator is 1 if $\theta \ge X_{(n)}$ and 0 otherwise, so the MLE is

$$\widehat{\theta} = \max(X_1, \ldots, X_n) = X_{(n)}.$$

By contrast, MoM gives $\widehat{\theta} = 2\bar{X}$; note it is possible that some $x_i > 2\bar{X}$. The two estimators differ markedly.

Invariance of MLEs (One-to-One Case) Suppose we reparametrize via $\lambda = h(\theta)$, one-to-one, with $\Lambda = h(\Theta)$

Example 4.11. Exponential distribution: rate λ vs. mean $\mu = 1/\lambda$.

Theorem 4.1 (Invariance of MLE (Injective h))**.** *If $\widehat{\theta}$ is the MLE of θ and h is one-to-one, then the MLE of $h(\theta)$ is $h(\widehat{\theta})$.*

Example 4.12. For Exponential(λ), $\widehat{\lambda} = 1/\bar{X}$; thus the MLE of $\mu = 1/\lambda$ is $\widehat{\mu} = \bar{X}$.

Invariance of MLEs (General Case) When h is not one-to-one (e.g., $h(x) = x^2$), multiple θ map to the same $\eta = h(\theta)$.

Definition 4.9 (Induced likelihood)**.** Define the *induced likelihood*

$$L^*(\eta \mid \boldsymbol{x}) = \sup_{\{\theta : h(\theta) = \eta\}} L(\theta \mid \boldsymbol{x}),$$

and call any maximizer $\widehat{\eta}$ an MLE of η.

Theorem 4.2 (Invariance of MLE (General))**.** *If $\widehat{\theta}$ is the MLE of θ, then for any h, an MLE of $\eta = h(\theta)$ is $h(\widehat{\theta})$.*

Proof.

$$L^*(\widehat{\eta} \mid \boldsymbol{x}) = \sup_{\eta} \sup_{\{h(\theta)=\eta\}} L(\theta \mid \boldsymbol{x}) = \sup_{\theta} L(\theta \mid \boldsymbol{x}) = L(\widehat{\theta} \mid \boldsymbol{x}) = L^*\big(h(\widehat{\theta}) \mid \boldsymbol{x}\big). \qquad \square$$

Example 4.13. If $\bar{X}$ is the MLE for θ, then $\bar{X}^2$ is the MLE for θ^2; if $\widehat{\sigma}$ is the MLE for the standard deviation, then $\widehat{\sigma}^2$ is the MLE for the variance.

Bias–Variance Trade-off: Normal Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then $\bar{X}$ and S^2 are unbiased for μ and σ^2, while the MLEs are $\widehat{\mu} = \bar{X}$ and $\widehat{\sigma}^2 = \frac{n-1}{n} S^2$.

Since $\mathrm{Var}(\chi^2_{n-1}) = 2(n-1)$ and S^2 is unbiased,

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

For $\widehat{\sigma}^2$, by the bias–variance decomposition,

$$\mathrm{MSE}(\widehat{\sigma}^2) = \mathbb{E}\big[(\widehat{\sigma}^2 - \sigma^2)^2\big] = \frac{(n-1)^2}{n^2}\,\mathrm{Var}(S^2) + \left(\frac{\sigma^2}{n}\right)^2 = \frac{2n-1}{n^2}\sigma^4 < \frac{2\sigma^4}{n-1}.$$

Thus **trading variance for bias** improves MSE. Both S^2 and $\widehat{\sigma}^2$ are asymptotically unbiased and consistent; their asymptotic relative efficiency is 1.

Bias–Variance Trade-off: Uniform Suppose $X_1, \dots, X_n \sim \mathrm{Uniform}(0, \theta)$.

Estimator 1 (MoM): $d_1(\boldsymbol{X}) = 2\bar{X}$.

Estimator 2 (MLE): $d_2(\boldsymbol{X}) = \max_i X_i$. Compare:

$$\mathrm{Bias}_\theta(d_1) = 0, \qquad \mathrm{MSE}_\theta(d_1) = \mathrm{Var}(d_1) = 4\,\mathrm{Var}(\bar{X}) = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

For d_2, it is an order statistic with

$$f_{d_2}(x) = \frac{n}{\theta^n} x^{n-1} \mathbb{1}_{0 \le x \le \theta}, \quad \mathbb{E}[d_2] = \frac{n}{n+1}\theta, \quad \mathrm{Var}(d_2) = \frac{n\theta^2}{(n+2)(n+1)^2}.$$

Hence

$$\mathrm{Bias}_\theta(d_2) = -\frac{\theta}{n+1}, \qquad \mathrm{MSE}_\theta(d_2) = \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)} < \frac{\theta^2}{3n}.$$

Comparison of d_1 and d_2 By the CLT,

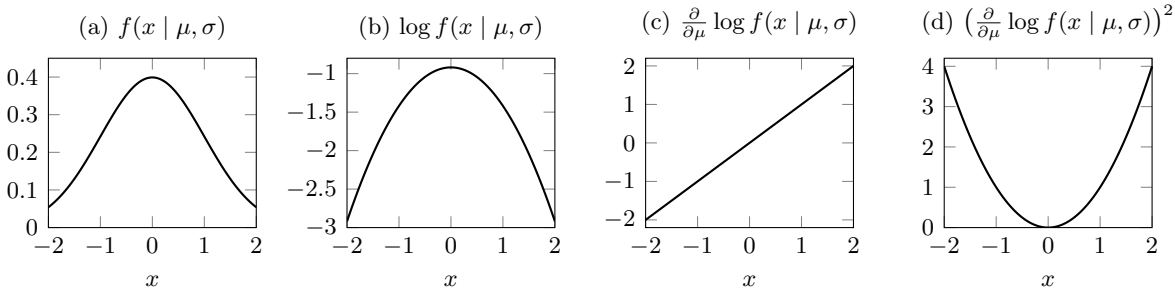
$$2\bar{X} \approx \mathcal{N}\big(\theta, \theta^2/[3n]\big),$$

and from order-statistic theory (Topic 2),

$$X_{(n)} \approx \theta - \theta\mathrm{Exp}(1)/n.$$

Thus:

- d_1 is unbiased; $\mathrm{MSE}(d_1) = \theta^2/(3n)$; consistent (WLLN).
- d_2 is negatively biased; $\mathrm{MSE}(d_2) = 2\theta^2/[(n+1)(n+2)]$; consistent and asymptotically unbiased.
- Asymptotically, d_2 is more efficient than d_1: the leading error of d_2 is $O(n^{-2})$ vs. $O(n^{-1})$ for d_1. MSEs that decrease on the order of n^{-2} are sometimes called *super-efficient*; in contrast, the sample mean's MSE typically decreases at rate n^{-1}.
- Unbiasedness alone can mislead: $d_3(\boldsymbol{X}) = (n+1)X_{(1)}$ is unbiased ($\mathbb{E}[d_3] = \theta$) but $\mathrm{Var}(d_3) = \frac{n}{n+2}\theta^2$, so it is not even consistent.



4.3 Fisher Information and the Cramér–Rao Bound

Best Unbiased Estimators If an estimator is unbiased, it is correct on average. For unbiased estimators, minimizing variance is the same as minimizing MSE. Since variance (and MSE) generally depends on θ, a natural comparison is uniform: if $\mathrm{Var}_\theta(W_1) \leq \mathrm{Var}_\theta(W_2)$ for all $\theta \in \Theta$, then W_1 is **uniformly better** than W_2.

Definition 4.10 (UMVUE)**.** An unbiased estimator W^* is a uniform minimum-variance unbiased estimator (UMVUE) if it is uniformly better than every other unbiased estimator. UMVUEs do not always exist.

Example 4.14 (Poisson)**.** If $\boldsymbol{X}$ is a sample from Poisson(λ), then both $\bar{X}$ and S^2 are unbiased:

$$\mathbb{E}_\lambda[\bar{X}] = \lambda, \qquad \mathbb{E}_\lambda[S^2] = \lambda.$$

We have shown that $\mathrm{Var}_\lambda(\bar{X}) \leq \mathrm{Var}_\lambda(S^2)$, so S^2 cannot be UMVUE for λ.

Fisher Information: Intuition Finding an UMVUE directly can be difficult because there may be infinitely many unbiased estimators. For instance, in the Poisson example, any convex combination

$$W_a(\bar{X}) = a\bar{X} + (1-a)S^2$$

is unbiased for any $a \in [0, 1]$. Before discussing constructive methods, it is helpful to ask what is *fundamentally* achievable with a fixed sample size.

Remark 4.5 (Question). Are we fundamentally limited by finite data, or can we achieve arbitrarily small variance with samples of fixed size?

A random sample carries information about θ. The natural question is: *how much* information can a sample provide? If the amount of information is intrinsically upper bounded, then the efficiency of any estimator is also limited. Fisher information makes this intuition precise.

Example 4.15 (Normal log-likelihood as a function of μ)**.** Consider $X \sim \mathcal{N}(\mu, \sigma^2)$. The figure below shows the log-likelihood $\log f(X \mid \mu)$ as a function of μ.

Observation

$$\begin{aligned} \text{If } \theta \text{ is the true value } &\Rightarrow\ L(\theta \mid X) \triangleq f(X \mid \theta) \text{ is large} \\ &\Rightarrow\ \theta \text{ lies near the peak of } l(\theta \mid X) \triangleq \log f(X \mid \theta) \\ &\Rightarrow\ \frac{\partial}{\partial \theta} l(\theta \mid X) \text{ is close to 0.} \end{aligned}$$

Definition 4.11 (Score)**.** The gradient of the log-likelihood is the score:

$$s(\theta \mid X) \triangleq \frac{\partial \log L(\theta \mid X)}{\partial \theta} = \frac{\partial l(\theta \mid X)}{\partial \theta}.$$

The score is the instantaneous slope of the log-likelihood in θ.

Sensitivity and Information If $|s(\mu \mid X)|$ is large, then X is highly informative about μ: the slope of $l(\mu \mid X)$ w.r.t. μ is large, indicating high sensitivity of the likelihood to changes in μ.

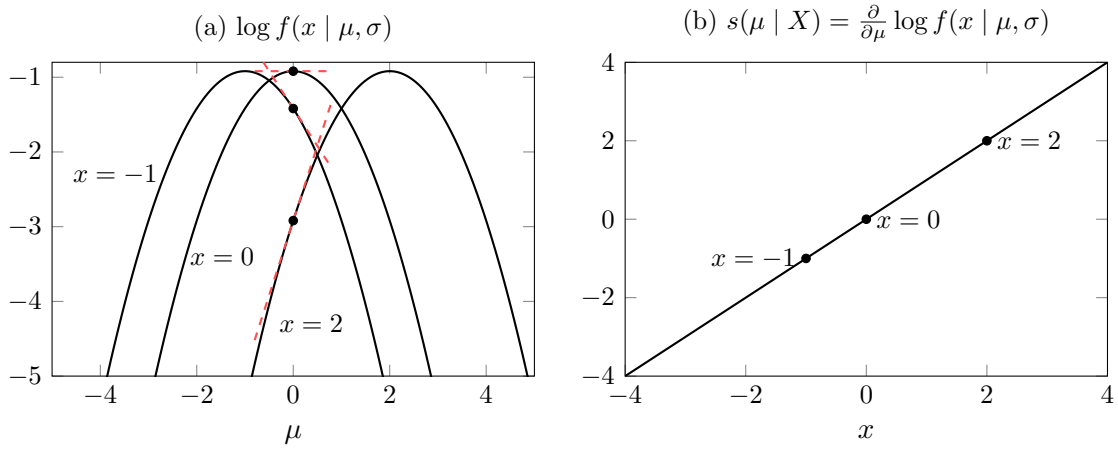


Figure 4: Normal with variance 1.

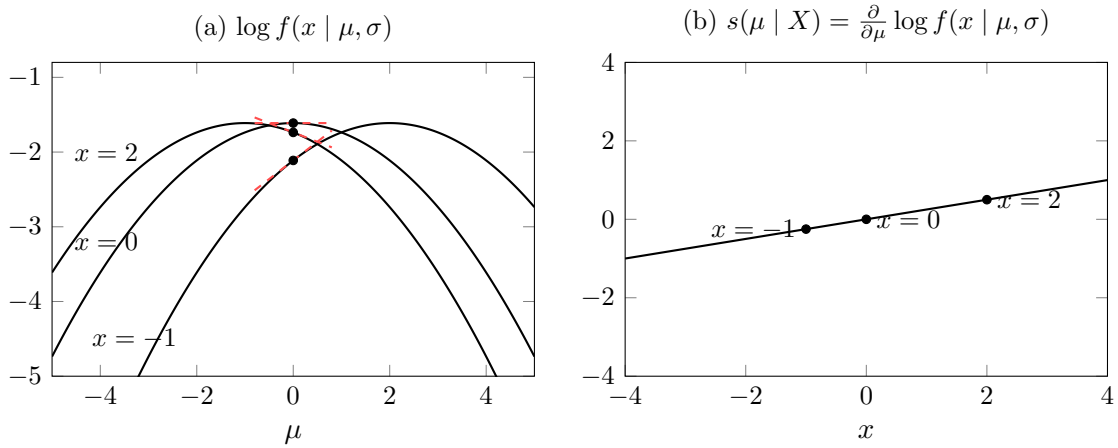


Figure 5: Normal with variance 4.

Definition To quantify information, consider

$$\mathrm{Var}_{\mu^*}\left(s(\mu^* \mid X)\right) = \mathrm{Var}_{X \sim \mathcal{N}(\mu^*, \sigma^2)}\left(\frac{\partial}{\partial \mu} l(\mu \mid X)\Big|_{\mu=\mu^*}\right).$$

A larger variance means the score is often large in magnitude, i.e., more information.

Definition 4.12 (Fisher information)**.** The Fisher information is

$$\mathcal{I}(\theta) \triangleq \mathrm{Var}_\theta\left(s(\theta \mid X)\right) = \mathrm{Var}_\theta\left(\frac{\partial}{\partial \theta} l(\theta \mid X)\right).$$

Assume f is differentiable in θ and differentiation and integration can be interchanged:

$$\int \frac{\partial}{\partial \theta} f(x \mid \theta) \mathrm{d}x = \frac{\partial}{\partial \theta} \int f(x \mid \theta) \mathrm{d}x = 0.$$

Hence

$$\mathbb{E}_\theta\big[s(\theta \mid X)\big] = \mathbb{E}_\theta\left[\frac{\partial}{\partial \theta} l(\theta \mid X)\right] = \int \frac{\partial \log f(x \mid \theta)}{\partial \theta} f(x \mid \theta) \mathrm{d}x = 0.$$

Under the true θ, its mean is zero. But sample to sample the score fluctuates; its variance is larger when the likelihood is more sensitive to θ.

Information is expected sensitivity of the log-likelihood at θ.

Proposition 4.3 (Alternative expression of Fisher information I)**.**

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right] = \int\left(\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right)^2 f(x \mid \theta)\mathrm{d}x.$$

Assume further that f is twice differentiable in θ and we may again interchange operations:

$$\int \frac{\partial^2}{\partial\theta^2} f(x \mid \theta)\mathrm{d}x = \frac{\partial^2}{\partial\theta^2}\int f(x \mid \theta)\mathrm{d}x = 0.$$

Also,

$$\frac{\partial^2}{\partial\theta^2} l(\theta \mid x) = \frac{\partial}{\partial\theta}\left[\frac{\partial_\theta f(x \mid \theta)}{f(x \mid \theta)}\right] = \frac{\partial^2_{\theta^2} f(x \mid \theta)}{f(x \mid \theta)} - \left(\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right)^2.$$

Proposition 4.4 (Alternative expression of Fisher information II)**.**

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} l(\theta \mid X)\right] = -\int \frac{\partial^2}{\partial\theta^2}\log f(x \mid \theta) f(x \mid \theta)\mathrm{d}x.$$

This second form is often the most convenient in calculations.

Interpretation

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} l(\theta \mid X)\right].$$

The quantity $-\frac{\partial^2}{\partial\theta^2} l(\theta \mid X)$ is the curvature of the log-likelihood at θ. Higher expected curvature means more information about θ.

Fisher information at a parameter value θ measures how sharply the likelihood moves when you nudge θ. A sharper (more curved) log-likelihood means you can localize θ more tightly from the same data.

Remarks—Random Sample For a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, define

$$\mathcal{I}_n(\theta) = \mathrm{Var}_\theta\left(s(\theta \mid \boldsymbol{X})\right) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} l(\theta \mid \boldsymbol{X})\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} l(\theta \mid \boldsymbol{X})\right].$$

With i.i.d. data, $L_n(\theta \mid \boldsymbol{X}) = \prod_{i=1}^n f(X_i \mid \theta)$ and $l(\theta \mid \boldsymbol{X}) = \sum_{i=1}^n l(\theta \mid X_i)$, so

Remark 4.6.

$$\mathbb{E}_\theta\left[s(\theta \mid \boldsymbol{X})\right] = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} l(\theta \mid \boldsymbol{X})\right] = \sum_{i=1}^n \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} l(\theta \mid X_i)\right] = 0, \quad \mathcal{I}_n(\theta) = n\mathcal{I}(\theta).$$

Example 4.16 ($\mathcal{N}(\mu, \sigma^2)$, unknown $\theta = \mu$, known σ)**.**

$$-\frac{\partial^2}{\partial\mu^2}\log f(x \mid \mu) = \frac{1}{\sigma^2} \;\Rightarrow\; \mathcal{I}(\mu) = \frac{1}{\sigma^2}, \quad \mathcal{I}_n(\mu) = \frac{n}{\sigma^2}.$$

Example 4.17 ($\mathcal{N}(\mu, \sigma^2)$, unknown $\theta = \sigma$, known μ)**.**

$$-\frac{\partial^2}{\partial\sigma^2}\log f(x \mid \sigma) = -\frac{1}{\sigma^2} + 3\frac{(x-\mu)^2}{\sigma^4} \;\Rightarrow\; \mathcal{I}(\sigma) = \frac{2}{\sigma^2}.$$

Example 4.18 ($\mathcal{N}(\mu, \sigma^2)$, unknown $\theta = \sigma^2$, known μ)**.**

$$-\frac{\partial^2}{\partial(\sigma^2)^2}\log f(x \mid \sigma^2) = -\frac{1}{2(\sigma^2)^2} + \frac{(x-\mu)^2}{(\sigma^2)^3} \;\Rightarrow\; \mathcal{I}(\sigma^2) = \frac{1}{2\sigma^4}.$$

Fisher Information can be extended to vector parameters.

Definition 4.13 (Fisher information (matrix form))**.** For $\boldsymbol{\theta} \in \mathbb{R}^k$,

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta} \mid X)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta} \mid X)\right)^{\top}\right] = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta} \mid X)\right],$$

where the Hessian $\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta} \mid X) = \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\boldsymbol{\theta} \mid X)\right] \in \mathbb{R}^{k \times k}$.

Cramér–Rao Bound We now return to point estimators.

Fisher information characterizes how informative the sample is.

With finite information, we cannot estimate arbitrarily well: the variance of any estimator must be lower bounded by a function of the information. This heuristic is formalized by the Cramér–Rao inequality.

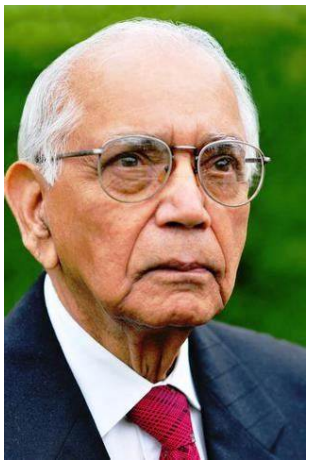


Figure 6: C. R. Rao (1920–2023), a student of R. A. Fisher.

Derivation via Cauchy–Schwarz The Fisher information is the variance of the score at the true θ:

$$\mathcal{I}(\theta) = \mathrm{Var}_\theta\left(s(\theta \mid X)\right).$$

If we know $\mathrm{Cov}_\theta(\widehat{\theta}, s(\theta \mid \boldsymbol{X}))$, then Cauchy–Schwarz yields

Proposition 4.5 (Cauchy–Schwarz inequality)**.**

$$\left(\mathrm{Cov}_\theta(\widehat{\theta}, s(\theta \mid \boldsymbol{X}))\right)^2 \le \mathrm{Var}_\theta(\widehat{\theta})\,\mathrm{Var}_\theta(s(\theta \mid \boldsymbol{X})) = \mathrm{Var}_\theta(\widehat{\theta}) n\mathcal{I}(\theta).$$

At first glance this is unhelpful because the covariance may depend on $\widehat{\theta}$. However, if $m(\theta) = \mathbb{E}_\theta[\widehat{\theta}]$, then (after justifying interchange of differentiation and integration)

$$\begin{aligned}
\mathrm{Cov}_\theta\left(\widehat{\theta}, s(\theta \mid \boldsymbol{X})\right) &= \mathbb{E}_\theta\left[\widehat{\theta} s(\theta \mid \boldsymbol{X})\right] \\
&= \int \widehat{\theta}(\boldsymbol{x}) \frac{\partial l(\theta \mid \boldsymbol{x})}{\partial \theta} \cdot f(\boldsymbol{x} \mid \theta) \mathrm{d}\boldsymbol{x} \\
&= \int \widehat{\theta}(\boldsymbol{x}) \frac{\partial}{\partial \theta} f(\boldsymbol{x} \mid \theta) \mathrm{d}\boldsymbol{x} = \frac{\mathrm{d}}{\mathrm{d}\theta} \int \widehat{\theta}(\boldsymbol{x}) f(\boldsymbol{x} \mid \theta) \mathrm{d}\boldsymbol{x} = m'(\theta).
\end{aligned}$$

Theorem 4.3 (Cramér–Rao inequality)**.** *Let $\widehat{\theta}(\boldsymbol{X})$ be any estimator with $m(\theta) = \mathbb{E}_\theta[\widehat{\theta}]$ and $\mathrm{Var}_\theta(\widehat{\theta}) < \infty$. Under the assumption that*

$$m'(\theta) = \int \widehat{\theta}(\boldsymbol{x}) \frac{\partial}{\partial \theta} f(\boldsymbol{x}|\theta) \mathrm{d}\boldsymbol{x}, \quad \textit{and} \quad \mathrm{Var}_\theta\, \widehat{\theta} < \infty,$$

we have

$$\mathrm{Var}_\theta(\widehat{\theta}) \;\geq\; \frac{[m'(\theta)]^2}{n\mathcal{I}(\theta)} \qquad \textit{for all } \theta.$$

Unbiased Case For an unbiased estimator $T(\boldsymbol{X})$, we have $m'(\theta) = 1$; hence, under suitable conditions,

Corollary 4.1 (Cramér–Rao inequality for unbiased estimators)**.**

$$\mathrm{Var}_\theta\left(T(\boldsymbol{X})\right) \;\geq\; \frac{1}{n\mathcal{I}(\theta)} \qquad \textit{for all } \theta.$$

Remarks Cramér and Rao developed the inequality independently in the 1940s.

As $\mathcal{I}(\theta)$ increases, the lower bound on variance decreases; thus it is also called the information bound.

$\big(n\mathcal{I}(\theta)\big)^{-1}$ is the *Cramér–Rao lower bound* (CRLB): no unbiased estimator based on n i.i.d. observations can have variance below it (under the regularity conditions).

Corollary 4.2 (Implication)**.** *If an unbiased estimator attains the CRLB, then it is the UMVUE.*

Example 4.19 (Bernoulli(p))**.** Bernoulli(p). Since

$$\frac{\partial^2}{\partial p^2} \log f(x \mid p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}, \quad \mathcal{I}(p) = \frac{n}{p(1-p)}.$$

For $\widehat{\theta} = \bar{X}$, $\mathbb{E}[\bar{X}] = p$ and $\mathrm{Var}(\bar{X}) = p(1-p)/n$, so $\bar{X}$ attains the CRLB and is UMVUE.

Example 4.20 (Poisson(λ))**.** Poisson(λ). With $\log f(x \mid \lambda) = -\lambda + x \log \lambda - \log x!$,

$$\mathcal{I}(\lambda) = \frac{n}{\lambda}.$$

Thus, for any unbiased W, $\mathrm{Var}_\lambda(W) \geq \lambda/n$. Since $\mathrm{Var}(\bar{X}) = \lambda/n$, $\bar{X}$ is UMVUE.

Example 4.21 (Normal $\mathcal{N}(\mu, \sigma^2)$)**.** Normal $\mathcal{N}(\mu, \sigma^2)$.

For μ, CRLB is σ^2/n; $\bar{X}$ attains it and is UMVUE.

For σ^2, with known μ, CRLB is $2\sigma^4/n$ and $\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ is UMVUE.

With unknown μ, $\mathrm{Var}(S^2) = 2\sigma^4/(n-1)$; the CRLB is not attained in this case.

Attainment of the CRLB From the proof, equality in Cauchy–Schwarz (for unbiased $\widehat{\theta}$) occurs iff

$$\widehat{\theta} = \theta + a(\theta) s(\theta \mid \boldsymbol{X}),$$

with nonrandom $a(\theta)$. In the normal example for σ^2 (with known μ),

$$s(\sigma^2 \mid \boldsymbol{X}) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^4}.$$

When μ is unknown, no unbiased estimator can be written in the required linear form, so the CRLB cannot be attained.

When the Inequality Fails: Uniform If $\boldsymbol{X}$ is a sample from Uniform$(0, \theta)$, then

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

This suggests the CRLB $\mathrm{Var}_\theta(W) \geq \theta^2/n$ for any unbiased W. Yet we saw that $\mathbb{E}[d_2(\boldsymbol{X})] = \frac{n}{n+1}\theta$, so $\widehat{\theta} = \frac{n+1}{n} d_2(\boldsymbol{X})$ is unbiased with

$$\mathrm{Var}_\theta(\widehat{\theta}) = \frac{(n+1)^2}{n^2} \cdot \frac{n\theta^2}{(n+2)(n+1)^2} = \frac{\theta^2}{n(n+2)} \leq \frac{\theta^2}{n}.$$

The issue is that the support of f depends on θ, so the required interchange of integration and differentiation fails.

4.4 Sufficiency, Completeness, and Unbiased Estimators

We now combine sufficiency and completeness to characterize optimal unbiased estimators.

4.4.1 Sufficiency

Using Sufficient Statistics to Improve Estimators Let W estimate $\tau(\theta)$ and let T be a sufficient statistic for θ. The statistic $\phi(T) \triangleq \mathbb{E}[W \mid T]$ is called the *Rao–Blackwell estimator* of $\tau(\theta)$.

Theorem 4.4 (Rao–Blackwell)**.** *If W is* unbiased *for $\tau(\theta)$, then*

$$\mathbb{E}_\theta[\phi(T)] = \tau(\theta), \qquad \mathrm{Var}_\theta(\phi(T)) \leq \mathrm{Var}_\theta(W).$$

Proof. Unbiasedness follows from the tower property:

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

For variance,

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

□

Why Sufficiency Is Needed For any statistic Y, one may consider $\mathbb{E}[W \mid Y]$, which remains unbiased and has no larger variance than W. Does this always improve an estimator? Not necessarily: with $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$, $\bar{X} = (X_1 + X_2)/2$ is sufficient, but $\phi'(X_1) \coloneqq \mathbb{E}[\bar{X} \mid X_1] = (X_1 + \theta)/2$ is not a statistic (it depends on θ). Sufficiency ensures $\mathbb{E}[W \mid T]$ is indeed a statistic (it is free of θ given T).

4.4.2 Completeness

Completeness and UMVUE

Theorem 4.5 (Lehmann–Scheffé)**.** *If T is* complete *and* sufficient *for θ, then for any statistic $\phi(T)$, the estimator $\phi(T)$ is UMVUE for $\mathbb{E}_\theta[\phi(T)]$.*

Proof. Let W be any unbiased estimator of $\mathbb{E}_\theta[\phi(T)]$. Then $g(T) \coloneqq \phi(T) - \mathbb{E}[W \mid T]$ satisfies $\mathbb{E}_\theta[g(T)] = 0$. Completeness implies $g(T) = 0$ a.s., hence $\phi(T) = \mathbb{E}[W \mid T]$. By Rao–Blackwell, $\mathrm{Var}_\theta(\phi(T)) \leq \mathrm{Var}_\theta(W)$ for all unbiased W, so $\phi(T)$ is UMVUE. □

Uniqueness of the Best Unbiased Estimator

Theorem 4.6. *If a UMVUE exists, it is unique.*

Proof. Suppose W_1 and W_2 are both UMVUEs and set $W^* = (W_1+W_2)/2$. By Cauchy–Schwarz,

$$\operatorname{Var}_\theta(W^*) \le \frac{1}{4}\operatorname{Var}_\theta(W_1) + \frac{1}{4}\operatorname{Var}_\theta(W_2) + \frac{1}{2}\sqrt{\operatorname{Var}_\theta(W_1)\operatorname{Var}_\theta(W_2)} = \operatorname{Var}_\theta(W_1).$$

Strict inequality would contradict the optimality of W_1, so equality holds and thus $W_1 = a(\theta)W_2 + b(\theta)$ with $a(\theta) = 1$ and $b(\theta) = 0$ by unbiasedness, i.e., $W_1 = W_2$. □

Example—Direct Application of L–S Finding UMVUEs is straightforward when a complete sufficient statistic is available.

Remark 4.7 (Method 1). Find a function ϕ such that $\mathbb{E}[\phi(T)] = \theta$, where T is complete and sufficient.

Example 4.22 (Bernoulli sample variance)**.** Bernoulli sample variance. For $X_i \sim \text{Bernoulli}(p)$, $\bar{X}$ is complete and sufficient. Since $X_i^2 = X_i$,

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) = \frac{n}{n-1}\bar{X}(1-\bar{X}),$$

which is unbiased for $p(1-p)$; hence S^2 is the UMVUE.

When CRLB Is Unattainable or Inapplicable

Example 4.23 ($\mathcal{N}(\mu, \sigma^2)$ with unknown μ)**.** S^2 is UMVUE for σ^2 because it is unbiased and a function of the complete sufficient pair $(\bar{X}, S^2)$.

Example 4.24 (Uniform$(0,\theta)$)**.** Let $Y = X_{(n)}$. The CRLB conditions fail here, but Y is complete and sufficient and $\frac{n+1}{n}Y$ is unbiased, hence UMVUE.

Uniform$(0,\theta)$: General Target If $\tilde{\theta} = g(\theta)$ with differentiable g, recall

$$f_{X_{(n)}}(x) = n\theta^{-n}x^{n-1}\mathbb{1}_{0\le x\le\theta}.$$

If $h(X_{(n)})$ is unbiased for $\tilde{\theta}$, then

$$\theta^n g(\theta) = n\int_0^\theta h(x)x^{n-1}\mathrm{d}x \quad (\theta > 0).$$

Differentiating,

$$n\theta^{n-1}g(\theta) + \theta^n g'(\theta) = nh(\theta)\theta^{n-1} \;\Rightarrow\; h(X_{(n)}) = g(X_{(n)}) + \frac{1}{n}X_{(n)}g'(X_{(n)}),$$

which is therefore UMVUE.

Example—The Conditioning Method We now present a second method to find UMVUEs by conditioning.

Remark 4.8 (Method 2). Find any unbiased estimator W; then $\mathbb{E}[W \mid T]$ is UMVUE (for complete sufficient T).

This works routinely in minimal full-rank exponential families, though $\mathbb{E}[W \mid T]$ may not have a closed form.

Example 4.25 (Conditioning method for Binomial(k, θ))**.** Consider Binomial(k, θ) and the target

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

Let $X_1, \dots, X_n$ be i.i.d. samples; then $T = \sum_{i=1}^n X_i \sim \text{Binomial}(kn, \theta)$ is complete and sufficient.

Find an unbiased W: take $h(X_1) = \mathbb{1}_{X_1=1}$ so that $\mathbb{E}_\theta[h(X_1)] = k\theta(1-\theta)^{k-1}$.

Compute $\phi(T) = \mathbb{E}[h(X_1) \mid T]$, i.e., the conditional probability of $X_1 = 1$ given T. Conditioning on $T = t$,

$$\phi(t) = \mathbb{P}(X_1 = 1 \mid T = t) = \frac{\mathbb{P}_\theta(X_1 = 1)\mathbb{P}_\theta\left(\sum_{i=2}^n X_i = t-1\right)}{\mathbb{P}_\theta\left(\sum_{i=1}^n X_i = t\right)} = \frac{\binom{k}{1}\binom{k(n-1)}{t-1}}{\binom{kn}{t}},$$

which does not depend on θ (by sufficiency). Therefore $\phi(T)$ is the UMVUE.

A Nonparametric Example—Empirical CDF Let $X_1, \dots, X_n$ be i.i.d. from an unknown CDF F. For fixed t, the empirical CDF $F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \le t}$ satisfies $nF_n(t) \sim \text{Binomial}(n, F(t))$, so $F_n(t)$ is unbiased for $F(t)$ with MSE $= F(t)(1-F(t))/n$. Hence $U(\boldsymbol{X}) = 1 - F_n(t)$ is unbiased for $\theta = 1 - F(t)$ with the same MSE.

When Θ is the set of all distributions with a density, the vector of order statistics is complete and sufficient; in this setting the empirical CDF is sufficient and complete, so $U(\boldsymbol{X})$ is UMVUE for θ, and F_n is UMVUE for F.

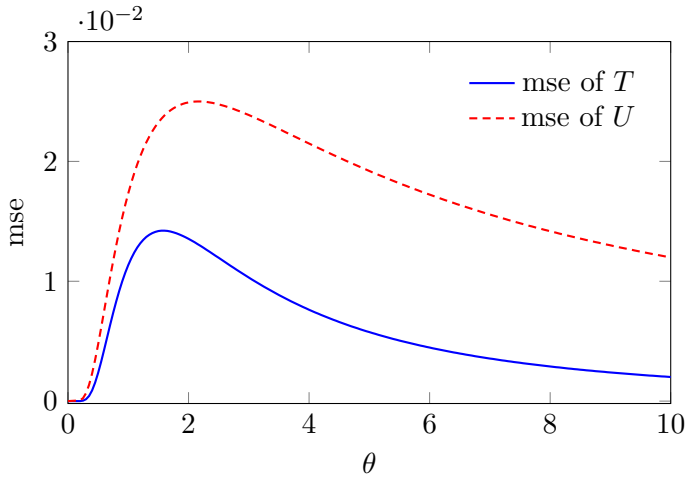


Figure 7: Cover illustration of Jun Shao's book.

Using Model Structure to Improve Nonparametric Estimators If we know F is Exp$(1/\theta)$, then $\bar{X}$ is complete and sufficient for θ (hence for $e^{-t/\theta} = 1 - F(t)$). By Lehmann–Scheffé, $T(\boldsymbol{X}) = \mathbb{E}[U(\boldsymbol{X}) \mid \bar{X}]$ is unbiased and is UMVUE, improving the MSE relative to $U(\boldsymbol{X})$ in this parametric submodel.

Summary (UMVUE Toolkit)

1. Check whether an estimator attains the CRLB (often not constructive, but useful for benchmarking).
2. Solve directly for ϕ with $\mathbb{E}[\phi(T)] = \theta$ using a complete sufficient T (algebraic tricks may help).
3. Condition any unbiased estimator on a complete sufficient T (choose W to simplify the conditional expectation).

Reading Materials *Same level*

Robert W. Keener, *Theoretical Statistics*, Chapter 4 (see §4.5 for another introduction to Fisher information).

Casella and Berger, *Statistical Inference*, Chapter 7. *Advanced*

Jun Shao, *Mathematical Statistics*, §§2.3, 2.4.1, 3.1 (see §3.1.2 for the case where a complete sufficient statistic is not available).

5 Hypothesis Testing

What is a hypothesis? A *hypothesis* is a scientific claim. For example, "smoking damages health" or "the coin is not fair." A *statistical hypothesis* is a mathematically precise version of such a claim. For instance, one may posit two life-expectancy distributions F (non-smokers) and G (smokers) with $\mu_F - \mu_G > 0$, or a Bernoulli head probability p for the coin with $p \neq 0.5$.

A scientific claim becomes *statistical* when we specify a model that could, in principle, have generated our data. The move from "smoking damages health" to a statement about distributions F and G introduces two crucial ingredients: (i) a probabilistic description of variability (people live different lengths of time even under the same conditions), and (ii) a *numerical* consequence that can be checked (a positive mean difference). This translation from words to math is what makes a hypothesis *testable*.

Definition 5.1 (Statistical hypothesis)**.** A *statistical hypothesis* is a hypothesis that is *testable* on the basis of a sample of data.

Remark 5.1. "Testable" means that the hypothesis makes predictions about the distribution of observable quantities. If repeated sampling from the postulated model would frequently contradict what we actually see, the hypothesis can be rejected. This is akin to Popper's notion of *falsifiability*: the hypothesis exposes itself to potential refutation by specifying what would count as incompatible data.

Null and Alternative Hypotheses In hypothesis testing, we ask whether there is sufficient statistical evidence to reject a presumed null hypothesis in favor of a conjectured alternative hypothesis. We observe a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ as the "evidence." The null hypothesis H_0 represents a status-quo claim one seeks to challenge (e.g., smoking does not affect health), while the alternative hypothesis H_1 represents the claim we suspect (e.g., smoking affects health).

The roles are asymmetric by design: H_0 is given the benefit of the doubt and is rejected only when the data would be *unlikely* if H_0 were true. This mirrors the legal principle "innocent until proven guilty." The test does not prove H_1; rather, it shows that H_0 is implausible given the data.

Definition 5.2 (Generic formulation)**.** For a parametric family indexed by θ, we typically write

$$H_0 : \ \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \ \theta \in \Theta_1,$$

where $\Theta_0 \cap \Theta_1 = \emptyset$ and often $\Theta_1 = \Theta_0^c$.

Remark 5.2. The decomposition $\Theta = \Theta_0 \cup \Theta_1$ expresses competing pictures of the world. In many applications $\Theta_1 = \Theta_0^c$, but sometimes scientific or regulatory constraints lead to a more nuanced Θ_1. For instance, testing bioequivalence involves two one-sided alternatives excluding a region of *practical* equivalence.

Types of Hypotheses A hypothesis is **simple** if it specifies a single distribution, e.g.,

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

It is **composite** if it allows multiple distributions, e.g.,

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

Example 5.1 (One-sided hypotheses)**.** One-sided hypotheses:

$$H_0 : \theta \leq \theta_0 \text{ vs. } H_1 : \theta > \theta_0 \qquad \text{or} \qquad H_0 : \theta \geq \theta_0 \text{ vs. } H_1 : \theta < \theta_0.$$

Example 5.2 (Two-sided hypothesis)**.** Two-sided hypothesis:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0.$$

Remark 5.3. Whether a hypothesis is simple or composite has deep consequences. The Neyman–Pearson lemma provides exact optimal tests for simple-vs-simple problems. For composite alternatives, optimality typically requires additional structure (such as a monotone likelihood ratio), and truly optimal tests may not exist for two-sided problems; one then turns to likelihood-ratio, unbiasedness, or invariance principles.

Tests and Critical Regions Because the sample $\boldsymbol{X}$ is the only available statistical evidence, our decision has to be based on $\boldsymbol{X}$.

Definition 5.3 (Nonrandomized test)**.** A *nonrandomized* test of H_0 versus H_1 is specified by a critical (rejection) region R: reject H_0 if $\boldsymbol{X} \in R$ and do not reject H_0 if $\boldsymbol{X} \notin R$.

Equivalently, a test is a statistic $\varphi(\boldsymbol{X}) \in \{0, 1\}$ with rejection region $R = \{\boldsymbol{x} : \varphi(\boldsymbol{x}) = 1\}$.

A critical region is a rule that partitions all potential datasets into two sets: those that count as "too surprising under H_0" and those that do not. In practice, we almost always implement R by thresholding a *test statistic* (such as a t- or z-statistic), thereby turning complex data into a single number whose tail behavior is known under H_0.

5.1 Characterizing Tests

Error Types Given a test for H_0 versus H_1 with rejection region R,

$$\text{do not reject } H_0 \iff \boldsymbol{X} \notin R, \qquad \text{reject } H_0 \iff \boldsymbol{X} \in R.$$

How can we say that this test is good, i.e., the critical region R is reasonable? We aim to reduce false decisions:

	Decision of the test	
State of nature	**Do not reject H_0**	**Reject H_0**
H_0 true	Correct	**Type I error**
H_0 false	**Type II error**	Correct

Remark 5.4. Type I and II errors are also called *false positives* and *false negatives*. Which error is more serious depends on context: in clinical safety studies, a false positive (approving a harmful drug) is worse; in screening for rare diseases, false negatives may be more consequential. This asymmetry motivates fixing a small bound on Type I error.

Probabilities of Error Let $\varphi(\boldsymbol{X}) \in \{0, 1\}$ be the test function and $R = \{\boldsymbol{x} : \varphi(\boldsymbol{x}) = 1\}$. If $\theta \in \Theta_0$ (null is true), the **Type I error probability** is

$$\alpha(\theta) = \mathbb{E}_\theta[\varphi(\boldsymbol{X})] = \mathbb{P}_\theta(\boldsymbol{X} \in R).$$

For simple H_0 this is a single number; for composite H_0 it is a function over Θ_0. If $\theta \in \Theta_1$ (alternative is true), the **Type II error probability** is

$$1 - \beta(\theta) = \mathbb{P}_\theta(\boldsymbol{X} \notin R) = 1 - \mathbb{E}_\theta[\varphi(\boldsymbol{X})].$$

Both errors cannot be minimized simultaneously: shrinking R lowers Type I error but raises Type II error, and vice versa. For example, if R is the entire sample space, Type II error is 0 but Type I error is 1.

The size of R controls a fundamental trade-off: a wide net catches more true effects (higher chance to reject under H_1) but also hauls in more false alarms (higher chance to reject under H_0). Good tests calibrate this trade-off explicitly rather than implicitly.

Significance Level and Power In the mathematical formulation above, H_0 and H_1 appear symmetric. In applications they are not: H_0 generally represents the status quo, or what someone would believe about θ without compelling evidence to the contrary. For this reason, attention is often focused on tests that have a small chance of falsely rejecting H_0 (Type I error).

Definition 5.4 (Significance level / size)**.** The *level of significance* is the worst-case Type I error:

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\boldsymbol{X} \in R).$$

In words, the level α is the worst chance of falsely rejecting H_0.

Remark 5.5. In some texts, α is called the *size* of the test, while the *level* of the test is any number $\geq \alpha$. Some authors distinguish *level-*α tests (size $\leq \alpha$) from *size-*α tests (size $= \alpha$). In continuous models, the supremum is often attained at a boundary point of Θ_0; in discrete models, exact size α may be impossible without *randomization* (see below).

Recall that the probability of Type II error is

$$1 - \beta(\theta) = 1 - \mathbb{E}_\theta[\varphi(\boldsymbol{X})] = \mathbb{P}_\theta(\varphi(\boldsymbol{X}) = 0) = \mathbb{P}_\theta(\boldsymbol{X} \notin R).$$

Definition 5.5 (Power)**.** For $\theta \in \Theta_1$, the *power* is

$$\beta(\theta) = \mathbb{E}_\theta[\varphi(\boldsymbol{X})] = \mathbb{P}_\theta(\boldsymbol{X} \in R).$$

The power is the ability to reject when the alternative hypothesis is correct. It depends on the specific value of θ; for composite alternatives, it is a family of probabilities $\{\beta(\theta) : \theta \in \Theta_1\}$.

Power typically increases with larger departures from H_0, larger sample sizes n, and lower noise levels. Thus, beyond the *significance* of a result lies its *detectability*: small but real effects require more data to be found with high probability.

We typically fix α (e.g., 0.05) and seek tests that are as powerful as possible.

Definition 5.6 (Uniformly more powerful)**.** A test φ_1 is *uniformly more powerful* than φ_2 if

$$\mathbb{E}_\theta[\varphi_1(\boldsymbol{X})] \geq \mathbb{E}_\theta[\varphi_2(\boldsymbol{X})] \quad \text{for every } \theta \in \Theta_1.$$

Remark 5.6. Pointwise comparisons at a single θ can be misleading; uniform dominance across *all* $\theta \in \Theta_1$ is a much stronger, and often unattainable, requirement. Existence of uniformly most powerful (UMP) tests hinges on structural properties of the model (e.g., a monotone likelihood ratio).

The choice of α is usually somewhat subjective. In some applications, the tolerance for Type I error is much tighter, e.g. in criminal-justice systems (presumption of innocence).

Power Function The significance level and the power can be summarized into a single function called the *power function*, defined as the chance of rejecting H_0 as a function of $\theta \in \Theta$.

Definition 5.7 (Power function)**.** The *power function* is

$$\beta(\theta) \triangleq \mathbb{E}_\theta[\varphi(\boldsymbol{X})] = \mathbb{P}_\theta(\boldsymbol{X} \in R) = \begin{cases} \text{Type I error probability}, & \theta \in \Theta_0, \\ 1 - (\text{Type II error probability}), & \theta \in \Theta_1. \end{cases}$$

In particular, $\beta(\theta) = \alpha(\theta)$ when $\theta \in \Theta_0$, and $1 - \beta(\theta)$ is the Type II error when $\theta \in \Theta_1$. The size is $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$, and larger $\beta(\theta)$ for $\theta \in \Theta_1$ indicates a more powerful test.

Remark 5.7. Plotting $\beta(\theta)$ gives the classical *operating characteristic* of a test. For one-sided tests, the power function is typically increasing with θ under standard regularity conditions; for two-sided tests, it often has a valley near θ_0 and rises symmetrically in either direction.

Example 5.3 (Binomial power function)**.** Let $X \sim \mathrm{Bin}(5, \theta)$ and test $H_0 : \theta \le 0.5$ vs. $H_1 : \theta > 0.5$.

$$\begin{aligned} \text{Test 1: } R = \{5\} \quad &\Rightarrow \quad \beta_1(\theta) = \mathbb{P}_\theta(X = 5) = \theta^5, \\ \text{Test 2: } R = \{3, 4, 5\} \quad &\Rightarrow \quad \beta_2(\theta) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta) + \theta^5. \end{aligned}$$

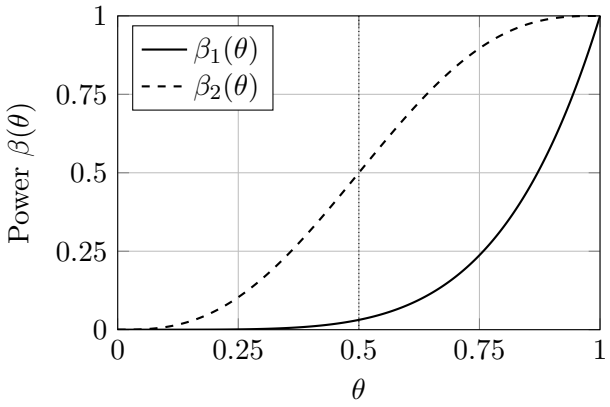


Figure 8: Illustration of two power functions for different rejection regions.

Remark 5.8. At $\theta = 0.5$, Test 1 has size $\beta_1(0.5) = (0.5)^5 = 1/32 \approx 0.03125$, while Test 2 has size $\beta_2(0.5) = \mathbb{P}(X \ge 3; n{=}5, p{=}0.5) = 0.5$. Thus Test 2 is far more powerful but violates a typical $\alpha = 0.05$ constraint. In discrete problems, an exact level like 0.05 is often unattainable with a nonrandomized rule; one then either accepts a conservative test (like Test 1) or uses boundary randomization to calibrate the size (see "Randomized Tests" below).

p-values Consider a family of tests with rejection region R_α at significance level α. As α decreases, R_α shrinks, so a fixed observation $\boldsymbol{x}$ will eventually fall outside the rejection region. The p-value is the smallest level α at which we would reject H_0.

Definition 5.8 (p-value)**.** Given a family of tests with rejection regions R_α at level α, the *p-value* of an observed sample $\boldsymbol{x}$ is the smallest α such that $\boldsymbol{x} \in R_\alpha$. Under H_0,

$$\mathbb{P}_0\{p(\boldsymbol{X}) \le \alpha\} = \alpha.$$

When $\alpha = p(\boldsymbol{x})$, the sample $\boldsymbol{x}$ lies on the boundary of R_α. Intuitively, $p(\boldsymbol{x})$ is the probability of an outcome "as or more extreme than" $\boldsymbol{x}$ under H_0.

The p-value is a calibrated surprise index under H_0: smaller values indicate outcomes that would have been rarer if H_0 were true. It is *not* the probability that H_0 is true, nor the error rate after seeing the data. Its precise definition depends on what counts as "more extreme," which is determined by the alternative (one-sided vs. two-sided) and the chosen test statistic.

Example 5.4. In a one-sided z-test for a mean with known variance, the test statistic $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ is $N(0,1)$ under H_0. Observing $Z = z_{\mathrm{obs}}$ yields $p = 1 - \Phi(z_{\mathrm{obs}})$. For a two-sided alternative one uses $p = 2\min\{\Phi(z_{\mathrm{obs}}),\, 1 - \Phi(z_{\mathrm{obs}})\}$.

Sufficiency and Tests When a sufficient statistic T is available, we may restrict our attention to tests based on T.

Theorem 5.1. *If $T(\boldsymbol{X})$ is sufficient for θ, then for any test φ the test $\psi(T) = \mathbb{E}[\varphi(\boldsymbol{X}) \mid T]$ has the same power function:*

$$\mathbb{E}_\theta[\varphi(\boldsymbol{X})] = \mathbb{E}_\theta[\psi(T(\boldsymbol{X}))] \quad \textit{for all } \theta.$$

Sufficiency ensures we can base tests on T without loss of power. This is the testing analogue of Rao–Blackwellization: conditioning on a sufficient statistic preserves all information about θ. Any decision rule can be "smoothed" by averaging over the irrelevant parts of the data (those not captured by T) without changing its operating characteristics. Note, however, that sufficiency alone does not guarantee better power; it only guarantees no loss when passing from $\boldsymbol{X}$ to T.

Randomized Tests When we want to achieve a specific level α, we may need to randomize at the boundary of R.

Definition 5.9 (Randomized test / critical function)**.** A *randomized* test uses a function $\varphi : \mathcal{X} \to [0, 1]$, where $\varphi(\boldsymbol{x})$ is the probability of rejecting H_0 given $\boldsymbol{X} = \boldsymbol{x}$.

The power is still

$$\beta(\theta) = \mathbb{P}_\theta(\text{reject } H_0) = \mathbb{E}_\theta\left[\mathbb{P}_\theta(\text{reject } H_0 \mid \boldsymbol{X})\right] = \mathbb{E}_\theta[\varphi(\boldsymbol{X})].$$

A nonrandomized test corresponds to $\varphi = \mathbb{1}_R$, and convex combinations of randomized tests are randomized tests.

In discrete models, probabilities jump in steps, so one cannot always hit a target size exactly using a simple threshold. Randomization at boundary outcomes "dithers" the decision to achieve the exact level. In continuous models, ties have probability zero and randomization is rarely needed.

Example 5.5. Continuing the binomial example with $n = 5$ and testing $H_0 : \theta \le 0.5$ at level $\alpha = 0.05$, start from $R = \{5\}$ (size $1/32 \approx 0.03125$). Add outcome $x = 4$ with probability γ when it occurs. At $\theta = 0.5$,

$$\alpha = \mathbb{P}(X = 5) + \gamma\,\mathbb{P}(X = 4) = \tfrac{1}{32} + \gamma\,\tfrac{5}{32} = 0.05 \;\Rightarrow\; \gamma = \tfrac{0.05 \cdot 32 - 1}{5} = 0.12.$$

Randomizing at $x = 4$ with probability 0.12 yields an exact level-0.05 test.

Uniformly Most Powerful (UMP) Tests Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. We aim to find the best test in $\mathcal{C}$.

Definition 5.10 (UMP)**.** Within a class $\mathcal{C}$ of tests (e.g., all level-α tests), a test with power $\beta(\cdot)$ is *uniformly most powerful* if

$$\beta(\theta) \ge \beta'(\theta) \quad \text{for all } \theta \in \Theta_1,$$

for every competitor with power $\beta'(\cdot)$ in $\mathcal{C}$.

If $\mathcal{C}$ denotes the class of level-α tests, then such a test is called a *UMP level-α test*. UMP tests are considered to be the best, but they may not always exist.

Remark 5.9. In the following sections, we will see that the Neyman–Pearson lemma characterizes UMP tests for simple vs. simple hypotheses as likelihood-ratio tests. For one-sided problems in one-parameter exponential families with a monotone likelihood ratio, UMP level-α tests exist and are again likelihood-ratio (threshold) tests. For most two-sided problems, no UMP test exists; one then uses likelihood-ratio tests for their asymptotic optimality, or restricts attention to unbiased tests or tests invariant under symmetries of the model.

5.2 Simple Hypotheses

Simple vs. Simple With $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, a test φ has

$$\alpha = \mathbb{E}_{\theta_0}[\varphi(\boldsymbol{X})], \qquad \beta = \mathbb{E}_{\theta_1}[\varphi(\boldsymbol{X})].$$

We consider the constrained maximization problem of maximizing the power β among all tests φ with significance level at most $\alpha = \mathbb{E}_{\theta_0}[\varphi(\boldsymbol{X})]$. In other words, for a simple alternative, finding a UMP test amounts to maximizing $\beta = \mathbb{E}_{\theta_1}[\varphi(\boldsymbol{X})]$ subject to the level constraint.

This is a classical constrained optimization: we want as many rejections under H_1 as possible while capping the false rejections under H_0 at α. Geometrically, we are choosing a region R in the sample space that has $\mathbb{P}_{\theta_0}(R) \leq \alpha$ but maximizes $\mathbb{P}_{\theta_1}(R)$. The optimal region places all available "Type I error budget" where the likelihood of H_1 relative to H_0 is largest.

Lemma 5.1. *Suppose $k \geq 0$ and φ^* maximizes*

$$\mathbb{E}_{\theta_1}[\varphi(\boldsymbol{X})] - k\mathbb{E}_{\theta_0}[\varphi(\boldsymbol{X})]$$

over all test functions, with $\mathbb{E}_{\theta_0}[\varphi^*(\boldsymbol{X})] = \alpha$. Then φ^* maximizes $\mathbb{E}_{\theta_1}[\varphi(\boldsymbol{X})]$ among all tests with level at most α.

Proof. For any φ with $\mathbb{E}_{\theta_0}[\varphi] \leq \alpha$,

$$\mathbb{E}_{\theta_1}[\varphi] \leq \mathbb{E}_{\theta_1}[\varphi] - k\mathbb{E}_{\theta_0}[\varphi] + k\alpha \leq \mathbb{E}_{\theta_1}[\varphi^*] - k\mathbb{E}_{\theta_0}[\varphi^*] + k\alpha = \mathbb{E}_{\theta_1}[\varphi^*]. \qquad \square$$

Remark 5.10. We need only consider tests with level/size exactly α: if the constraint is not tight, we can enlarge the rejection region slightly to increase power while keeping Type I error $\leq \alpha$. The lemma above is a Lagrange-multiplier argument in disguise: k plays the role of the multiplier that prices Type I error.

Likelihood Ratio Tests (LRT) for Simple vs. Simple To maximize the power, note that

$$\begin{aligned}
\mathbb{E}_{\theta_1}[\varphi] - k\mathbb{E}_{\theta_0}[\varphi] &= \int [f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)]\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int_{f(\boldsymbol{x}|\theta_1) > kf(\boldsymbol{x}|\theta_0)} |f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)|\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad - \int_{f(\boldsymbol{x}|\theta_1) < kf(\boldsymbol{x}|\theta_0)} |f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)|\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Clearly, any test φ^* maximizing this expression must have

$$\varphi^*(\boldsymbol{x}) = \begin{cases} 1, & f(\boldsymbol{x}|\theta_1) > kf(\boldsymbol{x}|\theta_0), \\ 0, & f(\boldsymbol{x}|\theta_1) < kf(\boldsymbol{x}|\theta_0). \end{cases}$$

Equivalently, in terms of the likelihood ratio

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_1 \mid \boldsymbol{x})}{L(\theta_0 \mid \boldsymbol{x})},$$

we reject for large $\lambda(\boldsymbol{x})$. The proof for the discrete case follows exactly the same lines.

The integrand is positive precisely where H_1 explains $\boldsymbol{x}$ much better than H_0. An optimal rule therefore uses the entire Type I error budget on such points and never wastes it where H_0 fits relatively well. Thresholding the likelihood ratio $\lambda(\boldsymbol{x})$ implements exactly this idea.

When $\lambda(\boldsymbol{x}) = k$ on a set with positive H_0-probability (typical in discrete models), no nonrandomized test can hit exact size α. Randomization at the boundary is then necessary to calibrate the size.

Definition 5.11 (Likelihood ratio test (simple vs. simple))**.** A likelihood ratio test has

$$\varphi(\boldsymbol{x}) = \begin{cases} 1, & \lambda(\boldsymbol{x}) > k, \\ \gamma \in [0,1], & \lambda(\boldsymbol{x}) = k, \\ 0, & \lambda(\boldsymbol{x}) < k, \end{cases}$$

with k, γ chosen to attain level α.

Example 5.6 (Exponential)**.** $X \sim \text{Exp}(\theta)$, $f_\theta(x) = \theta e^{-\theta x}\mathbb{1}_{x\geq 0}$. Test $H_0 : \theta = 1$ vs. $H_1 : \theta = \theta_1 > 1$.

$$\lambda(x) = \frac{\theta_1 e^{-\theta_1 x}}{1 \cdot e^{-x}} = \theta_1 e^{-(\theta_1 - 1)x}.$$

Reject for small x: $\lambda(x) > k \iff x < k' = \frac{\log(\theta_1/k)}{\theta_1 - 1}$. (When $X = k'$ the test can take any value in $[0,1]$, but the choice will not affect any power calculations since $\mathbb{P}_\theta(X = k') = 0$.) The level is

$$\alpha = \mathbb{P}_{H_0}(X < k') = 1 - e^{-k'} \quad \Rightarrow \quad k' = -\log(1-\alpha).$$

Thus the level-α UMP test is: reject if $X < -\log(1-\alpha)$.

Hence, if

$$\varphi_\alpha(x) = \begin{cases} 1, & X < -\log(1-\alpha), \\ 0, & X > -\log(1-\alpha), \end{cases}$$

then φ_α maximizes $\mathbb{E}_{\theta_1}[\varphi(X)]$ over all φ with level at most α.

Here larger θ makes small X more likely (the distribution is stochastically smaller as the rate increases). The optimal rejection region therefore collects small x. A pleasant surprise is that the critical value depends only on α, not on the particular $\theta_1 > 1$: this foreshadows the UMP result for composite one-sided tests under monotone likelihood ratio.

Remark 5.11. In continuous models such as the exponential, $\mathbb{P}_{\theta_0}(X = k') = 0$, so randomization at $X = k'$ is immaterial. In discrete models (next example), we must randomize at the boundary to attain an arbitrary target size.

Example 5.7 (Binomial example)**.** $X \sim \text{Bin}(2, \theta)$. Test $H_0 : \theta = \frac{1}{2}$ vs. $H_1 : \theta = \frac{3}{4}$. Under H_0, we have

$$\lambda(x) = \frac{\binom{2}{x}(3/4)^x(1/4)^{2-x}}{\binom{2}{x}(1/2)^x(1/2)^{2-x}} = \frac{3^x}{4} = \begin{cases} \frac{1}{4}, & x = 0, \\ \frac{3}{4}, & x = 1, \\ \frac{9}{4}, & x = 2. \end{cases}$$

For $\alpha = 0.05$, any nonrandomized choice with $x = 2$ in R yields $\alpha \geq \mathbb{P}_{\theta_0}(X = 2) = \frac{1}{4} > 0.05$. To attain $\alpha = 0.05$ we must randomize at $x = 2$: set $\varphi(2) = 1/5$, $\varphi(0) = \varphi(1) = 0$.

Previously, when we derived that the likelihood ratio test is optimal, we did not specify $\mathbb{E}_{\theta_0}[\varphi] = \alpha$. As we see from the binomial example, a randomized test is needed to achieve level α in a typical discrete situation.

Calibrating α here solves $\alpha = \mathbb{P}_{\theta_0}(X = 2) \cdot \gamma = \frac{1}{4}\gamma$, giving $\gamma = 0.20$. Randomization at the most extreme outcome allocates just enough probability mass to reach the target size.

Remark 5.12. Operationally, work with the H_0-distribution of $\Lambda = \lambda(\boldsymbol{X})$. Choose k as a $(1-\alpha)$-quantile of Λ when possible; otherwise, interpolate stochastically by randomizing at the atom at k.

This construction procedure yielding a level-α LRT is always possible, leading to the existence of UMP tests in simple-vs-simple problems.

Algorithm 1 Procedure to find an LRT with level α (simple vs. simple)

1: Compute the likelihood ratio $\Lambda = \lambda(\boldsymbol{X})$ under H_0.
2: Choose k so that $\mathbb{P}_{\theta_0}(\Lambda > k) \le \alpha \le \mathbb{P}_{\theta_0}(\Lambda \ge k)$ (i.e., α is bracketed by the jump at k).
3: If $\mathbb{P}_{\theta_0}(\Lambda = k) > 0$, choose $\gamma \in [0,1]$ such that $\mathbb{P}_{\theta_0}(\Lambda > k) + \gamma\,\mathbb{P}_{\theta_0}(\Lambda = k) = \alpha$.
4: Reject when $\Lambda > k$; randomize with probability γ when $\Lambda = k$; do not reject when $\Lambda < k$.

Theorem 5.2 (Neyman–Pearson Lemma: Existence)**.** *For any $\alpha \in [0,1]$ there exists a likelihood ratio test with level α, and any such LRT maximizes power among all tests with level at most α.*

Theorem 5.3 (Neyman–Pearson Lemma: Uniqueness)**.** *Fix α. Let φ_α be a level-α LRT with critical value k and $B = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \neq k\}$. If φ^* also maximizes power at level α, then φ^* and φ_α agree on B almost surely.*

Remark 5.13. Thus, for simple vs. simple, an optimal test must be an LRT (up to randomization on a null set). If the measure of B is 0 (usually the case for continuous distributions), then there is a unique nonrandomized test; otherwise, a randomized test is necessary to achieve the desired α.

Proof. Assume $k < \infty$ and let $B_1 = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) > k\}$ and $B_2 = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) < k\}$.

Let φ_α be the UMP likelihood ratio test. Then $\mathbb{E}_{\theta_1}[\varphi^*] = \mathbb{E}_{\theta_1}[\varphi_\alpha]$. Since φ_α maximizes $\mathbb{E}_{\theta_1}[\varphi] - k\mathbb{E}_{\theta_0}[\varphi]$, we have

$$k\mathbb{E}_{\theta_0}[\varphi_\alpha] = k\alpha \le k\mathbb{E}_{\theta_0}[\varphi^*] \quad \Rightarrow \quad \mathbb{E}_{\theta_0}[\varphi^*] = \alpha.$$

Hence,

$$\mathbb{E}_{\theta_1}[\varphi_\alpha] - k\mathbb{E}_{\theta_0}[\varphi_\alpha] = \mathbb{E}_{\theta_1}[\varphi^*] - k\mathbb{E}_{\theta_0}[\varphi^*].$$

Recall that

$$\mathbb{E}_{\theta_1}[\varphi_\alpha] - k\mathbb{E}_{\theta_0}[\varphi_\alpha] = \int \mathbb{1}_{B_1} |f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)| \mathrm{d}\boldsymbol{x}.$$

Therefore,

$$\int \mathbb{1}_{B_1} |f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)|(1 - \varphi^*(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} + \int \mathbb{1}_{B_2} |f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)|\varphi^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0.$$

Since the integrands are nonnegative, they must be 0 almost surely. □

The Neyman–Pearson lemma (1933) established the likelihood ratio as the canonical form of most powerful tests in simple-vs-simple settings. Its influence extends to composite problems via monotone likelihood ratios and, asymptotically, to likelihood-ratio tests in general models.

Corollary 5.1. *If $T(\boldsymbol{X})$ is sufficient with density $g(t \mid \theta_i)$, $i = 0, 1$, then the LRT based on T,*

$$\varphi_{T,\alpha}(t) = \mathbb{1}_{\lambda(t)>k} \text{ and possibly randomize at } \lambda(t) = k, \qquad \lambda(t) = \frac{g(t \mid \theta_1)}{g(t \mid \theta_0)},$$

is also UMP among level-α tests.

Proof. Factorization theorem implies that

$$f(\boldsymbol{x}|\theta_i) = g(T(\boldsymbol{x})|\theta_i)h(\boldsymbol{x}), \quad i = 0, 1.$$ □

Sufficiency means T captures all information about θ. Replacing the full data by T cannot degrade a likelihood-ratio comparison between H_0 and H_1; the optimal rejection region can therefore be drawn on the T-axis without loss.

5.3 UMP for One-Sided Tests

Previously, we studied UMP tests for simple hypotheses. The case where H_0 and H_1 are both simple is mainly of theoretical interest; when a hypothesis is not simple, it is called composite. We already saw an example with an exponential distribution (Example 5.6) where a UMP likelihood-ratio test for a composite H_1 exists. Unsurprisingly, likelihood-ratio ideas remain central for composite hypotheses.

Remark 5.14. For composite alternatives, "UMP" is a strong requirement: the test must dominate all competitors for every θ in the alternative. This is rarely possible without additional structure; the key property that rescues one-sided problems is a *monotone likelihood ratio*.

Composite Hypotheses and MLR Recall that a test φ^* with level α is called uniformly most powerful if

$$\beta^*(\theta) = \mathbb{E}_\theta[\varphi^*] \geq \beta(\theta) = \mathbb{E}_\theta[\varphi], \quad \forall \theta \in \Theta_1,$$

for all φ with level at most α.

When H_1 is composite, UMP tests generally only arise when the parameter of interest is univariate, $\theta \in \Theta \subset \mathbb{R}$, and the hypotheses are of the form $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is a fixed constant. In addition, the family of densities needs to have an appropriate structure.

Recall the exponential example (Example 5.6), where UMP exists for composite hypotheses. We now extend that result to a class of parametric problems in which the likelihood functions have a special property.

Definition 5.12 (Monotone likelihood ratio (MLR))**.** A family $f(\boldsymbol{x} \mid \theta)$, $\theta \in \Theta \subset \mathbb{R}$, has *monotone likelihood ratio in* a statistic $T(\boldsymbol{X})$ if for $\theta_2 > \theta_1$, the ratio $f(\boldsymbol{x} \mid \theta_2)/f(\boldsymbol{x} \mid \theta_1)$ is nondecreasing in $T(\boldsymbol{x})$ (on the set where at least one density is positive).

Remark 5.15. We assume the distributions are identifiable: $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$ whenever $\theta_1 \neq \theta_2$. Natural conventions concerning division by zero are used, with the likelihood ratio interpreted as $+\infty$ when $f(x \mid \theta_2) > 0$ and $f(x \mid \theta_1) = 0$. On the null set where both densities are zero, the likelihood ratio is not defined and monotonic dependence on T is not required.

MLR says that, as θ increases, the likelihood shifts monotonically with T: larger T's systematically favor larger θ's. This aligns the direction of evidence so that a single-threshold rule on T can serve as the most powerful test for every point in the alternative.

Example 5.8 (Exponential family)**.** If

$$f(\boldsymbol{x} \mid \theta) = \Big(\prod_i h(x_i)\Big) c(\theta)^n \exp\Big(\eta(\theta) \sum_i t(x_i)\Big),$$

with strictly increasing $\eta(\theta)$, then for $\theta_2 > \theta_1$ the likelihood ratio

$$\frac{f(\boldsymbol{x}|\theta_2)}{f(\boldsymbol{x}|\theta_1)} = \frac{c^n(\theta_2)}{c^n(\theta_1)} \exp\left((\eta(\theta_2) - \eta(\theta_1)) \sum_i t(x_i)\right)$$

is increasing in $T = \sum_i t(x_i)$ (hence MLR in T). Typical examples include binomial, Poisson, negative binomial, normal (known variance), exponential, gamma, beta, etc.

Example 5.9 (Uniform$(0, \theta)$)**.** For $\theta_2 > \theta_1$,

$$\frac{f(\boldsymbol{x} \mid \theta_2)}{f(\boldsymbol{x} \mid \theta_1)} = \frac{\theta_1^n}{\theta_2^n} \cdot \frac{\mathbb{1}_{x_{(n)} \leq \theta_2}}{\mathbb{1}_{x_{(n)} \leq \theta_1}},$$

which is nondecreasing in $x_{(n)}$ (on the relevant support).

Theorem 5.4 (UMP for one-sided tests under MLR)**.** *Consider $H_0 : \theta \le \theta_0$ vs. $H_1 : \theta > \theta_0$. If $f(\boldsymbol{x} \mid \theta)$ has MLR in $T(\boldsymbol{X})$, then the test*

$$\varphi^*(\boldsymbol{x}) = \begin{cases} 1, & T(\boldsymbol{x}) > c, \\ \gamma \in [0,1], & T(\boldsymbol{x}) = c, \\ 0, & T(\boldsymbol{x}) < c, \end{cases}$$

with c, γ chosen to have level α, is UMP among level-α tests.

Lemma 5.2. *If $f(\boldsymbol{x} \mid \theta)$ has MLR in T and $\psi(T)$ is nondecreasing, then $g(\theta) = \mathbb{E}_\theta[\psi(T)]$ is nondecreasing in θ.*

Proof. Let $\theta_1 < \theta_2$. Define

$$A = \{\boldsymbol{x} : f(\boldsymbol{x}|\theta_1) > f(\boldsymbol{x}|\theta_2)\}, \qquad B = \{\boldsymbol{x} : f(\boldsymbol{x}|\theta_1) < f(\boldsymbol{x}|\theta_2)\},$$

and let

$$a = \sup_{\boldsymbol{x} \in A} \psi(T(\boldsymbol{x})), \qquad b = \inf_{\boldsymbol{x} \in B} \psi(T(\boldsymbol{x})).$$

Then $b \ge a$ because of MLR and because ψ is nondecreasing. Now

$$\begin{aligned} g(\theta_2) - g(\theta_1) &= \int \psi(T(\boldsymbol{x}))(f(\boldsymbol{x}|\theta_2) - f(\boldsymbol{x}|\theta_1))\mathrm{d}\boldsymbol{x} \\ &\ge a \int_A (f(\boldsymbol{x}|\theta_2) - f(\boldsymbol{x}|\theta_1))\mathrm{d}\boldsymbol{x} + b \int_B (f(\boldsymbol{x}|\theta_2) - f(\boldsymbol{x}|\theta_1))\mathrm{d}\boldsymbol{x} \\ &= (b-a) \int_B (f(\boldsymbol{x}|\theta_2) - f(\boldsymbol{x}|\theta_1))\mathrm{d}\boldsymbol{x} \\ &\ge 0. \end{aligned}$$

□

Proof of Theorem 5.4. Consider simple hypotheses $\theta = \theta_0$ versus $\theta = \theta_1 > \theta_0$. By the Neyman–Pearson lemma, a UMP test for this simple-vs-simple problem is a likelihood-ratio test

$$\varphi(\boldsymbol{x}) = \begin{cases} 1, & \lambda(\boldsymbol{x}) > c, \\ \gamma, & \lambda(\boldsymbol{x}) = c, \\ 0, & \lambda(\boldsymbol{x}) < c. \end{cases}$$

Since the family of densities has monotone likelihood ratio, this UMP test can be chosen to depend on the data only through $T(\boldsymbol{x})$; call it $\varphi^*(\boldsymbol{x}) = \varphi^*(T(\boldsymbol{x}))$. This test does not depend on the particular $\theta_1 > \theta_0$ (because $T(\boldsymbol{X})$ is a statistic), hence it is UMP for $\theta = \theta_0$ versus the composite alternative $H_1 : \theta > \theta_0$.

Moreover, φ^* is nondecreasing in T, so the lemma above implies that $\beta(\theta) = \mathbb{E}_\theta[\varphi^*(T(\boldsymbol{X}))]$ is nondecreasing in θ. Therefore, for all $\theta \le \theta_0$ we have $\beta(\theta) \le \beta(\theta_0) = \alpha$, and φ^* is a UMP level-α test for H_0 versus H_1. □

Remark 5.16. The essence is: MLR aligns evidence so that the same tail of T is most informative against H_0 for all $\theta > \theta_0$. Thresholding T therefore simultaneously solves all simple subproblems H_0 vs. $\theta_1 > \theta_0$, yielding UMP for the composite alternative.

Corollary 5.2 (Exponential families)**.** *If X is in a one-parameter exponential family with strictly monotone $\eta(\theta)$, then: (i) if η is strictly increasing, the UMP for $H_0 : \theta \le \theta_0$ vs. $H_1 : \theta > \theta_0$ rejects for large T; (ii) if η is strictly decreasing, or for $H_0 : \theta \ge \theta_0$ vs. $H_1 : \theta < \theta_0$, the UMP rejects for small T.*

The sign of $\eta'(\theta)$ dictates which tail of T carries evidence against H_0. "Large-T" and "small-T" tests are simply mirrors of each other across this sign change.

Example 5.10 (Normal mean, σ known)**.** Test $H_0 : \mu \le \mu_0$. Here $T(\boldsymbol{X}) = \bar{X}$ and MLR holds. The UMP is

$$\text{reject if } \bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \quad \text{(nonrandomized)}.$$

Remark 5.17. Under H_0 with equality $\mu = \mu_0$, $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$, so the critical value is the $(1-\alpha)$-quantile $z_{1-\alpha}$. No randomization is needed because the distribution is continuous.

Example 5.11 (Bernoulli p)**.** $H_0 : p \le p_0$. Take $T(\boldsymbol{X}) = \sum_i X_i$ with strictly increasing $\eta(p) = \log\left(\frac{p}{1-p}\right)$. A (typically randomized) critical value on T yields a level-α UMP.

Example 5.12 (Poisson θ)**.** $H_0 : \theta \le \theta_0$. Take $T(\boldsymbol{X}) = \sum_i X_i$, $\eta(\theta) = \log\theta$ increasing. Randomization is typically needed for exact level α in the discrete case.

Example 5.13 (Uniform$(0,\theta)$)**.** MLR in $X_{(n)}$. The UMP rejects for large $X_{(n)}$ with

$$\alpha = \mathbb{E}_{\theta_0}[\varphi^*] = n\theta_0^{-n} \int_c^{\theta_0} x^{n-1}\mathrm{d}x = 1 - \left(\frac{c}{\theta_0}\right)^n,$$

so $c = \theta_0(1-\alpha)^{1/n}$ (nonrandomized).

For Uniform$(0,\theta)$, the maximum $X_{(n)}$ is sufficient and stochastically increases with θ. Evidence against $H_0 : \theta \le \theta_0$ accumulates in unusually large maxima, hence a right-tail test on $X_{(n)}$ is UMP.

5.4 Two-Sided Tests

Now we consider the two-sided hypothesis

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \ne \theta_0.$$

Remark 5.18. Typographical note: in a two-sided problem, the alternative is usually written as $H_1 : \theta \ne \theta_0$.

Definition 5.13 (Two-sided test)**.** A test φ is two-sided if there exist $t_1 \le t_2$ such that

$$\varphi(\boldsymbol{x}) = \begin{cases} 1, & T(\boldsymbol{x}) < t_1 \text{ or } T(\boldsymbol{x}) > t_2, \\ 0, & T(\boldsymbol{x}) \in (t_1, t_2), \end{cases}$$

and it is not equivalent to a one-sided test.

A two-sided test flags evidence on both tails of a statistic T: values of T that are unusually small or unusually large under H_0 constitute evidence against H_0. Symmetry of the rejection region is often motivated by symmetry in the model (e.g. normal mean with known variance), but the central message is that departures in either direction are relevant.

Nonexistence of UMP Two-Sided Tests Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$ for a population with a one-parameter exponential family, i.e., with density

$$c(\theta)h(x)e^{\eta(\theta)T(x)}, \quad \theta \in \Theta.$$

Assume that $\eta(\theta)$ is strictly increasing. Decompose the two-sided alternative into two one-sided problems: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \ge \theta_0$ and $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \le \theta_0$. There are two level-$\alpha$ UMP tests for these one-sided alternatives,

$$\varphi_+(\boldsymbol{x}) = \begin{cases} 1, & T(\boldsymbol{x}) > c_+, \\ \gamma, & T(\boldsymbol{x}) = c_+, \\ 0, & T(\boldsymbol{x}) < c_+, \end{cases} \qquad \text{and} \qquad \varphi_-(\boldsymbol{x}) = \begin{cases} 1, & T(\boldsymbol{x}) < c_-, \\ \gamma, & T(\boldsymbol{x}) = c_-, \\ 0, & T(\boldsymbol{x}) > c_-. \end{cases}$$

If $\theta_- < \theta_0 < \theta_+$, then φ_+ has the maximal power at θ_+, and φ_- has the maximal power at θ_-. Decomposing further into $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_-$ or $H_1 : \theta = \theta_+$, the uniqueness part of the Neyman–Pearson lemma shows that there cannot be a UMP level-α two-sided test.

Two-sided problems must be powerful in both directions. But the one-sided UMP tests φ_+ and φ_- point in opposite directions, and each is tailored to a different alternative. A single test cannot simultaneously dominate both, hence no UMP exists in general. This is why extra criteria (e.g. unbiasedness) are introduced to select a "best" two-sided test.

Example 5.14 (One-sided vs. two-sided power)**.** Consider three tests for a normal mean (known variance): Test 1 with rejection region $R_1 = (-\infty, \theta_0 - \sigma z(\alpha)/\sqrt{n})$ and power function β_1; Test 2 with rejection region $R_2 = (\theta_0 + \sigma z(\alpha)/\sqrt{n}, \infty)$ and power function β_2; and the two-sided Test 3 with $R_3 = R_1 \cup R_2$ and power function β_3.

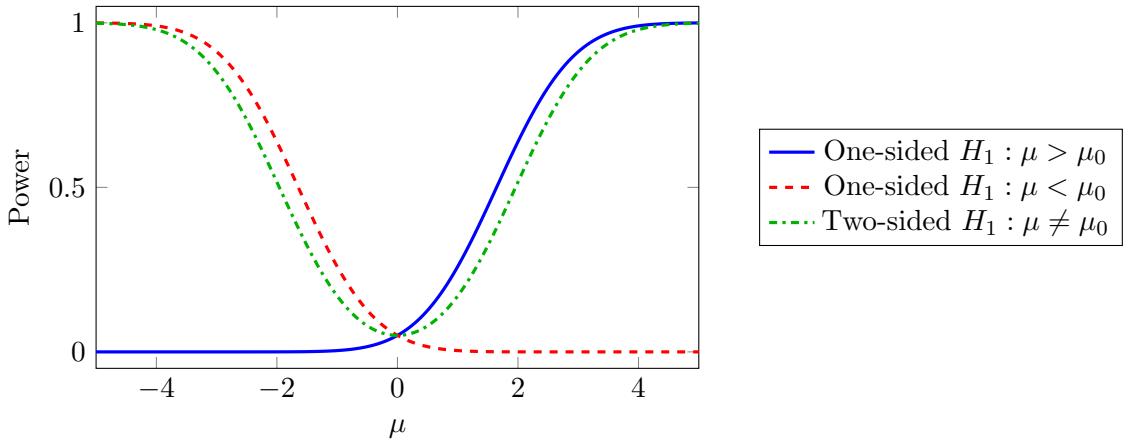


Figure 9: Illustrative power functions: the two-sided test cannot dominate both one-sided tests uniformly.

Unbiasedness and UMPU When a UMP test does not exist, we impose reasonable restrictions on the class of tests, then find the best one among the restricted class. One such restriction is that the test should be at least as good as a "silly guess" that rejects H_0 with probability α, independent of the observation.

Definition 5.14 (Unbiased test)**.** A level-α test with power $\beta(\theta)$ is *unbiased* if $\beta(\theta) \geq \alpha$ for all $\theta \in \Theta_1$ and $\beta(\theta) \leq \alpha$ for all $\theta \in \Theta_0$.

When UMP does not exist, one seeks a uniformly most powerful unbiased (UMPU) test. A UMP test (if it exists) is always unbiased, so UMPU is mainly relevant when UMP fails.

Unbiasedness rules out perverse two-sided tests that, near θ_0, have power lower than the level α. In exponential families, an elegant characterization emerges: among level-α tests, those whose power curve is flat to first order at θ_0 (i.e. $\beta'(\theta_0) = 0$) are the UMPU tests.

Theorem 5.5 (UMPU in exponential families (without proof))**.** *Consider $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ in a one-parameter exponential family $c(\theta)h(x)\exp(\eta(\theta)T(x))$ with differentiable strictly increasing $\eta(\theta)$ and $0 < \eta'(\theta_0) < \infty$. There exists a two-sided level-α test φ^* with $\beta'_{\varphi^*}(\theta_0) = 0$, and any such test is UMPU.*

Remarks UMPU tests in normal families (where UMP two-sided does not exist) include one-sample two-sided z-tests, one-sample two-sided t-tests, two-sided χ^2 tests (with certain unequal tails), and one- and two-sided two-sample F-tests and two-sample t-tests.

Remark 5.19. In these classical settings, UMPU tests are obtained by LRTs together with symmetry and unbiasedness. For the t-test, sufficiency and a pivot (the t statistic) deliver exact finite-sample calibration and the UMPU property.

5.5 Likelihood Ratio Tests

Previously, we introduced likelihood ratio tests for simple hypotheses. We now consider general hypotheses, possibly composite.

General LRT (Composite Hypotheses) Let Θ be the full parameter space and $\Theta_0 \subset \Theta$ the null set. The *likelihood ratio statistic* is

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid \boldsymbol{x})}{\sup_{\theta \in \Theta} L(\theta \mid \boldsymbol{x})}.$$

The denominator is the MLE over Θ (the maximizer of $L(\theta \mid \boldsymbol{x})$ over Θ), and the numerator is the constrained MLE over Θ_0. We reject for small $\lambda(\boldsymbol{x})$.

Definition 5.15 (LRT rule)**.** An LRT rejects for small $\lambda(\boldsymbol{x})$:

$$\varphi(\boldsymbol{x}) = \begin{cases} 1, & \lambda(\boldsymbol{x}) < c, \\ 0, & \lambda(\boldsymbol{x}) > c, \end{cases} \qquad \text{(randomize if } \lambda(\boldsymbol{x}) = c\text{)}.$$

The LRT compares the best possible fit under the null to the best possible fit overall. If the null's best fit is much worse, the data lean away from H_0. In regular problems, LRTs are natural, often optimal, and enjoy powerful asymptotic guarantees.

Example 5.15 (Normal, variance known)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\theta, 1)$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

$$\begin{aligned} \lambda(\boldsymbol{x}) &= \frac{(\sqrt{2\pi})^{-n} \exp\left[-\frac{1}{2}\sum_{i=1}^n (x_i - \theta_0)^2\right]}{(\sqrt{2\pi})^{-n} \exp\left[-\frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2\right]} \\ &= \exp\left\{-\tfrac{1}{2}\left[\sum (x_i - \theta_0)^2 - \sum (x_i - \bar{x})^2\right]\right\} = \exp\left[-\tfrac{n}{2}(\bar{x} - \theta_0)^2\right]. \end{aligned}$$

The rejection region $\{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$ for some $c \in [0, 1]$ is

$$\{\boldsymbol{x} : |\bar{x} - \theta_0| \geq \sqrt{-2\log(c)/n}\}.$$

This has the same form as the classical one-sample z-test for a normal mean, with a one-to-one correspondence between c and α. Similarly, the one-sided tests are also LRTs. Note that the test depends on the sample only through the sufficient statistic $\bar{x}$, and it is a UMPU test.

Remark 5.20. The algebra uses $\sum (x_i - \theta_0)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \theta_0)^2$. Thus $-2\log\lambda = n(\bar{x} - \theta_0)^2$, which under H_0 is χ^2_1 when σ^2 is known and scaled appropriately. This aligns the LRT with the familiar z-test.

Example 5.16 (Shifted exponential)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $f(x|\theta) = e^{-(x-\theta)}$, $x \geq \theta$. Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

The likelihood function is

$$L(\theta|\boldsymbol{X}) = \begin{cases} e^{-\sum x_i + n\theta}, & \theta \leq x_{(1)}, \\ 0, & \theta > x_{(1)}. \end{cases} \quad \Rightarrow \quad \widehat{\theta} = x_{(1)}.$$

So

$$\lambda(\boldsymbol{X}) = \begin{cases} 1, & x_{(1)} \leq \theta_0, \\ e^{-n(x_{(1)} - \theta_0)}, & x_{(1)} > \theta_0. \end{cases}$$

The rejection region is

$$\{\boldsymbol{x} : x_{(1)} \geq \theta_0 - \frac{\log(c)}{n}\}.$$

The test depends on the sample only through the sufficient statistic $x_{(1)}$.

Remark 5.21. Here the minimal order statistic $x_{(1)}$ is sufficient and increases stochastically with θ. The LRT is therefore a simple threshold on $x_{(1)}$, agreeing with the UMP result under MLR.

If $T(\boldsymbol{X})$ is a sufficient statistic for θ, we can obtain its likelihood function (pdf/pmf) $L^*(\theta|t) = g(t|\theta)$, which yields an LRT statistic $\lambda^*(t)$.

Theorem 5.6 (Sufficiency and LRT)**.** *If $T(\boldsymbol{X})$ is sufficient and $\lambda^*(t)$ and $\lambda(\boldsymbol{x})$ are the LRT statistics based on T and $\boldsymbol{X}$, then*

$$\lambda^*(T(\boldsymbol{x})) = \lambda(\boldsymbol{x}) \quad \textit{for every } \boldsymbol{x}.$$

Proof. Apply the factorization theorem:

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\Theta} L(\theta|\boldsymbol{x})} = \frac{\sup_{\Theta_0} f(\boldsymbol{x}|\theta)}{\sup_{\Theta} f(\boldsymbol{x}|\theta)} = \frac{\sup_{\Theta_0} g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{\sup_{\Theta} g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})} = \lambda^*(T(\boldsymbol{x})).$$

The statistic $\lambda^*(\cdot)$ depends on the sample only through $T(\cdot)$, as in the previous two examples. □

Remark 5.22. (Asymptotic calibration – Wilks' phenomenon.) Under standard regularity conditions, $-2\log\lambda(\boldsymbol{X}) \xrightarrow{d} \chi^2_d$ with $d = \dim(\Theta) - \dim(\Theta_0)$. This provides large-sample critical values for general composite tests without needing the exact finite-sample distribution.

Example 5.17 (Normal, σ unknown)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma)$. Consider testing $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$.

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\mu\le\mu_0,\sigma^2\ge 0} L(\mu,\sigma|\boldsymbol{x})}{\sup_{\mu\in\mathbb{R},\sigma^2\ge 0} L(\mu,\sigma|\boldsymbol{x})} = \begin{cases} 1, & \text{if } \widehat{\mu} \le \mu_0, \\ \frac{L(\mu_0,\widehat{\sigma}_0^2|\boldsymbol{x})}{L(\widehat{\mu},\widehat{\sigma}^2|\boldsymbol{x})}, & \text{if } \widehat{\mu} > \mu_0. \end{cases}$$

Recall that $(\widehat{\mu}, \widehat{\sigma}^2) = (\bar{X}, \frac{n-1}{n}S^2)$ is the MLE of (μ, σ^2). For the constrained problem, using the method of Lagrange multipliers, the MLE yields $(\mu_0, \widehat{\sigma}_0^2)$ for $\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2$ when $\widehat{\mu} > \mu_0$. The rejection region of the LRT is

$$R = \left\{ \frac{S^2}{\sum_{i=1}^n (X_i - \mu_0)^2} < c \right\}.$$

It has the same form as the classical one-sample t-test,

$$R = \left\{ \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} > t_{n-1}(1-\alpha) \right\},$$

using the algebraic identity.

Remark 5.23. Using $\sum(X_i - \mu_0)^2 = (n-1)S^2 + n(\bar{X} - \mu_0)^2$, one rewrites the ratio in the LRT as a monotone function of $T = \sqrt{n}(\bar{X} - \mu_0)/S$. Hence the LRT is equivalent to the classical t-test and enjoys UMPU in the one-sided normal-mean problem with unknown variance.

5.6 Sequential Testing

In all previous tests, we have fixed the sample size n in advance. But in many applications (e.g., clinical trials), we often observe the data sequentially: we continue to gather samples until a confident conclusion can be made. This idea goes back to Wald (1945), and is referred to as the *sequential probability ratio test* (SPRT).

Wald showed that, for given error probabilities, the SPRT minimizes the expected sample size among all (possibly sequential) tests—a striking optimality that explains its central role in quality control, biostatistics, and A/B testing.

Algorithm 2 SPRT decision rule

1: Choose thresholds $0 < c_0 < c_1 < \infty$.
2: At time n, compute $\lambda_n(\boldsymbol{X}_n)$.
3: **if** $\lambda_n \geq c_1$ **then**
4: Reject H_0.
5: **else if** $\lambda_n \leq c_0$ **then**
6: Accept H_0.
7: **else**
8: Continue sampling.
9: **end if**

Sequential Probability Ratio Test (SPRT) Instead of a fixed n, observe data sequentially. For testing

$$H_0 : X_i \overset{\text{i.i.d.}}{\sim} f_0 \quad \text{versus} \quad H_1 : X_i \overset{\text{i.i.d.}}{\sim} f_1,$$

define the running likelihood ratio

$$\lambda_n(\boldsymbol{X}_n) = \prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)} = \frac{L_1(\boldsymbol{X}_n)}{L_0(\boldsymbol{X}_n)}.$$

The running likelihood ratio accumulates evidence multiplicatively. Crossing the upper boundary indicates strong cumulative evidence for H_1; crossing the lower boundary indicates evidence for H_0; values in between call for more data. Plotting $\log \lambda_n$ turns this into a random walk with two absorbing barriers.

Thresholds and Error Control The thresholds c_0 and c_1 are chosen to control the Type I error $\alpha = \mathbb{P}_0(\lambda_n(\boldsymbol{X}_n) \geq c_1)$ and the Type II error $1 - \beta = \mathbb{P}_1(\lambda_n(\boldsymbol{X}_n) \leq c_0)$. Let us express the two types of error:

$$\begin{aligned}\beta \geq \mathbb{P}_1(\lambda_n(\boldsymbol{X}_n) \geq c_1) &= \int_{\lambda_n(\boldsymbol{x}) \geq c_1} \frac{L_1(\boldsymbol{x})}{L_0(\boldsymbol{x})} L_0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_{\lambda_n(\boldsymbol{x}) \geq c_1} \lambda_n(\boldsymbol{x}) L_0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\geq c_1 \int_{\lambda_n(\boldsymbol{x}) \geq c_1} L_0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = c_1 \alpha.\end{aligned}$$

Similarly,

$$\begin{aligned}1 - \alpha \geq \mathbb{P}_0(\lambda_n(\boldsymbol{X}_n) \leq c_0) &= \int_{\lambda_n(\boldsymbol{x}) \leq c_0} \frac{L_0(\boldsymbol{x})}{L_1(\boldsymbol{x})} L_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_{\lambda_n(\boldsymbol{x}) \leq c_0} \lambda_n^{-1}(\boldsymbol{x}) L_0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\geq c_0^{-1} \int_{\lambda_n(\boldsymbol{x}) \leq c_0} L_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = c_0^{-1}(1 - \beta).\end{aligned}$$

Now we have

$$\beta \geq c_1 \alpha, \quad \text{and} \quad 1 - \alpha \geq c_0^{-1}(1 - \beta).$$

This implies that

$$c_1 \leq \frac{\beta}{\alpha}, \quad \text{and} \quad c_0 \geq \frac{1 - \beta}{1 - \alpha}.$$

So we set

$$c_1 = \frac{\beta}{\alpha}, \quad \text{and} \quad c_0 = \frac{1 - \beta}{1 - \alpha},$$

to guarantee both Type I and Type II errors are controlled at level α and β, respectively.

Remark 5.24 (Conventions vary). Some authors write the error constraints as $\mathbb{P}_0(\text{reject } H_0) \leq \alpha$ and $\mathbb{P}_1(\text{accept } H_0) \leq \beta_{\text{err}}$ (where β_{err} is the *Type II error*, not the power). Under that parameterization, the classical Wald approximations use

$$A = \frac{1-\beta_{\text{err}}}{\alpha} \quad \text{and} \quad B = \frac{\beta_{\text{err}}}{1-\alpha}$$

for the upper and lower likelihood-ratio boundaries. The bounds derived above ($\beta \geq c_1\alpha$, $1-\alpha \geq c_0^{-1}(1-\beta)$) are one-sided inequalities; equalities hold only approximately, and the final choice of constants must be aligned with the notation in use.

Expected Stopping Time and Wald's Identity Let us consider the expectation of the log-likelihood ratio statistic at a fixed time n:

$$\mathbb{E}_1[\log \lambda_n] = \mathbb{E}_1\left[\sum_{i=1}^{n} \log \frac{f_1(X_i)}{f_0(X_i)}\right] = \sum_{i=1}^{n} \mathbb{E}_1\left[\log \frac{f_1(X_i)}{f_0(X_i)}\right] = n \cdot \mathrm{KL}(f_1 \| f_0),$$

where $\mathrm{KL}(f_1\|f_0) = \mathbb{E}_1\left[\log \frac{f_1(X_i)}{f_0(X_i)}\right]$ is the Kullback–Leibler divergence between f_1 and f_0. Similarly,

$$\mathbb{E}_0[\log \lambda_n] = -n \cdot \mathrm{KL}(f_0 \| f_1).$$

We have one complication: we do not know when the test will stop, i.e.,

$$\tau = \inf\{n \geq 0 : \lambda_n \geq c_1 \text{ or } \lambda_n \leq c_0\}$$

is a (random) stopping time. More importantly, τ is not independent of λ_n.

Luckily, we have Wald's identity.

Theorem 5.7 (Wald's identity)**.** *Let $Y_1, Y_2, \ldots$ be i.i.d. with mean μ. Let τ be a random variable such that $\mathbb{E}[\tau] < \infty$ and the event $\{\tau = t\}$ is determined by $Y_1, \ldots, Y_t$ and independent of Y_i for $i > t$. Then*

$$\mathbb{E}\left[\sum_{i=1}^{\tau} Y_i\right] = \mu \mathbb{E}[\tau].$$

By Wald's identity, if Y_i are i.i.d. with mean μ and $\mathbb{E}[\tau] < \infty$, then $\mathbb{E}_\theta[\sum_{i=1}^{\tau} Y_i] = \mu\mathbb{E}[\tau]$. Applying this to $Y_i = \log \frac{f_1(X_i)}{f_0(X_i)}$, we obtain

$$\mathbb{E}_1[\log \lambda_\tau] = \mathbb{E}_1[\tau] \cdot \mathrm{KL}(f_1\|f_0) \quad \Rightarrow \quad \mathbb{E}_1[\tau] = \frac{\mathbb{E}_1[\log \lambda_\tau]}{\mathrm{KL}(f_1\|f_0)},$$

and

$$\mathbb{E}_0[\log \lambda_\tau] = -\mathbb{E}_0[\tau] \cdot \mathrm{KL}(f_0\|f_1) \quad \Rightarrow \quad \mathbb{E}_0[\tau] = \frac{\mathbb{E}_0[\log \lambda_\tau]}{-\mathrm{KL}(f_0\|f_1)}.$$

Approximations using the thresholds give

$$\mathbb{E}_0[\tau] \approx \frac{\alpha \log(\beta/\alpha) + (1-\alpha)\log((1-\beta)/(1-\alpha))}{-\mathrm{KL}(f_0\|f_1)},$$

$$\mathbb{E}_1[\tau] \approx \frac{\beta \log(\beta/\alpha) + (1-\beta)\log((1-\beta)/(1-\alpha))}{\mathrm{KL}(f_1\|f_0)}.$$

Hence, expected sample size grows as errors tighten (smaller α or $1-\beta$) or as the models f_0, f_1 become harder to distinguish (smaller KL). Wald's SPRT is optimal: it minimizes $\mathbb{E}[\tau]$ among tests with the same (α, β).

KL divergence measures the per-observation information separating f_0 and f_1. The average sample number is roughly "boundary height divided by information per step." When f_0 and f_1 are similar (small KL), evidence accumulates slowly and more data are needed.

Remark 5.25. Under H_0, $(\lambda_n)_{n\geq 1}$ is a nonnegative martingale with $\mathbb{E}_0[\lambda_n] = 1$. Optional stopping (under suitable integrability) explains why one can bound error probabilities in terms of the thresholds c_0, c_1 and why the SPRT is efficient.

Reading Materials on Sequential Testing

- Keywords to search: sequential testing, always-valid inference, p-value, e-value, Ville's inequality, nonnegative martingale.
- Papers:
  - Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. The Annals of Mathematical Statistics, 41 (5), 1397–1409.
  - Shafer, G., Shen, A., Vereshchagin, N., & Vovk, V. (2011). Test martingales, Bayes factors and p-values.
  - Ramdas, A., Ruf, J., Larsson, M., & Koolen, W. M. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv preprint arXiv:2009.03167.
  - Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. Operations Research, 70(3), 1806–1821.
  - Ly, A., Boehm, U., Grünwald, P., Ramdas, A., & van Ravenzwaaij, D. (2025). A Tutorial on Safe Anytime-Valid Inference: Practical Maximally Flexible Sampling Designs for Experiments Based on e-Values.
- A lecture note by Aaditya Ramdas, especially L13, L14, L18: `https://www.stat.cmu.edu/~aramdas/martingales18/`
- More specific to adaptive sampling in bandit models:
  - Kaufmann, E., & Koolen, W. M. (2021). Mixture martingales revisited with applications to sequential tests and confidence intervals. Journal of Machine Learning Research, 22(246), 1–44.

Reading Materials *Same level*

- Casella and Berger, Statistical inference, Chapter 8.
- Lehmann and Romano, Testing statistical hypotheses, Chapter 3.1–3.4, 3.7.
- Keener, Theoretical statistics, Chapter 12.1–12.3, 12.6.

Advanced

- Jun Shao, Mathematical statistics, Chapter 6.1, 6.2, 6.4.1–6.4.3.

6 Confidence Set

6.1 Introduction

Point estimation summarizes the data by a single number. Confidence sets keep a measure of *uncertainty* explicit. The goal is frequentist: under repeated sampling, a $(1-\alpha)$ confidence set contains the fixed, unknown parameter with probability at least $1-\alpha$. It does not assign a posterior probability to θ (that is Bayesian).

Point estimation typically starts from a statistic (an estimator) for the unknown parameter. For each realized sample, that statistic returns a single value, so a point estimate by itself provides little information about accuracy. Confidence sets (especially confidence intervals in one dimension) address this by producing a *random set* that is designed to bracket the true parameter with high probability.

Confidence sets Let $\boldsymbol{\theta} \in \Theta$ be a k-vector of unknown parameters of an unknown population $\mathbb{P} \in \mathcal{P}$.

Definition 6.1 (Confidence set)**.** Let $C(\boldsymbol{X}) \subset \Theta$ be a measurable set depending only on the sample $\boldsymbol{X}$. If

$$\inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(\boldsymbol{\theta} \in C(\boldsymbol{X})) \geq 1-\alpha,$$

where $\alpha \in (0,1)$ is fixed, then $C(\boldsymbol{X})$ is called a *confidence set* for $\boldsymbol{\theta}$ with (at least) confidence level $1-\alpha$.

Remark 6.1 (Terminology). Many authors reserve "significance level" for tests and say a confidence set has *confidence level* $1-\alpha$. The definition above uses a *uniform* (worst-case) coverage over $\mathcal{P}$, which is stronger than pointwise coverage at a single $\mathbb{P}$.

A confidence set is a random element that covers the unknown $\boldsymbol{\theta}$ with the stated probability. The quantity $\inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(\boldsymbol{\theta} \in C(\boldsymbol{X}))$ is called the *confidence coefficient* of $C(\boldsymbol{X})$, i.e., the worst-case *coverage probability*. This parallels hypothesis testing: confidence level and confidence coefficient play roles analogous to the test level and size.

Definition 6.2 (Confidence interval)**.** For a real-valued parameter θ, if $C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})]$, then $C(\boldsymbol{X})$ is called a *confidence interval* for θ. If $C(\boldsymbol{X}) = (-\infty, \overline{\theta}(\boldsymbol{X})]$ or $C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \infty)$, then $C(\boldsymbol{X})$ is called an *upper* (or *lower*) confidence bound for θ.

A confidence set (or interval) is also called a set (or interval) estimator of θ.

A confidence interval balances two desiderata: *coverage* (probability to include the true θ) and *precision* (short expected length). Widening the interval trivially improves coverage, but useful procedures aim for high coverage at the *shortest possible* lengths.

Example 6.1 (Confidence interval for a normal mean with known σ^2)**.** Since $\bar{X}$ is sufficient for μ when σ^2 is known, it is natural to consider endpoints of the form $\underline{\theta}(\bar{X})$ and $\overline{\theta}(\bar{X})$. Consider intervals

$$[\bar{X} - c,\ \bar{X} + c],$$

where $c > 0$ is a constant. Because $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, we have

$$\begin{aligned}\mathbb{P}(\mu \in [\bar{X}-c,\ \bar{X}+c]) &= \mathbb{P}\big(|\bar{X}-\mu| \leq c\big) = \mathbb{P}\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \leq \frac{\sqrt{n}\,c}{\sigma}\right) \\ &= 2\Phi\left(\frac{\sqrt{n}\,c}{\sigma}\right) - 1 = 1 - 2\Phi\left(-\frac{\sqrt{n}\,c}{\sigma}\right),\end{aligned}$$

which is independent of μ. In particular, choosing $c = z_{\alpha/2}\sigma/\sqrt{n}$ gives a $(1-\alpha)$ confidence interval.

Two cautions are worth keeping in mind. First, we can make the confidence coefficient arbitrarily close to 1 by letting $c \to \infty$, but the resulting interval may be so wide that it is practically useless. Second, if σ^2 is *unknown* and we still use a fixed-width interval $[\bar{X} - c, \bar{X} + c]$, then the worst-case coverage over σ is 0 (letting $\sigma \to \infty$), so such a procedure is not suitable for uniform inference over (μ, σ^2).

Remark 6.2 (Frequentist interpretation). For fixed μ and σ, the random interval covers μ with the stated probability across repeated samples; after observing data, we do *not* assign a probability to the fixed event $\{\mu \in C(\boldsymbol{X})\}$. This subtlety often causes confusion.

The example above suggests a standard design principle. First, choose a target confidence level $1 - \alpha \in (0, 1)$ and construct a confidence set with that level. Second, among all confidence intervals with the desired coverage, prefer those that are more precise. For a two-sided interval $C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})]$, the (random) length is $\overline{\theta}(\boldsymbol{X}) - \underline{\theta}(\boldsymbol{X})$. For one-sided bounds, the length is infinite, so one typically evaluates precision through the distance from the bound to the true value (e.g., $\overline{\theta}(\boldsymbol{X}) - \theta$ for an upper bound).

Example 6.2 (Normal model with $\boldsymbol{\theta} = (\mu, \sigma^2)$ at a given level $1 - \alpha$)**.** In the normal model, $(\bar{X}, S^2)$ is sufficient for (μ, σ^2), so we focus on confidence sets of the form $C(\bar{X}, S^2)$. Recall the standard pivots

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1), \qquad W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \qquad \bar{X} \perp S^2.$$

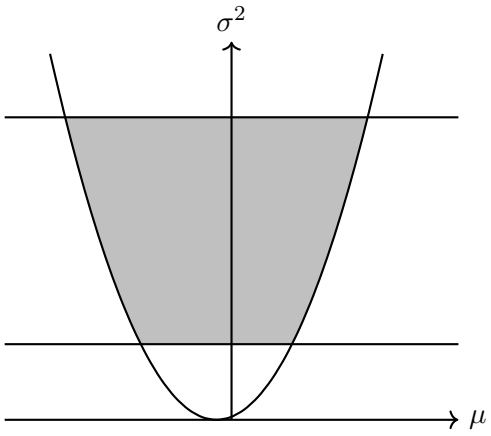
Choose constants $\tilde{c}_\alpha$, $c_{1,\alpha}$, and $c_{2,\alpha}$ such that

$$\mathbb{P}(-\tilde{c}_\alpha \le Z \le \tilde{c}_\alpha) = \sqrt{1-\alpha}, \qquad \mathbb{P}(c_{1,\alpha} \le W \le c_{2,\alpha}) = \sqrt{1-\alpha}.$$

By independence of $\bar{X}$ and S^2 (equivalently, of Z and W),

$$\mathbb{P}\left(-\tilde{c}_\alpha \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha,\ c_{1,\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2,\alpha}\right) = 1 - \alpha, \qquad \forall\, (\mu, \sigma^2).$$

Geometrically, this produces a rectangular $(1 - \alpha)$ confidence set in (μ, σ^2)-space.



Remark 6.3 (Typographical note). In the normal model, the standard facts are

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1), \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \qquad \bar{X} \perp S^2.$$

The construction above leverages independence to build a rectangular $(1 - \alpha)$ set in (μ, σ^2)-space.

Definition 6.3 (Asymptotic significance level)**.** Let $\boldsymbol{\theta} \in \Theta$ be a k-vector of unknown parameters of an unknown population $\mathbb{P} \in \mathcal{P}$, and let $C_n(\boldsymbol{X})$ be a confidence set for $\boldsymbol{\theta}$ constructed from a sample of size n. If

$$\liminf_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\big(\boldsymbol{\theta} \in C_n(\boldsymbol{X})\big) \;\geq\; 1-\alpha,$$

then we call $1-\alpha$ an *asymptotic significance level* (equivalently, asymptotic confidence level) of $C_n(\boldsymbol{X})$. If the limit

$$\lim_{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\big(\boldsymbol{\theta} \in C_n(\boldsymbol{X})\big)$$

exists, it is called the *limiting confidence coefficient* of $C_n(\boldsymbol{X})$.

Asymptotic guarantees justify approximate intervals built from asymptotically pivotal statistics (via CLT or the delta method). They assure that, for large n, coverage is near the target uniformly over the model.

Example 6.3 (Uniform$(0,\theta)$)**.** For a sample $X_1, \ldots, X_n$ from Uniform$(0,\theta)$, $Y = \max_i X_i$ is the MLE. Then

$$\mathbb{P}_\theta\big(\theta \in [aY, bY]\big) = \mathbb{P}_\theta\left(\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right) = \frac{1}{a^n} - \frac{1}{b^n},$$
$$\mathbb{P}_\theta\big(\theta \in [Y+c, Y+d]\big) = \mathbb{P}_\theta\left(1 - \frac{d}{\theta} \leq \frac{Y}{\theta} \leq 1 - \frac{c}{\theta}\right) = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n.$$

What is the confidence coefficient and the limiting confidence coefficient for each case?

Remark 6.4. For $[aY, bY]$ (constants independent of θ), the coverage does not depend on θ, so the confidence coefficient equals $(1/a^n - 1/b^n)$. If $a = a_n$ and $b = b_n$ are chosen to make this equal to $1-\alpha$, then the confidence coefficient is exactly $1-\alpha$ for every n. For $[Y+c, Y+d]$ (fixed $c < d$), the infimum over θ is 0 because, as $\theta \to \infty$, the coverage tends to 0; hence both the confidence coefficient and the limiting coefficient are 0.

6.2 Construction of Confidence Sets

We next develop constructive methods for deriving confidence sets with controlled coverage.

6.2.1 Pivotal Quantity

A common method for constructing confidence sets is based on *pivotal quantities*.

Definition 6.4 (Pivotal quantity)**.** A known (Borel) function Q of $(\boldsymbol{X}, \boldsymbol{\theta})$ is called a *pivotal quantity* if and only if the distribution of $Q(\boldsymbol{X}, \boldsymbol{\theta})$ does not depend on $\mathbb{P} \in \mathcal{P}$.

A pivotal quantity may depend on $\mathbb{P}$ through $\boldsymbol{\theta}$, but its *distribution* is fixed once the model family is fixed. A pivot is usually not a statistic (because it involves unknown parameters), even though its distribution is known. In applications, we typically want a pivot that depends on the data and the *target* parameter, but not on nuisance parameters; this is different from an ancillary statistic.

Pivots are "calibrated rulers": we transform data and the parameter into a quantity with a fixed reference distribution. Inverting tail probabilities of this fixed law yields confidence sets with exact (or approximate) coverage.

Given a pivot $Q(\boldsymbol{X}, \boldsymbol{\theta})$, choose constants a and b so that

$$\mathbb{P}\left(a \leq Q(\boldsymbol{X}, \boldsymbol{\theta}) \leq b\right) \geq 1-\alpha.$$

Then

$$C(\boldsymbol{X}) = \big\{\boldsymbol{\theta} \in \Theta :\; a \leq Q(\boldsymbol{X}, \boldsymbol{\theta}) \leq b\big\}$$

is a level $1-\alpha$ confidence set.

If $Q(\boldsymbol{X},\boldsymbol{\theta})$ has a continuous distribution, one can typically choose a and b so that $\mathbb{P}(a \le Q \le b) = 1-\alpha$, giving confidence coefficient $1-\alpha$. If Q is discrete, exact equal-tailed choices may be impossible; coverage is then conservative (at least $1-\alpha$) unless one allows randomization.

Computing $C(\boldsymbol{X})$ Once $Q(\boldsymbol{X},\boldsymbol{\theta})$ and (a,b) are chosen, we compute $C(\boldsymbol{X})$ by *inverting* the inequality $a \le Q(\boldsymbol{X},\boldsymbol{\theta}) \le b$ for $\boldsymbol{\theta}$ at the observed sample. When θ is real-valued and, for fixed $\boldsymbol{X}$, the map $\theta \mapsto Q(\boldsymbol{X},\theta)$ is monotone, the inversion yields an interval

$$C(\boldsymbol{X}) = [\underline{\theta}(\boldsymbol{X}), \overline{\theta}(\boldsymbol{X})].$$

If monotonicity fails, the set can become a union of several intervals, which is harder to interpret. For multivariate $\boldsymbol{\theta}$, inversion can be algebraically complicated and often requires numerical computation.

Example 6.4 (Location-scale families)**.** Consider a location-scale model where $X_i = \mu + \sigma Z_i$ with a known distribution of Z_i.

(i) μ unknown, σ^2 known (target $\theta = \mu$). Then $\bar{X} - \mu$ (equivalently $(\bar{X}-\mu)/(\sigma/\sqrt{n})$) is pivotal, and

$$C(\boldsymbol{X}) = \{\mu :\ c_1 \le \bar{X} - \mu \le c_2\} = [\bar{X} - c_2,\ \bar{X} - c_1].$$

The choice of (c_1, c_2) is not unique; an often-used choice is the equal-tailed $c_1 = -c_2$.

(ii) σ^2 unknown, μ known (target $\theta = \sigma^2$ or σ). Many functions of the standardized variables $(X_i - \mu)/\sigma$ have distributions independent of σ; for example, S/σ, $(\bar{X}-\mu)/\sigma$, and $\prod_{i=1}^n((X_i-\mu)/\sigma)$ are pivots. For instance,

$$C(\boldsymbol{X}) = \{\sigma :\ c_1 \le S/\sigma \le c_2\} = [S/c_2,\ S/c_1],$$

and similarly,

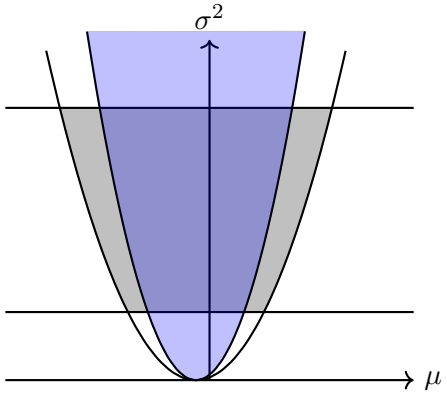
$$C(\boldsymbol{X}) = \{\sigma :\ c_1 \le (\bar{X}-\mu)/\sigma \le c_2\} = [(\bar{X}-\mu)/c_2,\ (\bar{X}-\mu)/c_1].$$

(iii) σ^2 unknown, μ unknown (target $\theta = \sigma^2$). A useful pivot is S/σ; note that quantities such as $(\bar{X}-\mu)/\sigma$ or $\prod_{i=1}^n((X_i-\mu)/\sigma)$ involve the nuisance parameter μ and are therefore not directly useful for a confidence set for σ^2 alone.

(iv) $\boldsymbol{\theta} = (\mu, \sigma^2)$ both unknown. Under normality, $\sqrt{n}(\bar{X}-\mu)/S$ has a t distribution and $\sqrt{n}(\bar{X}-\mu)/\sigma$ is standard normal; these can be used to build joint (typically unbounded) confidence sets such as

$$C(\boldsymbol{X}) = \big\{(\mu,\sigma^2) :\ c_1 \le (\bar{X}-\mu)/\sigma \le c_2\big\}.$$

If the model is normal, the blue region below illustrates an unbounded set compared with a bounded confidence set.



Remark 6.5 (Equal-tailed vs. shortest intervals). Equal-tailed choices ($c_1 = -c_2$) yield symmetric intervals but may be suboptimal in skewed settings. Shortest (highest-density) intervals place endpoints at equal density and include the mode (see the theorem on shortest intervals below).

Pivoting the CDF

Lemma 6.1. *Let $\boldsymbol{T}(\boldsymbol{X}) = (T_1(\boldsymbol{X}), \ldots, T_s(\boldsymbol{X}))$ be independent statistics. Suppose each T_i has a continuous CDF $F_{T_i,\boldsymbol{\theta}}$ indexed by $\boldsymbol{\theta}$. Then*

$$Q(\boldsymbol{T}, \boldsymbol{\theta}) = \prod_{i=1}^{s} F_{T_i,\boldsymbol{\theta}}\big(T_i(\boldsymbol{X})\big)$$

is a pivotal quantity.

Proof. For each i, the probability integral transform gives $U_i = F_{T_i,\boldsymbol{\theta}}(T_i(\boldsymbol{X})) \sim \text{Uniform}(0,1)$. Independence of the T_i implies the U_i are independent, hence $Q = \prod_{i=1}^{s} U_i$ has a distribution that depends only on s (not on $\boldsymbol{\theta}$). □

A naive choice is $T(\boldsymbol{X}) = X_1$ ($s = 1$) or $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{X}$ ($s = n$), but that is usually not ideal; one typically seeks a low-dimensional and informative $\boldsymbol{T}$ (often a sufficient statistic).

Corollary 6.1. *Suppose θ and T in the lemma are real-valued, and let $\alpha_1, \alpha_2 > 0$ satisfy $\alpha_1 + \alpha_2 = \alpha \leq 1/2$. Then the set*

$$C(T) = \{\theta : \ \alpha_1 \leq Q(T, \theta) \leq 1 - \alpha_2\}, \qquad Q(T, \theta) = F_{T,\theta}(T),$$

has probability $1 - \alpha$.

Proof. Because $F_{T,\theta}(T)$ is Uniform$(0,1)$ when $F_{T,\theta}$ is continuous, we have $\mathbb{P}(\alpha_1 \leq F_{T,\theta}(T) \leq 1 - \alpha_2) = 1 - \alpha$. □

There is no guarantee that $C(T)$ is an interval. However, when $\theta \mapsto F_{T,\theta}(t)$ is monotone for each fixed t, we can invert the inequalities to obtain a (possibly one-sided) confidence interval.

Theorem 6.1 (Inverting a monotone CDF)**.** *Assume the setting of Corollary 6.1. Suppose that, for each t, the map $\theta \mapsto F_{T,\theta}(t)$ is monotone.*
(a) If $F_{T,\theta}(t)$ is nonincreasing in θ for each t, define

$$\underline{\theta} = \inf\{\theta : \ F_{T,\theta}(T-) \leq 1 - \alpha_2\}, \qquad \overline{\theta} = \sup\{\theta : \ F_{T,\theta}(T) \geq \alpha_1\}.$$

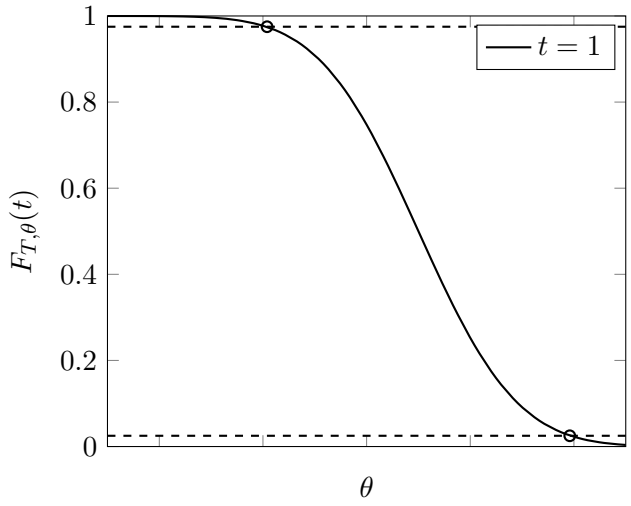
(b) If $F_{T,\theta}(t)$ is nondecreasing in θ for each t, define

$$\underline{\theta} = \inf\{\theta : \ F_{T,\theta}(T) \geq \alpha_1\}, \qquad \overline{\theta} = \sup\{\theta : \ F_{T,\theta}(T-) \leq 1 - \alpha_2\}.$$

Then $[\underline{\theta}, \overline{\theta}]$ is a $(1 - \alpha)$ confidence interval.

Proof. In the special case where $F_{T,\theta}(t)$ is continuous and strictly monotone in θ, the set $\{\theta : \alpha_1 \leq F_{T,\theta}(T) \leq 1 - \alpha_2\}$ is an interval and the endpoints are obtained by inversion of the two inequalities; the coverage is $1 - \alpha$ by Corollary 6.1. The general version (allowing discontinuities and non-strict monotonicity) uses left limits $T-$ and is treated in detail in standard references (e.g., Shao, Theorem 7.1). □

Illustration



Remarks This method works even when the CDF $F_{T,\theta}(t)$ is not continuous, and it does not require strict monotonicity (see Theorem 7.1 of Jun Shao). Moreover, when the parametric family has a monotone likelihood ratio in T, the CDF is monotone in θ, so CDF inversion produces valid $(1-\alpha)$ confidence intervals.

MLR guarantees that larger T favors larger θ (or the reverse). Thus, inverting CDF inequalities produces contiguous intervals rather than disjoint unions.

Example 6.5 (Location exponential)**.** If $X_1, \ldots, X_n$ is a sample from

$$f(x|\mu) = e^{-(x-\mu)}\mathbb{1}_{x\geq\mu},$$

then $Y = \min_i X_i$ is sufficient for μ with density

$$f_Y(y|\mu) = ne^{-n(y-\mu)}\mathbb{1}_{y\geq\mu}.$$

The CDF is

$$F_{Y,\mu}(y) = 1 - e^{-n(y-\mu)}, \qquad y \geq \mu,$$

which is strictly decreasing in μ. For fixed α, solve the equal-tailed equations

$$F_{Y,\mu_U(y)}(y) = \frac{\alpha}{2}, \qquad 1 - F_{Y,\mu_L(y)}(y) = \frac{\alpha}{2}.$$

Equivalently,

$$1 - e^{-n(y-\mu_U(y))} = \frac{\alpha}{2}, \qquad e^{-n(y-\mu_L(y))} = \frac{\alpha}{2},$$

so

$$\mu_U(y) = y + \frac{1}{n}\log(1-\alpha/2), \qquad \mu_L(y) = y + \frac{1}{n}\log(\alpha/2).$$

Thus, a $(1-\alpha)$ confidence interval is

$$\left[Y + \frac{1}{n}\log(\alpha/2),\ Y + \frac{1}{n}\log(1-\alpha/2)\right].$$

Remark 6.6. Because $Y - \mu$ is exponential with rate n, the interval above is exact and has constant length on the natural log scale, reflecting the memoryless property.

6.2.2 Inverting Acceptance Regions of Tests

Another popular method of constructing confidence sets uses the duality between confidence sets and hypothesis tests. The idea is simple: for each point null $H_0 : \theta = \theta_0$, design a level-α test; then collect all θ_0 that are *not rejected*. The resulting set is a $(1-\alpha)$ confidence set. Thus, good tests often generate good confidence intervals.

For a (nonrandomized) test φ, the set $\{\boldsymbol{x} : \varphi(\boldsymbol{x}) = 0\}$ is called the *acceptance region*. (For randomized tests, one should interpret acceptance in terms of the test function; the inversion principle still works but the set language becomes less literal.)

Theorem 6.2 (Invert tests to get confidence sets)**.** *For each $\theta_0 \in \Theta$, let φ_{θ_0} be a test for $H_0 : \theta = \theta_0$ (versus some H_1) with significance level at most α, and let $A(\theta_0) = \{\boldsymbol{x} : \varphi_{\theta_0}(\boldsymbol{x}) = 0\}$ be its acceptance region. For each $\boldsymbol{x}$, define*

$$C(\boldsymbol{x}) = \{\theta_0 : \ \boldsymbol{x} \in A(\theta_0)\}.$$

Then $C(\boldsymbol{X})$ is a level $1-\alpha$ confidence set for θ. If φ_{θ_0} has size α for every θ_0, then $C(\boldsymbol{X})$ has confidence coefficient $1-\alpha$.

Proof. For any $\theta_0 \in \Theta$,

$$\mathbb{P}_{\theta_0}\big(\theta_0 \in C(\boldsymbol{X})\big) = \mathbb{P}_{\theta_0}\big(\boldsymbol{X} \in A(\theta_0)\big) = 1 - \mathbb{E}_{\theta_0}\big[\varphi_{\theta_0}(\boldsymbol{X})\big] \geq 1 - \alpha.$$

Taking the infimum over θ_0 gives the stated confidence level. If $\mathbb{E}_{\theta_0}[\varphi_{\theta_0}] = \alpha$ for every θ_0, then the inequality holds as equality. □

Theorem 6.3 (Invert confidence sets to get tests)**.** *Let $C(\boldsymbol{X})$ be a confidence set for θ with confidence coefficient $1-\alpha$. Fix $\theta_0 \in \Theta$ and define*

$$A(\theta_0) = \{\boldsymbol{x} : \ \theta_0 \in C(\boldsymbol{x})\}.$$

Then the test $\varphi(\boldsymbol{X}) = 1 - \mathbb{1}_{A(\theta_0)}(\boldsymbol{X})$ has significance level at most α for testing $H_0 : \theta = \theta_0$ versus some H_1.

Proof. Under H_0 (i.e., under $\theta = \theta_0$),

$$\mathbb{E}_{\theta_0}[\varphi(\boldsymbol{X})] = \mathbb{P}_{\theta_0}(\theta_0 \notin C(\boldsymbol{X})) \leq \alpha,$$

which is exactly the level requirement. □

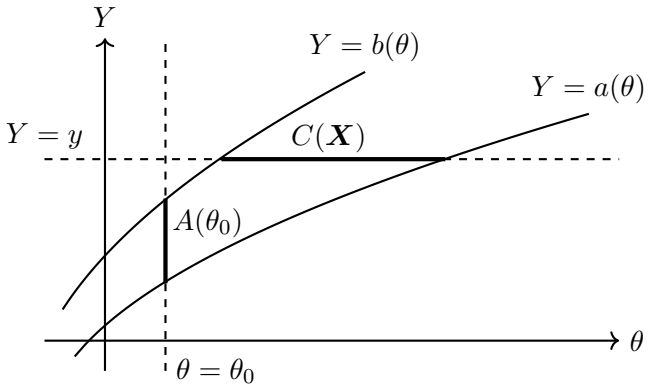
Three quick remarks: it is often easier to construct a level-α test than a confidence set directly; in general there is no guarantee that the inverted set is an interval; and tests with good power properties typically lead to confidence sets with good precision properties.

Example 6.6 (Normal mean with known σ)**.** For a sample $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$ with known σ, consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. A size-α test rejects when $|\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}$. Inverting the acceptance region yields

$$\begin{aligned} &\mathbb{P}\left(\bar{X} - \tfrac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + \tfrac{z_{\alpha/2}\sigma}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha, \qquad \forall\, \mu_0, \\ \Rightarrow\ &\mathbb{P}_\mu\left(\bar{X} - \tfrac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \tfrac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) = 1 - \alpha. \end{aligned}$$

One-sided tests analogously give upper or lower confidence bounds.

Illustration Suppose $A(\theta_0) = \{Y : \ a(\theta_0) \le Y \le b(\theta_0)\}$ for some real-valued statistic $Y(\boldsymbol{X})$ and some nondecreasing functions $a(\theta)$ and $b(\theta)$. Then the inverted confidence set $C(\boldsymbol{X})$ is an interval obtained by slicing the acceptance band at $Y = y$.



Binomial one-sided Let $X_1, \dots, X_n$ be i.i.d. Bernoulli(p) and consider testing $H_0 : p = p_0$ versus $H_1 : p > p_0$. The sufficient statistic is $T = \sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$. The binomial family has a monotone likelihood ratio in T, so a UMP level-α test rejects for large T and has an acceptance region of the form $A(p_0) = \{t : \ t \le m(p_0)\}$, where $m(p_0)$ is chosen so that

$$\sum_{y=m(p_0)+1}^{n} \binom{n}{y} p_0^y (1-p_0)^{n-y} \le \alpha \ < \sum_{y=m(p_0)}^{n} \binom{n}{y} p_0^y (1-p_0)^{n-y}.$$

As p varies, $m(p)$ is an integer-valued, nondecreasing step function. Given an observed t, define

$$\underline{p} = \inf\{p : \ m(p) \ge t\}.$$

Then $(\underline{p}, 1]$ is a level $1 - \alpha$ confidence interval for p. (The confidence coefficient may exceed $1 - \alpha$; randomized intervals remedy this, but we omit that discussion here.)

Remark 6.7. Inverting exact binomial tests yields the classical Clopper–Pearson intervals, which are conservative (coverage $\ge 1 - \alpha$). Score or Wilson-type intervals trade small coverage oscillations for shorter expected length.

Good tests and good confidence intervals The test–CI duality suggests that better tests should give better confidence intervals. Concretely, let $C(\boldsymbol{X})$ be a confidence set constructed from a UMP test φ^* for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$, and let $C'(\boldsymbol{X})$ be a competing confidence set. Define a test $\varphi'(\boldsymbol{X}) = 1$ if $\theta_0 \notin C'(\boldsymbol{X})$ and $\varphi'(\boldsymbol{X}) = 0$ otherwise. Then φ' has level at most α. Since φ^* is UMP, for any $\theta > \theta_0$,

$$\mathbb{P}_\theta\big(\theta_0 \notin C(\boldsymbol{X})\big) = \mathbb{E}_\theta[\varphi^*] \ge \mathbb{E}_\theta[\varphi'] = \mathbb{P}_\theta\big(\theta_0 \notin C'(\boldsymbol{X})\big),$$

which implies that $C(\boldsymbol{X})$ has a smaller chance of covering the incorrect value θ_0 when the true parameter lies to the right:

$$\mathbb{P}_\theta\big(\theta_0 \in C(\boldsymbol{X})\big) \le \mathbb{P}_\theta\big(\theta_0 \in C'(\boldsymbol{X})\big).$$

Better power against $\theta > \theta_0$ translates (via inversion) into a smaller chance that the CI contains the false value θ_0. This is one mechanism linking test optimality to CI precision.

Expected length via Fubini In practice we care about the (expected) length of a confidence interval, and the quantity $\mathbb{P}_\theta(\theta_0 \in C(\boldsymbol{X}))$ may seem indirect. Fubini's theorem makes the connection precise. Let $\theta \in \mathbb{R}$ and let $\lambda(A)$ denote the length of a set A. Then

$$\begin{aligned}
\mathbb{E}_\theta\big[\lambda(C(\boldsymbol{X}) \cap (-\infty, \theta])\big] &= \mathbb{E}_\theta\left[\int_{-\infty}^{\theta} \mathbb{1}_{C(\boldsymbol{X})}(\theta_0)\,\mathrm{d}\theta_0\right] \\
&= \int\int_{-\infty}^{\theta} \mathbb{1}_{C(\boldsymbol{x})}(\theta_0)\,\mathrm{d}\theta_0\,\mathrm{d}\mathbb{P}_\theta(\boldsymbol{x}) \\
&= \int_{-\infty}^{\theta} \mathbb{P}_\theta\big(\theta_0 \in C(\boldsymbol{X})\big)\,\mathrm{d}\theta_0 \\
&\le \int_{-\infty}^{\theta} \mathbb{P}_\theta\big(\theta_0 \in C'(\boldsymbol{X})\big)\,\mathrm{d}\theta_0 = \mathbb{E}_\theta\big[\lambda(C'(\boldsymbol{X}) \cap (-\infty, \theta])\big].
\end{aligned}$$

Remark 6.8. This integral identity formalizes the heuristic: *tests with greater power produce CIs with shorter expected length*, at least on one side of θ. Similar calculations apply to the upper tail.

Shortest confidence interval of the form $[a, b]$

Definition 6.5 (Unimodal)**.** A function f is *unimodal* if there exists x^* such that f is nondecreasing on $(-\infty, x^*]$ and nonincreasing on $[x^*, \infty)$.

Theorem 6.4 (Shortest fixed-mass interval under unimodality)**.** *Let f be a unimodal pdf with a mode x^*. Suppose $a < b$ satisfy*

$$\int_a^b f(x)\,\mathrm{d}x = 1 - \alpha, \qquad f(a) = f(b) > 0, \qquad a \le x^* \le b.$$

Then $[a, b]$ has the shortest length among all intervals I such that $\int_I f(x)\,\mathrm{d}x = 1 - \alpha$.

For symmetric densities such as the standard normal and the t distribution, the shortest fixed-mass interval is the familiar central equal-tailed interval, using $\pm z_{\alpha/2}$ or $\pm t_{n-1,\alpha/2}$.

Among all intervals with fixed mass $1 - \alpha$ under a unimodal density, the shortest is the *highest-density* one: endpoints have equal height and the interval straddles the mode.

Optimizing expected length For a normal population with unknown σ, one may use the pivot $T = (\bar{X} - \mu)/(S/\sqrt{n}) \sim t_{n-1}$. Consider confidence intervals of the form

$$\bar{X} - b\frac{S}{\sqrt{n}} \le \mu \le \bar{X} - a\frac{S}{\sqrt{n}},$$

where $a < b$ are constants chosen so that $\mathbb{P}(a \le T \le b) = 1 - \alpha$. The interval length is $(b - a)S/\sqrt{n}$, so minimizing expected length reduces to minimizing $b - a$ under the coverage constraint. Because the t density is unimodal and symmetric, the minimum is achieved at the symmetric choice $a = -b = t_{n-1,\alpha/2}$.

6.3 Asymptotic Confidence Sets

Using the MLE to construct confidence intervals The MLE is often a good point estimator of a parameter θ. To turn it into a confidence interval, two questions are central: is $\widehat{\theta}_n$ consistent (i.e., $\widehat{\theta}_n \xrightarrow{p} \theta$), and what is the limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$?

Suppose $\sqrt{n}(\widehat{\theta}_n - \theta) \Rightarrow \mathbb{P}$ where the limit law does not depend on θ. Choose c_1, c_2 so that

$$\mathbb{P}\left(c_1 \le \sqrt{n}(\widehat{\theta}_n - \theta) \le c_2\right) \approx 1 - \alpha,$$

and invert the inequalities to obtain an approximate CI for θ.

The MLE often behaves like a noisy version of the true parameter with noise $\approx N(0, \mathcal{I}(\theta)^{-1}/n)$. A $(1-\alpha)$ CI is obtained by backing out the set of θ for which the observed $\widehat{\theta}_n$ is *not too surprising* under that approximation.

Assumption 6.1 (Technical conditions for MLE consistency)**.** Assume:

1. **Strong identifiability.** For every $\epsilon > 0$,

$$\inf_{\tilde{\theta}:|\tilde{\theta}-\theta|\geq\epsilon} \mathrm{KL}(\theta, \tilde{\theta}) > 0, \qquad \mathrm{KL}(\theta, \tilde{\theta}) = \mathbb{E}_\theta \left[\log\left(\frac{f_\theta(X)}{f_{\tilde{\theta}}(X)}\right)\right].$$

2. **Uniform LLN.** Let $R_n(\theta, \tilde{\theta}) = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f_\theta(X_i)}{f_{\tilde{\theta}}(X_i)}\right)$. Then

$$\sup_{\tilde{\theta}} \left| R_n(\theta, \tilde{\theta}) - \mathrm{KL}(\theta, \tilde{\theta}) \right| \xrightarrow{p} 0.$$

Remark 6.9. Strong identifiability keeps the true θ separated (in KL) from alternatives; the uniform LLN lets empirical log-likelihood ratios concentrate around their expectations uniformly in $\tilde{\theta}$. Together they ensure the population KL maximizer at θ is mimicked by the sample likelihood maximizer $\widehat{\theta}_n$.

Theorem 6.5 (Consistency of the MLE)**.** *If Assumption 6.1 holds, then the MLE is consistent.*

Consider the log-likelihood ratio $R_n(\theta, \tilde{\theta})$. Its population version equals $-\mathrm{KL}(\theta, \tilde{\theta})$, which is uniquely maximized at $\tilde{\theta} = \theta$. Uniform convergence transfers that maximizer from population to sample, giving $\widehat{\theta}_n \to \theta$.

Inconsistency of the MLE The MLE can fail to be consistent if the model is not strongly identifiable, or if the uniform LLN fails (e.g., the parameter space is too large).

Example 6.7 (Neyman–Scott incidental parameters)**.** Suppose $\{Y_{i,1}, Y_{i,2} \overset{i.i.d.}{\sim} N(\mu_i, \sigma^2)\}_{i=1}^n$. We want to estimate σ^2.

$$l(\sigma^2, \mu) = -n\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n [(Y_{i,1}-\mu_i)^2 + (Y_{i,2}-\mu_i)^2].$$

The MLE for μ_i is $\widehat{\mu}_i = (Y_{i,1}+Y_{i,2})/2$, and the MLE for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^n [(Y_{i,1}-\widehat{\mu}_i)^2 + (Y_{i,2}-\widehat{\mu}_i)^2] = \frac{1}{4n}\sum_{i=1}^n (Y_{i,1}-Y_{i,2})^2 \xrightarrow{p} \frac{\sigma^2}{2}.$$

Remark 6.10. This is the classic incidental parameters problem: there are n nuisance means μ_i with only two observations each. Information about σ^2 is lost to estimating many nuisance parameters, and the MLE of σ^2 is biased and inconsistent.

Limiting distribution of the MLE

Example 6.8 (Bernoulli(p) MLE)**.** The MLE is $\widehat{p}_n = \bar{X}_n$. By the CLT,

$$\sqrt{n}\frac{\widehat{p}_n - p}{\sqrt{p(1-p)}} \Rightarrow N(0,1).$$

Since the Fisher information for one observation is $\mathcal{I}(p) = 1/[p(1-p)]$,

$$\sqrt{n}(\widehat{p}_n - p) \Rightarrow N\big(0, [\mathcal{I}(p)]^{-1}\big).$$

Asymptotic normality of MLEs typically follows from a Taylor expansion of the score around the true parameter and the CLT:

$$0 = \dot{\ell}(\widehat{\theta}) \approx \dot{\ell}(\theta) + \ddot{\ell}(\theta)(\widehat{\theta} - \theta).$$

Wald intervals for p can behave poorly near the boundary 0 or 1 (variance estimates degenerate). Score-based (Wilson) or exact (Clopper–Pearson) intervals have superior coverage in those regimes.

Limiting distribution of the MLE (counterexample)

Example 6.9 (Uniform$(0, \theta)$)**.** The MLE $\widehat{\theta}_n = X_{(n)}$ satisfies

$$n(\widehat{\theta}_n - \theta) \Rightarrow -\text{Exp}(1/\theta),$$

which is not normal. This is a *non-regular* problem: regularity conditions fail because the support depends on θ, and Fisher information is not defined in the usual way.

Remark 6.11. Non-regular problems require different asymptotics (often extreme-value or boundary limit theory). Likelihood-ratio methods still apply but can have non-χ^2 limits.

Assumption 6.2 (Sufficient conditions for asymptotic normality of the MLE)**.** Assume:

1. The parameter dimension d is fixed (does not grow with n).
2. $f(x|\boldsymbol{\theta})$ is smooth (thrice differentiable) in $\boldsymbol{\theta}$.
3. Conditions in Cramér–Rao hold.
4. The parameter $\boldsymbol{\theta}$ is identifiable.
5. If $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, then $\boldsymbol{\theta}$ lies in the interior of Θ.

Theorem 6.6 (Limiting distribution of the MLE)**.** *Under Assumption 6.2,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \Rightarrow N\left(0, [\mathcal{I}(\boldsymbol{\theta})]^{-1}\right).$$

Remark 6.12. Interior points avoid boundary effects; positive definite $\mathcal{I}(\boldsymbol{\theta})$ ensures local curvature of the log-likelihood. A standard proof uses a mean-value expansion of the score and the continuous mapping theorem. For a rigorous treatment (under more general conditions), see Section 9.3 of Keener, *Theoretical Statistics*.

Asymptotic confidence set Assume (for scalar $\theta \in \mathbb{R}$) that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \Rightarrow N\left(0, [\mathcal{I}(\theta)]^{-1}\right).$$

Then

$$\sqrt{n\mathcal{I}(\theta)}(\widehat{\theta}_n - \theta) \Rightarrow N(0, 1),$$

which is an *approximate pivot*. Hence,

$$\mathbb{P}_\theta\left(\sqrt{n\mathcal{I}(\theta)}|\widehat{\theta}_n - \theta| \leq z_{\alpha/2}\right) \to 1 - \alpha,$$

and a $(1 - \alpha)$ *asymptotic confidence set* is

$$C(\boldsymbol{X}) = \left\{\theta \in \Theta : \ \sqrt{n\mathcal{I}(\theta)}|\widehat{\theta}_n - \theta| \leq z_{\alpha/2}\right\}.$$

This is the population-information version of a Wald set: keep those θ at (information-weighted) distance at most $z_{\alpha/2}/\sqrt{n}$ from $\widehat{\theta}_n$.

Asymptotic confidence interval The set

$$C(\boldsymbol{X}) = \left\{\theta \in \Theta : \ \sqrt{n\mathcal{I}(\theta)}|\widehat{\theta}_n - \theta| \leq z_{\alpha/2}\right\}$$

need not be an interval because $\mathcal{I}(\theta)$ depends on θ. If $\mathcal{I}(\cdot)$ is continuous and bounded away from 0, then by the continuous mapping theorem,

$$\sqrt{\mathcal{I}(\widehat{\theta}_n)/\mathcal{I}(\theta)} \xrightarrow{p} 1.$$

By Slutsky's theorem,

$$\sqrt{n\mathcal{I}(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta) \Rightarrow N(0,1),$$

which leads to the *Wald interval*

$$\left(\widehat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{n\mathcal{I}(\widehat{\theta}_n)}},\ \widehat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{n\mathcal{I}(\widehat{\theta}_n)}}\right).$$

Wald intervals are not invariant to reparameterization and can misbehave near boundaries or under skewness. Likelihood-ratio or score-based intervals are often more robust in finite samples.

Observed information Our asymptotic CI uses the expected Fisher information $\mathcal{I}(\theta) = -\mathbb{E}_\theta[\ell''(\theta|X)]$. Since we may expect $-\ell''(\widehat{\theta}|\boldsymbol{X})/n \xrightarrow{p} \mathcal{I}(\theta)$, another common choice replaces $\mathcal{I}(\widehat{\theta}_n)$ by the *observed Fisher information*:

$$\left(\widehat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{-\ell''(\widehat{\theta}|\boldsymbol{X})}},\ \widehat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{-\ell''(\widehat{\theta}|\boldsymbol{X})}}\right).$$

Remark 6.13. Observed information adapts to the local curvature of the realized likelihood and can improve finite-sample behavior relative to using $\mathcal{I}(\widehat{\theta}_n)$.

Profile confidence interval Can we use the full likelihood shape rather than just curvature at a point? Taylor expand $\ell(\theta|\boldsymbol{X})$ about $\widehat{\theta}_n$ (where $\ell'(\widehat{\theta}_n|\boldsymbol{X}) = 0$):

$$2\ell(\widehat{\theta}_n|\boldsymbol{X}) - 2\ell(\theta|\boldsymbol{X}) \approx \left[\sqrt{-\ell''(\widehat{\theta}_n|\boldsymbol{X})}(\widehat{\theta}_n - \theta)\right]^2.$$

Under regularity, we expect

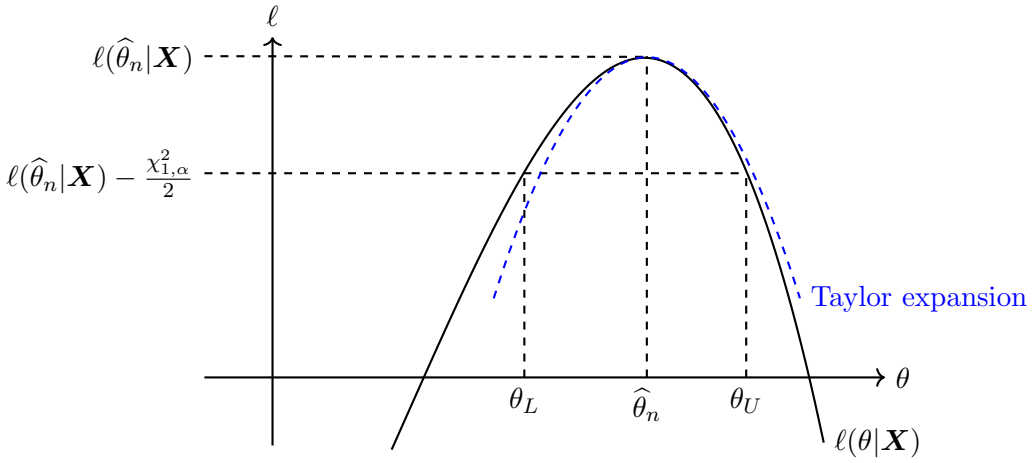
$$2\ell(\widehat{\theta}_n|\boldsymbol{X}) - 2\ell(\theta|\boldsymbol{X}) \Rightarrow Z^2 \sim \chi^2_1.$$

This motivates the *profile likelihood-ratio* confidence set

$$C(\boldsymbol{X}) = \left\{\theta \in \Theta : \ 2\ell(\widehat{\theta}_n|\boldsymbol{X}) - 2\ell(\theta|\boldsymbol{X}) \leq \chi^2_{1,\alpha}\right\}.$$

Remark 6.14. (Wilks phenomenon.) Under regularity, $-2\{\ell(\theta) - \ell(\widehat{\theta})\} \xrightarrow{d} \chi^2_1$ (often even in the presence of nuisance parameters). Profile-LR intervals inherit invariance and typically outperform Wald intervals in finite samples.

Profile confidence interval (illustration)



Poisson population Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(θ). Then

$$\ell(\theta|\boldsymbol{X}) = n\bar{X}\log\theta - n\theta - \log\left(\prod_{i=1}^{n} X_i!\right).$$

The MLE is $\widehat{\theta}_n = \bar{X}$ and the Fisher information (per observation) is $\mathcal{I}(\theta) = 1/\theta$.

Three common large-sample intervals are:

$$C_1(\boldsymbol{X}) = \left\{\theta > 0 : \ \sqrt{n/\theta}\,|\widehat{\theta}_n - \theta| < z_{\alpha/2}\right\} = \left\{\theta > 0 : \ (\widehat{\theta}_n - \theta)^2 < z^2_{\alpha/2}\,\frac{\theta}{n}\right\}$$

$$= \left\{\theta > 0 : \ \theta^2 - 2\widehat{\theta}_n\theta + \widehat{\theta}_n^2 < \frac{z^2_{\alpha/2}}{n}\theta\right\},$$

$$C_2(\boldsymbol{X}) = \left(\bar{X} - \frac{z_{\alpha/2}}{\sqrt{n/\bar{X}}},\ \bar{X} + \frac{z_{\alpha/2}}{\sqrt{n/\bar{X}}}\right),$$

$$C_3(\boldsymbol{X}) = \left\{\theta > 0 : \ 2\ell(\widehat{\theta}_n|\boldsymbol{X}) - 2\ell(\theta|\boldsymbol{X}) \le \chi^2_{1,\alpha}\right\} = \left\{\theta > 0 : \ \theta - \bar{X} - \bar{X}\log(\theta/\bar{X}) < \frac{\chi^2_{1,\alpha}}{2n}\right\}.$$

The endpoints of C_1 and C_3 are always positive; the Wald interval C_2 can have a negative lower endpoint when $\bar{X}$ is close to 0. The LR interval C_3 is computed numerically.

Remark 6.15. Solving the quadratic inequality in C_1 yields roots

$$\theta_\pm = \frac{2\widehat{\theta}_n + z^2_{\alpha/2}/n \ \pm \ \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{4\widehat{\theta}_n + z^2_{\alpha/2}/n}}{2},$$

so $C_1 = (\theta_-, \theta_+)$. In small samples or for $\bar{X}$ near 0, C_3 (profile LR) typically attains better coverage than the Wald interval C_2.

Higher dimension Under regularity, the MLE is asymptotically normal:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \Rightarrow N\left(0, [\mathcal{I}(\boldsymbol{\theta})]^{-1}\right).$$

If we are interested in a scalar functional $g(\boldsymbol{\theta})$, the delta method provides the asymptotic distribution.

Theorem 6.7 (Multivariate delta method)**.** *If $g : \Theta \to \mathbb{R}$ is differentiable at $\boldsymbol{\theta}$, $\mathcal{I}(\boldsymbol{\theta})$ is positive definite, and $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \Rightarrow N(0, [\mathcal{I}(\boldsymbol{\theta})]^{-1})$, then*

$$\sqrt{n}\big(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\big) \Rightarrow N\left(0, \nu^2(\boldsymbol{\theta})\right), \qquad \nu^2(\boldsymbol{\theta}) = (\nabla g(\boldsymbol{\theta}))^\top \mathcal{I}(\boldsymbol{\theta})^{-1} \nabla g(\boldsymbol{\theta}).$$

Ellipsoidal sets and χ^2 If ν_n is a consistent estimator of $\nu(\boldsymbol{\theta})$, then an asymptotic CI for $g(\boldsymbol{\theta})$ is

$$\left(g(\widehat{\boldsymbol{\theta}}_n) - \frac{z_{\alpha/2}\nu_n}{\sqrt{n}},\ g(\widehat{\boldsymbol{\theta}}_n) + \frac{z_{\alpha/2}\nu_n}{\sqrt{n}} \right).$$

For $\boldsymbol{\theta}$ itself, quadratic forms lead to χ^2 confidence sets.

Theorem 6.8. *For $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with positive definite Σ,*

$$(\boldsymbol{Y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_p.$$

Proof. Let $\Sigma = U\Lambda U^\top$ be an eigendecomposition. Then $\boldsymbol{Y} \overset{d}{=} \boldsymbol{\mu} + U\Lambda^{1/2}\boldsymbol{Z}$ with $\boldsymbol{Z} \sim N(\boldsymbol{0}, I)$, so

$$(\boldsymbol{Y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{Z}^\top \Lambda^{1/2} U^\top (U\Lambda U^\top)^{-1} U\Lambda^{1/2}\boldsymbol{Z} = \boldsymbol{Z}^\top \boldsymbol{Z} \sim \chi^2_p. \qquad \square$$

Remark 6.16. In practice, a $(1-\alpha)$ ellipsoidal CI for $\boldsymbol{\theta}$ is

$$\left\{ \vartheta :\ n(\widehat{\boldsymbol{\theta}} - \vartheta)^\top \widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}})\, (\widehat{\boldsymbol{\theta}} - \vartheta) \ \leq\ \chi^2_{p,\,1-\alpha} \right\}.$$

Componentwise intervals can be derived by Bonferroni if needed.

Remarks In constructing asymptotic CIs, several approximations are made: the MLE is treated as asymptotically normal; $(\mathcal{I}(\widehat{\theta}_n))^{-1}$ is used as a proxy for $(\mathcal{I}(\theta))^{-1}$; and, for $g(\boldsymbol{\theta})$, $g(\widehat{\boldsymbol{\theta}}_n)$ is linearized by a Taylor expansion. Under the stated conditions these are valid as $n \to \infty$, but finite-sample accuracy is not guaranteed, so coverage is often checked by simulation.

Poisson coverage (simulation)

Example 6.10 (Poisson(θ))**.** We have $\sqrt{n}(\widehat{\theta} - \theta) \to N(0, \theta)$. Using the Wald interval

$$\left\{ -z_{\alpha/2} \leq \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sqrt{\widehat{\theta}}} \leq z_{\alpha/2} \right\},$$

simulated coverages (desired 90% and 95%) are:

	Desired 90%			Desired 95%		
	$\theta = 0.1$	$\theta = 1$	$\theta = 5$	$\theta = 0.1$	$\theta = 1$	$\theta = 5$
$n = 10$	0.63	0.91	0.90	0.63	0.93	0.95
$n = 30$	0.79	0.89	0.90	0.80	0.93	0.95
$n = 100$	0.91	0.90	0.90	0.93	0.94	0.95

These are based on Monte Carlo: simulate n i.i.d. Poisson samples, compute the CI, check coverage, repeat many times, and report the fraction covered.

Remark 6.17. Under-coverage at small θ reflects skewness and boundary effects. Profile-LR or variance-stabilizing transforms (e.g. Anscombe's $2\sqrt{X + 3/8}$) typically improve coverage.

6.4 Bootstrap

We now move to nonparametric inference, where we treat the data-generating distribution $\mathbb{P}$ itself as unknown (not necessarily indexed by a finite-dimensional θ). In this setting, many targets are naturally expressed as *functionals* of $\mathbb{P}$.

Statistical functionals

Definition 6.6 (Statistical functional)**.** A *statistical functional* is a map ψ that sends a distribution $\mathbb{P}$ to a real number (or vector).

Examples include the mean $\psi(\mathbb{P}) = \int x \, \mathrm{d}F(x)$, variance, median, quantiles, and many other features. In parametric models, a functional is simply $\psi(\theta)$.

Plug-in estimators Given data $\boldsymbol{X}$, define the empirical CDF $F_n(t) = \frac{1}{n}\sum_i \mathbb{1}_{X_i \le t}$. For any function g,

$$\widehat{g} = \int g(x) \, \mathrm{d}F_n(x) = \frac{1}{n}\sum_i g(X_i).$$

Definition 6.7 (Plug-in estimator)**.** Let $\mathbb{P}_n$ be the empirical distribution based on $\boldsymbol{X}$. The plug-in estimator of $\psi(\mathbb{P})$ is

$$\widehat{\psi}_n = \psi(\mathbb{P}_n).$$

Plug-in estimator examples

Example 6.11 (Plug-in estimator examples)**.** The sample mean $\bar{X}$, the sample variance $\frac{1}{n}\sum(X_i - \bar{X})^2$, the sample covariance $\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$, and $\frac{1}{n}\sum_i g(X_i)$ (estimating $\mathbb{E}[g(X)]$) are all plug-in estimators. *Question:* how do we get confidence intervals for functionals based on plug-in estimators?

What do people do in practice? A common workflow is: find an estimator $\widehat{\psi}$ for ψ; estimate its standard deviation $\widehat{\sigma}$ (the standard error); assume $(\widehat{\psi} - \psi)/\widehat{\sigma} \to N(0,1)$ (in a nonparametric setting there is no Fisher information); and then report a $(1-\alpha)$ CI, e.g. for $\alpha = 0.05$,

$$[\widehat{\psi} - 1.96\widehat{\sigma}, \ \widehat{\psi} + 1.96\widehat{\sigma}].$$

A rough rule of thumb is "2σ from the mean $\Rightarrow$ significance."

The normal approximation may be poor for small n, heavy tails, or skewed functionals (e.g. quantiles). Bootstrap offers a data-driven alternative to approximate variability and quantiles of $\widehat{\psi}$.

Monte Carlo: a naive approach $\widehat{\psi}_n$ is a function of $\boldsymbol{X} = (X_1, \ldots, X_n)$; its variance is itself a functional on $\mathbb{R}^n$-valued distributions. A direct Monte Carlo approach would: draw n i.i.d. samples and compute $\widehat{\psi}_n$ to obtain $\widehat{\psi}_{n,1}$; repeat B times to get $\widehat{\psi}_{n,1}, \ldots, \widehat{\psi}_{n,B}$; and estimate the variance by

$$\frac{1}{B}\sum_{i=1}^{B}\big(\widehat{\psi}_{n,i} - \overline{\widehat{\psi}_n}\big)^2.$$

But this requires $n \times B$ fresh draws from $\mathbb{P}$, which is infeasible when only one dataset is available.

Definition 6.8 (Bootstrap sample)**.** Suppose $X_1, \cdots, X_n$ is a sample from CDF F. A *bootstrap sample* is $(\tilde{X}_1, \cdots, \tilde{X}_n)$, where $\tilde{X}_i$ are i.i.d. from the empirical distribution

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i \le x}.$$

Equivalently, draw with replacement uniformly from $\{X_1, \ldots, X_n\}$.

Bootstrap replicates Consider B bootstrap samples:

$$\begin{aligned}(X_1^{(1)}, \cdots, X_n^{(1)}) &\to \widehat{\psi}_n^{(1)},\\ (X_1^{(2)}, \cdots, X_n^{(2)}) &\to \widehat{\psi}_n^{(2)},\\ &\vdots\\ (X_1^{(B)}, \cdots, X_n^{(B)}) &\to \widehat{\psi}_n^{(B)}.\end{aligned}$$

Then $\widehat{\psi}_n^{(1)}, \ldots, \widehat{\psi}_n^{(B)}$ estimate the sampling distribution; the bootstrap SE is their sample SD.

Remark 6.18. The bootstrap approximates $\mathcal{L}(\widehat{\psi}_n|F)$ by $\mathcal{L}(\widehat{\psi}_n|F_n)$. When $F_n \to F$ and $\widehat{\psi}_n$ is smooth in F, this works remarkably well even for complex functionals.

Failure of bootstrap

Example 6.12 (Uniform$(0,\theta)$)**.** Estimate the distribution of $n\big(\theta - X_{(n)}\big)$, known to be Exp$(\theta)$. The nonparametric bootstrap uses $n\big(X_{(n)} - X_{(n)}^B\big)$. Since $X_{(n)}^B$ equals $X_{(n)}$ with probability $1-(1-1/n)^n \approx 0.63$, the bootstrap distribution has a ≈ 0.63 point mass at 0, which can be far from Exp(θ) near 0. In such settings, one often prefers a parametric bootstrap if the model (Uniform) is credible.

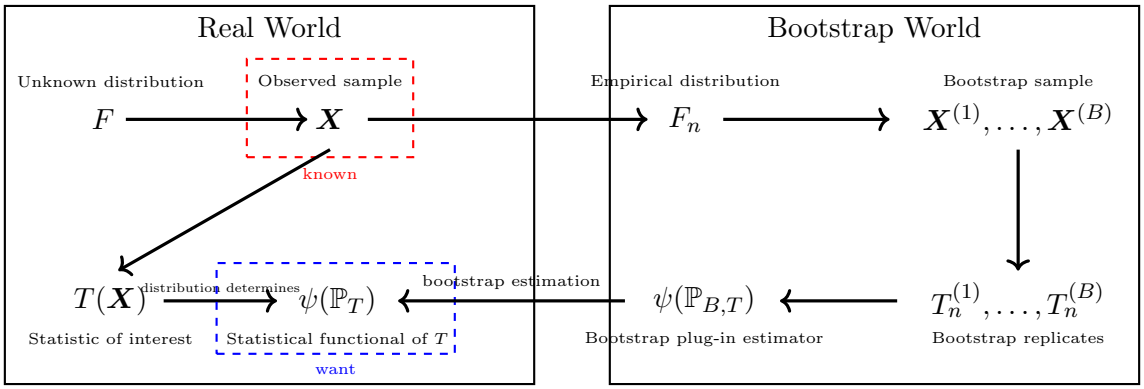
Nonparametric bootstrap struggles with statistics driven by sample extremes (max/min) or by very sparse events. Parametric bootstrap (if the model is trustworthy) or alternative resampling (e.g. m-out-of-n bootstrap) can help.

Parametric bootstrap If $\mathbb{P}_{\boldsymbol{\theta}}$ is known up to $\boldsymbol{\theta}$, the bootstrap can be made parametric: estimate $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}_n$ from the observed sample, and then resample from $\mathbb{P}_{\widehat{\boldsymbol{\theta}}_n}$. This is useful when the parametric family is well justified but one cannot collect new samples from the true distribution.

Definition 6.9 (Parametric bootstrap sample)**.** Given $X_1, \cdots, X_n$, a *parametric bootstrap sample* is $(\tilde{X}_1, \cdots, \tilde{X}_m)$ with $\tilde{X}_i \overset{i.i.d.}{\sim} \mathbb{P}_{\widehat{\boldsymbol{\theta}}_n}$.

The rest of the algorithm is as before. Nonparametric bootstrap is useful when the family is uncertain (model misspecification).

Why does bootstrap work?



Two asymptotic regimes matter: $B \to \infty$ and $n \to \infty$. As $B \to \infty$, Monte Carlo error vanishes and we obtain the exact functional of the empirical distribution. As $n \to \infty$, the bootstrap distribution should converge to the true sampling distribution. The theory is involved; see Efron's paper and follow-ups: B. Efron, *Better Bootstrap Confidence Intervals*, JASA, 1984.

Bootstrap confidence intervals Two basic constructions are widely used. If $\widehat{\theta}_n$ is asymptotically normal, then a $(1-\alpha)$ CI is

$$\widehat{\theta}_n \pm z_{\alpha/2}\widehat{\sigma}_n,$$

where $\widehat{\sigma}_n$ is the bootstrap SE. Alternatively, the *percentile interval* is

$$\left(\widehat{\theta}_{(\alpha/2)},\ \widehat{\theta}_{(1-\alpha/2)}\right),$$

where $\widehat{\theta}_{(\tau)}$ is the τ-quantile of the B bootstrap values of $\widehat{\theta}_n$.

Remark 6.19. Other popular choices include the *basic* bootstrap interval, the *studentized* (bootstrap-t) interval, and Efron's *BCa* interval which corrects for bias and skewness. These often have better coverage than the raw percentile method, especially for skewed statistics.

6.5 Bayesian Intervals

Bayes estimation: introduction In Bayesian inference, the parameter θ itself is treated as random with a prior distribution $\pi(\theta)$. After observing data $\boldsymbol{X}$, we update to the posterior $\pi(\theta|\boldsymbol{X})$ via Bayes' rule. Bayesian point estimation and interval estimation are both functions of the posterior.

The Bayes estimator depends on a loss function, and the posterior mean is optimal for squared-error loss. The Bayesian credible interval (or credible set) is defined by posterior probability content, in contrast to frequentist coverage under repeated sampling. The key conceptual difference is that Bayesian statements are conditional on the observed data, while frequentist coverage is a statement about the procedure across repeated samples.

Bayes estimators If the loss is quadratic, the Bayes estimator is the posterior mean:

$$\widehat{\theta}_{\text{Bayes}} = \mathbb{E}(\theta|\boldsymbol{X}).$$

Other loss functions give different Bayes estimators; for example, the posterior median is optimal under absolute-error loss, and the posterior mode is optimal under 0–1 loss (MAP estimation).

Example 6.13 (A toy example)**.** Suppose $X \sim \text{Bernoulli}(\theta)$ and we place a $\text{Uniform}(0,1)$ prior on θ, i.e., $\theta \sim \text{Beta}(1,1)$. Then the posterior is

$$\theta|X = x \sim \text{Beta}(1+x, 2-x),$$

so for $x = 1$ we have $\theta|X = 1 \sim \text{Beta}(2,1)$ and the posterior mean is

$$\mathbb{E}(\theta|X=1) = \frac{2}{3}.$$

Bayes estimator—binomial Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(p) and let $\sum X_i = y$. With a $\text{Beta}(\alpha,\beta)$ prior on p, the posterior is

$$p|\boldsymbol{X} \sim \text{Beta}(y+\alpha, n-y+\beta),$$

and the posterior mean is

$$\widehat{p} = \frac{y+\alpha}{\alpha+\beta+n} = \frac{n}{\alpha+\beta+n}\cdot\frac{y}{n} + \frac{\alpha+\beta}{\alpha+\beta+n}\cdot\frac{\alpha}{\alpha+\beta}.$$

Thus the posterior mean is a convex combination of the MLE y/n and the prior mean $\alpha/(\alpha+\beta)$, with weights depending on n and the prior "sample size" $\alpha+\beta$.

Bayes estimator—normal Suppose $X_1, \cdots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with known σ^2, and place a normal prior $\mu \sim N(a, b^2)$. Then the posterior is normal with mean

$$\mathbb{E}(\mu|\boldsymbol{X}) = \frac{nb^2}{nb^2+\sigma^2}\bar{X} + \frac{\sigma^2}{nb^2+\sigma^2}a,$$

and variance $\frac{\sigma^2 b^2}{\sigma^2+nb^2}$. Again, the posterior mean shrinks $\bar{X}$ toward the prior mean a.

Conjugate family

Definition 6.10 (Conjugate family)**.** If the posterior distribution is in the same family as the prior, then the prior family is called *conjugate*.

The advantage of conjugacy is analytic tractability: posterior updates reduce to updating a few hyperparameters. The disadvantage is reduced flexibility: conjugate priors may not reflect genuine prior beliefs.

Conjugate family—exponential family Suppose the likelihood belongs to a full exponential family:

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

A conjugate prior takes the form

$$f(\eta|\tau, n_0) \propto \exp\big(\eta^\top \tau - n_0 A(\eta)\big),$$

where τ and n_0 are hyperparameters. If $X_1, \cdots, X_N$ are observed, then the posterior is in the same family with updated parameters:

$$\eta|X_1, \cdots, X_N \sim f(\eta|\tau + \sum_{n=1}^{N} T(X_n),\ n_0 + N).$$

Bayesian intervals A credible set is defined by posterior probability. A $1-\alpha$ credible set $C(\boldsymbol{X})$ satisfies

$$\mathbb{P}(\theta \in C(\boldsymbol{X}) \mid \boldsymbol{X}) = 1 - \alpha,$$

where the probability is under the posterior distribution.

Poisson credible set Let $X_1, \cdots, X_n$ be i.i.d. Poisson(λ) and $Y = \sum_i X_i$. Take a conjugate Gamma prior $\lambda \sim \Gamma(a, b)$ (shape a, scale b). Then the posterior is

$$\lambda|\boldsymbol{X} \sim \Gamma\left(a + y,\ \frac{b}{nb+1}\right).$$

Equivalently,

$$\frac{2(nb+1)}{b}\lambda \ \sim\ \chi^2_{2(a+y)}.$$

Therefore an equal-tailed $(1-\alpha)$ credible interval can be written as

$$[\lambda_L, \lambda_U] = \left[\frac{b}{2(nb+1)}\chi^2_{2(a+y),\,1-\alpha/2},\ \frac{b}{2(nb+1)}\chi^2_{2(a+y),\,\alpha/2}\right],$$

where (in the common "upper-tail" notation) $\chi^2_{\nu,\gamma}$ denotes the value c such that $\mathbb{P}(\chi^2_\nu \geq c) = \gamma$.

Bayesian optimality A natural Bayesian analogue of "shortest CI" is to minimize interval length subject to posterior probability content. One seeks $C(\boldsymbol{X})$ such that

$$\int_{C(\boldsymbol{X})} \pi(\theta|\boldsymbol{X})\,\mathrm{d}\theta = 1-\alpha$$

and $\lambda(C(\boldsymbol{X}))$ is minimized, where $\lambda(\cdot)$ denotes length.

Corollary 6.2. *If the posterior density $\pi(\theta|\boldsymbol{X})$ is unimodal, then the shortest $1-\alpha$ credible interval is the* highest posterior density *(HPD) set:*

$$C(\boldsymbol{X}) = \{\theta:\ \pi(\theta|\boldsymbol{X}) \geq k\},$$

where k is chosen so that $\int_{\pi(\theta|\boldsymbol{X})\geq k} \pi(\theta|\boldsymbol{X})\,\mathrm{d}\theta = 1-\alpha$.

Remark 6.20. HPD sets are invariant under one-to-one transformations only up to reparameterization effects on density; in practice, one should interpret HPD intervals with the chosen parameterization in mind.

Reading materials

- Casella and Berger, *Statistical Inference*, Chapter 9.
- Keener, *Theoretical Statistics*, Chapters 9.1–9.5, 12.4.
- Jun Shao, *Mathematical Statistics*, Sections 2.4.3, 7.1–7.2.2, 7.3.1–7.3.3.

7 Regression Models

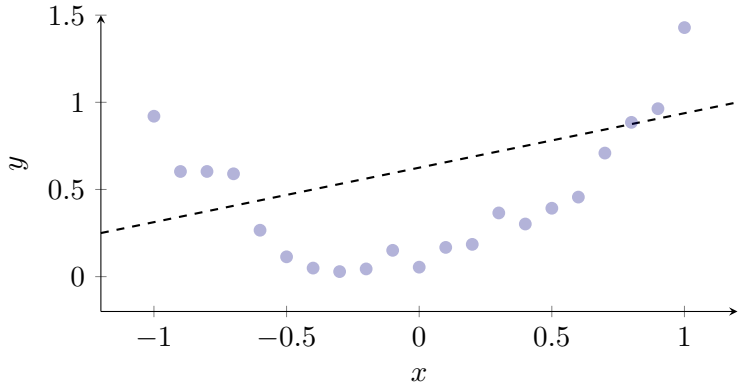
7.1 Multiple Linear Regression

Introduction Simple linear regression studies the linear association between a single *predictor* x and a *response* Y:

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Two complementary tasks follow: *Estimation* (which line best predicts Y from x?) and *Inference* (how uncertain are the estimated slope and intercept?). We shall see that, although the formulas arise from algebra, the geometry is central: ordinary least squares (OLS) projects the data vector $y = (Y_1, \ldots, Y_n)^\top$ onto the span of the predictors.

What if the relationship is not linear? A linear function of x may miss systematic curvature. The panel below illustrates data generated from a quadratic relation; a simple linear fit (dashed) fails to capture the pattern.

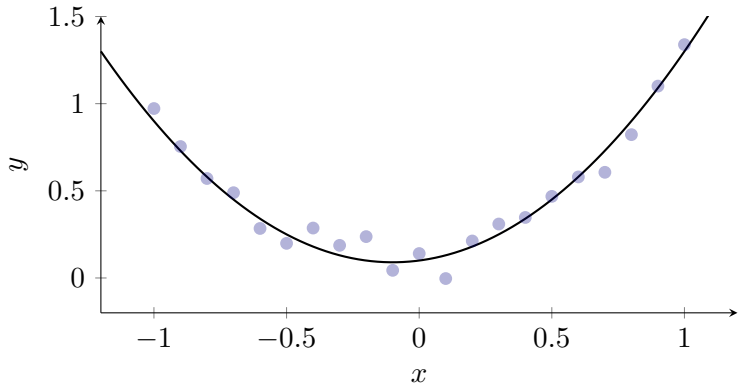


$Y = \alpha + \beta x + \varepsilon$? No!

$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.

Conventional explanation: Fit a parabola to data.

"Nonlinear regression" can still be *linear in the parameters*. Quadratic, interaction, and spline regressions are all *linear models* in this sense; only models nonlinear in *parameters* (e.g., $Y = \alpha x^\beta$) are *nonlinear least squares* unless transformed.



Enlarge the Feature Space A natural remedy is to expand the set of predictors.

augment x with x^2:

$$x_1 = x, \qquad x_2 = x^2,$$

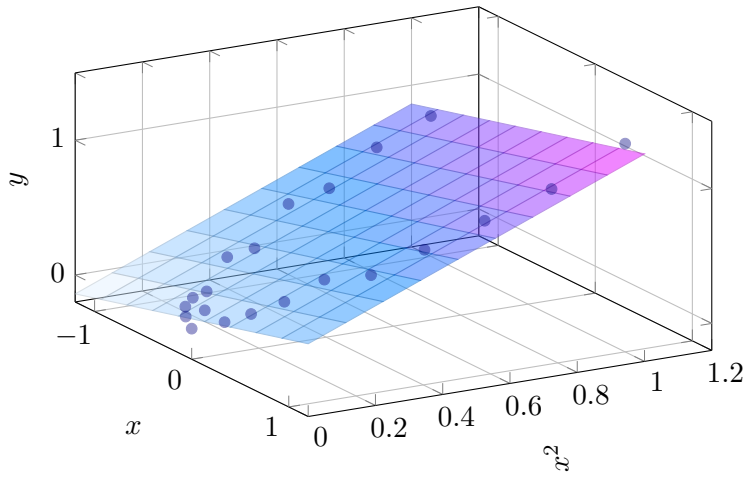
and fit

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

or include interactions such as $x_1 x_2$ in higher dimensions.

These are still *linear models* because they are linear in the unknown parameters β; they remain solvable by OLS using an expanded design matrix (new columns for x^2, $x_1 x_2$, splines, etc.). Bottom line: although Y is quadratic in x, it is linear in the unknown parameters.

Feature engineering turns curvature into linearity in coefficients. Practical tips: center x before adding x^2 to reduce collinearity; prefer orthogonal polynomial bases for numerical stability; beware of wild extrapolation with high-degree polynomials.



More generally, use a non-linear feature map $\boldsymbol{\phi}(\boldsymbol{x})$ and assume

$$Y = \boldsymbol{\beta}^\top \boldsymbol{\phi}(\boldsymbol{x}) + \varepsilon.$$

If $\boldsymbol{\phi}(\boldsymbol{x}) = (1, x, x^2, \ldots, x^d)$ is the vector of polynomial basis functions, this is **polynomial regression**.

Remark 7.1 (Choice of basis and orthogonality). The monomial basis $(1, x, \ldots, x^d)$ yields a Vandermonde design that can be ill-conditioned. A numerically stable alternative is to work with an *orthogonal polynomial* basis. If $\{\phi_j\}_{j=0}^d$ are (approximately) orthonormal, then $X^\top X$ is (approximately) diagonal and the OLS coefficients *decouple*. Orthogonality reduces multicollinearity and improves numerical stability without changing the model class.

Multiple linear regression Multiple linear regression extends the idea of simple linear regression to p predictors. With $x_i = (x_{i1}, \ldots, x_{ip})^\top$ and a column of ones for the intercept.

Definition 7.1 (Multiple linear regression model)**.** For $i = 1, \ldots, n$,

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

In vector form,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} | & | & & | \\ \mathtt{x}_1 & \mathtt{x}_2 & \ldots & \mathtt{x}_p \\ | & | & & | \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y} \in \mathbb{R}^{n \times 1}$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$.

$\boldsymbol{X}$ is the *design matrix*, $\boldsymbol{Y}$ is the response.

The j-th column of $\boldsymbol{X}$, $\mathtt{x}_j = x_{\cdot j} \in \mathbb{R}^n$, is the vector of the j-th predictor.

$\boldsymbol{\beta}$ is the *unknown parameter*; ε_i is the *error* for observation i.

We usually set $x_{i1} = 1$ to include an intercept. **Goal**: find parameter values that fit the data best. Think of $\boldsymbol{X}$ as p arrows in $\mathbb{R}^n$, each being a column vector of the design matrix. Fitting regression means expressing $\boldsymbol{Y}$ as closely as possible as a linear combination of those p arrows. More on this geometric interpretation later.

Examples of the design matrix $\boldsymbol{X}$ Location model: $p = 1$, $\boldsymbol{X} = (1, \cdots, 1)^\top$, $\beta_1 = \mu$:

$$Y = \mu + \varepsilon.$$

Two-sample model: $p = 2$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \dots & 0 \\ 0 & \cdots & 0 & 1 & \dots & 1 \end{bmatrix}^\top, \qquad \boldsymbol{\beta} = (\mu_1, \mu_2)^\top.$$

One-way (simple) ANOVA, compare the mean of k groups: $p = k$, $\boldsymbol{X} = ?$ In ANOVA, $\boldsymbol{X}$ typically uses one-hot (dummy) coding for groups. Include an intercept *or* use k dummies with a sum-to-zero constraint, but not both (avoid the "dummy-variable trap").

Simple linear regression: $Y = \alpha + \beta x + \varepsilon$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \boldsymbol{\beta} = (\alpha, \beta)^\top.$$

Higher-order regression: $Y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}, \qquad \boldsymbol{\beta} = (\alpha, \beta_1, \beta_2)^\top.$$

This is powerful together with the Weierstrass theorem.

Theorem 7.1 (Weierstrass)**.** *For any continuous f on a compact interval $[a, b]$ and any $\varepsilon > 0$, there exists a polynomial p with*

$$\sup_{x \in [a,b]} |f(x) - p(x)| < \varepsilon.$$

Thus, increasing the degree d allows polynomial regression to approximate the regression function on $[a, b]$ arbitrarily well in principle. In finite samples, however, higher d increases variance and can overfit; choose d by model selection or regularization (e.g., cross-validation, ridge), or use localized bases such as splines when appropriate. Alternatively, we will use *kernel methods* later.

Multiplicative error: $Y = \alpha x^\beta e^\varepsilon$. Taking logs,

$$\log(Y) = \log(\alpha) + \beta \log(x) + \varepsilon, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & \log x_1 \\ \vdots & \vdots \\ 1 & \log x_n \end{bmatrix}, \quad \boldsymbol{\beta} = (\log \alpha, \beta)^\top.$$

After a log transform, inference targets the *log-scale mean.* If we are modeling additive error with log-scale mean, predicting the original Y requires bias correction (e.g., multiply by $\exp(\widehat{\sigma}^2/2)$ under normal errors on the log scale).

Remark 7.2 (Linear in the parameters vs. nonlinear regression)*.* The phrase "nonlinear regression" refers to models that are *nonlinear in the parameters*, e.g.

$$Y_i \;=\; \alpha\, x_i^\beta + \varepsilon_i.$$

Such models generally require iterative *nonlinear least squares* unless a transformation yields a linear form with an appropriate error structure (e.g., if $Y_i = \alpha x_i^\beta e^{\varepsilon_i}$ with multiplicative log-normal errors, then $\log Y_i = \log \alpha + \beta \log x_i + \varepsilon_i'$ is linear in the parameters).

7.2 Least Squares Estimator

Method of Least Squares Consider

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

For a candidate $\widehat{\boldsymbol{\beta}}$, fitted values $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.
Residuals $\widehat{\boldsymbol{\varepsilon}}(\widehat{\boldsymbol{\beta}}) \triangleq \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.
Estimate $\boldsymbol{\beta}$ by minimizing the residual sum of squares (RSS):

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

(RSS is also called SSE or SSR.)

Least Squares Estimator From

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2,$$

differentiate $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$ and set to zero:

$$-2\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{X}^\top\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{Y}.$$

Proposition 7.1 (Normal equation)**.**

$$\boldsymbol{X}^\top\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{Y}.$$

The normal equations encode orthogonality: at the minimizer, the residual vector is orthogonal to every column of $\boldsymbol{X}$. No further reduction in RSS is possible by moving in any predictor direction.

Matrix differentiation Let $f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))^\top \in \mathbb{R}^m$ for $\boldsymbol{x} \in \mathbb{R}^n$, and write

$$\frac{\partial f}{\partial \boldsymbol{x}} \triangleq \left[\partial f_i/\partial x_j\right] \in \mathbb{R}^{m\times n} \quad \text{(the Jacobian)}.$$

Theorem 7.2 (Matrix differentiation)**.** *If $f(\boldsymbol{x}) = A\boldsymbol{x}$ with $A \in \mathbb{R}^{m\times n}$, then $\frac{\partial f}{\partial \boldsymbol{x}} = A$. If $g(\boldsymbol{x}) = \boldsymbol{x}^\top B\boldsymbol{x}$ with $B \in \mathbb{R}^{n\times n}$, then*

$$\frac{\partial g}{\partial \boldsymbol{x}} = \left[(B + B^\top)\boldsymbol{x}\right]^\top.$$

Proof. (1) For any direction $\boldsymbol{h} \in \mathbb{R}^n$,

$$Df(\boldsymbol{x})[\boldsymbol{h}] = \lim_{t\to 0} \frac{f(\boldsymbol{x} + t\boldsymbol{h}) - f(\boldsymbol{x})}{t} = \lim_{t\to 0} \frac{A(\boldsymbol{x} + t\boldsymbol{h}) - A\boldsymbol{x}}{t} = A\boldsymbol{h}.$$

Thus the Jacobian (the linear map $\boldsymbol{h} \mapsto Df(\boldsymbol{x})[\boldsymbol{h}]$) is A, independent of $\boldsymbol{x}$.

(2) Let $g(\boldsymbol{x}) = \boldsymbol{x}^\top B\,\boldsymbol{x}$. For any $\boldsymbol{h}$,

$$g(\boldsymbol{x} + t\boldsymbol{h}) = (\boldsymbol{x} + t\boldsymbol{h})^\top B(\boldsymbol{x} + t\boldsymbol{h}) = \boldsymbol{x}^\top B\boldsymbol{x} + t\,\boldsymbol{h}^\top B\boldsymbol{x} + t\,\boldsymbol{x}^\top B\boldsymbol{h} + t^2\,\boldsymbol{h}^\top B\boldsymbol{h}.$$

Hence

$$Dg(\boldsymbol{x})[\boldsymbol{h}] = \frac{\mathrm{d}}{\mathrm{d}t} g(\boldsymbol{x} + t\boldsymbol{h})\bigg|_{t=0} = \boldsymbol{h}^\top B\boldsymbol{x} + \boldsymbol{x}^\top B\boldsymbol{h} = \boldsymbol{h}^\top(B + B^\top)\boldsymbol{x}.$$

By the Riesz representation for directional derivatives, $Dg(\boldsymbol{x})[\boldsymbol{h}] = \boldsymbol{h}^\top\nabla g(\boldsymbol{x})$ for all $\boldsymbol{h}$, so $\nabla g(\boldsymbol{x}) = (B + B^\top)\boldsymbol{x}$ and therefore $\dfrac{\partial g}{\partial \boldsymbol{x}} = [(B + B^\top)\boldsymbol{x}]^\top$. □

Remark 7.3. Only the symmetric part $\frac{1}{2}(B + B^\top)$ affects $g(\boldsymbol{x})$ and its gradient; if B is symmetric, $\nabla g(\boldsymbol{x}) = 2B\boldsymbol{x}$, and if B is skew-symmetric, then $g(\boldsymbol{x}) \equiv 0$ and $\nabla g(\boldsymbol{x}) \equiv \boldsymbol{0}$.

Solving the normal equation Assume $n \geq p$ and $\boldsymbol{X}$ has full column rank so $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible.

Then the least squares estimator (LSE)

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \qquad \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Remark 7.4 (Collinearity). Collinearity arises when one predictor is (nearly) a linear combination of others. Then $\boldsymbol{X}^\top \boldsymbol{X}$ has very small eigenvalues (near singular), making parameter estimates unstable to small perturbations. Remedies include centering/scaling predictors, dropping redundant columns, or adding a ridge penalty $\lambda \|\boldsymbol{\beta}\|_2^2$ (which replaces $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ by $(\boldsymbol{X}^\top \boldsymbol{X} + \lambda I)^{-1}$). More on this later.

Least Squares—Computational considerations For a tall design $X \in \mathbb{R}^{n \times p}$ with $n \geq p$, the normal equations $X^\top X \hat{\beta} = X^\top y$ are typically solved by either *Cholesky on* $X^\top X$ or a *QR factorization of* X.

Cholesky (normal equations). Factor $X^\top X = LL^\top$ with L lower triangular; then solve

$$Lz = X^\top y \quad \text{(forward substitution)}, \qquad L^\top \hat{\beta} = z \quad \text{(back substitution)}.$$

Computation: forming $X^\top X$ costs $O(np^2)$; Cholesky costs $O(\frac{1}{3}p^3)$; triangular solves are $O(p^2)$.

QR (direct least squares). Compute a (thin) Householder QR, $X = QR$, with $Q^\top Q = I$ and $R \in \mathbb{R}^{p \times p}$ upper triangular; then

$$R\,\hat{\beta} = Q^\top y \quad \Rightarrow \quad \hat{\beta} \text{ by back substitution.}$$

Computation: QR costs about $2np^2 - \frac{2}{3}p^3$ flops ($\approx 2np^2$ when $n \gg p$).

Remark 7.5. Forming $X^\top X$ *squares the condition number* (the ratio of the largest to smallest singular value, the larger the more unstable to inversion). QR (and especially SVD) avoids this and is more numerically stable, which matters with nearly collinear predictors or large p. When rank deficiency is suspected, use *column-pivoted QR* or the *SVD* to obtain a reliable (minimum-norm) solution.

Example 7.1 (Simple Linear Regression)**.** Assume $Y_i = \alpha + \beta x_i + \varepsilon_i$ with x not constant so that $S_{xx} > 0$. Write

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \quad S_{xY} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}).$$

Proposition 7.2 (Least Squares Estimators)**.**

$$\widehat{\beta} = \frac{S_{xY}}{S_{xx}}, \qquad \widehat{\alpha} = \bar{Y} - \widehat{\beta}\,\bar{x}, \qquad RSS = \sum_{i=1}^{n} \big(Y_i - \widehat{\alpha} - \widehat{\beta} x_i\big)^2 = \frac{S_{xx} S_{YY} - S_{xY}^2}{S_{xx}}.$$

$\widehat{\beta}$ is the sample covariance of (x, Y) divided by the variance of x; it is the slope that makes residuals uncorrelated with x.

Proof. Let $\mathbf{1} \in \mathbb{R}^n$ be the all-ones vector, $H \triangleq I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ (the centering matrix), and

$$\tilde{\boldsymbol{x}} = H\boldsymbol{x} = \big(x_i - \bar{x}\big)_{i=1}^n, \qquad \tilde{\boldsymbol{y}} = H\boldsymbol{y} = \big(Y_i - \bar{Y}\big)_{i=1}^n.$$

Then

$$S_{xx} = \|\tilde{\boldsymbol{x}}\|^2 = \tilde{\boldsymbol{x}}^\top \tilde{\boldsymbol{x}}, \quad S_{xY} = \tilde{\boldsymbol{x}}^\top \tilde{\boldsymbol{y}}, \quad S_{YY} = \|\tilde{\boldsymbol{y}}\|^2 = \tilde{\boldsymbol{y}}^\top \tilde{\boldsymbol{y}}.$$

OLS minimizes $\|\boldsymbol{y} - \alpha\mathbf{1} - \beta\boldsymbol{x}\|^2$. Using $H\mathbf{1} = \mathbf{0}$,

$$\min_{\alpha,\beta} \|\boldsymbol{y} - \alpha\mathbf{1} - \beta\boldsymbol{x}\|^2 = \min_{\beta} \|H(\boldsymbol{y} - \beta\boldsymbol{x})\|^2 = \min_{\beta} \|\tilde{\boldsymbol{y}} - \beta\tilde{\boldsymbol{x}}\|^2.$$

Because minimizing over α just projects onto the span of $\mathbf{1}$.[4]

The (scalar) normal equation is $\tilde{\boldsymbol{x}}^\top(\tilde{\boldsymbol{y}} - \beta\tilde{\boldsymbol{x}}) = 0$, so

$$\widehat{\beta} = \frac{\tilde{\boldsymbol{x}}^\top\tilde{\boldsymbol{y}}}{\tilde{\boldsymbol{x}}^\top\tilde{\boldsymbol{x}}} = \frac{S_{xY}}{S_{xx}}.$$

From the first normal equation $\mathbf{1}^\top(\boldsymbol{y} - \widehat{\alpha}\mathbf{1} - \widehat{\beta}\boldsymbol{x}) = 0$ we get $\widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{x}$. Finally,

$$RSS = \|\boldsymbol{y} - \widehat{\beta}\boldsymbol{x}\|^2 = \boldsymbol{y}^\top\boldsymbol{y} - 2\widehat{\beta}\boldsymbol{x}^\top\boldsymbol{y} + \widehat{\beta}^2\boldsymbol{x}^\top\boldsymbol{x} = S_{YY} - \frac{S_{xY}^2}{S_{xx}} = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}. \qquad \square$$

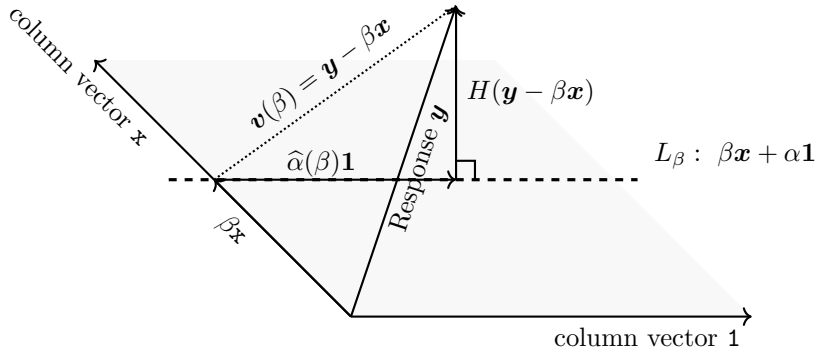


Figure 10: For fixed β, the minimizer over α is the orthogonal projection of $\boldsymbol{y}$ onto $L_\beta = \{\beta\boldsymbol{x} + \alpha\mathbf{1}\}$, yielding $\widehat{\alpha}(\beta)\mathbf{1}$; the remaining vector is $H(\boldsymbol{y} - \beta\boldsymbol{x})$.

Geometric Interpretation Let $\mathtt{x}_j \in \mathbb{R}^n$ denote the j-th column of $\boldsymbol{X}$:

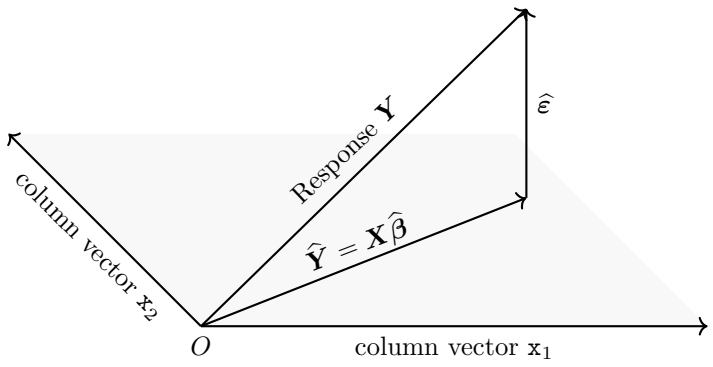
$$\boldsymbol{X} = \begin{bmatrix} | & | & & | \\ \mathtt{x}_1 & \mathtt{x}_2 & \cdots & \mathtt{x}_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} — & \boldsymbol{x}_1 & — \\ & \vdots & \\ — & \boldsymbol{x}_n & — \end{bmatrix} \in \mathbb{R}^{n\times p}.$$

Definition 7.2 (Column space of $\boldsymbol{X}$)**.** A p-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{X}$:

$$\boldsymbol{X}\boldsymbol{\beta} = \beta_1\mathtt{x}_1 + \cdots + \beta_p\mathtt{x}_p, \quad \forall\,\boldsymbol{\beta} \in \mathbb{R}^p.$$

Projection view The fitted value $\widehat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$ lies in the column space.

Normal equations imply $(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top\boldsymbol{X} = 0$: residuals are orthogonal to the column space.



[4]See the geometric interpretation below.

$\widehat{\boldsymbol{Y}} = P\boldsymbol{Y}$ is the **projection** of $\boldsymbol{Y}$ onto the column space, where $P \triangleq \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.
$\widehat{\boldsymbol{\varepsilon}} = Q\boldsymbol{Y}$ where $Q = I - P$; $\widehat{\boldsymbol{Y}}^\top\widehat{\boldsymbol{\varepsilon}} = 0$. **Note that** $PQ = QP = 0$**.**

Remark 7.6. P is symmetric and idempotent ($P^2 = P$). Its diagonal entries h_{ii} are the *leverages*; $\frac{1}{n}\sum_i h_{ii} = \frac{p}{n}$. Large h_{ii} signals high influence potential.

Probabilistic Model Assume
Exogeneity: $\mathbb{E}[\varepsilon_i] = 0$.
Homoscedasticity: $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.
ε_i are independent normal random variables. Likelihood:

$$L(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-n} \exp\left(-\frac{\sum_{i=1}^n (Y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2}{2\sigma^2}\right) = \sigma^{-n}\exp\left(-\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2}\right).$$

MLE of $\boldsymbol{\beta}$ coincides with LSE (independent of σ^2).

Normality turns least squares into maximum likelihood. Even without normality, OLS remains unbiased and variance-optimal among linear estimators (next theorem).

7.2.1 Gauss-Markov Theorem

Best Linear Unbiased Estimator (BLUE) Assume the fixed-design linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\operatorname{rank}(\boldsymbol{X}) = p$, exogeneity $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$, and homoskedastic uncorrelated errors $\operatorname{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Let

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

Theorem 7.3 (Gauss–Markov)**.** *1.* ***Unbiasedness.***

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbb{E}[\boldsymbol{Y}] = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\beta}.$$

2. ***Covariance.***

$$\operatorname{Cov}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\operatorname{Cov}(\boldsymbol{Y})\,\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}.$$

3. ***Best linear unbiased estimator (BLUE).*** *For any given $\boldsymbol{a}$, $\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}$ is a linear[5] unbiased estimator of $\theta = \boldsymbol{a}^\top\boldsymbol{\beta}$. Moreover, among all linear unbiased estimators $\boldsymbol{b}^\top\boldsymbol{Y}$ of θ, its variance is minimal.*

Proof of BLUE. Fix $\boldsymbol{a} \in \mathbb{R}^p$ and consider estimating $\theta = \boldsymbol{a}^\top\boldsymbol{\beta}$. The OLS plug-in $\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}$ is linear and unbiased with

$$\operatorname{Var}(\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}) = \sigma^2\,\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}.$$

Let $\boldsymbol{b}^\top\boldsymbol{Y}$ be any *linear unbiased* estimator of θ; unbiasedness is $\boldsymbol{b}^\top\boldsymbol{X} = \boldsymbol{a}^\top$. Then

$$\operatorname{Var}(\boldsymbol{b}^\top\boldsymbol{Y}) = \sigma^2\,\boldsymbol{b}^\top\boldsymbol{b}, \qquad \operatorname{Var}(\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}) = \sigma^2\,\boldsymbol{a}^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a} = \sigma^2\,\boldsymbol{b}^\top\boldsymbol{P}\,\boldsymbol{b},$$

where the last equality uses $\boldsymbol{a} = \boldsymbol{X}^\top\boldsymbol{b}$. Hence

$$\operatorname{Var}(\boldsymbol{b}^\top\boldsymbol{Y}) - \operatorname{Var}(\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}) = \sigma^2\,\boldsymbol{b}^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{b} = \sigma^2\,\boldsymbol{b}^\top\boldsymbol{Q}\,\boldsymbol{b} = \sigma^2\,\|\boldsymbol{Q}\boldsymbol{b}\|^2 \;\geq\; 0,$$

with equality iff $\boldsymbol{Q}\boldsymbol{b} = \boldsymbol{0}$, i.e., $\boldsymbol{b} \in \operatorname{col}(\boldsymbol{X})$. Solving $\boldsymbol{b} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}$ yields $\boldsymbol{b}^\top\boldsymbol{Y} = \boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}$, so $\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}}$ has minimum variance among all linear unbiased estimators (BLUE). □

The Gauss–Markov theorem predates widespread use of normal models in regression. Aitken later extended it to GLS under correlated/heteroscedastic errors.

[5] An estimator of the form $\widehat{\beta} = A\boldsymbol{Y} + \boldsymbol{\mu}$ for some matrix A and vector $\boldsymbol{\mu}$.

Properties of LSE

1. $\mathbb{E}[\widehat{\boldsymbol{Y}}] = \boldsymbol{X}\boldsymbol{\beta}$.
2. $\mathrm{Cov}[\widehat{\boldsymbol{Y}}] = \sigma^2 P$.
3. $\mathrm{Cov}[\widehat{\boldsymbol{\varepsilon}}] = \sigma^2 Q$.
4. $\mathrm{Cov}[\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{\varepsilon}}] = 0$ since $PQ = 0$.

Proposition 7.3 (Unbiased estimator of σ^2)**.**

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{n-p} = \frac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{n-p}.$$

Dividing by $n-p$ (not n) corrects for the p degrees of freedom used to fit $\widehat{\boldsymbol{Y}}$; $\mathrm{rank}(Q) = n-p$.

Proof. Note that $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = Q\boldsymbol{Y}$. Since $Q\boldsymbol{X} = \boldsymbol{0}$ and $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we have $Q\boldsymbol{Y} = Q\boldsymbol{\varepsilon}$. Therefore

$$\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y}^\top Q^\top Q \boldsymbol{Y} = \boldsymbol{\varepsilon}^\top Q^\top Q \boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top Q \boldsymbol{\varepsilon}.$$

Using the cyclic property $\mathrm{tr}(AB) = \mathrm{tr}(BA)$,

$$\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \mathrm{tr}(\boldsymbol{\varepsilon}^\top Q \boldsymbol{\varepsilon}) = \mathrm{tr}(Q \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top).$$

Taking expectations and using $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2 I_n$,

$$\mathbb{E}\left[\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2\right] = \mathrm{tr}\Big(Q\,\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\Big) = \sigma^2\,\mathrm{tr}(Q).$$

To compute $\mathrm{tr}(Q)$, recall that Q is symmetric and idempotent, so its eigenvalues are either 0 or 1. Hence $\mathrm{tr}(Q)$ equals the number of eigenvalues equal to 1, i.e., $\mathrm{tr}(Q) = \mathrm{rank}(Q)$. Under the full-rank assumption $\mathrm{rank}(\boldsymbol{X}) = p$, we have $\mathrm{rank}(P) = p$, so $\mathrm{rank}(Q) = \mathrm{rank}(I - P) = n - p$. Therefore $\mathrm{tr}(Q) = n - p$ and

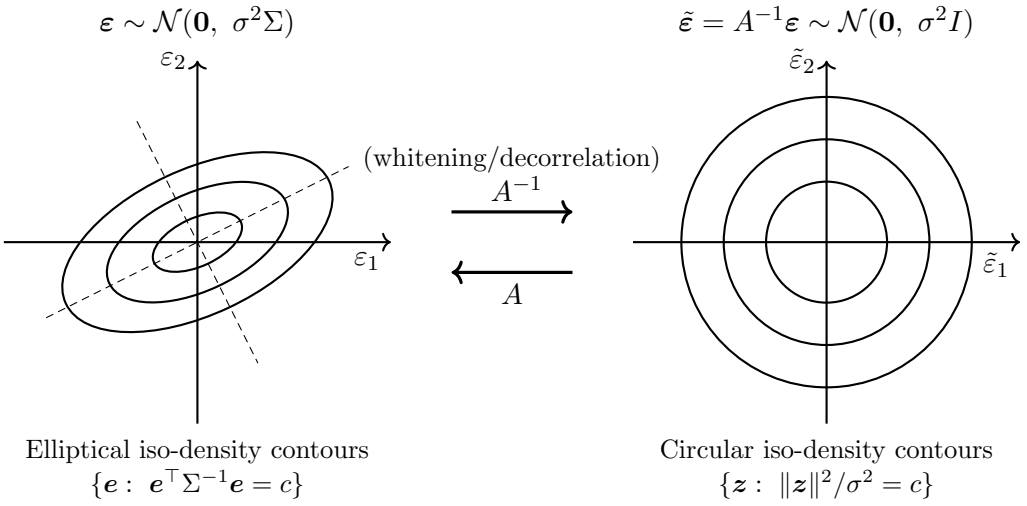
$$\mathbb{E}\left[\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2\right] = (n-p)\sigma^2 \qquad \Longrightarrow \qquad \mathbb{E}[\widehat{\sigma}^2] = \sigma^2. \qquad \square$$

Generalized Least Squares Suppose $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\Sigma)$ with positive definite $\Sigma = AA^\top$ known (Cholesky decomposition). Define $\tilde{\boldsymbol{Y}} = A^{-1}\boldsymbol{Y} = A^{-1}\boldsymbol{X}\boldsymbol{\beta} + A^{-1}\boldsymbol{\varepsilon}$. Then with $\tilde{\boldsymbol{X}} = A^{-1}\boldsymbol{X}$ and $\tilde{\boldsymbol{\varepsilon}} = A^{-1}\boldsymbol{\varepsilon}$, we have $\mathbb{E}[\tilde{\boldsymbol{\varepsilon}}] = \boldsymbol{0}$ and $\mathrm{Cov}[\tilde{\boldsymbol{\varepsilon}}] = \sigma^2 I$.

New independent variables $\tilde{X} = A^{-1}X$, new i.i.d. error $\tilde{\boldsymbol{\varepsilon}} = A^{-1}\boldsymbol{\varepsilon}$.

Check:

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\varepsilon}}] &= \mathbb{E}[A^{-1}\boldsymbol{\varepsilon}] = A^{-1}\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}, \\
\mathrm{Cov}[\tilde{\boldsymbol{\varepsilon}}] &= A^{-1}\,\mathrm{Cov}[\boldsymbol{\varepsilon}](A^{-1})^\top = A^{-1}\sigma^2(AA^\top)(A^{-1})^\top = \sigma^2 I.
\end{aligned}$$



GLS Remarks If $\Sigma = \mathrm{diag}(v_1, \ldots, v_n)$, minimizing $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top\Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ is **weighted least squares**.

In practice Σ is unknown: one may run OLS, estimate $\widehat{\Sigma}$ under structure, then apply GLS (Cochrane–Orcutt).

If Σ is not positive definite (only semi-definite), one can use the Gram-Schmidt process (Schmidt orthonormalization) to find a set of standard normal random variables and the matrix that transform them to the original error terms.

Think about the geometric interpretation of covariance and use the fact that uncorrelated means independence for normal random variables.

Read the wikipedia page for Gram-Schmidt process.

Remark 7.7. As an alternative to modeling Σ, heteroscedasticity-robust (Huber–White) standard errors leave $\widehat{\boldsymbol{\beta}}$ unchanged but adjust its estimated covariance, improving inference under mild model misspecification.

7.3 Statistical Inference

Distributions under normal errors Assume $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$. Recall that

$$P = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I)$.
$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$.
$\widehat{\boldsymbol{Y}} = P\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + P\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 P)$.
$\widehat{\boldsymbol{\varepsilon}} = Q\boldsymbol{Y} = Q\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 Q)$ and is independent of $\widehat{\boldsymbol{Y}}$ because $\mathrm{Cov}(\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{\varepsilon}}) = \sigma^2 PQ = 0$.
Since $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{Y}}$, $\widehat{\boldsymbol{\beta}}$ is independent of $\widehat{\sigma}^2 = \|\widehat{\boldsymbol{\varepsilon}}\|_2^2/(n-p)$. (Cf. sample mean is independent of sample variance for normal data.)

Using a standard lemma (below),

$$(n-p)\widehat{\sigma}^2/\sigma^2 = \frac{\boldsymbol{\varepsilon}^\top Q\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2_{n-p}.$$

Lemma 7.1. *Let A be symmetric idempotent ($A^2 = A$) with $r = \mathrm{tr}(A)$. If $X \sim \mathcal{N}(0, \sigma^2 I)$, then*

$$\frac{X^\top AX}{\sigma^2} \sim \chi^2_r.$$

7.3.1 Hypothesis test and Confidence Interval

Inference for a single coefficient Using independence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$,

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{ii}}} \sim t_{n-p}.$$

Proposition 7.4 (Hypothesis Test for β_i)**.**

H_0	H_1	*Test Stat.*	*Level* α	*p-value*
$\beta_i = 0$	$\beta_i \neq 0$	$\dfrac{\widehat{\beta}_i}{\widehat{\sigma}\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{ii}}}$	*Reject if* $\lvert TS\rvert > t_{n-p}(1-\alpha/2)$	$2\,\mathbb{P}\{T_{n-p} > \lvert TS\rvert\}$

The $1-\alpha$ **confidence interval** for β_i is

$$\left[\widehat{\beta}_i - t_{n-p}(1-\alpha/2)\,\widehat{\sigma}\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{ii}},\ \widehat{\beta}_i + t_{n-p}(1-\alpha/2)\,\widehat{\sigma}\sqrt{(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{ii}}\right].$$

Remark 7.8. For any contrast $c^\top \boldsymbol{\beta}$, replace $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{ii}$ by $c^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} c$. Testing many coefficients inflates false positives; adjust for multiplicity or report CIs.

Example 7.2 (Simple Linear Regression (inference))**.** $Y = \alpha + \beta x + \varepsilon$.

Test for β under $H_0 : \beta = \beta_0$:

$$\sqrt{\frac{S_{xx}}{\widehat{\sigma}^2}}(\widehat{\beta} - \beta_0) \sim t_{n-2}.$$

$(1-\gamma)$ CI:

$$\left(\widehat{\beta} - t_{n-2}(1-\gamma/2)\frac{\widehat{\sigma}}{\sqrt{S_{xx}}},\ \widehat{\beta} + t_{n-2}(1-\gamma/2)\frac{\widehat{\sigma}}{\sqrt{S_{xx}}}\right).$$

Test for α under $H_0 : \alpha = \alpha_0$:

$$\sqrt{\frac{S_{xx} n}{\widehat{\sigma}^2 \sum_i x_i^2}}(\widehat{\alpha} - \alpha_0) \sim t_{n-2}.$$

$(1-\gamma)$ CI:

$$\left(\widehat{\alpha} - \sqrt{\frac{\widehat{\sigma}^2 \sum_i x_i^2}{S_{xx} n}}\, t_{n-2}(1-\gamma/2),\ \widehat{\alpha} + \sqrt{\frac{\widehat{\sigma}^2 \sum_i x_i^2}{S_{xx} n}}\, t_{n-2}(1-\gamma/2)\right).$$

Inference for $\boldsymbol{\beta}$ (global F test) To test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ vs. $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$,

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{p\,\widehat{\sigma}^2} = \frac{\frac{\boldsymbol{\varepsilon}^\top P \boldsymbol{\varepsilon}}{\sigma^2}/p}{\frac{\boldsymbol{\varepsilon}^\top Q \boldsymbol{\varepsilon}}{\sigma^2}/(n-p)} \sim F_{p,n-p} \quad \text{(under } H_0\text{)}.$$

Reject for large values. The $(1-\alpha)$ confidence set is the ellipsoid

$$\left\{\boldsymbol{\beta} :\ (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq p\,\widehat{\sigma}^2\, F_{p,n-p}(1-\alpha)\right\}.$$

When $p = 1$, this F test reduces to the squared t test. Geometrically, it compares the energy of the fitted component ($P\boldsymbol{Y}$) to the residual energy ($Q\boldsymbol{Y}$).

True location of the hyperplane For any $\boldsymbol{x}_0$, let $Y_0 = \boldsymbol{x}_0^\top \boldsymbol{\beta}$ and $\widehat{Y}_0 = \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}$. Then

$$\frac{\widehat{Y}_0 - Y_0}{\widehat{\sigma}\sqrt{\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0}} \sim t_{n-p},$$

yielding a pointwise $(1-\alpha)$ confidence band:

$$\mathbb{P}\Big(\boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}} - w(\boldsymbol{x}_0) \le \boldsymbol{x}_0^\top \boldsymbol{\beta} \le \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}} + w(\boldsymbol{x}_0)\Big) = 1 - \alpha.$$

A **simultaneous** band over all $\boldsymbol{x}$ uses Cauchy–Schwarz:

$$\begin{aligned} |\boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0^\top \boldsymbol{\beta}|^2 &= |\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})|^2 = \big\langle (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0,\ (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \big\rangle^2 \\ &\le (\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0) \Big((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \Big) \end{aligned}$$

and the latter term leads to $F_{p,n-p}$

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{p \widehat{\sigma}^2} \sim F_{p,n-p}$$

giving

$$|\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} - \boldsymbol{x}^\top \boldsymbol{\beta}| \le \sqrt{\boldsymbol{x}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}}\ \sqrt{p\, \widehat{\sigma}^2\, F_{p,n-p}(1-\alpha)}.$$

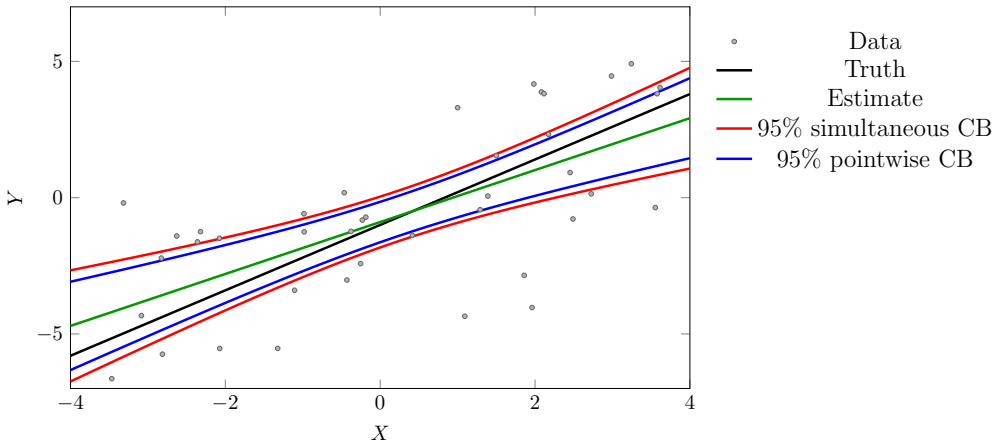
Example 7.3 (Simple Linear Regression (confidence bands))**.**

$$\text{CI for mean at } x_0: \quad \widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2}(1-\gamma/2) \cdot \sqrt{\frac{RSS}{(n-2)} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}.$$

$$\text{Simultaneous band for the line:} \quad \widehat{\alpha} + \widehat{\beta} x_0 \pm \sqrt{2F_{2,n-2}(1-\gamma)} \cdot \sqrt{\frac{RSS}{(n-2)} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}.$$

Note: if $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$.

Remark 7.9. The pointwise interval uses t_{n-2} at a fixed x_0. The *Working–Hotelling* band is simultaneous over all x_0 in the design range, hence wider.



Prediction Interval Confidence bands above are for $\mathbb{E}[Y] = \boldsymbol{x}^\top \boldsymbol{\beta}$. For a *new* response $Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon$ at $\boldsymbol{x}_0$,

$$\frac{\widehat{Y}_0 - Y_0}{\widehat{\sigma}\sqrt{1 + \boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0}} \sim t_{n-p}.$$

For simple linear regression:

$$\text{Mean: } \widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2}(1-\alpha/2) \sqrt{\frac{RSS}{(n-2)} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]},$$

$$\text{Response: } \widehat{\alpha} + \widehat{\beta} x_0 \pm t_{n-2}(1-\alpha/2) \sqrt{\frac{RSS}{(n-2)} \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}.$$

Prediction intervals add the irreducible noise variance on top of parameter uncertainty, hence are wider than intervals for the mean.

7.4 Testing of the Nested Models

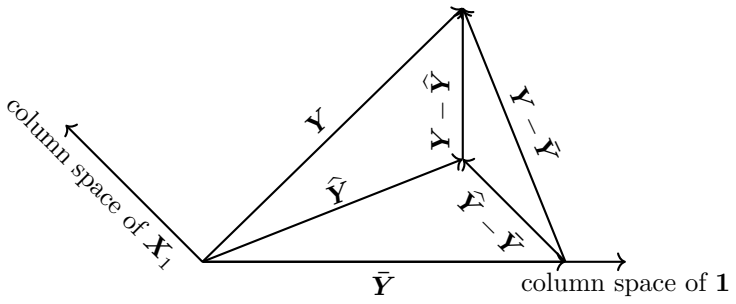
Nested Model Test—F-statistic Let

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix} = (\mathbf{1}, \boldsymbol{X}_1), \qquad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top.$$

Test $H_0 : \beta_2 = \cdots = \beta_p = 0$ vs. H_1 unrestricted. The restricted fit is $\bar{\boldsymbol{Y}}$. Define

$$F = \frac{\|\bar{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}\|_2^2 / (p-1)}{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2 / (n-p)} \sim F_{p-1, n-p}.$$

F compares the *reduction in RSS* achieved by adding predictors (numerator) to the *leftover RSS* (denominator), scaled by degrees of freedom. Equivalently, it tests whether the partial R^2 is significantly larger than zero.



Nested Models H_1: full model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

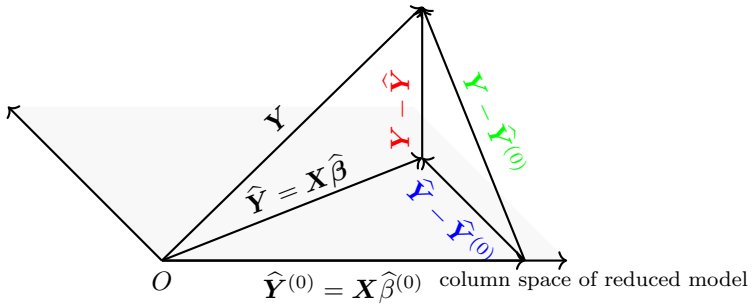
H_0: reduced model $B\boldsymbol{\beta} = \boldsymbol{b}$ (e.g., set some coefficients to zero). A typical case: B is $(p-q) \times p$ selecting $p - q$ coefficients to test against zero.

For example, B is $(p-q) \times p$, $\boldsymbol{b} = \mathbf{0}$. Test if $p - q$ of the coefficients are zero.

$$B = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \end{pmatrix}$$

Can we represent the relationship by a simpler model?

Nested Model Test (geometry) If H_0 holds, the restricted column space is a subspace of the full space.



Define $RSS = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2$ under H_1 and $RSS_0 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^{(0)}\|_2^2$ under H_0.

Recall the properties of LSE: $RSS/(n-p)$ is an unbiased estimator of σ^2, under both hypotheses. $(RSS_0 - RSS)/(p-q)$ is an unbiased estimator of σ^2 under H_0. More importantly, they are orthogonal!

Pythagoras theorem: $\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^{(0)}\|_2^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2 + \|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^{(0)}\|_2^2$. Thus under H_0,

$$\frac{(\,RSS_0 - RSS\,)/(p-q)}{RSS/(n-p)} = \frac{\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^{(0)}\|_2^2/(p-q)}{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2/(n-p)} \sim F_{p-q,n-p},$$

and we reject the reduced model for large values.

ANOVA as a nested test ANOVA is a special case:

$$\boldsymbol{X} = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{pmatrix}^\top, \quad \boldsymbol{\beta} = (\mu_1, \ldots, \mu_k)^\top.$$

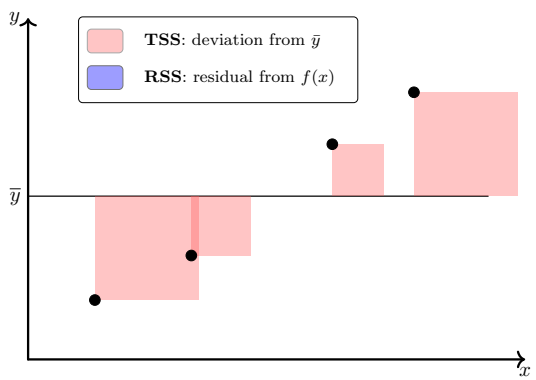
$H_0 : \mu_1 = \cdots = \mu_k$ compares group means.

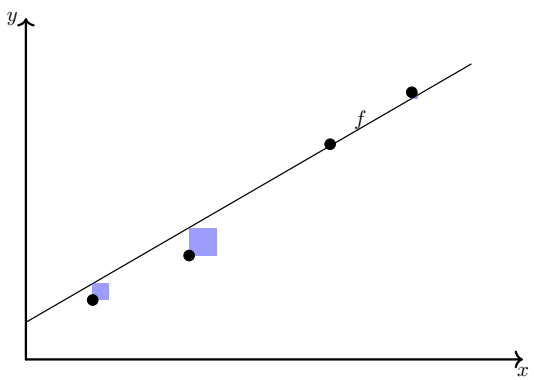
$\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}^{(0)}\|_2^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2 + \|\widehat{\boldsymbol{Y}}^{(0)} - \widehat{\boldsymbol{Y}}\|_2^2 \Rightarrow SST = SSB + SSW.$

Definition 7.3 (Coefficient of Determination)**.**

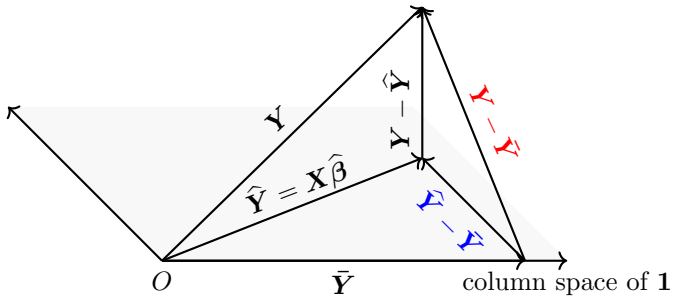
$$R^2 \triangleq \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \quad TSS = \|\boldsymbol{Y} - \bar{\boldsymbol{Y}}\|_2^2.$$

TSS is variability not explained by the mean; RSS is variability not captured by the model. When $R^2 \approx 1$, the fit is nearly perfect; when $R^2 \approx 0$, the model is no better than the sample mean.





Also a F distribution? No!



Remark 7.10. Use *adjusted* $R^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$ to compare models with different p, and prefer out-of-sample measures (cross-validation) for predictive performance.

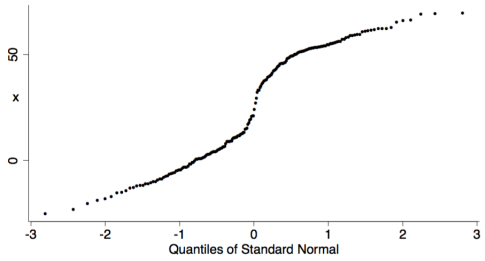
7.5 Box-Cox

We now examine diagnostic checks and response transformations for improving model adequacy.

7.5.1 Diagnostics

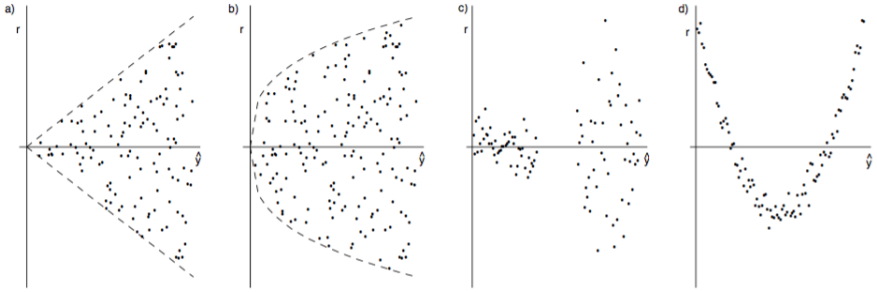
Normality Assumption: ε_i are normal.

Normal probability plot: plot $\Phi^{-1}((i-0.5)/n)$ vs. standardized order statistics. Residuals $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$ should be roughly normal for large n.



Homoscedasticity Assumption: i.i.d. errors.

Plot $\widehat{Y}_i$ vs. $\widehat{\varepsilon}_i$ to diagnose non-constant variance or structure.



Remark 7.11. Other useful displays: scale–location plot (absolute residuals versus fitted), leverage–residual plot, and studentized residuals. Patterns suggest missing nonlinearity or heteroscedasticity; outliers/leverage point to influential cases.

7.5.2 Box-Cox

Transformation Often we transform Y before fitting. For $Y \propto x_1^{\beta_1} \cdots x_p^{\beta_p}$,

$$\log(Y) = \sum_{i=1}^{p} \beta_i \log(x_i) + \varepsilon.$$

Box and Cox (1964) propose a systematic approach.

Remark 7.12. To some extent, Box–Cox transformations can mitigate violations of normality or homoscedasticity.

Box–Cox Transformation Let

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log Y, & \lambda = 0. \end{cases}$$

Definition 7.4 (Box–Cox model)**.**

$$Y^{(\lambda)} = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

Box–Cox requires $Y > 0$. If responses can be nonpositive, shift by a constant c so that $Y + c > 0$ before transforming; report results on the original scale.

MLE for Box–Cox The likelihood is

$$L(\lambda, \boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{\| \boldsymbol{Y}^{(\lambda)} - \boldsymbol{X}\boldsymbol{\beta} \|^2}{2\sigma^2} \right) \cdot \prod_{i=1}^{n} |Y_i|^{\lambda - 1},$$

where we used transformation of variables with Jacobian $J(\lambda, \boldsymbol{Y}) = \prod_{i=1}^n |d\boldsymbol{Y}^{(\lambda)}/d\boldsymbol{Y}| = \prod_{i=1}^n |Y_i|^{\lambda-1}$.

Remark 7.13 (Profile over λ). For fixed λ:

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}^{(\lambda)}, \qquad \widehat{\sigma}^2(\lambda) = \frac{1}{n} \| \boldsymbol{Y}^{(\lambda)} - \widehat{\boldsymbol{Y}}^{(\lambda)} \|^2.$$

Plug in:

$$\log L(\lambda) = (\lambda - 1) \sum_{i=1}^{n} \log |Y_i| - \frac{n}{2} \log \widehat{\sigma}^2(\lambda) - \frac{n}{2},$$

and set $\widehat{\lambda}_{\mathrm{MLE}} = \arg\max_\lambda \log L(\lambda)$.

Remark 7.14. Plotting the profile log-likelihood over λ with a $\pm \frac{1}{2}\chi^2_{1,1-\alpha}$ reference yields a CI for λ. If 0 lies in the CI, log is a defensible choice.

7.6 Spline Regression

Nonlinear regression Recall polynomial regression

$$Y = \boldsymbol{\beta}^\top \boldsymbol{\phi}(\boldsymbol{x}) + \varepsilon, \quad \boldsymbol{\phi}(\boldsymbol{x}) = (1, x, \ldots, x^d).$$

Theorem 7.4 (Weierstrass Theorem)**.** *Any continuous f on $[0, 1]$ can be uniformly approximated by polynomials.*

Spline regression Polynomials may struggle with spatially varying smoothness.

Splines: piecewise polynomials of degree d with $d-1$ continuous derivatives; discontinuities in higher derivatives occur at *knots*.

Example 7.4 (Linear Splines)**.** Linear spline with knots $\tau_1 < \tau_2$:

$$f(x) = \begin{cases} \beta_0 + \beta_1 x, & x \le \tau_1, \\ \beta_0 + \beta_1 x + \beta_2 (x-\tau_1)^+, & \tau_1 < x \le \tau_2, \\ \beta_0 + \beta_1 x + \beta_2 (x-\tau_1)^+ + \beta_3 (x-\tau_2)^+, & x > \tau_2. \end{cases}$$

The basis functions

$$B_0(x) = 1, B_1(x) = x, B_2(x) = (x-\tau_1)^+, B_3(x) = (x-\tau_2)^+.$$

Example 7.5 (Cubic Splines)**.** Cubic spline with knots $\tau_1 < \tau_2$:

$$f(x) = \begin{cases} \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, & x \in (-\infty, \tau_1], \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 ((x-\tau_1)^+)^3, & x \in (\tau_1, \tau_2], \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 ((x-\tau_1)^+)^3 + \beta_5 ((x-\tau_2)^+)^3, & x \in (\tau_2, \infty). \end{cases}$$

The basis functions

$$B_0(x) = 1, B_1(x) = x, B_2(x) = x^2, B_3(x) = x^3, B_4(x) = ((x-\tau_1)^+)^3, B_5(x) = ((x-\tau_2)^+)^3.$$

Definition 7.5 (Spline regression)**.**

$$\boldsymbol{Y} = \boldsymbol{\beta}^\top \boldsymbol{B}(\boldsymbol{x}) + \varepsilon.$$

This is a multiple linear regression in the spline basis. Cubic splines are widely used; later one may study fully nonparametric methods (e.g., kernel ridge regression).

Remark 7.15. Knot placement controls flexibility. Natural splines enforce linear tails outside boundary knots. Penalized splines (P-splines) place many knots and shrink curvature via a roughness penalty, selecting smoothness by cross-validation or information criteria.

7.7 Robust Methods

Influence of observations Cook's distance for observation i:

$$D_i = \frac{(\widehat{\boldsymbol{y}}^{(-i)} - \widehat{\boldsymbol{y}})^\top (\widehat{\boldsymbol{y}}^{(-i)} - \widehat{\boldsymbol{y}})}{p\,\widehat{\sigma}^2} = \frac{(\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{X}^\top \boldsymbol{X})(\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}})}{p\,\widehat{\sigma}^2}.$$

Large D_i indicates influential points (potential outliers). Robust methods mitigate sensitivity to heavy tails.

Influence blends *leverage* (unusual x_i, via h_{ii}) and *residual size*.

P is symmetric and idempotent ($P^2 = P$). Its diagonal entries h_{ii} are the *leverages*; $\frac{1}{n}\sum_i h_{ii} = \frac{p}{n}$. Large h_{ii} signals high influence potential. A point with high leverage can strongly affect $\widehat{\boldsymbol{\beta}}$ even with a modest residual.

Robust Regression (least absolute deviations) OLS solves

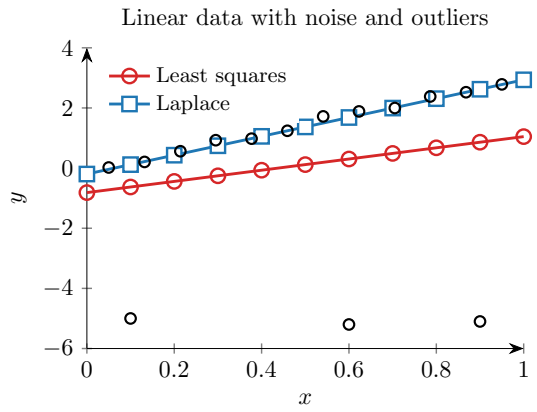
$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2.$$

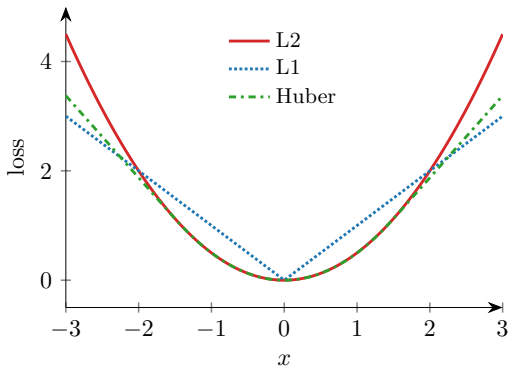
Quadratic loss over-weights outliers. An alternative is

$$\widehat{\boldsymbol{\beta}}_{\text{Robust}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |Y_i - \widehat{Y}_i| = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_1,$$

which corresponds to a Laplace error model:

$$p(Y \mid \boldsymbol{X}, \boldsymbol{\beta}, b) = \text{Laplace}(Y|\boldsymbol{X}, \boldsymbol{\beta}, b) \propto \exp(-|Y - \boldsymbol{x}^\top \boldsymbol{\beta}|/b).$$





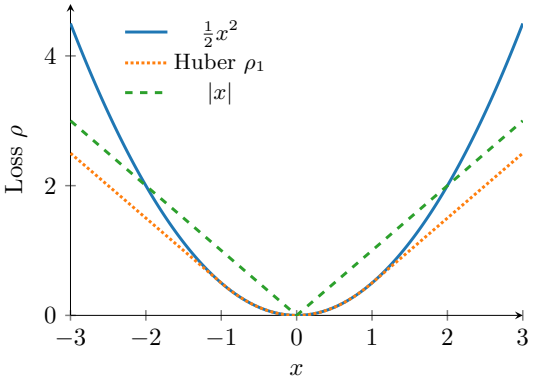
LAD via linear programming Introduce $\varepsilon_i = \varepsilon_i^+ - \varepsilon_i^-$ with $\varepsilon_i^+, \varepsilon_i^- \geq 0$. Then LAD is equivalent to

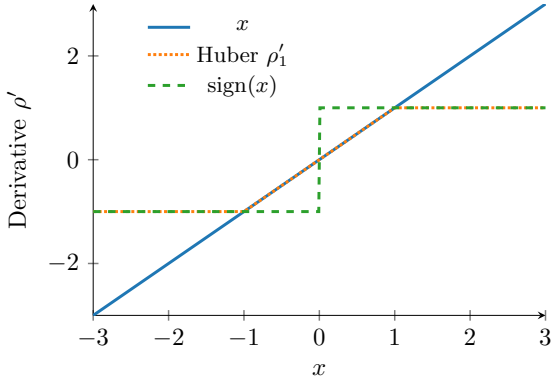
$$\min_{\boldsymbol{\beta}, \boldsymbol{\varepsilon}^+, \boldsymbol{\varepsilon}^-} \sum_{i=1}^{n} (\varepsilon_i^+ + \varepsilon_i^-) \quad \text{s.t.} \quad \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i^+ - \varepsilon_i^- = Y_i, \quad \varepsilon_i^\pm \geq 0,$$

a linear program (dimension $p + 2n$).

Huber Regression A smooth compromise between OLS and LAD: use the Huber loss

$$\rho_\delta(\varepsilon) = \begin{cases} \varepsilon^2/2, & |\varepsilon| \leq \delta, \\ \delta|\varepsilon| - \delta^2/2, & |\varepsilon| > \delta. \end{cases} \tag{7.1}$$





Remark 7.16. Huber M-estimation can be fit by iteratively reweighted least squares (IRLS), downweighting outliers while keeping quadratic efficiency near the center. Choice of δ tunes robustness versus efficiency.

Example 7.6 (Robustness of the sample mean)**.** If $X_i \sim N(\mu, \sigma^2)$, $\bar{X}$ is optimal with $\mathrm{Var}(\bar{X}) = \sigma^2/n$. Under contamination

$$X_i = \begin{cases} N(\mu, \sigma^2), & 1 - \delta, \\ f(x), & \delta, \end{cases}$$

with mean θ and variance τ^2, then

$$\mathrm{Var}(\bar{X}) = (1 - \delta)\frac{\sigma^2}{n} + \frac{\tau^2}{n} + \frac{\delta(1 - \delta)}{n}(\theta - \mu)^2.$$

Heavy tails (e.g., Cauchy) can make $\mathrm{Var}(\bar{X})$ infinite.

Median vs. mean The sample median solves $\arg\min_a \sum_i |x_i - a|$ and is more robust. If θ is the population median and f the density at θ,

$$\sqrt{n}\,(X_{(n/2)} - \theta) \xrightarrow{d} N\left(0,\ \frac{1}{[2f(\theta)]^2}\right).$$

The median is the obtained as the estimator of $\beta_1 = \theta$ in the robust regression for the location model $p = 1$, $X = (1, \cdots, 1)^\top$, $\beta_1 = \theta$. Robust regression is more robust than least square linear regression.

Trade-off: under normality, the median is less efficient than the mean since $1/[2f(\theta)]^2 = \pi\sigma^2/2 > \sigma^2$.

Something in-between (Huber estimator) Minimize $\sum \rho(x_i - a)$ with Huber ρ in (7.1). Tuning δ interpolates between mean-like and median-like behavior; asymptotically normal with variance depending on δ.

References

- Jianqing Fan et al., *Statistical Foundation of Data Science*. `https://orfe.princeton.edu/~jqfan/fan/classes/525/chapters1-3.pdf`
- Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press.
- John Fox, *Applied Regression Analysis, Linear Models, and Related Methods*. Sage.
- Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*. Springer.
- Lecture notes by Profs. Künsch and Meinshausen.

8 Model Selection and Regularization

8.1 Introduction

Model Selection and the Variance-Bias Trade-Off

- In previous topics, we are given a set of data and our goal is to find a model that fits the data well. (LRT, F statistic, R^2, nested model tests, etc.)
- This set of data is usually called the **training set**, and the error of the model is called the **training error**.
- In statistical learning, the more important task is to predict the outcome under a future scenario beyond the training set.
- To assess a model, people use a **test set** that is independent of the training dataset, but that follows the same probability distribution as the training dataset. The error over the test set is called **testing error**.

The central tension in model building is between how well a model explains the data we already have and how reliably it will behave on data we have not yet seen. Good training performance can be achieved by memorizing the quirks of a particular dataset; generalization requires identifying the *stable* structure that will persist for new draws from the same data-generating process. Throughout this chapter we repeatedly translate that qualitative idea into quantitative criteria that trade off fit against complexity.

Model Selection and the Variance-Bias Trade-Off To make this more concrete, consider a regression setting:

Example 8.1 (Regression models)**.**

$$Y = f(\boldsymbol{X}) + \varepsilon,$$

where ε are i.i.d. with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$.

- From the data $(\boldsymbol{Y}, \boldsymbol{X})$, we obtain a regression fit $\widehat{f}$.
- For a future input point $\boldsymbol{x}_0$, the squared-error loss is

  $$\mathrm{Err}(\boldsymbol{x}_0) = \mathbb{E}\big[(Y_0 - \widehat{f}(\boldsymbol{x}_0))^2\big] = \sigma_\varepsilon^2 + \mathrm{Bias}^2(\widehat{f}(\boldsymbol{x}_0)) + \mathrm{Var}(\widehat{f}(\boldsymbol{x}_0)).$$

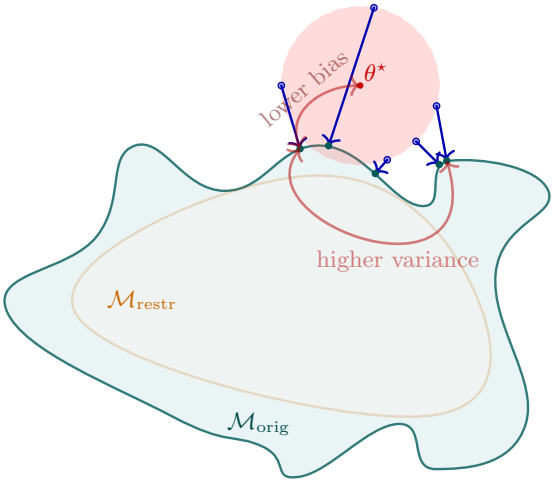
  The bias–variance decomposition!

For the linear regression model, the OLS regression fit is unbiased. We shall see a new *ridge regression* that is biased, but will have smaller variance.

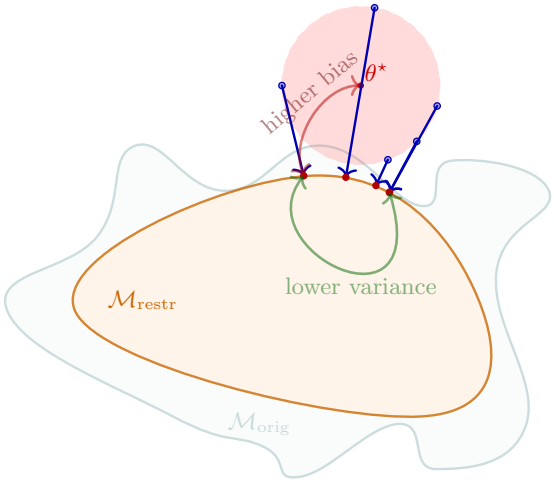
Remarks

The decomposition says: expected prediction error at $\boldsymbol{x}_0$ is the *noise floor* σ_ε^2 plus two model-dependent terms. Increasing flexibility (more parameters or richer function classes) usually *reduces bias* but *increases variance*. Regularization does the opposite. The optimal test error occurs where these two forces balance. In linear models with OLS, $\widehat{f}$ is unbiased, so all improvements must come from variance reduction—this is exactly what ridge seeks.

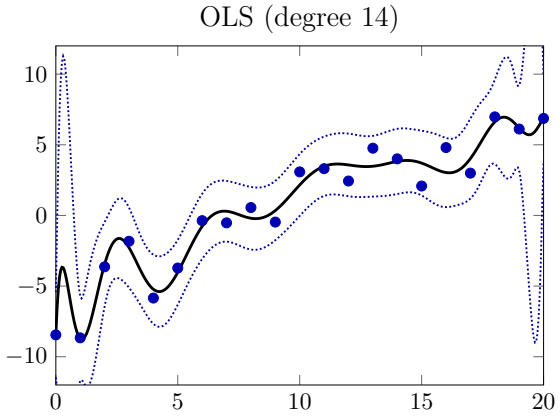
- Typically we would like to choose our model complexity to trade bias off with variance in such a way as to minimize the test error.

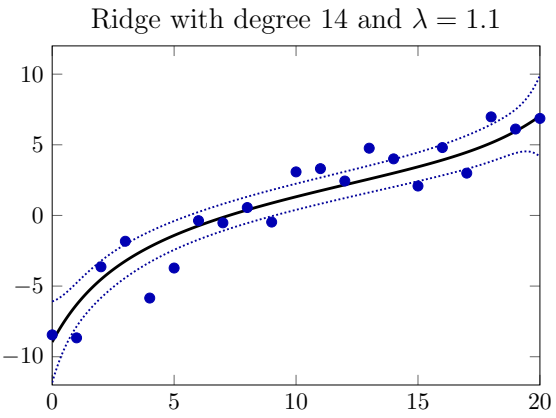


Original model space

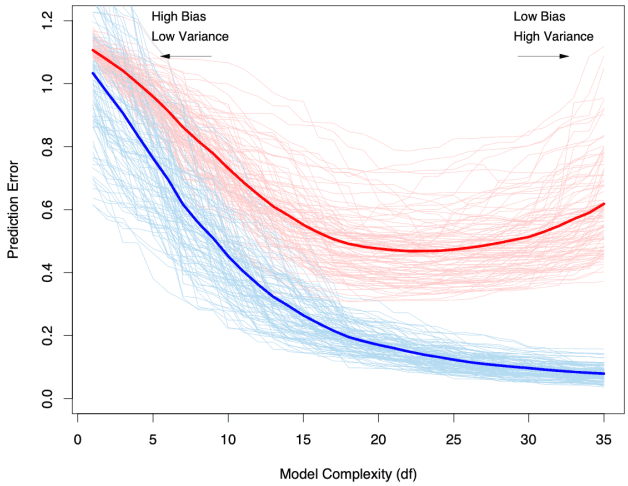


Restricted model space (convex)





- Training error is usually smaller than test error, and it does not properly account for **model complexity** (overfitting).



Choose a model as simple as possible while keeping certain prediction accuracy.

Remarks

Training error keeps decreasing as we make the model more flexible because the model can always fit idiosyncrasies of the sample. Test error follows a U-shape: too simple underfits (high bias), too complex overfits (high variance). The goal is not maximal fit but *sufficient* flexibility to capture signal while suppressing noise.

- Not all existing input features are important for prediction.
- Keeping redundant inputs in a model can lead to poor prediction and poor interpretation.
- We consider two ways of variable/model selection:
  - Subset selection.
  - Shrinkage/regularization methods.

Subset selection tries to *remove* unhelpful variables; shrinkage methods keep all variables but *reduce* their influence, often driving some to zero (lasso). Subset selection is combinatorial; shrinkage trades combinatorics for continuous penalties that are computationally tractable and statistically stabilizing.

8.2 Subset Selection

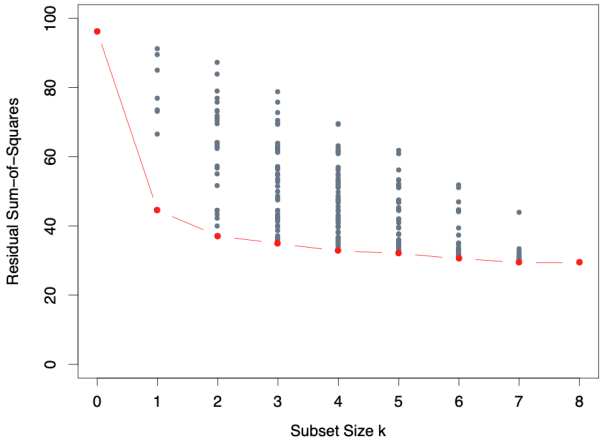
Subset Selection for Linear Models For a regression model with linear predictor $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$,

- Nested model tests can help us determine whether a particular simpler model is good enough.
- We have a lot of independent variables; which variables should we include in the model?

Even when nested tests (e.g., F-tests) say a larger model fits *significantly* better, the associated increase in variance may hurt prediction. Subset selection therefore aims to control complexity directly by limiting the number of active coefficients while keeping predictive performance high.

Best Subset

- A natural idea: for each $k \leq p$, find the subset of $\{x_j\}_{j=1}^p$ with size k that gives the smallest RSS.



How to find the best size-k subset?

- There are in total 2^{p-1} possible models. It can be very time consuming!

- The best subset selection problem is nonconvex and is known to be NP-hard.
- Efficient algorithms make this problem feasible for p as large as 30–40.

How to choose the model complexity?

- Typically we choose the smallest model that minimizes an estimate of the expected prediction error.
- Cross-validation, Mallows' C_p, AIC, BIC. (More later.)

Exhaustive search gives an oracle answer for each k, but the search space doubles with every new variable. In practice we either limit p (pre-screening) or use heuristics (stepwise) and then pick k by a criterion that approximates test error. The point is not to find the *unique* true subset (which may not exist) but a parsimonious model with strong out-of-sample performance.

Stepwise Method Stepwise method: a greedy algorithm that produces nested models.

- **Forward**: starting from $Y = \beta_0 + \varepsilon$. In each step, **add** the (one) variable x_i that most significantly improves the fit:
  - the largest F-statistic, or
  - the smallest deviance, or
  - the smallest LRT statistic.
- **Backward**: starting from the saturated model. In each step, **remove** the (one) variable that is least significant:
  - the smallest F-statistic, or
  - the largest deviance, or
  - the largest LRT statistic.
- Stopping rule:
  - Forward: Stop until all remaining variables are not significant.
  - Backward: Stop until all remaining variables in the model are significant.
  - Alternatively, use cross-validation, Mallows' C_p, AIC, BIC.

Remarks

- Ease of computation.
- The stopping rule doesn't necessarily give us the best model, because it is a greedy method.
- Lower variance but perhaps more bias.
- Forward and backward can be combined, using two significance levels and going back and forth.
- The order of removing or including variables doesn't imply the rank of importance.
- Backward and forward may give very different solutions.

Discussion

Greedy procedures look only one step ahead and may get trapped in local optima. Their selections also depend on the order in which correlated variables enter. Moreover, p-values used during selection do *not* have their usual interpretation afterwards (post-selection inference). These caveats explain why penalization methods, which optimize a *single* global objective, are so popular.

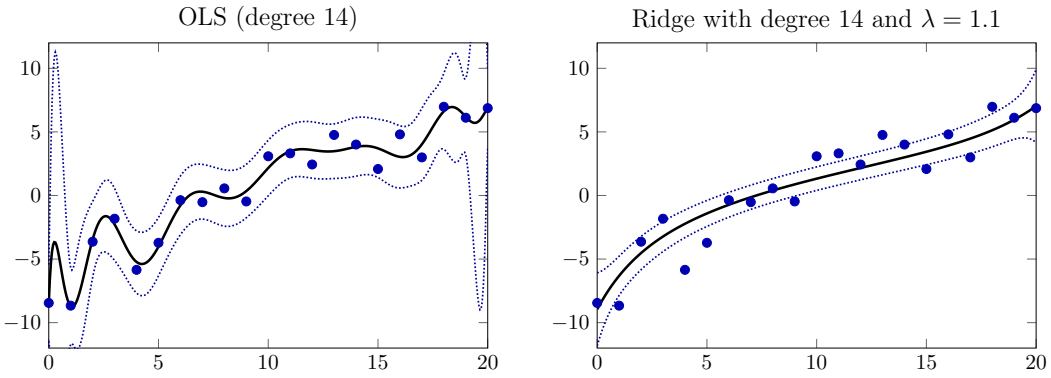
8.3 Shrinkage Methods

Introduction

- MLE picks the parameter values that are the best for fitting the **training data**.
- For this reason, MLE can result in **overfitting**.
- If the data are noisy and we are using lots of parameters, we usually end up with complex functions.

Unregularized least squares minimizes training error with no penalty for complexity. When predictors are numerous or correlated, the fitted coefficients can be large in magnitude and unstable across samples. Shrinkage modifies the objective so that simpler explanations are preferred unless the data *strongly* support complexity.

Example 8.2 (Overfitting)**.**



- Fitting $n = 21$ data points using a degree $p = 14$ polynomial, while the ground truth is a degree 6 polynomial.
- MLE gives the following estimators:

$$2.15, 18.63, -30.56, -323.75, 655.28, 2707.96, -5281.54, -9834.40,$$

$$18116.19, 17246.41, -29883.94, -14378.91, 23535.62, 4571.71, -7113.99.$$

- Many large values in the MLE.
- Small changes in data will result in huge changes in the estimation (high variance).

Shrinkage Methods

- To address overfitting, we usually encourage the parameters to be **small**.
- Methods that are developed to achieve this goal are usually called **shrinkage methods**.

Penalties quantify our preference for small coefficients. An l_2 penalty pulls all coefficients toward zero smoothly (ridge). An l_1 penalty applies a constant force toward zero and can *stick* some coefficients exactly at zero (lasso), thereby performing variable selection and estimation simultaneously.

8.3.1 Ridge Regression

Ridge Regression Ridge regression shrinks the estimators by imposing *penalization for large values*:

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}.$$

- l_2 norm as penalization on $\boldsymbol{\beta}$.
- $\lambda > 0$ is a complexity parameter. The larger the value of λ, the greater the amount of shrinkage.

Intuition

Ridge behaves like adding λ pseudo-observations that encourage coefficients to be small. Geometrically, it replaces the unconstrained least-squares solution with the point where elliptical RSS contours first touch the l_2 ball $\{\|\boldsymbol{\beta}\|_2^2 \leq t\}$. As λ grows, the touchpoint moves toward the origin, reducing variance at the expense of bias.

Ridge Solution The ridge regression is solved by

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \mathbf{I})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

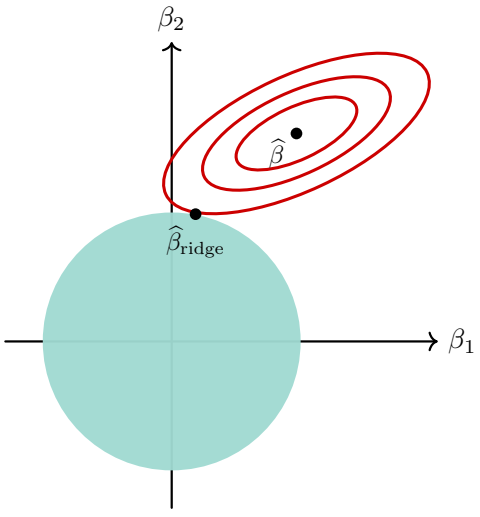
- Compared with OLS, ridge regression adds a positive constant to the diagonal of $\boldsymbol{X}^\top \boldsymbol{X}$ before inversion.
- **Ridge penalty handles collinearity**: It makes the matrix nonsingular, even if $\boldsymbol{X}^\top \boldsymbol{X}$ is not of full rank (multicollinearity).
- This was the main motivation for ridge regression when it was first introduced.

Intuition

Collinearity makes $\boldsymbol{X}^\top \boldsymbol{X}$ ill-conditioned: small data perturbations cause large coefficient swings. The ridge term $\lambda \boldsymbol{I}$ *separates* nearly equal eigenvalues, stabilizing the inversion much like adding friction to a slippery system.

Alternative Derivation: Option 1 Ridge regression can be obtained, using the Lagrange multiplier method, by

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \qquad \text{such that} \qquad \|\boldsymbol{\beta}\|_2^2 \leq t.$$



Proof. Let $f(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ and $g(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$. Both f and g are closed, proper, convex and differentiable. For $t > 0$, the constraint set $\{\boldsymbol{\beta} : g(\boldsymbol{\beta}) \leq t\}$ has nonempty interior, so Slater's condition holds and strong duality applies to

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \;\text{ s.t. }\; g(\boldsymbol{\beta}) \leq t.$$

Its Lagrangian is $\mathcal{L}(\boldsymbol{\beta}, \mu) = f(\boldsymbol{\beta}) + \mu(g(\boldsymbol{\beta}) - t)$, $\mu \geq 0$. KKT conditions at an optimum $(\boldsymbol{\beta}^\star, \mu^\star)$ are

$$\underbrace{\nabla f(\boldsymbol{\beta}^\star) + \mu^\star \nabla g(\boldsymbol{\beta}^\star)}_{=2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}^\star - \boldsymbol{Y}) + 2\mu^\star\boldsymbol{\beta}^\star} = \mathbf{0}, \quad g(\boldsymbol{\beta}^\star) \leq t, \quad \mu^\star \geq 0, \quad \mu^\star(g(\boldsymbol{\beta}^\star) - t) = 0.$$

If the constraint is inactive, $\mu^\star = 0$ and $\boldsymbol{\beta}^\star$ is the OLS solution. If active, then $\mu^\star > 0$ and the stationarity condition is exactly the first-order condition of the *penalized* problem with $\lambda = \mu^\star$:

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \lambda g(\boldsymbol{\beta}) \quad \Longleftrightarrow \quad 2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y}) + 2\lambda\boldsymbol{\beta} = \mathbf{0}.$$

Thus any constrained optimum with active constraint solves the penalized problem for $\lambda = \mu^\star$; conversely, any penalized minimizer $\widehat{\boldsymbol{\beta}}(\lambda)$ solves the constrained problem with $t = \|\widehat{\boldsymbol{\beta}}(\lambda)\|_2^2$ (by complementary slackness). If $\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p \succ 0$ (e.g., $\lambda > 0$), the solution is unique and

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{Y},$$

so the correspondence $\lambda \mapsto t(\lambda) = \|\widehat{\boldsymbol{\beta}}(\lambda)\|_2^2$ is well-defined and nonincreasing, yielding a one-to-one mapping between $\lambda \geq 0$ and the active $t \geq 0$. This establishes the equivalence. □

Alternative Derivation: Option 2 Ridge regression can also be obtained by a Bayesian approach.

- Consider $N(0, \tau^2\boldsymbol{I})$ as the prior distribution for $\boldsymbol{\beta}$.
- Let the likelihood of Y_i given $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ be $N(\boldsymbol{x}_i^\top\boldsymbol{\beta}, \sigma^2)$.
- The connection between λ and τ is $\lambda = \sigma^2/\tau^2$.
- One can check that $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ maximizes the posterior density (MAP estimator). It is also the posterior mean due to the symmetry of the normal distribution.

Intuition

The Bayesian view encodes the belief that *a priori* large coefficients are unlikely. Data that strongly contradict this belief can still pull coefficients away from zero, but weak signals are damped, improving predictive stability.

Proof. Assume the Gaussian linear model $Y \mid \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ and prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau^2\boldsymbol{I}_p)$. The posterior density is

$$p(\boldsymbol{\beta} \mid \boldsymbol{Y}) \;\propto\; p(\boldsymbol{Y} \mid \boldsymbol{\beta})p(\boldsymbol{\beta}) \;\propto\; \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{2\tau^2}\|\boldsymbol{\beta}\|_2^2\right\}.$$

Maximizing $p(\boldsymbol{\beta} \mid \boldsymbol{Y})$ over $\boldsymbol{\beta}$ is equivalent to minimizing the negative log-posterior (dropping constants independent of $\boldsymbol{\beta}$):

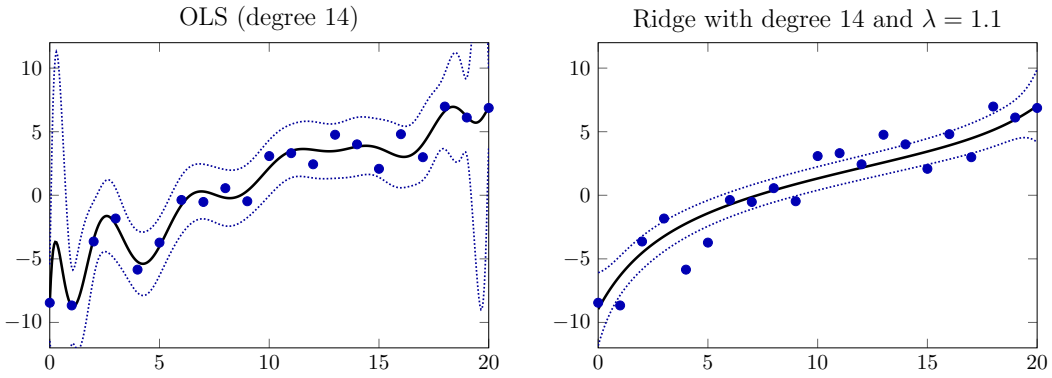
$$\min_{\boldsymbol{\beta}} \; \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\tau^2}\|\boldsymbol{\beta}\|_2^2.$$

Multiplying by $2\sigma^2$ does not change the minimizer, so with $\lambda := \sigma^2/\tau^2$ this is exactly the ridge criterion

$$\min_{\boldsymbol{\beta}} \; \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

The first-order condition gives $2\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Y}) + 2\lambda\boldsymbol{\beta} = \mathbf{0}$, i.e. $(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{Y}$, whose unique solution (for $\lambda > 0$) is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$, the ridge estimator. □

Ridge Regression—Bias–Variance Trade-off



- For $\lambda = 1.1$, the ridge estimates are

$$1.32, 4.36, -1.09, 1.67, -0.47, 0.92,$$
$$-0.26, 0.56, -0.19, 0.32, -0.13, 0.14, -0.09, -0.003, -0.038$$

The above observation is confirmed by the following.

Bias–variance trade-off

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{ridge}}) = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \sigma^2,$$
$$\mathrm{Bias}(\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{ridge}}) = -\lambda (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\beta}.$$

One can show that

- $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{ridge}}) \preccurlyeq (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sigma^2$ for all $\lambda > 0$.
- $\left. \dfrac{d\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{ridge}})}{d\lambda} \right|_{\lambda=0} < 0$, so for sufficiently small λ, ridge always improves OLS.

Intuition

Variance shrinks quickly because directions with small singular values (ill-determined by the data) are heavily penalized. Bias grows linearly in λ but is often tiny along well-determined directions. This is why a small amount of ridge can reduce total MSE even when OLS is unbiased.

Proof. Work in the fixed-design model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$. Let $\boldsymbol{A} := \boldsymbol{X}^\top \boldsymbol{X}$ (assume $\boldsymbol{A} \succ 0$ so OLS exists). The ridge estimator is

$$\widehat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{A} \boldsymbol{\beta} + (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}.$$

Hence

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda] = (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{A} \boldsymbol{\beta} = \Big(\boldsymbol{I} - \lambda (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \Big) \boldsymbol{\beta},$$

so

$$\mathrm{Bias}(\widehat{\boldsymbol{\beta}}_\lambda) = \mathbb{E}[\widehat{\boldsymbol{\beta}}_\lambda] - \boldsymbol{\beta} = -\lambda (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\beta}.$$

Also, using $\mathrm{Var}(\boldsymbol{X}^\top \boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{A}$,

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_\lambda) = (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \sigma^2 \boldsymbol{A} (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} = \sigma^2 (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{A} (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}.$$

1. **Variance domination.** Write

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_\lambda)/\sigma^2 = \boldsymbol{A}^{-1/2} \underbrace{\left(\boldsymbol{A}^{1/2}(\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}^{1/2}\right)^2}_{=:\,\boldsymbol{B}^2} \boldsymbol{A}^{-1/2}.$$

Since $\boldsymbol{A}+\lambda\boldsymbol{I} \succ \boldsymbol{A}$ for $\lambda > 0$, the Loewner order and inversion give $(\boldsymbol{A}+\lambda\boldsymbol{I})^{-1} \prec \boldsymbol{A}^{-1}$, hence $\boldsymbol{B} = \boldsymbol{A}^{1/2}(\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}^{1/2} \prec \boldsymbol{I}$ with $\boldsymbol{B} \succeq \boldsymbol{0}$. Thus $\boldsymbol{B}^2 \preccurlyeq \boldsymbol{I}$ and

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_\lambda) \;\preccurlyeq\; \sigma^2\boldsymbol{A}^{-1} = \mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}), \qquad \forall\, \lambda > 0.$$

2. **Initial MSE decrease.** The mean-squared error is

$$\begin{aligned}\mathrm{MSE}(\lambda) = \mathbb{E}\|\widehat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_2^2 &= \mathrm{tr}\left[\mathrm{Var}(\widehat{\boldsymbol{\beta}}_\lambda)\right] + \|\mathrm{Bias}(\widehat{\boldsymbol{\beta}}_\lambda)\|_2^2 \\ &= \sigma^2\,\mathrm{tr}\left[\boldsymbol{A}(\boldsymbol{A}+\lambda\boldsymbol{I})^{-2}\right] + \lambda^2\boldsymbol{\beta}^\top(\boldsymbol{A}+\lambda\boldsymbol{I})^{-2}\boldsymbol{\beta}.\end{aligned}$$

Using $\frac{d}{d\lambda}(\boldsymbol{A}+\lambda\boldsymbol{I})^{-2} = -2(\boldsymbol{A}+\lambda\boldsymbol{I})^{-3}$,

$$\mathrm{MSE}'(\lambda) = -2\sigma^2\,\mathrm{tr}\left[\boldsymbol{A}(\boldsymbol{A}+\lambda\boldsymbol{I})^{-3}\right] + 2\lambda\boldsymbol{\beta}^\top(\boldsymbol{A}+\lambda\boldsymbol{I})^{-2}\boldsymbol{\beta} - 2\lambda^2\boldsymbol{\beta}^\top(\boldsymbol{A}+\lambda\boldsymbol{I})^{-3}\boldsymbol{\beta}.$$

At $\lambda = 0$ the last two terms vanish, giving

$$\mathrm{MSE}'(0) = -2\sigma^2\,\mathrm{tr}(\boldsymbol{A}^{-2}) < 0,$$

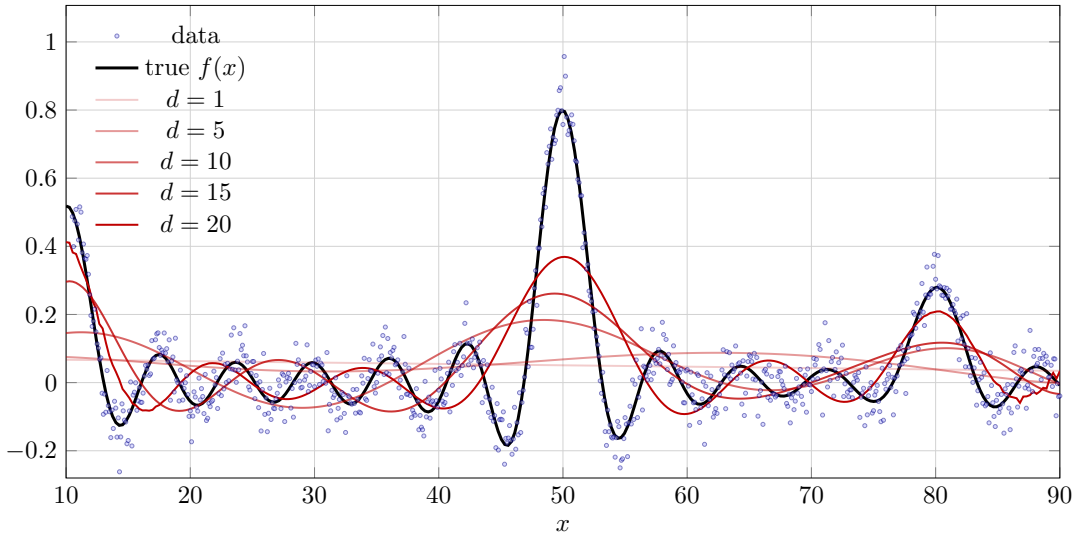
since $\boldsymbol{A}^{-2} \succ 0$. Therefore, for sufficiently small $\lambda > 0$, ridge strictly improves the OLS MSE. □

8.3.2 Kernel Ridge Regression

Example 8.3 (Motivation)**.** Consider the true model as a sum of shifted *sinc* functions (shifted and rescaled $\sin(x)/x$):

$$f(x) = 0.5\frac{\sin(x-10)}{x-10} + 0.8\frac{\sin(x-50)}{x-50} + 0.3\frac{\sin(x-80)}{x-80}.$$

Polynomial regression fits on sinc mixture ($N = 1000$, $\sigma = 0.05$)



Remarks

The Weierstrass approximation theorem states that for every continuous function, there exists a sequence of polynomial functions that converges uniformly to it. However, the theorem only states that a set of polynomial functions exists, without providing a general method of finding one. See Runge's phenomenon for the failure of a high-degree polynomial to approximate a function with equally spaced points (or Gibbs phenomenon for Fourier series).

- The true model has various degrees of smoothness, hence polynomial regression is not economical.
- We may use spline regressions, but need to work out the knots.
- Is there a method that does this automatically?

A single global basis (e.g., polynomials) is poorly matched to signals that vary in smoothness across the domain.[6] Kernel methods build predictions by measuring similarity to training points; with the right kernel, the method adapts locally without the analyst hand-choosing basis functions or knots.

Kernel Ridge Regression We now introduce a nonparametric generalization of ridge regression.

- It generalizes the idea of *enlarging the feature space.*
- In particular, it includes penalized polynomial regression as a special case.

Recall the ridge solution

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Remark

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \boldsymbol{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \boldsymbol{I})^{-1} \boldsymbol{Y}, \quad \widehat{\boldsymbol{Y}}^{\text{ridge}} = \mathbf{X}\mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \boldsymbol{I})^{-1} \boldsymbol{Y}.$$
$$\widehat{Y}^{\text{ridge}}(\boldsymbol{x}) = \mathbf{x}\mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \boldsymbol{I})^{-1} \boldsymbol{Y}.$$

Proof. Because

$$\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}) = \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{X}^\top = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}) \boldsymbol{X}^\top. \qquad \square$$

Note that $\boldsymbol{X}\boldsymbol{X}^\top$ appears in our new expression of the ridge solution.

- The ij-th entry of $\boldsymbol{X}\boldsymbol{X}^\top \in \mathbb{R}^{n \times n}$ is $\boldsymbol{x}_i^\top \boldsymbol{x}_j = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$, the *inner product* between two feature vectors.
- The *predicted values* depend only on $\boldsymbol{x}^\top \boldsymbol{X}^\top \in \mathbb{R}^n$, and the j-th entry is $\boldsymbol{x}^\top \boldsymbol{x}_j = \langle \boldsymbol{x}, \boldsymbol{x}_j \rangle$, the *inner product* between the new observation and the j-th training datum.

Core observation

Prediction by ridge regression boils down to computing inner products between feature vectors. This is the foundation of the so-called **kernel trick**.

[6] A *global* polynomial of degree m has a single complexity knob (m) that applies uniformly over the entire domain. If f is smoother on some subregions and less smooth (e.g., has a kink or rapid curvature) on others, a single global degree must be chosen to handle the *least* smooth part, thereby overfitting elsewhere.

Kernel Trick Suppose we use another "inner product" to replace the usual one.

- Replace $\langle \cdot, \cdot \rangle$ by a similarity measure $K(\cdot, \cdot)$, called a *kernel*.
- A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **positive definite kernel** on $\mathcal{X}$ if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$$

  holds for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{R}$.

- For ridge regression, we have the *linear kernel*

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \boldsymbol{x}_1^\top \boldsymbol{x}_2.$$

- Then

$$\boldsymbol{x}\boldsymbol{X}^\top \to (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_n)), \qquad \boldsymbol{X}\boldsymbol{X}^\top \to \boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{1 \leq i,j \leq n}.$$

- The fitted vector is now

$$\widehat{\boldsymbol{Y}} = \boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Y},$$

  and the prediction for a new observation $\boldsymbol{x}$ is

$$\widehat{y} = (K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_n))(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Y}. \tag{8.1}$$

This is the so-called **kernel ridge regression**.

Example 8.4 (Commonly used kernels)**.**

- Linear kernel:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = \boldsymbol{x}_i^\top \boldsymbol{x}_j.$$

- Polynomial kernel of degree *up to* d:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i^\top \boldsymbol{x}_j)^d.$$

  - This is equivalent to enlarging the feature space to include all polynomials with degree up to d, hence polynomial regression of degree d.

- Polynomial kernel of degree *exactly* d:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j)^d.$$

- Gaussian (radial basis function) kernel:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2).$$

  - **Local behavior**: The kernel decreases exponentially fast in the distance of two feature vectors. Points far away have little effect in regression.
  - The corresponding feature space is implicit and infinite-dimensional.

Instead of the linear model, we now consider

$$y = \sum_{j=1}^{n} \alpha_j K(\boldsymbol{x}, \boldsymbol{X}_j) + \varepsilon.$$

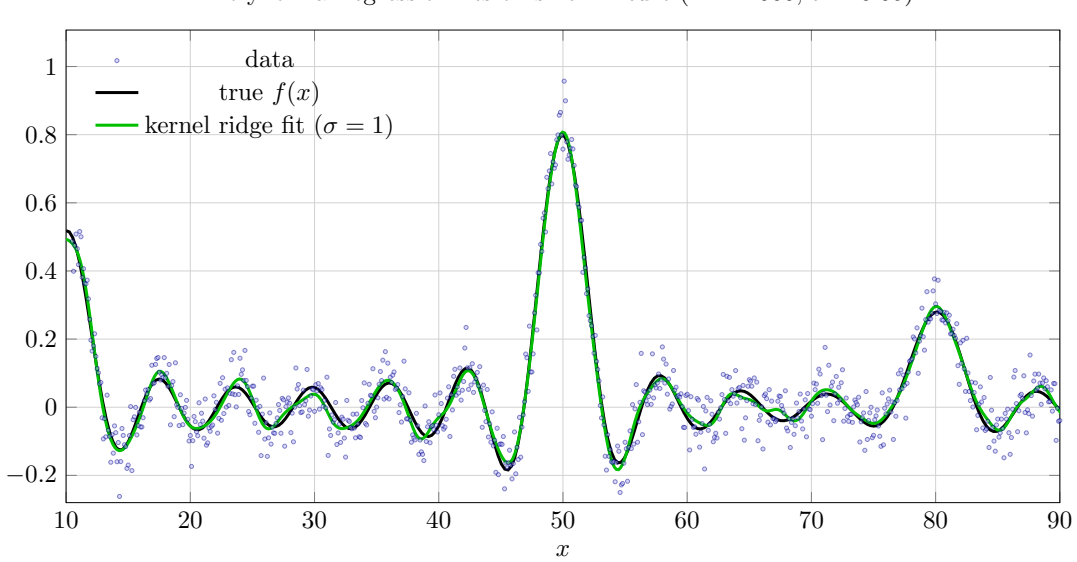
- Kernel ridge regression approximates multivariate regression using the kernel functions $\{K(\cdot, \boldsymbol{X}_j)\}_{j=1}^n$.
- For the polynomial kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i^\top \boldsymbol{x}_j)^d$, this is equivalent to polynomial regression.

Remarks

The representer theorem (8.1) explains why the prediction uses only $K(\boldsymbol{x}, \boldsymbol{x}_j)$: among all functions in the reproducing-kernel Hilbert space, the penalized risk minimizer lives in the span of kernel evaluations at the training points. γ controls locality for the Gaussian kernel (smaller γ = smoother fits); λ controls regularization strength. They should be tuned together.

Example 8.5 (Gaussian kernel versus polynomial kernel)**.** Consider again

$$f(x) = 0.5\frac{\sin(x-10)}{x-10} + 0.8\frac{\sin(x-50)}{x-50} + 0.3\frac{\sin(x-80)}{x-80}.$$



Remarks

Polynomials extrapolate globally and can oscillate wildly between knots; Gaussian kernels perform localized averaging, adapting to regions of different smoothness. When the signal has localized features, kernel ridge with a suitable γ can dramatically reduce variance without inflating bias.

8.3.3 Lasso Regression

Another very important shrinkage method is the **least absolute selection and shrinkage operator (LASSO)**.

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}.$$

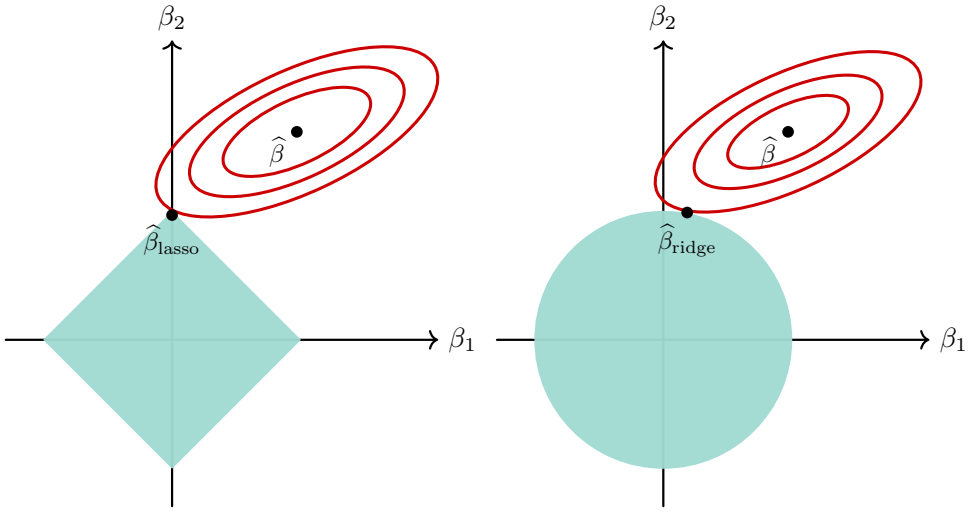
- The key difference between lasso and ridge regressions is that lasso uses l_1-norm (sum of absolute values) instead of l_2-norm (sum of squares).

Comparing Lasso with Ridge Regressions

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \ \|\boldsymbol{Y} - (\beta_0 + X\boldsymbol{\beta})\|_2^2 \quad \text{such that} \quad \|\boldsymbol{\beta}\|_1 \leq t$$

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \ \|\boldsymbol{Y} - (\beta_0 + X\boldsymbol{\beta})\|_2^2 \quad \text{such that} \quad \|\boldsymbol{\beta}\|_2^2 \leq t$$

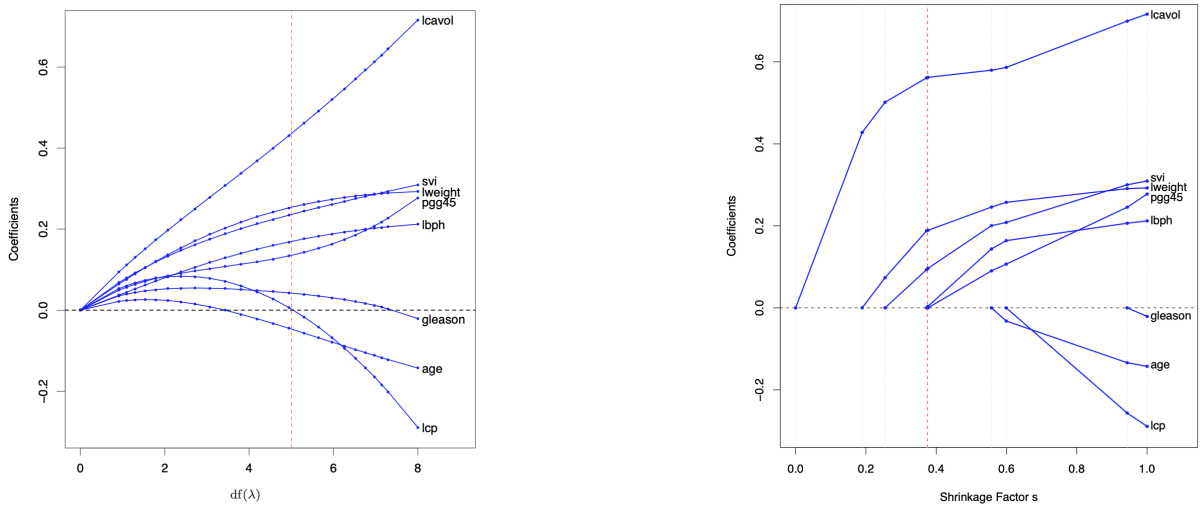
- There is a one-to-one correspondence between λ and t.



Intuition

The l_1 ball has corners on the coordinate axes; when the elliptical RSS contours expand to touch the constraint, they often meet at a corner, setting some coefficients *exactly* to zero. Thus lasso both *selects* variables and *shrinks* estimates.

Ridge and LASSO Solution Paths



$\text{df}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$, where d_j are the singular values of X. $s = t / \sum_{j=1}^{p} |\widehat{\beta}_j|$.

Effective degrees of freedom

The "effective degrees of freedom" quantifies model flexibility after shrinkage. For ridge it interpolates between 0 (all coefficients shrunk to 0) and p (OLS). For lasso, df is well approximated by the number of nonzero coefficients along the path, which is why the lasso path is often plotted against sparsity level. More on this later.

Something in Between

- Lasso results in a "sparse" solution, so that some of the estimators are exactly 0. This is a type of **model/variable selection**.
- Lasso estimates do not need to be unique if covariates are collinear. Hence it can be unstable for high-dimensional data.
- Ridge regression handles multicollinearity.
- *Elastic net* (Zou and Hastie, 2005) regularization uses a convex combination of both:

$$\arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\boldsymbol{Y} - X\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\}.$$

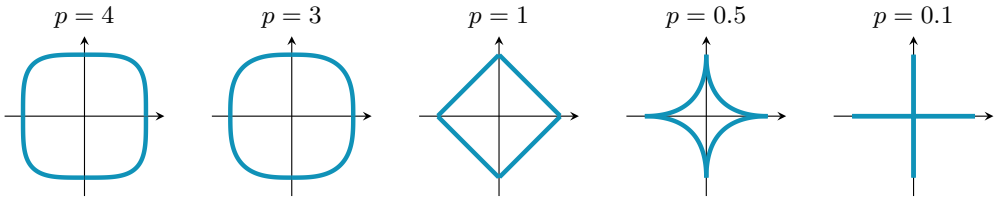
Elastic net inherits sparsity from l_1 and stability against collinearity from l_2. When groups of correlated predictors exist, elastic net tends to keep or drop them together, improving interpretability over lasso alone.

Discussion: Subset Selection, Ridge Regression and the Lasso Recall that

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - X^\top \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_p^p \right\}.$$

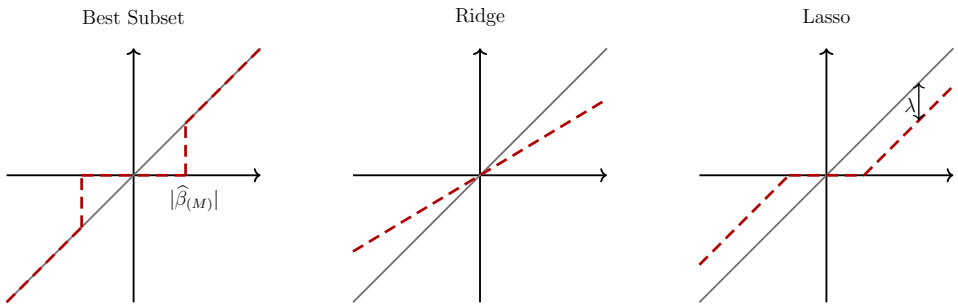
- Ridge: $p = 2$.
- Lasso: $p = 1$.
- Best subset: $p = 0$ (assuming $0^0 = 0$).

The contour plot of the penalties are shown below.



Illustrations with orthogonal designs Assuming that $\{\boldsymbol{x}_j\}$ are *orthogonal*, the three procedures have explicit solutions. Let $\widehat{\boldsymbol{\beta}}$ denote the OLS estimator. The estimators from best subset, ridge and lasso are given in the following table:

Estimater	Formula
Best subset (size M)	$\widehat{\beta}_j^{\text{subset}} = \widehat{\beta}_j \cdot \mathbb{1}\big(\lvert\widehat{\beta}_j\rvert \geq \lvert\widehat{\beta}_{(M)}\rvert\big)$
Ridge	$\widehat{\beta}_j^{\text{ridge}} = \widehat{\beta}_j / (1 + \lambda)$
Lasso	$\widehat{\beta}_j^{\text{lasso}} = \text{sign}\big(\widehat{\beta}_j\big)(\lvert\widehat{\beta}_j\rvert - \lambda)_+$



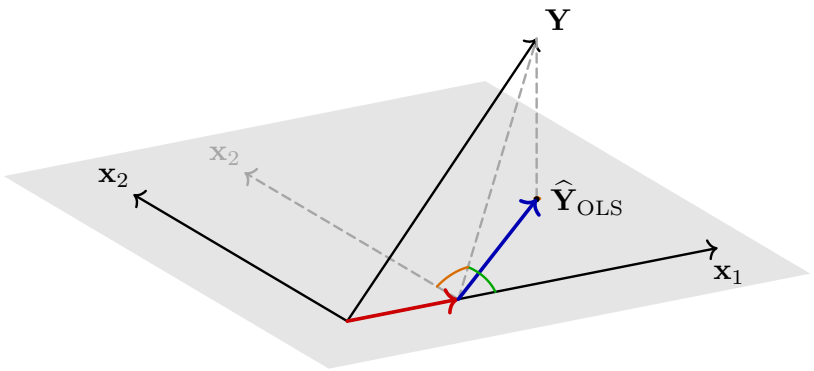
In an *orthogonal* design, these penalties lead to closed-forms: ridge applies multiplicative shrinkage; lasso applies soft-thresholding; subset selection applies hard-thresholding.

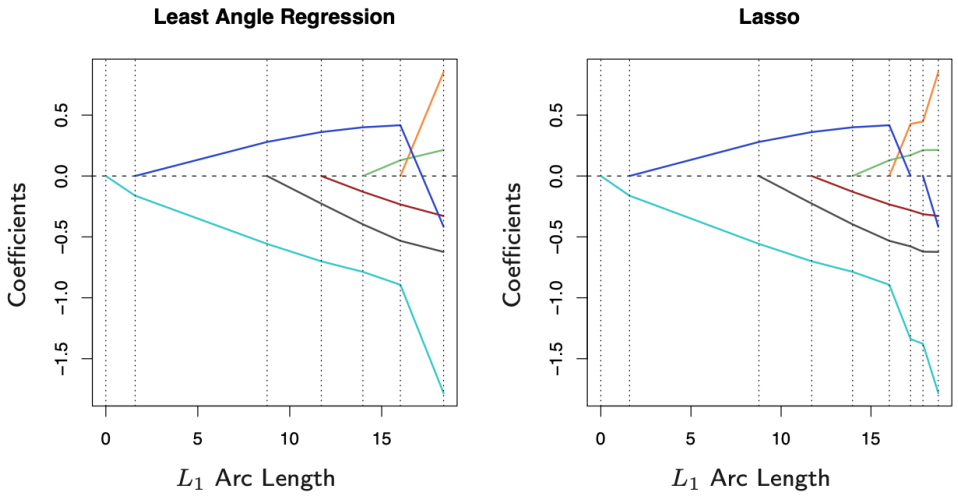
Least Angle Regression We have not talked about how to compute the lasso solution.

- The lasso problem is formulated as a quadratic programming (QP) problem, so usual QP solvers can be applied.
- Alternatively, we now discuss an extremely efficient (and intuitive) algorithm to solve for the *entire lasso path*, i.e., all solutions with respect to λ.
- Efron et al. (2004), *Least angle regression*.

Least Angle Regression (LAR) starts with all coefficients at zero and moves in the direction most correlated with the residuals. When another predictor becomes equally correlated, the path turns to move *equiangularly* between them. With a simple modification, this path exactly traces the lasso solutions as λ decreases.

Graphical Illustration of Least Angle Regression





Algorithm 3 Least Angle Regression (LAR)

1: **Standardize** each predictor column x_j to have mean 0 and unit ℓ_2-norm, $\|x_j\|_2 = 1$.
2: **Initialize** the residual and coefficients: $r \leftarrow y - \bar{y}$ (equivalently $r \leftarrow y$ if y is centered), $\beta_1, \beta_2, \ldots, \beta_p \leftarrow 0$.
3: **Find the most-correlated predictor:** $j \leftarrow \arg\max_\ell \left|\langle x_\ell, r\rangle\right|$.
4: **Univariate move (enter x_j):** increase β_j from 0 toward its least-squares value (in the direction $\operatorname{sign}\langle x_j, r\rangle$) *until* there exists a competitor $k \neq j$ with $\left|\langle x_k, r\rangle\right| = \left|\langle x_j, r\rangle\right|$.
5: **Equiangular move (active set $\{j, k\}$):** move β_j and β_k in the direction defined by their joint least-squares coefficients of the current residual on (x_j, x_k) (i.e., along the equiangular direction) *until* some other competitor x_i achieves $\left|\langle x_i, r\rangle\right|$ equal to the current active correlation.
6: **Drop inactive variables:** If a nonzero coefficient hits zero, drop its variable from the active set of variables, and recompute the current joint least-square direction.
7: **Repeat** the "find most-correlated" and "equiangular move" steps, adding one variable at each tie, *until* all p predictors have entered (or the process has taken $\min(N-1, p)$ steps).
8: **Output:** the piecewise-linear solution path $\beta(\cdot)$; after $\min(N-1, p)$ steps the path reaches the full least-squares solution.

8.4 Model Selection Criteria

For a given data set and some candidate models, a model selection criterion gives each model a score, and we can pick the model with the highest score.

- Usually the criterion rewards good fit, but penalizes complexity.
- Here we look at
  - Adjusted R^2: adjusts R^2, taking account of the degrees of freedom.
  - Mallows' C_p: based on the MSE of the estimations.
  - Akaike Information Criterion (AIC): based on log-likelihood.
  - Bayesian Information Criterion (BIC): based on Bayesian statistics.
- The general idea is to "estimate" the testing error by making adjustments to the training error.

Discussion & intuition

Each criterion is a proxy for out-of-sample performance. They all reduce to "fit − penalty", differing mainly in how they define the penalty. No single score is uniformly best; understanding their assumptions helps you choose appropriately.

8.4.1 Adjusted R^2

Recall the coefficient of determination R^2 from the linear model

$$R^2 = 1 - \frac{\text{RSS}}{\text{SST}}.$$

- R^2 reflects the *training error*.
- A model with larger R^2 is not necessarily better than another model with smaller R^2 when we consider *test error*!

Instead of directly maximizing R^2, we maximize a penalized version of R^2.

Adjusted R^2

The adjusted R^2, taking into account the degrees of freedom, is

$$\text{adjusted } R^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{SST}/(n-1)}.$$

- With more inputs, R^2 always increases, but the adjusted R^2 could decrease since more inputs are penalized by the smaller degrees of freedom of the residuals.
- Maximizing adjusted R^2 is equivalent to

$$\min\{\text{RSS}/(n-p-1)\}.$$

Discussion & intuition

Adjusted R^2 is simple and fast, but it implicitly assumes homoskedastic Gaussian errors and OLS fitting. It is useful for quick screening, less so when models are misspecified or estimated outside least squares.

8.4.2 Mallows' C_p

We consider the fixed–design linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}, \quad \text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n,$$

so that $\mathbb{E}(Y_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ and $\text{Var}(Y_i) = \sigma^2$. While least squares minimizes the *training* criterion $\text{RSS} = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2$, our ultimate target is prediction accuracy at the design points. A natural measure is the *in-sample prediction risk*

$$\text{MSE}(\widehat{Y}_i) = \mathbb{E}\big[(\widehat{Y}_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2\big].$$

A simple algebraic rearrangement gives, for each i,

$$(\widehat{Y}_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 = (Y_i - \widehat{Y}_i)^2 - (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + 2(\widehat{Y}_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Taking expectations and using $\mathbb{E}[(Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2] = \sigma^2$ yields

$$\text{MSE}(\widehat{Y}_i) = \mathbb{E}\big[(Y_i - \widehat{Y}_i)^2\big] - \sigma^2 + 2\,\text{Cov}(\widehat{Y}_i, Y_i).$$

Summing over i and dividing by σ^2,

$$\sum_{i=1}^{n} \frac{\text{MSE}(\widehat{Y}_i)}{\sigma^2} = \mathbb{E}\left[\frac{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2}{\sigma^2}\right] - n + 2\sum_{i=1}^{n} \frac{\text{Cov}(\widehat{Y}_i, Y_i)}{\sigma^2}. \tag{8.2}$$

Intuition. Equation (8.2) compares what we can compute (the expected training error $\mathbb{E}[\text{RSS}]$) to what we really care about (the expected prediction error of $\widehat{\boldsymbol{Y}}$ against the noise-free signal $\boldsymbol{X}\boldsymbol{\beta}$). The last term is the *optimism correction*: because we used $\boldsymbol{Y}$ to construct $\widehat{\boldsymbol{Y}}$, the two are positively correlated, so the training error tends to underestimate the true prediction error. The size of this downward bias is governed by $\sum_i \text{Cov}(\widehat{Y}_i, Y_i)$.

Linear smoothers and effective degrees of freedom. Many estimators used in regression are linear in the response:

Linear smoother

A fitting method is called a linear smoother if we can write

$$\widehat{\boldsymbol{Y}} = \boldsymbol{S}\boldsymbol{Y},$$

for some $n \times n$ "smoother" matrix $\boldsymbol{S}$. For ordinary least squares in a given model, $\boldsymbol{S} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is the hat matrix.

In this case

$$\text{Cov}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) = \text{Cov}(\boldsymbol{S}\boldsymbol{Y}, \boldsymbol{Y}) = \boldsymbol{S}\,\text{Var}(\boldsymbol{Y}) = \sigma^2\boldsymbol{S},$$

so

$$\sum_{i=1}^{n} \text{Cov}(\widehat{Y}_i, Y_i) = \sigma^2 \text{tr}(\boldsymbol{S}).$$

Substituting into (8.2) yields the fundamental identity

$$\sum_{i=1}^{n} \frac{\text{MSE}(\widehat{Y}_i)}{\sigma^2} = \mathbb{E}\left[\frac{\text{RSS}}{\sigma^2}\right] - n + 2\text{tr}(\boldsymbol{S}). \tag{8.3}$$

This motivates the definition

$$\text{df}(\widehat{\boldsymbol{Y}}) \equiv \text{tr}(\boldsymbol{S}),$$

the (effective) *degrees of freedom* of the fit. For least squares in a p-parameter model (including the intercept), $\boldsymbol{S} = \boldsymbol{P}$ is an idempotent projection with $\text{tr}(\boldsymbol{P}) = p$.

Degrees of freedom is generally a useful concept because it allows us to put different procedures on equal footing.

Method	$\boldsymbol{S}$	$\text{tr}(\boldsymbol{S})$
Multiple linear regression	$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$	p
Ridge regression	$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top$	$\sum_{i=1}^{n} \frac{d_i^2}{d_i^2+\lambda}$
Kernel ridge regression	$\boldsymbol{K}(\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}$	$\sum_{i=1}^{n} \frac{\gamma_i}{\gamma_i+\lambda}$

d_i's are the singular values of $\boldsymbol{X}$ and γ_i's are the eigenvalues of $\boldsymbol{K}$.

Relationship to prediction risk

For a linear smoother, (8.3) says:

$$\text{prediction risk} \approx \text{training error} + 2\sigma^2 \times \text{df} - n\sigma^2.$$

The term $2\sigma^2\text{df}$ is the price we pay for flexibility. A smoother that uses more degrees of freedom (larger $\text{tr}(\boldsymbol{S})$) fits the training data better but requires a bigger optimism correction.

Mallows' C_p. Identity (8.3) suggests an estimator of the (scaled) prediction risk: replace $\mathbb{E}[\text{RSS}/\sigma^2]$ by its observed value RSS/σ^2.

Definition 8.1 (Mallows' C_p for a linear smoother)**.** For a fixed linear smoother $\widehat{\boldsymbol{Y}} = \boldsymbol{S}\boldsymbol{Y}$, define

$$C_p(\widehat{\boldsymbol{Y}}) \triangleq \frac{\mathrm{RSS}}{\sigma^2} - n + 2\mathrm{tr}(\boldsymbol{S}).$$

Then

$$\mathbb{E}\big[C_p(\widehat{\boldsymbol{Y}})\big] = \sum_{i=1}^{n} \frac{\mathrm{MSE}(\widehat{Y}_i)}{\sigma^2}.$$

In particular, for ordinary least squares in a p-parameter model,

$$C_p = \frac{\mathrm{RSS}}{\sigma^2} - n + 2p.$$

Thus minimizing C_p over a collection of candidate models approximately minimizes the in-sample prediction risk $\sum_i \mathrm{MSE}(\widehat{Y}_i)$.

Connection to out-of-sample prediction. Let $\boldsymbol{Y}^0 = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^0$ denote an independent replicate at the same design points. The expected *prediction* error at these points is

$$\mathbb{E}\big[\|\boldsymbol{Y}^0 - \widehat{\boldsymbol{Y}}\|_2^2\big] = \sum_{i=1}^{n} \mathbb{E}\big[(Y_i^0 - \widehat{Y}_i)^2\big] = \sum_{i=1}^{n} \mathrm{MSE}(\widehat{Y}_i) + n\sigma^2,$$

since $\mathrm{Var}(Y_i^0) = \sigma^2$ and Y_i^0 is independent of $\widehat{Y}_i$. Combining with (8.3),

$$\frac{1}{\sigma^2}\mathbb{E}\big[\|\boldsymbol{Y}^0 - \widehat{\boldsymbol{Y}}\|_2^2\big] = \mathbb{E}\left[\frac{\mathrm{RSS}}{\sigma^2}\right] + 2\mathrm{tr}(\boldsymbol{S}).$$

Up to the constant n, this is exactly the quantity that C_p estimates; the $-n$ in C_p is historical and chosen so that a correctly specified least-squares model has $\mathbb{E}[C_p] = p$ (see below). Minimizing C_p therefore aligns with minimizing expected out-of-sample error at the training design.

Example 8.6 (Mallows' C_p model setup)**.** Let $\boldsymbol{X}$ collect all candidate explanatory variables (columns). Let $\boldsymbol{X}_p$ denote the p columns used in the p-th candidate model; write $P_p := \boldsymbol{X}_p(\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^\top$ for the orthogonal projector onto $\mathrm{col}(\boldsymbol{X}_p)$. Assume the true mean is generated by some subset $\boldsymbol{X}_{p_0}$, i.e. $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_{p_0}\boldsymbol{\beta}_{p_0}$ with p_0 unknown. For model p, the residual is $\boldsymbol{\varepsilon}_p = (\boldsymbol{I}_n - P_p)\boldsymbol{Y}$, so

$$\mathrm{RSS}(p) = \boldsymbol{Y}^\top(\boldsymbol{I}_n - P_p)\boldsymbol{Y} = \boldsymbol{Y}^\top\Big(\boldsymbol{I}_n - \boldsymbol{X}_p(\boldsymbol{X}_p^\top\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^\top\Big)\boldsymbol{Y}.$$

Lemma 8.1 (Quadratic expectation)**.** *Let Σ be the covariance of $\boldsymbol{Y}$. Then for any fixed matrix A,*

$$\mathbb{E}[\boldsymbol{Y}^\top A\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{Y}]^\top A\mathbb{E}[\boldsymbol{Y}] + \mathrm{tr}(\Sigma A).$$

With $\Sigma = \sigma^2\boldsymbol{I}_n$ and $A = \boldsymbol{I}_n - P_p$,

$$\begin{aligned}\mathbb{E}[\mathrm{RSS}(p)] &= \mathbb{E}\big[\boldsymbol{Y}^\top(\boldsymbol{I}_n - P_p)\boldsymbol{Y}\big] = \boldsymbol{\beta}^\top\boldsymbol{X}^\top(\boldsymbol{I}_n - P_p)\boldsymbol{X}\boldsymbol{\beta} + \mathrm{tr}\big[(\boldsymbol{I}_n - P_p)\sigma^2\boldsymbol{I}_n\big] \\ &= \|(\boldsymbol{I}_n - P_p)\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \big(n - \mathrm{tr}(P_p)\big)\sigma^2 = \|(\boldsymbol{I}_n - P_p)\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + (n-p)\sigma^2,\end{aligned}$$

since P_p is a rank-p symmetric idempotent projector with $\mathrm{tr}(P_p) = p$. Let S^2 be an estimate of σ^2. Then Mallows' criterion

$$C_p(p) = \frac{\mathrm{RSS}(p)}{S^2} - n + 2p$$

satisfies

$$\mathbb{E}[C_p(p)] = \mathbb{E}\Big[\frac{\mathrm{RSS}(p)}{S^2}\Big] - n + 2p \approx \underbrace{\frac{\|(\boldsymbol{I}_n - P_p)\boldsymbol{X}\boldsymbol{\beta}\|_2^2}{\sigma^2}}_{\text{squared bias}} + \underbrace{p}_{\text{variance (df)}}.$$

where the approximation uses $S^2 \approx \sigma^2$.

Case A: Overspecified and correctly specified models, $p \geq p_0$. If the true signal lies in $\mathrm{col}(\boldsymbol{X}_p)$, i.e., $\mathrm{col}(\boldsymbol{X}_{p_0}) \subseteq \mathrm{col}(\boldsymbol{X}_p)$, then

$$(\boldsymbol{I}_n - P_p)\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \qquad \Rightarrow \qquad \mathbb{E}[\mathrm{RSS}(p)] = (n-p)\sigma^2,$$

and hence $\mathbb{E}[C_p(p)] \approx p$. Once all truly relevant directions are included, the projection error vanishes and only the variance term $\sigma^2(n-p)$ remains. Consequently, the C_p values for $p \geq p_0$ fluctuate around the 45° line $C_p \approx p$. Among such models, larger p buys little reduction in expected prediction error but incurs greater variability.

Case B: Underspecified models, $p < p_0$. If the model omits some truly relevant regressors, then the bias term is positive:

$$\|(\boldsymbol{I}_n - P_p)\boldsymbol{X}\boldsymbol{\beta}\|_2^2 > 0.$$

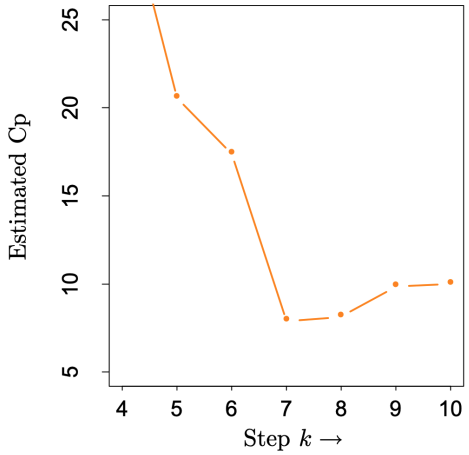
As p increases within a nested sequence ($\mathrm{col}(\boldsymbol{X}_p) \subset \mathrm{col}(\boldsymbol{X}_{p+1})$), the projector P_p grows and the residual projection $(\boldsymbol{I}_n - P_p)\boldsymbol{X}\boldsymbol{\beta}$ shrinks, so

$$0 \leq \|(\boldsymbol{I}_n - P_{p+1})\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \leq \|(\boldsymbol{I}_n - P_p)\boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

Thus $\mathbb{E}[C_p(p)]$ decreases with p until the true dimension p_0 is reached; for $p \geq p_0$ it levels off near the line $C_p \approx p$.

Implications for the C_p plot.

- For $p < p_0$, the bias term dominates and $\mathbb{E}[C_p(p)]$ decreases as relevant variables are added.
- For $p \geq p_0$, the bias term is 0 and $\mathbb{E}[C_p(p)] \approx p$; points lie near the 45° line.
- If the p-th model is (close to) the truth, its C_p tends to be close to, and sometimes slightly below, p (sampling and plug-in variability).
- In practice, inspect the plot of C_p versus p and choose the smallest model whose C_p is among the minimal values—often the *elbow* where the curve first meets the line $C_p \approx p$.



Practical guidance.

- *Model choice.* Compute C_p for each candidate model and prefer those with smaller C_p. A common diagnostic plot is C_p versus p; models with C_p close to p behave approximately unbiased at the design points.
- *Estimating σ^2.* In practice σ^2 is unknown. Mallows' original proposal uses an external estimate $\widehat{\sigma}^2$, typically from a "large" reference model believed to approximate the truth. Then

$$\widehat{C}_p = \frac{\mathrm{RSS}}{\widehat{\sigma}^2} - n + 2\mathrm{tr}(\boldsymbol{S}).$$

  The closer $\widehat{\sigma}^2$ is to σ^2, the more reliable the comparison of C_p across models.
- *Beyond least squares.* For any fixed linear smoother $\widehat{\boldsymbol{Y}} = \boldsymbol{S}\boldsymbol{Y}$, the same formula with $\mathrm{df} = \mathrm{tr}(\boldsymbol{S})$ applies. For example, ridge regression has

$$\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top$$

and

$$\mathrm{df}_{\mathrm{ridge}}(\lambda) = \mathrm{tr}(\boldsymbol{S}_\lambda) = \mathrm{tr}\big[(\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\big] = \sum_{j=1}^{p} \frac{\mu_j}{\mu_j + \lambda},$$

where $\{\mu_j\}$ are the eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$.

- *Optimism correction.* The expected gap between out-of-sample and training error is

$$\frac{1}{n}\mathbb{E}\big[\|\boldsymbol{Y}^0 - \widehat{\boldsymbol{Y}}\|_2^2 - \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|_2^2\big] = \frac{2\sigma^2}{n}\mathrm{tr}(\boldsymbol{S}) \geq 0,$$

quantifying precisely how reusing the data makes the training error overly optimistic by about $2\sigma^2$ per effective parameter.

8.4.3 Akaike's Information Criterion (AIC)

Mallows' C_p is primarily justified for linear models fitted by ordinary least squares. Akaike's Information Criterion (AIC) extends the same bias–variance trade-off to general likelihood-based models and therefore applies far beyond the OLS setting.

Kullback–Leibler motivation. Let $(\boldsymbol{Y}, \boldsymbol{X})$ denote the observed data and consider a family of candidate models for $Y \mid X$ indexed by a parameter $\boldsymbol{\beta} \in \mathbb{R}^P$, with likelihood (or density) $f_Y(y; X, \boldsymbol{\beta})$. Let $\boldsymbol{\beta}^*$ denote the (unknown) parameter that actually generates the data. AIC is derived by measuring the discrepancy between the candidate model $f_Y(\cdot; X, \boldsymbol{\beta})$ and the truth $f_Y(\cdot; X, \boldsymbol{\beta}^*)$ using the *Kullback–Leibler (KL) divergence*

$$D_{\mathrm{KL}}(\boldsymbol{\beta}\|\boldsymbol{\beta}^*) = \int \log\left(\frac{f_Y(y; X, \boldsymbol{\beta}^*)}{f_Y(y; X, \boldsymbol{\beta})}\right) f_Y(y; X, \boldsymbol{\beta}^*)\mathrm{d}y.$$

Because

$$D_{\mathrm{KL}}(\boldsymbol{\beta}\|\boldsymbol{\beta}^*) = \underbrace{\mathbb{E}_{\boldsymbol{\beta}^*}[\log f_Y(Y; X, \boldsymbol{\beta}^*)]}_{\text{constant in } \boldsymbol{\beta}} - \mathbb{E}_{\boldsymbol{\beta}^*}[\log f_Y(Y; X, \boldsymbol{\beta})],$$

minimizing $D_{\mathrm{KL}}(\boldsymbol{\beta}\|\boldsymbol{\beta}^*)$ is equivalent to maximizing the (unknown) *expected out-of-sample log-likelihood*

$$H(\boldsymbol{\beta}) \triangleq \mathbb{E}_{\boldsymbol{\beta}^*}[\log f_Y(Y; X, \boldsymbol{\beta})].$$

Bias correction and the AIC formula. Let

$$\ell(\boldsymbol{\beta}) \triangleq \log f_Y(\boldsymbol{y}_{\mathrm{obs}}; \boldsymbol{X}, \boldsymbol{\beta})$$

be the log-likelihood for the observed sample, and let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood estimator (MLE) in a given candidate model. A naive plug-in for $H(\boldsymbol{\beta}^*)$ is $\ell(\widehat{\boldsymbol{\beta}})$, but this is optimistic because the same data are used both to fit and to evaluate the model.

Under standard regularity conditions, Akaike showed that the leading-order optimism equals the number of free parameters p in the model (including, for example, variance parameters in Gaussian models), yielding the approximation

$$\mathbb{E}_{\boldsymbol{\beta}^*}\big[\ell(\widehat{\boldsymbol{\beta}}) - p\big] \approx H(\boldsymbol{\beta}^*).$$

Consequently, an approximately unbiased estimator of $-2H(\boldsymbol{\beta}^*)$ is

$$\mathrm{AIC} = -2\ell(\widehat{\boldsymbol{\beta}}) + 2p.$$

Here p denotes the number of estimated free parameters in the model

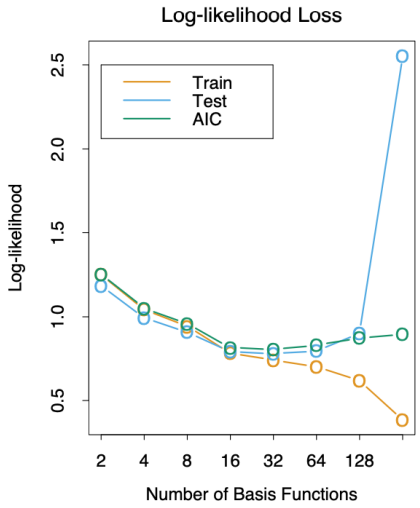
Interpretation and use. Smaller AIC indicates a model that is estimated to be closer (in KL divergence) to the data-generating process, hence better from a predictive point of view. AIC values are *relative*: only differences across models fitted to the same dataset are meaningful. For a collection of candidate models $\{M_i\}$, define

$$\Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min}.$$

The *relative likelihood* of model i is

$$\exp(-\Delta_i/2).$$

- **General applicability.** Unlike the likelihood ratio test (LRT), which compares only nested models, AIC imposes no nesting requirement and can rank any set of likelihood-based candidates.
- **Relation to C_p.** In the Gaussian linear model, AIC reduces (up to an additive constant) to Mallows' C_p; the factor of 2 in the penalty produces the familiar $2p$ complexity term.



8.4.4 Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC)

$$\text{BIC} = -2l(\widehat{\boldsymbol{\beta}}_p) + \log(n)p.$$

- Similar to AIC, applicable to models fitted by maximizing a log-likelihood.
- Motivated by a Bayesian approach to model selection.
- $\exp(-\text{BIC}/2)$ approximates the posterior probability of the current model, so small BIC means large posterior probability.

Compare AIC and BIC

- No clear choice between AIC and BIC.
- 2 vs. $\log(n)$: BIC is more conservative (more penalty to complicated models).
- BIC is asymptotically consistent: as the number of observations grows, the probability that BIC will choose the true model approaches 1.
- This is not the case for AIC, which tends to choose models that are too complex.
- For finite samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity.

8.5 Cross-Validation

K-fold cross-validation is a general-purpose method of prediction/test error estimation, especially in a data-rich scenario.

1	2	3	4	5
Train	Train	Validation	Train	Train

- Divide the data into K parts **(before doing anything else)**.
- For each part, fit the model using data from the other parts (training sets), and calculate the prediction error on the current part (validation set).

Discussion

Cross-validation directly mimics the train/test split that matters for prediction. It captures both bias and variance effects and applies uniformly across modeling families, including those where information criteria are hard to compute.

Our ultimate goal is to produce the best model with the best prediction accuracy.

- Suppose in the i-th run we obtain a test error MSE_i on the validation set.
- Average over the K estimates of the test errors, and obtain

$$\text{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K}\text{MSE}_i.$$

We then choose the model with the smallest CV error.

Discussion & intuition

Report not only the mean CV error but also its variability across folds. When differences are small relative to this variability, prefer the simpler model ("one-standard-error" rule).

Leave-One-Out Cross-Validation An interesting special case is the n-fold cross-validation, also known as the leave-one-out cross-validation (LOOCV).

- K-fold cross-validation is more biased. It estimates the performance of a model trained on a dataset of size $n\frac{K-1}{K}$. If we use a model trained on the full data, it will perform slightly better than the cross-validation estimate suggests.
- K-fold can also have higher variance if the sample size is smaller (so the training set is even smaller).
- LOOCV is approximately unbiased (because only one datum is left out).
- LOOCV has higher variance because in each iteration you are using essentially the same set of data.

One advantage of LOOCV is that it is computationally inexpensive for some models, as we will see next.

Leave-one-out CV error

The leave-one-out CV error (under quadratic loss) is

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}^{(-i)}(\boldsymbol{X}_i)\right)^2,$$

where $\widehat{Y}^{(-i)}(\boldsymbol{X}_i)$ is the predicted value at $\boldsymbol{X}_i$ computed using all the data except the i-th observation.

By definition, we need to repeat the fitting process n times to compute this CV error. However, we can avoid such computation in many popular regression models. In particular, many regression methods are *linear smoothers* with different $\boldsymbol{S}$ matrices.

Method	$\boldsymbol{S}$
Multiple linear regression	$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$
Ridge regression	$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top$
Kernel ridge regression	$\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}$

Discussion

Here $\boldsymbol{S}$ is the *hat matrix*. Its diagonal S_{ii} measures the **leverage** of observation i (See how ridge reduces the leverage of any observation). Points with large leverage influence the fit strongly and, as we will see, require a larger correction in LOOCV.

Assume that a linear smoother is fitted on $\{(X_i, Y_i)\}_{i=1}^n$. Let $\boldsymbol{x}$ be a new covariate vector and $\widehat{f}(\boldsymbol{x})$ its predicted value using the linear smoother. Augment the dataset by including $(\boldsymbol{x}, \widehat{f}(\boldsymbol{x}))$ and refit the linear smoother on this augmented dataset.

Definition 8.2 (Self-stable)**.** The linear smoother is said to be self-stable if the fit based on the augmented dataset is identical to the fit based on the original data regardless of $\boldsymbol{x}$.

- Multiple linear regression, ridge regression and kernel ridge regression are all self-stable.

For self-stable linear smoothers, leave-one-out cross-validation is particularly appealing because in many cases we have the following reduction.

Theorem 8.1.

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{f}^{(-i)}(\boldsymbol{X}_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(Y_i - \widehat{Y}_i\right)^2}{\left(1 - S_{ii}\right)^2}.$$

Proof. Apply the linear smoother to $\{(X_j, Y_j) : j \neq i\}$ to obtain $\widehat{f}^{(-i)}(\boldsymbol{X}_i)$. Apply the linear smoother to $(\boldsymbol{X}, \tilde{\boldsymbol{Y}}) \stackrel{\text{def}}{=} \{(X_j, Y_j), j \neq i, (\boldsymbol{X}_i, \widehat{f}^{(-i)}(\boldsymbol{X}_i))\}$. By the self-stable property,

$$\widehat{f}^{(-i)}(\boldsymbol{X}_i) = (\boldsymbol{S}\tilde{\boldsymbol{Y}})_i = S_{ii}\widehat{f}^{(-i)}(\boldsymbol{X}_i) + \sum_{j\neq i} S_{ij}Y_j.$$

Hence

$$\widehat{f}^{(-i)}(\boldsymbol{X}_i) = \frac{\sum_{j\neq i} S_{ij}Y_j}{1 - S_{ii}}.$$

On the other hand,

$$\widehat{Y}_i = (\boldsymbol{S}\boldsymbol{Y})_i = S_{ii}Y_i + \sum_{j\neq i} S_{ij}Y_j.$$

Thus,

$$Y_i - \widehat{f}^{(-i)}(\boldsymbol{X}_i) = Y_i - \frac{\sum_{j\neq i} S_{ij}Y_j}{1 - S_{ii}} = \frac{Y_i - (S_{ii}Y_i + \sum_{j\neq i} S_{ij}Y_j)}{1 - S_{ii}} = \frac{Y_i - \widehat{Y}_i}{1 - S_{ii}}.$$

Substituting into the definition of CV gives the result. □

Discussion

This is the classic *PRESS* (predicted residual error sum of squares) identity. High-leverage points (S_{ii} large) get a larger leave-one-out correction because the model has "listened" to them more during training.

Generalized Cross-Validation For some smoothers, $\mathrm{tr}(\boldsymbol{S})$ can be computed more easily than its diagonal elements. To take advantage of this, suppose that we approximate each diagonal element of $\boldsymbol{S}$ by their average, which equals $\mathrm{tr}(\boldsymbol{S})/n$.

Generalized cross-validation

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(Y_i - \widehat{Y}_i\right)^2}{\left(1 - S_{ii}\right)^2} \approx \frac{1}{n}\frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\left(1 - \mathrm{tr}(\boldsymbol{S})/n\right)^2} =: \mathrm{GCV}.$$

Parameter Tuning Using GCV Now we are ready to handle tuning parameter selection in the linear smoother. Write $\boldsymbol{S} = \boldsymbol{S}_\lambda$ and

$$\mathrm{GCV}(\lambda) = \frac{1}{n} \frac{\boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{S}_\lambda)^2 \boldsymbol{Y}}{\big(1 - \mathrm{tr}(\boldsymbol{S}_\lambda)/n\big)^2}.$$

According to GCV, the best λ is

$$\lambda^{\mathrm{GCV}} = \arg\min_\lambda \mathrm{GCV}(\lambda).$$

Remarks

- N-fold CV is generally preferred over a single validation split if computation allows.
- 5-fold or 10-fold CV generally works well.
- Except for the cases with linear smoothers, leave-one-out CV requires n additional trainings per model, which is not always feasible in practice.

Reading Materials

- Jianqing Fan et al., "Statistical Foundation of Data Science," `https://orfe.princeton.edu/~jqfan/fan/classes/525/chapters1-3.pdf`
- Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie, *The Elements of Statistical Learning*, Springer.
- Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani, *Least Angle Regression,* `https://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf`
- Rafael A. Irizarry, `http://rafalab.github.io/pages/754/section-09.pdf`

9 Generalized Linear Model

9.1 Introduction

General Linear Models Multiple linear regression models the conditional mean of a real-valued response as an affine function of predictors and adds a homoscedastic noise term. This is the classical baseline for parametric prediction and inference.

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Here the response Y_i is modeled by a linear function of explanatory variables x_{ij} plus an error term ε_i.

- Multiple linear regression is sometimes called a **general linear model**.

In this usage, **"general"** refers to dependence on potentially many explanatory variables ($p \geq 1$), as opposed to the **simple linear regression model**

$$Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i.$$

The word "linear" means *linear in the parameters* rather than "a straight line in x." A model may be linear in the *parameters* even if it is nonlinear in x:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon_i.$$

- The errors ε_i are typically assumed i.i.d. normal with mean 0 and common variance σ^2.
- In generalized least squares, we may allow a joint multivariate normal with a known covariance matrix.
- Normality underpins the exact small-sample inference of basic linear models (t- and F-tests).
- The linear model's power comes from orthogonality and quadratic loss: normality and homoscedasticity make least squares both efficient and algebraically simple.

Restrictions of General Linear Models General linear models are not appropriate when

- the range of Y is restricted (binary outcomes, counts, bounded proportions);
- the error distribution is markedly non-normal so that $\mathrm{Var}(Y)$ depends on $\mathbb{E}[Y]$.
- Poisson distributions suit arrival counts or misprint counts; their variance *equals* the mean.

Generalized linear models extend linear models to handle non-Gaussian responses and mean–variance relationships.

- We begin with two special cases beyond linear regression: **logistic regression** and **Poisson regression**.
- *Big picture.* GLMs separate three ingredients: a linear predictor, a response distribution from the exponential family, and a link function tying them together.

9.2 Logistic Regression

Motivation Binary outcomes are ubiquitous, e.g., success/failure, alive/dead, default/no default.

- Many applications involve a binary response $Y \in \{0, 1\}$ whose success probability $\mathbb{P}(Y = 1)$ depends on predictors.
- This viewpoint aligns with binary classification: given features, which label (0/1) should we expect?

Because $Y_i \in \{0, 1\}$, the mean $\mu_i = \mathbb{E}[Y_i] \in [0, 1]$. Plain linear regression can predict values outside $[0, 1]$ and imposes constant variance, so it is ill-suited here.

Logistic Regression Recall the linear model

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \text{with } \mu_i = \eta_i \triangleq \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

For binary Y, we instead posit

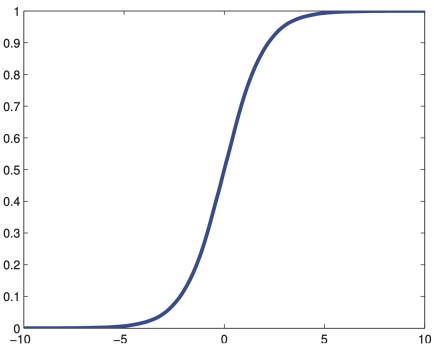
$$Y_i \sim \text{Bernoulli}(\mu_i(\eta_i)), \qquad \text{where } g(\mu_i) = \eta_i$$

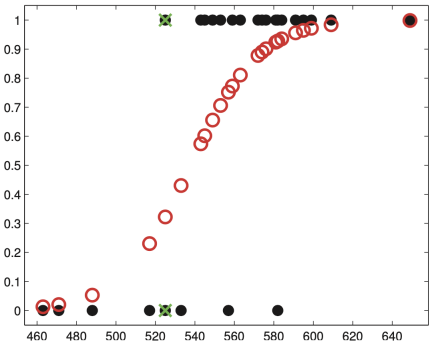
with linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i^\top \boldsymbol{\beta} \in \mathbb{R}$.

- The linear predictor spans $\mathbb{R}$; the mean must live in $[0, 1]$.
- A **link function** g transports $\mu \in (0, 1)$ to $\eta \in \mathbb{R}$; we assume g is smooth and invertible.

Logit link and the sigmoid We usually choose the **logit (log-odds)**:

$$\eta = g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \quad \Longleftrightarrow \quad \mu = \frac{1}{1+\exp(-\eta)}.$$





Prediction: $\mathbb{P}(Y = 1 \mid x) = \mu = 1/(1 + \exp(-\eta))$. A simple classification rule is $\widehat{Y} = 1$ iff $\mathbb{P}(Y = 1 \mid x) > 0.5$.

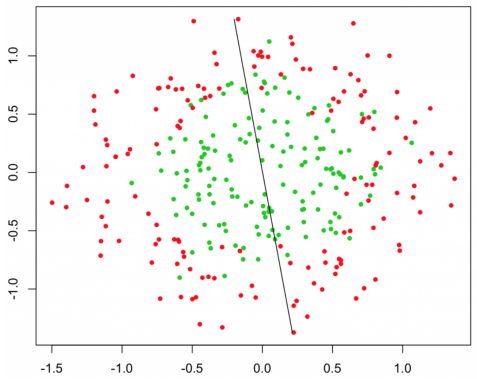
Logistic Regression as Classification At threshold 0.5,

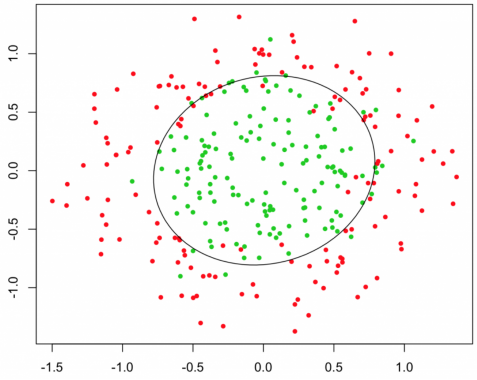
$$\mathbb{P}(Y = 1 \mid x) = 0.5 \iff \eta = 0 \iff \boldsymbol{\beta}^\top \boldsymbol{x} = 0,$$

so the decision boundary is a hyperplane. Using features $\boldsymbol{\phi}(\boldsymbol{x})$ (polynomials, interactions, kernels) yields nonlinear boundaries:

$$\mathbb{P}(Y = 1 \mid x) = 0.5 \iff \boldsymbol{\beta}^\top \boldsymbol{\phi}(\boldsymbol{x}) = 0.$$

Geometry. The log-odds are linear in features; curves in input space arise from nonlinear basis expansions.





- With $\boldsymbol{\phi}(\boldsymbol{x}) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$, the boundary is quadratic.

Estimation via MLE We estimate $\boldsymbol{\beta}$ by MLE. Using $\mu = 1/(1+\exp(-\eta))$ we have

$$\frac{\mu}{1-\mu} = e^{\eta}, \qquad \eta = \log\left(\frac{\mu}{1-\mu}\right).$$

Thus

$$f_{Y_i}(Y_i) = \mu_i^{Y_i}(1-\mu_i)^{1-Y_i} = \frac{\exp(Y_i\eta_i)}{1+\exp(\eta_i)} = \frac{\exp(Y_i\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}.$$

For n independent cases,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp(Y_i\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}.$$

The "canonical link" (logit) turns the Bernoulli likelihood into a convex optimization problem in $\boldsymbol{\beta}$.

The log-likelihood is

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i\boldsymbol{x}_i^\top\boldsymbol{\beta} - \sum_{i=1}^{n} \log\bigl(1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\bigr).$$

Its gradient is

$$\frac{\partial\ell}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n} Y_i\boldsymbol{x}_i - \sum_{i=1}^{n} \frac{\boldsymbol{x}_i\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})} = \sum_{i=1}^{n} (Y_i-\mu_i)\,\boldsymbol{x}_i = \boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{\mu}), \quad \mu_i = \frac{1}{1+\exp(-\boldsymbol{x}_i^\top\boldsymbol{\beta})}.$$

Analogy. This mimics the normal equation of linear regression but with nonlinearity hidden in $\mu(\boldsymbol{\beta})$.

Estimation Equations Setting the score to zero yields

$$\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{\mu}) = \mathbf{0}.$$

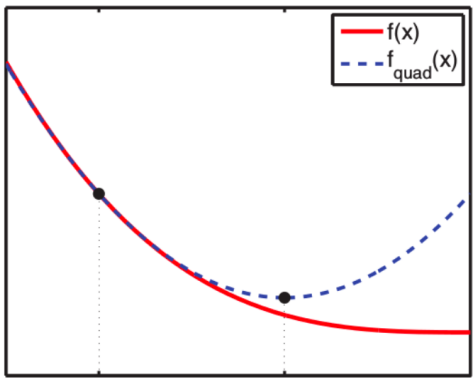
- $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ are the **fitted means**.
- $Y_i - \mu_i$ are **residuals**.
- Compare with linear regression: $\boldsymbol{X}^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0}$.
- Here the equations are **nonlinear** in $\boldsymbol{\beta}$; we require iterations (e.g., Newton–Raphson).

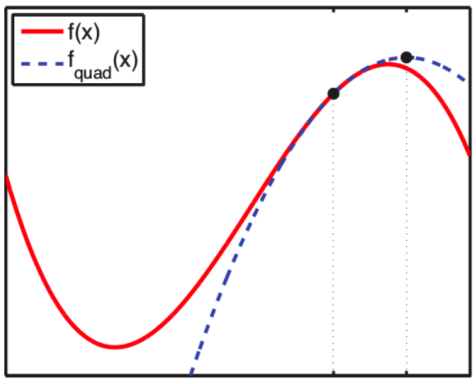
Intuition. We seek coefficients making residuals orthogonal to predictors, now measured on the *probability scale* via the link $\mu = g^{-1}(\eta)$.

Newton–Raphson Method (1D) Newton–Raphson maximizes a smooth $f(x)$ by repeatedly maximizing a quadratic Taylor approximation around x_0:

$$f(x_0+\mathrm{d}x) \approx f(x_0) + f'(x_0)\mathrm{d}x + \tfrac{1}{2}f''(x_0)(\mathrm{d}x)^2,$$

so the step is $\mathrm{d}x = -f'(x_0)/f''(x_0)$ and $x_1 = x_0 + \mathrm{d}x$. Iterate until successive iterates stabilize. *Heuristic.* Newton follows local curvature; when the quadratic approximation is faithful, convergence is rapid.





Newton–Raphson Method (N-D) For $f : \mathbb{R}^p \to \mathbb{R}$ twice differentiable, write the Taylor expansion at $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}_0 + \mathbf{d}\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^\top \mathbf{d}\boldsymbol{x} + \tfrac{1}{2}\mathbf{d}\boldsymbol{x}^\top \boldsymbol{H}(\boldsymbol{x}_0)\mathbf{d}\boldsymbol{x},$$

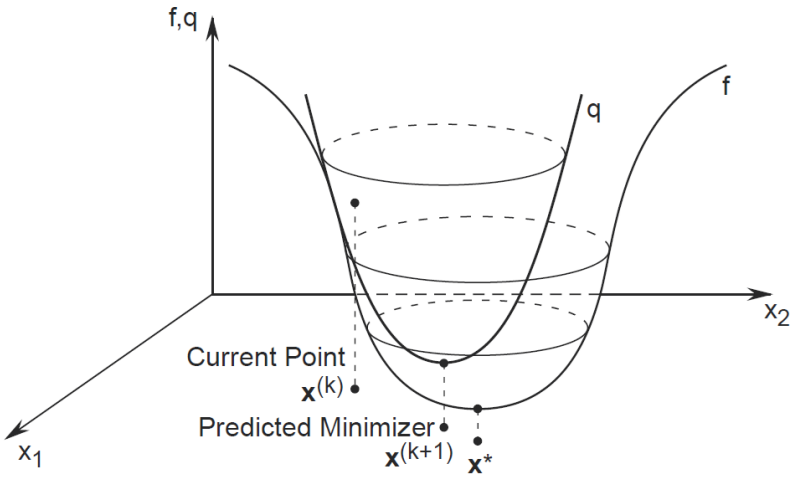
where $\boldsymbol{H}$ is the Hessian, i.e.,

$$(\boldsymbol{H}(\boldsymbol{x}_0))_{ij} = \frac{\partial f(\boldsymbol{x}_0)}{\partial x_i \partial x_j}.$$

The maximizer of the quadratic leads to

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \boldsymbol{H}(\boldsymbol{x}_0)^{-1}\nabla f(\boldsymbol{x}_0).$$

Iterate until successive iterates stabilize.



Numerically Solving the Estimation Equations—Hessian & concavity For logistic regression

$$\nabla \ell(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}), \qquad \boldsymbol{H}(\boldsymbol{\beta}) = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2} = -\boldsymbol{X}^\top W \boldsymbol{X}, \quad W = \operatorname{diag}\big(\mu_i(1-\mu_i)\big).$$

Proof. By the chain rule, $\partial\mu_i/\partial\boldsymbol{\beta} = \partial\mu_i/\partial\eta_i \times \partial\eta_i/\partial\boldsymbol{\beta}$, and

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2} = \frac{\partial}{\partial \boldsymbol{\beta}}\big[\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu})\big] = -\boldsymbol{X}^\top \operatorname{diag}\left(\tfrac{\partial \mu_i}{\partial \eta_i}\right)\boldsymbol{X} = -\boldsymbol{X}^\top \operatorname{diag}\big(\mu_i(1-\mu_i)\big)\boldsymbol{X}. \qquad \square$$

- Here, $\mu_i = 1/(1+\exp(-\boldsymbol{x}_i^\top\boldsymbol{\beta}))$ is the **fitted value**.
- Since W is positive semidefinite, $-\boldsymbol{X}^\top W \boldsymbol{X}$ is negative semidefinite; under full column rank of X, it is negative definite.
- Therefore $\ell(\boldsymbol{\beta})$ is *strictly concave* and has a unique global maximizer; Newton converges. (Concavity is the computational gift of the canonical link.)

With initial $\boldsymbol{\beta}_0$, Newton updates are

$$\boldsymbol{\beta}_{l+1} = \boldsymbol{\beta}_l + \big(\boldsymbol{X}^\top W_l \boldsymbol{X}\big)^{-1} \boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}_l),$$

where $\boldsymbol{\mu}_l = \mu(\boldsymbol{\beta}_l)$ and $W_l = \operatorname{diag}(\mu_{li}(1-\mu_{li}))$.

Pseudo-response and IRLS view Rewrite

$$\begin{aligned}\boldsymbol{\beta}_{l+1} &= \boldsymbol{\beta}_l + \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}_l) \\ &= \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\boldsymbol{\beta}_l + \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}_l)\right) \\ &= \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top W_l \left(X\boldsymbol{\beta}_l + W_l^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_l)\right) \\ &= \left(\boldsymbol{X}^\top W_l \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top W_l \boldsymbol{z}_l,\end{aligned}$$

where

$$\boldsymbol{z}_l = \boldsymbol{X}\boldsymbol{\beta}_l + W_l^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_l),$$

is the **pseudo-response**.

So $\boldsymbol{\beta}_{l+1}$ solves the **weighted least squares** problem

$$\boldsymbol{\beta}_{l+1} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} W_{lii}\big(z_{li} - \boldsymbol{\beta}^\top \boldsymbol{x}_i\big)^2.$$

Interpretation. Each iteration fits a linear model to a *pseudo-response* that corrects the current linear predictor by the scaled residuals.

Iteratively Reweighted Least Squares (IRLS)

- Initialize $\boldsymbol{\beta}_0$.
- At step l:

$$\mu_{li} = \frac{1}{1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta}_l)}, \quad W_l = \operatorname{diag}\big(\mu_{li}(1 - \mu_{li})\big),$$
$$\boldsymbol{z}_l = \boldsymbol{X}\boldsymbol{\beta}_l + W_l^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_l), \qquad \boldsymbol{\beta}_{l+1} = (\boldsymbol{X}^\top W_l \boldsymbol{X})^{-1} \boldsymbol{X}^\top W_l \boldsymbol{z}_l.$$

Binomial Data For grouped binomial observations (with n_i trials at common $\boldsymbol{x}_i$ and observed proportion Y_i),

$$\ell(\boldsymbol{\beta}) = c + \sum_{i=1}^{m} n_i Y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \sum_{i=1}^{m} n_i \log\big(1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\big),$$

which leads to the same estimating equations as the binary case. Grouping improves numerical stability when many identical covariates appear.

9.3 Poisson Regression

Poisson Regression for Counts Count responses arise in mortality studies, insurance claims, transportation, and sports. For counts, variance typically grows with the mean, contradicting homoscedasticity assumptions.

Assume

$$Y_i \sim \text{Poisson}\big(\mu(\eta_i)\big), \qquad \text{where } g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

- The canonical choice is $g(\mu) = \log \mu$, i.e., $\mu = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, mapping $\mathbb{R} \to (0, \infty)$.
- Hence Poisson regression is a **log-linear model**.
- Note that g maps positive real line to the entire real line.

Maximum Likelihood Estimation The log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{\beta} Y_i - \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \sum_{i=1}^{n} \log(Y_i!).$$

The score and likelihood equations are

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (Y_i - \mu_i)\boldsymbol{x}_i = \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}), \qquad \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{0},$$

with $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. *Parallel with logistic*: Same algebra, different mean–variance function.

Review: Contingency Table and Chi-Squared Test of Independence Given discrete pairs $(Y^{(1)}, Y^{(2)})$, test H_0: independence. Build a frequency table of counts: $\#\{Y^{(1)} = i, Y^{(2)} = j\} = y_{ij}$, and the marginal frequency $y_{\cdot j}$ and $y_{i\cdot}$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The test statistic is Pearson's chi-squared

$$V = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(y_{ij} - y_{i\cdot} y_{\cdot j}/n)^2}{y_{i\cdot} y_{\cdot j}/n} \overset{\text{large } n}{\approx} \chi^2_{(I-1)(J-1)}.$$

Classical view. Pearson's statistic compares observed counts to independence-based expected counts.

An Alternative Test of Independence With total n, the table is multinomial with probabilities $\pi_{ij} = \mathbb{P}(Y^{(1)} = i, Y^{(2)} = j)$ and marginals $\pi_{i\cdot}, \pi_{\cdot j}$. Test $H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ via MLE/LRT. The likelihood is

$$L(\boldsymbol{Y}) = \frac{n!}{\prod_{i,j} y_{ij}!} \prod_{i,j} \pi_{ij}^{y_{ij}}, \qquad \ell(\boldsymbol{Y}) = C + \sum_{i,j} y_{ij} \log \pi_{ij}.$$

- Under the alternative hypothesis (full model), we have a constraint $\sum_i \sum_j \pi_{ij} = 1$. Use Lagrange multiplier to obtain the MLE:
$$\widehat{\pi}_{ij} = \frac{y_{ij}}{n}$$
- Under the null hypothesis, we have constraints $\sum_i \pi_{i\cdot} = 1$ and $\sum_j \pi_{\cdot j} = 1$.

$$\widehat{\pi}_{i\cdot} = \frac{y_{i\cdot}}{n}, \quad \widehat{\pi}_{\cdot j} = \frac{y_{\cdot j}}{n} \quad \Rightarrow \quad \widehat{\mu}_{ij} = n\widehat{\pi}_{i\cdot}\widehat{\pi}_{\cdot j} \quad \text{(MLE of the expected counts)}.$$

Likelihood-ratio Test of Independence The LRT statistic simplifies to the **deviance**

$$D = 2 \sum_{i,j} y_{ij} \log \left(\frac{y_{ij}}{\widehat{\mu}_{ij}} \right) \overset{\text{large } n}{\approx} \chi^2_{(I-1)(J-1)}.$$

The deviance is always positive, and we reject H_0 for large D.

Test of Independence as Poisson Regression Model the IJ cell counts as independent Poisson variables $Y_{ij} \sim \text{Poisson}(\mu_{ij})$.

- Think of IJ independent Poisson processes of arrivals.
- Conditioning on total $n = \sum_{i,j} Y_{ij}$ yields a multinomial with probabilities $\pi_{ij} = \mu_{ij}/n$ (Poisson thinning).

Thus multinomial LRTs have a Poisson-regression counterpart.

- Independence model:
$$\log \mu_{ij} = \eta + \alpha_i + \beta_j \quad \Rightarrow \quad \mu_{ij} = e^{\eta} e^{\alpha_i} e^{\beta_j}.$$
- Saturated model:
$$\log \mu_{ij} = \eta + \gamma_{ij}.$$
- Conditioning on n recovers the same MLEs as the multinomial derivation; the LRT statistic is the same (details omitted):
$$D = 2 \sum_{i,j} Y_{ij} \log \left(\frac{Y_{ij}}{\widehat{\mu}_{ij}} \right).$$

Poisson Regression for Three-Way Tables Poisson regression extends naturally to higher-way tables and richer association patterns.

Example 9.1 (Three-way contingency table)**.** Let S denote social status ($I = 4$), E parental encouragement ($J = 2$), P college plans ($K = 2$).

Table 1: College plans by social stratum and parental encouragement

Social Stratum	Parental Encouragement	College Plans		Total
Lower	Low	749	35	784
	High	233	133	366
Lower Middle	Low	627	38	665
	High	330	303	633
Upper Middle	Low	420	37	457
	High	374	467	841
Higher	Low	153	26	179
	High	266	800	1066
Total		3152	1839	4991

Table 2: Three-way contingency table models, hypotheses, and Poisson regression forms

Model	Hypothesis	Poisson Regression
$S+E+P$	$H_0 : \pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$	$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k$
$SE+P$	$H_0 : \pi_{ijk} = \pi_{ij.}\pi_{..k}$	$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$
$SE+EP$	$H_0 : \pi_{ijk} = \pi_{ij.}\pi_{.jk}/\pi_{.j.}$	$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk}$
$SE+SP+EP$	???	$\log \mu_{ijk} = \eta + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$

Models

- $S+E+P$: mutual independence.
- $SE+P$: S associated with E; jointly independent of P.
- $SE+EP$: $S \perp P \mid E$ (pairwise associations but no SP given E).
- $SE+SP+EP$: pairwise associations without three-way interaction.

Deviance:

$$D = 2\sum_i \sum_j \sum_k Y_{ijk} \log\left(\frac{Y_{ijk}}{\widehat{\mu}_{ijk}}\right).$$

- For $SE+SP+EP$ no closed-form MLE exists; fit via iterative algorithms (e.g., IRLS).

9.4 Exponential Family

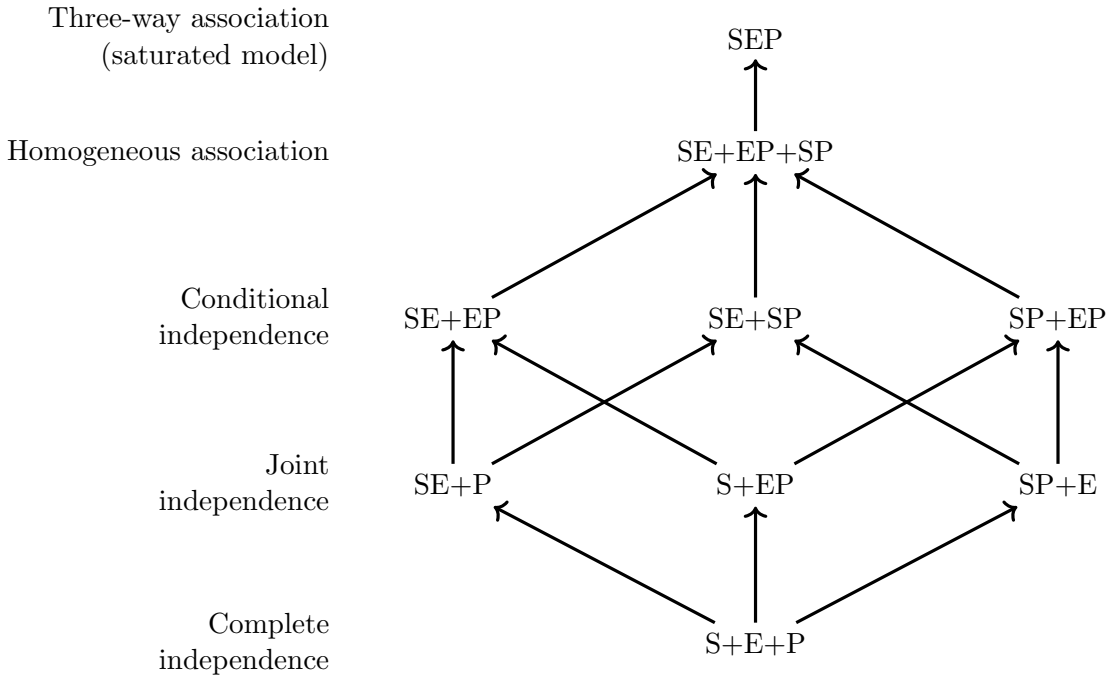
Motivation: Exponential Family and GLM Linear, logistic, and Poisson regression look very different on the surface, but they share a common algebraic backbone. In each case, the conditional distribution of the response belongs to an *exponential family*. This shared structure makes it possible to treat them in a unified way, especially when we study mean–variance relations and maximum likelihood estimation.

For the three canonical examples we have

$$\text{Linear: } Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad \text{Logistic: } Y_i \sim \text{Bernoulli}(\mu_i), \quad \text{Poisson: } Y_i \sim \text{Poisson}(\mu_i).$$

Model	MLE $\widehat{\mu}_{ijk}$	Deviance	d.f.	Sig. level 0.05
$S+E+P$	$Y_{i..}Y_{.j.}Y_{..k}/n^2$	2714.0	10	reject
$SE+P$	$Y_{ij.}Y_{..k}/n$	1877.4	7	reject
$SE+EP$	$Y_{ij.}Y_{.jk}/Y_{.j.}$	255.5	6	reject
$SE+SP+EP$	???	1.575	3	fail to reject

Figure 11: Illustration of nested models.



- Normal, Bernoulli, and Poisson distributions are all exponential families.
- Generalized linear models (GLMs) exploit this structure to build regression models for any response family of this type.

It is therefore worth isolating the key properties of exponential families once and for all.

Exponential Family

A family of densities (or pmfs) is said to be an *exponential family* if it can be written in the form

$$f(x) = h(x)c(\tilde{\boldsymbol{\theta}}) \exp\Big(\sum_{i=1}^{k} w_i(\tilde{\boldsymbol{\theta}})t_i(x)\Big) = h(x) \exp\big(\boldsymbol{\theta}^\top \boldsymbol{t}(x) - A(\boldsymbol{\theta})\big),$$

where we have absorbed the normalizing factor as $c(\tilde{\boldsymbol{\theta}}) = \exp(-A(\boldsymbol{\theta}))$ and set $\theta_i = w_i(\tilde{\boldsymbol{\theta}})$.

Note. The map from the original parameters $\tilde{\boldsymbol{\theta}}$ to the *natural* parameters $\boldsymbol{\theta}$ need not be one-to-one. Even if $\tilde{\boldsymbol{\theta}} \mapsto \boldsymbol{\theta}$ is not injective, the constant $c(\tilde{\boldsymbol{\theta}})$ can still be expressed purely in terms of $\boldsymbol{\theta}$, so that $A(\boldsymbol{\theta})$ and $c(\tilde{\boldsymbol{\theta}})$ are defined consistently on level sets of the mapping.

In this representation,

- $\boldsymbol{\theta}$ are the **natural (canonical) parameters**;
- $\boldsymbol{t}(x)$ are the **sufficient statistics**;
- $A(\boldsymbol{\theta})$ is the **log-partition (cumulant) function** that ensures f integrates (or sums) to one.

The strength of this form is that the key probabilistic quantities, such as means, variances, and covariances of the sufficient statistics, are encoded compactly in A.

Cumulants To anchor the notation, it is useful to list a few familiar families in canonical form:

Family	$\boldsymbol{\theta}$	$\boldsymbol{t}(x)$	$A(\boldsymbol{\theta})$
Gaussian	$(\mu/\sigma^2, -1/(2\sigma^2))$	(x, x^2)	$-\theta_1^2/(4\theta_2) - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi)$
Bernoulli	$\log(\mu/(1-\mu))$	x	$\log(1+e^\theta)$
Poisson	$\log(\lambda)$	x	e^θ

The function A is more than a normalizing constant: it encodes the entire moment structure of the sufficient statistics. Once we know A, its gradient and Hessian automatically yield the mean vector and covariance matrix of $\boldsymbol{t}(X)$.

Cumulants

For $\boldsymbol{Z}$, $K(\boldsymbol{s}) = \log \mathbb{E}[e^{\boldsymbol{s}^\top \boldsymbol{Z}}]$. Derivatives of $A(\boldsymbol{\theta})$ generate moments of $\boldsymbol{t}(X)$:

$$\frac{\partial A}{\partial \boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{t}(X)], \qquad \frac{\partial^2 A}{\partial \boldsymbol{\theta}^2} = \mathrm{Cov}(\boldsymbol{t}(X)).$$

Remark 9.1 (Cumulant function in exponential families). If $\boldsymbol{Z}$ is a random vector, its cumulant generating function (cgf) is

$$K(\boldsymbol{s}) = \log \mathbb{E}\big[e^{\boldsymbol{s}^\top \boldsymbol{Z}}\big],$$

the logarithm of the moment generating function. Its derivatives at $\boldsymbol{s} = \boldsymbol{0}$ give

$$\nabla_{\boldsymbol{s}} K(\boldsymbol{s})\big|_{\boldsymbol{s}=\boldsymbol{0}} = \mathbb{E}[\boldsymbol{Z}], \qquad \nabla^2_{\boldsymbol{s}} K(\boldsymbol{s})\big|_{\boldsymbol{s}=\boldsymbol{0}} = \mathrm{Cov}(\boldsymbol{Z}),$$

and higher derivatives at $\boldsymbol{0}$ yield higher-order cumulants.

Now consider a k-parameter exponential family

$$p_{\boldsymbol{\theta}}(x) = h(x) \exp\big\{\boldsymbol{\theta}^\top \boldsymbol{t}(x) - A(\boldsymbol{\theta})\big\}, \qquad \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k.$$

Under $p_{\boldsymbol{\theta}}$, define the cgf of the sufficient statistics $\boldsymbol{t}(X)$ as

$$K_{\boldsymbol{\theta}}(\boldsymbol{s}) = \log \mathbb{E}_{\boldsymbol{\theta}}\big[e^{\boldsymbol{s}^\top \boldsymbol{t}(X)}\big].$$

A direct calculation shows that, for all $\boldsymbol{s}$ with $\boldsymbol{\theta} + \boldsymbol{s} \in \Theta$,

$$K_{\boldsymbol{\theta}}(\boldsymbol{s}) = A(\boldsymbol{\theta} + \boldsymbol{s}) - A(\boldsymbol{\theta}).$$

Consequently,

$$\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \nabla_{\boldsymbol{s}} K_{\boldsymbol{\theta}}(\boldsymbol{s})\big|_{\boldsymbol{s}=\boldsymbol{0}} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{t}(X)],$$
$$\nabla^2_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{s}} K_{\boldsymbol{\theta}}(\boldsymbol{s})\big|_{\boldsymbol{s}=\boldsymbol{0}} = \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{t}(X)),$$

so $A(\boldsymbol{\theta})$ acts as a cumulant generating function for the sufficient statistics in the natural parameter $\boldsymbol{\theta}$.

Checking the first two cumulants Let us verify these identities directly. Starting from

$$A(\boldsymbol{\theta}) = \log \int h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x,$$

we differentiate under the integral sign (justified, for example, by dominated convergence) to obtain

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} A(\boldsymbol{\theta}) &= \frac{\int h(x) \frac{\partial}{\partial \boldsymbol{\theta}} \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x}{\int h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x} \quad \text{[interchange of limit and integral justified by DCT]} \\
&= \frac{\int \boldsymbol{t}(x) h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x}{\int h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x)) \mathrm{d}x} \\
&= \int \boldsymbol{t}(x) h(x) \exp\big(\boldsymbol{\theta}^\top \boldsymbol{t}(x) - A(\boldsymbol{\theta})\big) \mathrm{d}x \\
&= \mathbb{E}[\boldsymbol{t}(X)].
\end{aligned}$$

In other words, the gradient of A gives the mean of the sufficient statistics. Differentiating once more,

$$\begin{aligned}
\frac{\partial^2}{\partial \theta_i \partial \theta_j} A(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \int \boldsymbol{t}(x) h(x) \exp\big(\boldsymbol{\theta}^\top \boldsymbol{t}(x) - A(\boldsymbol{\theta})\big) \mathrm{d}x \\
&= \int \boldsymbol{t}(x) \left(\boldsymbol{t}(x) - \frac{\partial}{\partial \boldsymbol{\theta}} A(\boldsymbol{\theta})\right)^\top h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{t}(x) - A(\boldsymbol{\theta})) \mathrm{d}x \\
&= \mathbb{E}[\boldsymbol{t}(X) \boldsymbol{t}(X)^\top] - \mathbb{E}[\boldsymbol{t}(X)] \mathbb{E}[\boldsymbol{t}(X)]^\top \\
&= \mathrm{cov}(\boldsymbol{t}(X)).
\end{aligned}$$

Thus the Hessian of A is the covariance matrix of the sufficient statistics, as claimed.

Example 9.2 (Bernoulli)**.** For a Bernoulli(μ) variable written in canonical form, the log-partition function is

$$A(\theta) = \log(1 + e^{\theta}) \Rightarrow \mathbb{E}[X] = \frac{1}{1 + e^{-\theta}} = \mu, \quad \mathrm{Var}(X) = \mu(1 - \mu).$$

Example 9.3 (Poisson)**.** For a Poisson(λ) variable, $A(\theta) = e^{\theta} = \lambda$ gives $\mathbb{E}[X] = \lambda$ and $\mathrm{Var}(X) = \lambda$.

MLE Exponential families are particularly pleasant for likelihood-based inference. For i.i.d. observations x_i, the log-likelihood takes the form

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log h(x_i) + \boldsymbol{\theta}^{\top} \sum_{i=1}^{n} \boldsymbol{t}(x_i) - nA(\boldsymbol{\theta}),$$

so that

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \boldsymbol{t}(x_i) - n \frac{\partial A}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \boldsymbol{t}(x_i) - n\mathbb{E}[\boldsymbol{t}(X)].$$

Setting the score to zero, the maximum likelihood estimator satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{t}(x_i) = \mathbb{E}[\boldsymbol{t}(X)].$$

In words: the MLE equates empirical and model moments of the sufficient statistics. Exponential-family MLE is thus a method-of-moments estimator in disguise.

9.5 GLM

With the exponential-family foundation in place, we now formalize the generalized linear model framework.

9.5.1 Definition of GLM

Generalized linear models (GLMs) start from an exponential-family response and then introduce a regression structure: the mean of the response is linked to a linear predictor through a smooth, possibly nonlinear transformation. Formally, a GLM specifies

- a *random component*: a response density from an exponential family (so that mean and variance are linked via A);
- a *systematic component*: a linear predictor $\eta_i = \boldsymbol{x}_i^{\top} \boldsymbol{\beta}$;
- a *link function* g relating the mean $\mu_i = \mathbb{E}[Y_i \mid \boldsymbol{x}_i]$ to the predictor via $g(\mu_i) = \eta_i$ (equivalently, $\mu_i = g^{-1}(\eta_i)$).

GLMs are particularly useful when Y is constrained (binary, counts, proportions) or clearly non-Gaussian, but we still wish to exploit linear structure in the predictors.

Several familiar models fit neatly into this template:

Model	Family	Link	Range of Y_i	$\mathrm{Var}(Y_i)$
Linear regression	Gaussian	Identity	$(-\infty, \infty)$	$\phi = \sigma^2$
Logistic regression	Bernoulli	Logit	$\{0, 1\}$	$\mu_i(1 - \mu_i)$
Poisson regression	Poisson	Log	$\{0, 1, 2, \dots\}$	μ_i

Generalized Linear Model (GLM)

In summary, a GLM consists of:

- **Random component:** Y_i following an exponential-family distribution (e.g., Bernoulli, Poisson, Gamma) with mean μ_i and variance $\mathrm{Var}(Y_i) = \phi V(\mu_i)$;
- **Linear predictor:** $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i^{\top} \boldsymbol{\beta}$;
- **Link:** a smooth, invertible function g such that $g(\mu_i) = \eta_i$ and $\mu_i = g^{-1}(\eta_i)$.

Viewed this way, GLMs extend ordinary linear regression to a broad class of non-Gaussian responses while retaining much of its interpretability and computational machinery.

9.5.2 Exponential Family within GLM

Exponential Family (scalar form) To simplify notation, consider first a univariate exponential family with scalar canonical parameter θ:

$$Y \sim f(y \mid \theta, \phi) = \exp\Big[\frac{y\theta - A(\theta)}{\phi} + c(y, \phi)\Big],$$

where ϕ is a dispersion parameter (e.g., $\phi = \sigma^2$ for a normal family; $\phi = 1$ for Bernoulli and Poisson). This is just the scalar version of the general form discussed above.

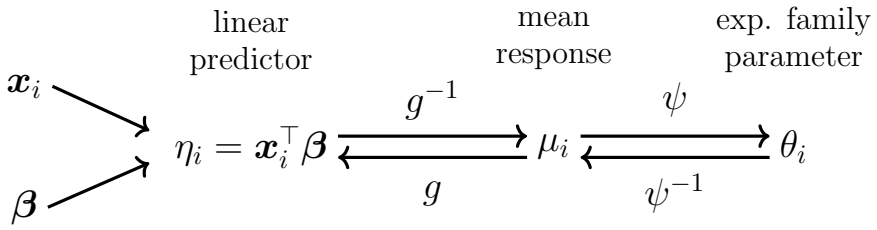
Exponential Family (mean–variance) The cumulant identities translate directly:

- $\mu_i \triangleq \mathbb{E}[Y_i \mid \theta_i, \phi] = A'(\theta_i), \quad \mathrm{Var}(Y_i \mid \theta_i, \phi) = A''(\theta_i)\phi.$
- Since $A'' > 0$, A' is strictly increasing, so the mapping $\mu \leftrightarrow \theta$ is one-to-one:

$$\theta_i = \psi(\mu_i) = (A')^{-1}(\mu_i), \qquad V(\mu_i) \triangleq A''(\theta_i) = \frac{\mathrm{Var}(Y_i)}{\phi}.$$

The function $V(\mu)$ is often called the *variance function*. The family-specific variance function $V(\mu)$ captures the *shape* of the mean–variance relation, while ϕ scales the overall noise level. In GLMs this unifies Gaussian models (where ϕ must be estimated) with Bernoulli/Poisson models (where we typically fix $\phi = 1$).

Figure 12: Overview of GLM Notations and Links



This diagram emphasizes that there are really three layers: the linear predictor η_i, the mean μ_i, and the canonical parameter θ_i. The link g connects η and μ, while the exponential-family structure connects μ and θ via ψ.

9.5.3 Maximum Likelihood Estimation

For GLMs, the log-likelihood can be written as a sum of independent contributions,

$$\ell(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}(\theta_i Y_i - A(\theta_i)) + \sum_{i=1}^{n} c(Y_i, \phi) = \sum_{i=1}^{n} \ell_i + \text{const.}$$

The dependence on $\boldsymbol{\beta}$ is through θ_i, which in turn is linked to μ_i and $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. Applying the chain rule carefully,

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \\ &= \frac{1}{\phi}(Y_i - A'(\theta_i))\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \\ &= \frac{1}{\phi}(Y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}.\end{aligned}$$

This is the basic *score formula* for GLMs; the only difference across specific models lies in the choices of A, g, and hence $\mu(\eta)$.

Exponential Family—Canonical Link Function Among all possible links, the *canonical link* plays a special role. It is defined by

$$\underbrace{\psi(\mu_i) = g(\mu_i)}_{\text{canonical link } g=\psi}, \quad \text{and hence} \quad \underbrace{\theta_i = \psi(\mu_i)}_{\text{def of } \psi} = \underbrace{g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}}_{\text{GLM assumptions}}.$$

In words, the canonical link chooses g so that *the linear predictor equals the canonical parameter*. For the standard GLMs, this reproduces the familiar links:

Model	Canonical link	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta) = A'(\theta)$
Linear	Identity	$\theta = \mu$	$\mu = \theta$
Logistic	Logit	$\theta = \log \frac{\mu}{1-\mu}$	$\mu = \frac{1}{1+e^{-\theta}}$
Poisson	Log	$\theta = \log \mu$	$\mu = e^{\theta}$

Concavity Canonical links greatly simplify derivatives, and in particular they make the log-likelihood concave in $\boldsymbol{\beta}$ for the usual response families, which is invaluable for optimization. To see this, note that $A''(\theta) = \mathrm{Var}(Y)/\phi > 0$, so the cumulant A is convex in θ. Under the canonical link, $\theta_i = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, so A is convex as a function of $\boldsymbol{\beta}$. Consequently, the log-likelihood $\ell(\boldsymbol{\beta})$ is concave in $\boldsymbol{\beta}$, and Newton–Raphson (or variants) is well behaved.

Maximum Likelihood with Canonical Link: score & Hessian With $\theta_i = \eta_i$, we have $\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \theta_i}{\partial \eta_i} = 1$, and the score and Hessian simplify to

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\sum_{i=1}^{n}(Y_i - \mu_i)\boldsymbol{x}_i = \frac{1}{\phi}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}).$$

The likelihood equation takes the exact same form as in the linear, logistic and Poisson regression models

$$\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{0}.$$

Since $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = A''(\theta_i) = V(\mu_i)$,

$$\boldsymbol{H} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2} = -\frac{1}{\phi}\sum_{i=1}^{n}\frac{\partial \mu_i}{\partial \eta_i}\boldsymbol{x}_i\boldsymbol{x}_i^\top = -\frac{1}{\phi}\boldsymbol{X}^\top W \boldsymbol{X},$$

where

$$W = \mathrm{diag}\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\theta_1}, \frac{\mathrm{d}\mu_2}{\mathrm{d}\theta_2}, \ldots, \frac{\mathrm{d}\mu_n}{\mathrm{d}\theta_n}\right), \quad \frac{\mathrm{d}\mu_i}{\mathrm{d}\theta_i} = (\psi^{-1})'(\theta_i) = A''(\theta_i) = V(\mu_i).$$

Note that W here (under canonical link) depend on $\boldsymbol{\beta}$ through $\boldsymbol{\theta} = \boldsymbol{\eta} = \boldsymbol{X}^\top\boldsymbol{\beta}$.

Newton and IRLS (canonical link) The Newton update can be written in two equivalent ways:

$$\boldsymbol{\beta}_{l+1} = \boldsymbol{\beta}_l + (\boldsymbol{X}^\top W_l \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{\mu}_l) = (\boldsymbol{X}^\top W_l \boldsymbol{X})^{-1}\boldsymbol{X}^\top W_l \boldsymbol{z}_l,$$

where

$$\boldsymbol{z}_l = \boldsymbol{\theta}_l + W_l^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_l), \qquad \boldsymbol{\theta}_l = \boldsymbol{X}\boldsymbol{\beta}_l, \quad \boldsymbol{\mu}_l = g^{-1}(\boldsymbol{\eta}_l).$$

The second form suggests an *iteratively reweighted least squares* (IRLS) algorithm: at each step, regress the pseudo-response $\boldsymbol{z}_l$ on $\boldsymbol{X}$ with weights W_l.

9.5.4 General Link and Fisher Scoring

General Link Function When g is not canonical, the algebra becomes slightly more involved but the basic ideas remain. For a general link,

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_j} &= \frac{1}{\phi}\sum_{i=1}^{n}(Y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \\
&= \frac{1}{\phi}\sum_{i=1}^{n}(Y_i - A'(\theta_i))\psi'(\mu_i)g'(\mu_i)^{-1}x_{ij} \\
&= \frac{1}{\phi}\sum_{i=1}^{n}(Y_i - \mu_i)\frac{1}{V(\mu_i)}\frac{1}{g'(\mu_i)}x_{ij}
\end{aligned}$$

This expression is obtained by applying the chain rule to a single contribution

$$\ell_i(\boldsymbol{\beta}) = \frac{1}{\phi}\big\{Y_i\theta_i - A(\theta_i)\big\}, \qquad \theta_i = \psi(\mu_i), \mu_i = g^{-1}(\eta_i), \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

and using $A'(\theta_i) = \mu_i$ and $\psi'(\mu_i) = 1/V(\mu_i)$. To compute $\psi'(\mu)$, differentiate the identity $\mu = A'(\theta)$ with respect to μ:

$$1 = \frac{\mathrm{d}\mu}{\mathrm{d}\mu} = \frac{\mathrm{d}}{\mathrm{d}\mu}A'(\theta) = A''(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}\mu} = A''(\theta)\psi'(\mu) \quad \Longrightarrow \quad \psi'(\mu) = \frac{1}{A''(\theta)} = \frac{1}{V(\mu)}.$$

Differentiating again and keeping track of the dependence on μ_i leads to

$$\begin{aligned} H_{jk} &= \frac{1}{\phi}\sum_{i=1}^{n}\left[\frac{\mathrm{d}\eta_i}{\mathrm{d}\beta_k}\frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i}\frac{\mathrm{d}}{\mathrm{d}\mu_i}\left((Y_i-\mu_i)\frac{1}{V(\mu_i)}\frac{1}{g'(\mu_i)}\right)x_{ij}\right] \\ &= \frac{1}{\phi}\sum_{i=1}^{n}\left[\left(x_{ik}\frac{1}{g'(\mu_i)}\right)\frac{\mathrm{d}}{\mathrm{d}\mu_i}\left(\frac{Y_i-\mu_i}{V(\mu_i)g'(\mu_i)}\right)x_{ij}\right], \end{aligned}$$

where

$$\frac{\mathrm{d}}{\mathrm{d}\mu_i}\left(\frac{Y_i-\mu_i}{g'(\mu_i)V(\mu_i)}\right) = -\frac{1}{g'(\mu_i)V(\mu_i)} + (Y_i-\mu_i)\frac{\mathrm{d}}{\mathrm{d}\mu_i}\left(\frac{1}{g'(\mu_i)V(\mu_i)}\right)$$

The inner derivative $\frac{\mathrm{d}}{\mathrm{d}\mu_i}(\cdot)$ accounts for how the weight factor changes as the mean μ_i moves along the link g. The exact expression is somewhat messy and, crucially, depends on the data Y_i.

General Link—Newton–Raphson The full Newton step would use this observed Hessian. However, because

$$\mathbb{E}\left[(Y_i-\mu_i)\frac{\mathrm{d}}{\mathrm{d}\mu_i}\left(\frac{1}{g'(\mu_i)V(\mu_i)}\right)\right] = 0,$$

its expectation simplifies considerably. This observation motivates replacing the stochastic Hessian by its expectation.

General Link—Fisher Scoring The expected Hessian is

$$\mathbb{E}[H] = -\frac{1}{\phi}\boldsymbol{X}^\top W \boldsymbol{X}, \qquad W = \operatorname{diag}\left(\frac{1}{g'(\mu_i)^2 V(\mu_i)}\right).$$

The **Fisher information** is then

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}[-H] = \frac{1}{\phi}\boldsymbol{X}^\top W \boldsymbol{X}.$$

Using $\mathbb{E}[H]$ instead of H gives the **Fisher scoring** update, which is algebraically identical to IRLS (with the appropriate W determined by g and V). Under canonical links, the observed and expected information coincide, so Newton and Fisher scoring are the same.

9.5.5 Dispersion Estimation

Estimation of the dispersion parameter An important feature of GLMs is that the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ does not depend on the dispersion parameter ϕ. However, ϕ enters the variance $\operatorname{Var}(Y_i) = V(\mu_i)\phi$ and therefore is needed for standard errors, tests, and confidence intervals. In practice, ϕ is usually estimated via the **Pearson chi-squared statistic**

$$\mathcal{X}^2 = \sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

where $\widehat{\mu}_i$ is the fitted mean under the GLM and $V(\mu)$ is the variance function of the chosen family. It is convenient to consider the *scaled* Pearson statistic

$$\mathcal{X}_s^2 = \frac{\mathcal{X}^2}{\phi}, \qquad \text{(recall that } \operatorname{Var}(Y_i) = V(\mu_i)\phi\text{)},$$

which, under the model and for large n, satisfies

$$\mathcal{X}_s^2 \approx \chi^2_{n-p},$$

with $n-p$ degrees of freedom, where p is the number of fitted parameters. This suggests the asymptotically unbiased estimator

$$\widehat{\phi} = \frac{\mathcal{X}^2}{n-p},$$

obtained by matching $\mathcal{X}^2$ to the mean of a χ^2_{n-p} distribution. This estimator reduces to the usual unbiased estimator of σ^2 in the Gaussian linear model and generalizes it to the GLM setting.

Example 9.4 (Linear regression)**.** For the normal model, $V(\mu_i) = 1$ and $\phi = \sigma^2$, so

$$\widehat{\phi} = \widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

is the familiar unbiased estimator of the error variance.

Example 9.5 (Logistic regression)**.** For Bernoulli data, $V(\mu_i) = \mu_i(1-\mu_i)$ and the model assumes $\phi = 1$. Nevertheless,

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i - \widehat{\mu}_i)^2}{\mu_i(1-\mu_i)}$$

is useful diagnostically: values far from 1 can indicate overdispersion or underdispersion relative to the nominal binomial variance.

9.6 Statistical Inference

Statistical Inference Once a GLM has been fitted, we still need to answer inferential questions: How variable is $\widehat{\boldsymbol{\beta}}$? Which predictors are important? Is the model adequate? For GLMs, these questions are typically addressed through large-sample (asymptotic) theory:

- asymptotic normality of the MLE;
- hypothesis testing for linear combinations of coefficients;
- construction of confidence intervals and regions.

The basic tools are the Wald test, the likelihood ratio test, and the score test.

General Linear Hypothesis Many hypotheses of interest can be written as linear constraints on $\boldsymbol{\beta}$:

- $H_0 : \beta_j = \beta_j^*, \quad H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$,
- $H_0 : \beta_i = \beta_j, \quad H_0 : \beta_1 = \beta_2 = \beta_3 = 0$,
- **General form:** $H_0 : C\boldsymbol{\beta} = \boldsymbol{r}$ with $C \in \mathbb{R}^{q\times p}$ full rank.

We consider three tests that provide different ways of assessing such hypotheses.

9.6.1 Wald's test

To introduce the Wald test, we first recall the large-sample distribution of the maximum likelihood estimator (MLE) in a general GLM. The MLE $\widehat{\boldsymbol{\beta}}$ is a p-dimensional random vector. Under suitable regularity conditions, it is asymptotically normal.

Asymptotic normality of the MLE

As $n \to \infty$, we have

$$\widehat{\boldsymbol{\beta}} \approx N\big(\boldsymbol{\beta}, \mathcal{I}^{-1}(\boldsymbol{\beta})\big),$$

where $\mathcal{I}^{-1}(\boldsymbol{\beta})$ is the inverse of the Fisher information matrix.

In practice, the unknown Fisher information $\mathcal{I}(\boldsymbol{\beta})$ is replaced by its plug-in estimate evaluated at the MLE. For a GLM with dispersion parameter ϕ, this takes the familiar form

$$\widehat{\mathcal{I}} = \mathcal{I}(\widehat{\boldsymbol{\beta}}) = \frac{1}{\phi} X^\top \widehat{W} X,$$

where $\widehat{W}$ is the working weight matrix evaluated at $\widehat{\boldsymbol{\beta}}$.

When we are interested in testing a *single* component of $\boldsymbol{\beta}$, the asymptotic normality of the corresponding marginal coordinate can be used directly. If the dispersion parameter ϕ is known, this leads to a large-sample z-test for that coordinate. If ϕ is unknown and estimated from the data, the resulting statistic is typically interpreted as approximately t-distributed with an appropriate number of degrees of freedom.

For testing *several* components of $\boldsymbol{\beta}$ simultaneously, however, the marginal approach is no longer sufficient. Instead we work with a quadratic form in the joint asymptotic distribution of $\widehat{\boldsymbol{\beta}}$, which leads to the Wald test.

Construction of the Wald test. Suppose we want to test a general linear hypothesis

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{r},$$

where C is a full-rank $q \times p$ matrix with $q < p$, and $\boldsymbol{r}$ is a given q-vector. The construction of the Wald statistic relies on a standard fact about quadratic forms of multivariate normal vectors.

Fact (quadratic form of a multivariate normal)

For a q-dimensional normal random vector $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, V)$ with nonsingular covariance matrix V, we have

$$(\boldsymbol{Y} - \boldsymbol{\mu})^\top V^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_q.$$

Under H_0, we have $C\boldsymbol{\beta} = \boldsymbol{r}$, and the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ implies that $C\widehat{\boldsymbol{\beta}}$ is approximately $N(C\boldsymbol{\beta}, C\mathcal{I}^{-1}(\boldsymbol{\beta})C^\top)$. Replacing the unknown matrix $\mathcal{I}(\boldsymbol{\beta})$ by its plug-in estimate $\widehat{\mathcal{I}}$, we define the **Wald test statistic**

$$W \;=\; (C\widehat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \left[C\widehat{\mathcal{I}}^{-1} C^\top \right]^{-1} (C\widehat{\boldsymbol{\beta}} - \boldsymbol{r}).$$

Under H_0 and for large n, this statistic satisfies

$$W \approx \chi^2_q.$$

We reject H_0 if the observed value of W is large, that is, if it exceeds the $(1-\alpha)$-quantile of the χ^2_q distribution.

The Fisher information $\mathcal{I}(\boldsymbol{\beta})$ may also depend on a dispersion parameter ϕ. In GLMs such as logistic regression and Poisson regression, the dispersion parameter is fixed at $\phi = 1$. In models with unknown dispersion, such as linear regression, one plugs in a consistent estimate $\widehat{\phi}$ of ϕ into $\widehat{\mathcal{I}}$; the resulting Wald statistic is still asymptotically χ^2_q.

Example 9.6 (Linear regression)**.** Consider the classical linear regression model

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

and suppose we wish to test the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.

In this case, we can write $C = I_p$ and $\boldsymbol{r} = \boldsymbol{\beta}_0$. The Fisher information is

$$\mathcal{I}(\boldsymbol{\beta}) = \frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{X}.$$

If the variance σ^2 is known, the Wald statistic becomes

$$W = \frac{1}{\sigma^2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{X}^\top \boldsymbol{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \sim \chi^2_p \quad \text{under } H_0.$$

In practice, σ^2 is usually unknown and is estimated by the residual mean square

$$\widehat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

The corresponding statistic

$$\frac{1}{\widehat{\sigma}^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \sim F_{p,n-p} \quad \text{under } H_0,$$

is the familiar F-test for a linear hypothesis in regression.

The difference between these two forms is subtle but important: when σ^2 is known, the natural large-sample reference distribution is χ^2_p; when σ^2 is unknown and estimated from the data, the exact finite-sample distribution is $F_{p,n-p}$. As n grows large, the $F_{p,n-p}$ distribution converges to χ^2_p/p in a suitable sense, so the two perspectives are asymptotically equivalent. More precisely, when we plug in the estimator $\widehat{\sigma}^2$ (which is proportional to a χ^2_{n-p} random variable under the model), we should use the $F_{p,n-p}$ reference distribution; for large n, the chi-squared approximation is typically adequate.

9.6.2 Likelihood ratio test

We now turn to the likelihood ratio test, another fundamental large-sample test for general linear hypotheses in GLMs. Again consider the hypothesis

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{r},$$

where C is a $q \times p$ matrix.

Let $\widehat{\boldsymbol{\beta}}$ denote the unconstrained MLE of the GLM under the full model, and let $\widehat{\boldsymbol{\beta}}_0$ denote the MLE under the restricted model satisfying $C\boldsymbol{\beta} = \boldsymbol{r}$. The likelihood ratio compares the maximized likelihoods under these two models.

Likelihood ratio statistic. The **likelihood ratio statistic** is given by

$$\Lambda = \frac{L(\widehat{\boldsymbol{\beta}}_0)}{L(\widehat{\boldsymbol{\beta}})}, \qquad -2\log\Lambda = 2\log\frac{L(\widehat{\boldsymbol{\beta}})}{L(\widehat{\boldsymbol{\beta}}_0)} = 2\big[l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}_0)\big].$$

Under H_0 and suitable regularity conditions, Wilks' theorem states that

$$2\big[l(\widehat{\boldsymbol{\beta}}) - l(\widehat{\boldsymbol{\beta}}_0)\big] \approx \chi^2_q.$$

Thus $-2\log\Lambda$ is compared to the χ^2_q distribution, and we reject H_0 when the statistic is large.

If the dispersion parameter ϕ is unknown, one must use the *same* consistent estimator $\widehat{\phi}$ in both log-likelihoods $l(\widehat{\boldsymbol{\beta}})$ and $l(\widehat{\boldsymbol{\beta}}_0)$, so that the nuisance parameter cancels in the likelihood ratio. Under standard assumptions, this yields the same χ^2_q limit.

9.6.3 Score test

A third classical approach is based on the score function, i.e., the gradient of the log-likelihood with respect to the parameter.

Score function and its large-sample distribution. For a GLM with general link function g, the score function is

$$s(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \qquad s(\boldsymbol{\beta}) = (s_1(\boldsymbol{\beta}), \ldots, s_p(\boldsymbol{\beta}))^\top,$$

where the jth component can be written as

$$s_j(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n} x_{ij}\frac{Y_i - \mu_i}{g'(\mu_i)V(\mu_i)}.$$

Properties of the score

For any sample size n, we have

$$\mathbb{E}[s(\boldsymbol{\beta})] = \mathbf{0} \quad \text{and} \quad \mathrm{Cov}\big(s(\boldsymbol{\beta})\big) = \mathcal{I}(\boldsymbol{\beta}).$$

Furthermore, as $n \to \infty$,

$$s(\boldsymbol{\beta}) \approx N\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta})\big),$$

so the quadratic form

$$s(\boldsymbol{\beta})^\top \mathcal{I}^{-1}(\boldsymbol{\beta}) s(\boldsymbol{\beta}) \approx \chi^2_p.$$

To test $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, we evaluate the score and information at the null value, forming

$$S = s(\boldsymbol{\beta}_0)^\top \mathcal{I}^{-1}(\boldsymbol{\beta}_0) s(\boldsymbol{\beta}_0),$$

which is approximately χ^2_p under H_0 for large n. A key feature of the score test is that it only requires fitting the model under the null hypothesis: $\widehat{\boldsymbol{\beta}}$ under the full model is not needed.

Comparison of Wald, likelihood ratio, and score tests The Wald, likelihood ratio (LR), and score tests are all large-sample procedures derived from the same underlying likelihood framework, and under standard regularity conditions they are *asymptotically equivalent*: for fixed-dimensional hypotheses and large n, they yield the same rejection regions up to $o(1)$ differences in size and power.

In finite samples, however, their behavior can differ. Empirically and theoretically, the likelihood ratio and score tests often perform better than the Wald test when the sample size is modest or when the parameter is near the boundary of the parameter space. Intuitively, both the LR and score tests exploit more of the global shape of the likelihood function (as in profile likelihood confidence intervals), whereas the Wald test relies heavily on a local quadratic approximation around $\widehat{\boldsymbol{\beta}}$.

From a practical standpoint, the LR test is particularly attractive when the likelihood function can be computed easily: it only requires maximizing the likelihood under the full and restricted models and does not require explicit computation of the Fisher information. The score test is convenient when fitting the full model is difficult but the null model is simple. The Wald test is often the easiest to implement when reliable standard errors for $\widehat{\boldsymbol{\beta}}$ are readily available.

9.6.4 Goodness-of-fit tests

We now turn from testing specific linear hypotheses to assessing the overall adequacy of a GLM. Suppose we have data $(Y_i, \boldsymbol{x}_i)$ for $i = 1, 2, \ldots, n$. Conceptually, we can compare three kinds of models:

- The **saturated model**, in which the number of parameters equals the number of observations so that each observation has its own mean parameter μ_i.
- The **null model**, which uses a single parameter, so that $\mu_i = \mu$ for all i.
- The **GLM of interest**, which has p parameters and specifies $\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$.

Our goal is to select a model that describes the data well while using as few parameters as possible. To quantify the discrepancy between the observed responses $\boldsymbol{Y}$ and the fitted means $\widehat{\boldsymbol{\mu}}$ from a candidate model, two commonly used measures are **Pearson's chi-squared statistic** and the **deviance**.

Pearson's chi-squared statistic. The Pearson chi-squared statistic is defined by

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

where $V(\mu)$ is the variance function of the GLM and $\widehat{\mu}_i$ is the fitted mean for observation i. This generalizes the residual sum of squares (RSS) in linear regression.

Example 9.7 (Normal)**.** For a normal linear model with constant variance, $V(\mu_i) = \sigma^2$ and

$$\mathcal{X}^2 = \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i)^2 = RSS.$$

Example 9.8 (Poisson)**.** For a Poisson model with $V(\mu_i) = \mu_i$, we obtain

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}.$$

Example 9.9 (Binomial)**.** For a binomial model with $V(\mu_i) = \mu_i(1-\mu_i)$, we obtain

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i(1-\widehat{\mu}_i)}.$$

If the GLM is correctly specified, then $\mathcal{X}^2/(n-p)$ provides an estimate of the dispersion parameter ϕ:

- For the normal model, $RSS/(n-p) \approx \sigma^2 = \phi$.
- For canonical binomial models, $\mathcal{X}^2/(n-p) \approx 1 = \phi$.
- For Poisson models, $\mathcal{X}^2/(n-p) \approx 1 = \phi$.

Large values of $\mathcal{X}^2$ relative to its degrees of freedom indicate lack of fit.

Deviance. To define the deviance, recall that in an exponential family, the natural parameter θ_i is linked to the mean parameter μ_i via

$$\mu_i = A'(\theta_i), \quad \theta_i = \psi(\mu_i),$$

where A is the cumulant function and ψ is the inverse mean–parameter mapping. The log-likelihood can thus be expressed as *a function of the mean parameters*:

$$l(\boldsymbol{\mu}) = \frac{1}{\phi}\sum_{i=1}^{n}\big(\theta_i Y_i - A(\theta_i)\big) = \frac{1}{\phi}\sum_{i=1}^{n}\big(\psi(\mu_i)Y_i - A(\psi(\mu_i))\big).$$

In the saturated model, the MLE for μ_i is simply $\mu_i = Y_i$ (each observation is fitted exactly). In the GLM, the MLE for μ_i is given by the fitted mean

$$\widehat{\mu}_i = g^{-1}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}).$$

The deviance compares the fitted model to the saturated model. It is defined as

$$D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}) = 2\big[l(\boldsymbol{Y}) - l(\widehat{\boldsymbol{\mu}})\big] = \frac{2}{\phi}\sum_{i=1}^{n}\Big((\tilde{\theta}_i - \widehat{\theta}_i)Y_i - A(\tilde{\theta}_i) + A(\widehat{\theta}_i)\Big),$$

where $\tilde{\theta}_i = \psi(Y_i)$ (the natural parameter in the saturated model) and $\widehat{\theta}_i = \psi(\widehat{\mu}_i)$ (the natural parameter under the fitted GLM).

- The deviance is always non-negative, since the saturated model provides the largest possible log-likelihood among all models that assign one parameter per observation.
- In regular situations, the deviance is finite, but in some edge cases it can diverge.

Deviance and the likelihood ratio statistic. The deviance plays a central role in model comparison. For two nested GLMs, say a full model and a restricted model satisfying $C\boldsymbol{\beta} = \boldsymbol{r}$, the likelihood ratio statistic can be written as the difference of their deviances. Specifically, if D_{full} and D_{res} denote the deviances of the full and restricted models, then

$$2\big[l(\widehat{\boldsymbol{\mu}}_{\text{full}}) - l(\widehat{\boldsymbol{\mu}}_{\text{res}})\big] = D_{\text{res}} - D_{\text{full}}.$$

Thus the LR statistic is equivalently the reduction in deviance when moving from the restricted to the full model, and is asymptotically χ^2_q when the restriction imposes q constraints.

Under additional restrictive conditions (roughly, when the number of parameters p is fixed and n is large, and some regularity assumptions hold), the deviance of a correctly specified GLM can itself be approximated by a chi-squared distribution with $n-p$ degrees of freedom:

$$D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}) \approx \chi^2_{n-p}.$$

However, this approximation is not universally valid, in part because the degrees of freedom grow with n and the usual fixed-parameter asymptotics can break down.

Using deviance for goodness-of-fit. When the chi-squared approximation for the deviance is reliable, we can use $D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}})$ as a global goodness-of-fit statistic. A GLM with a deviance much larger than the typical range of a χ^2_{n-p} random variable provides a poor fit to the data. Conversely, a deviance that is "too small" relative to its degrees of freedom may indicate overfitting or unmodeled overdispersion in ϕ. In practice, both Pearson's chi-squared statistic and the deviance are examined when diagnosing GLM fit.

9.6.5 Nested model tests

We now formalize the use of deviance differences to compare nested models. Recall that in linear regression we compared nested models via differences in residual sum of squares. The same principle extends to GLMs through the deviance.

For a model $\mathcal{M}$, let $\mathcal{B}$ denote the corresponding parameter space for $\boldsymbol{\beta}$.

Example 9.10 (Model with $\beta_1 = 0$)**.** For the model with $\beta_1 = 0$, we have $\mathcal{B} = \{\boldsymbol{\beta} : \beta_1 = 0\}$.

Example 9.11 (Model with $C\boldsymbol{\beta} = \boldsymbol{r}$)**.** For the model with $C\boldsymbol{\beta} = \boldsymbol{r}$, we have $\mathcal{B} = \{\boldsymbol{\beta} : C\boldsymbol{\beta} = \boldsymbol{r}\}$.

We write $\mathcal{M}_1 \subset \mathcal{M}_2$ if $\mathcal{B}_1 \subset \mathcal{B}_2$, that is, if model $\mathcal{M}_1$ is a special case of model $\mathcal{M}_2$.

Definition 9.1 (Nested model)**.** We say that a sequence of models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ is *nested* if

$$\mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots \subset \mathcal{B}_m.$$

Let D_i be the deviance of the ith model in a nested sequence. Because each successive model allows more flexibility, the deviance cannot increase as we move to a larger model.

Monotonicity of deviance

If $\mathcal{M}_1 \subset \mathcal{M}_2$, then $D_1 \geq D_2$.

We now focus on the common case where $\mathcal{B}_i$ is a linear subspace of $\mathbb{R}^p$ of dimension q_i, with

$$q_1 < q_2 < \cdots < q_m.$$

Example 9.12 (A special case)**.** In model i, let

$$\mathcal{B}_i = \{\boldsymbol{\beta} : \beta_{i+1} = \beta_{i+2} = \cdots = \beta_p = 0\},$$

so that we add parameters one at a time.

For nested models, the deviance of model i can be written as

$$D(\boldsymbol{Y}, \widehat{\boldsymbol{\mu}}_{(i)}) = 2\big[l(\boldsymbol{Y}) - l(\widehat{\boldsymbol{\mu}}_{(i)})\big],$$

where $\widehat{\boldsymbol{\mu}}_{(i)}$ is the vector of fitted means under model i.

Asymptotic distribution of deviance differences

For all $i < j$, the difference in deviance satisfies

$$D_i - D_j \approx \chi^2_{q_j - q_i},$$

for large samples.

Thus:

- The deviance D_i measures the lack of fit of model i relative to the saturated model.
- The difference $D_i - D_j$ measures the improvement in fit when moving from model i to model j, and its chi-squared approximation provides a natural test for whether the additional parameters in model j are needed.

Deviance-based hypothesis tests for nested models. Consider testing $H_0 : \mathcal{M}_i$ versus $H_1 : \mathcal{M}_j$, where $\mathcal{M}_i \subset \mathcal{M}_j$ and the corresponding parameter dimensions are p_i and p_j with $p_i < p_j$. The natural test statistic is the reduction in deviance

$$\Delta D = D_i - D_j = 2\big[l(\widehat{\boldsymbol{\mu}}_{(j)}) - l(\widehat{\boldsymbol{\mu}}_{(i)})\big] \approx \chi^2_{p_j - p_i}.$$

Here ΔD is exactly twice the log-likelihood ratio between models $\mathcal{M}_i$ and $\mathcal{M}_j$. Because the degrees of freedom $p_j - p_i$ are fixed (and typically small), the chi-squared approximation for ΔD is usually more accurate than for a single deviance D_i, whose degrees of freedom grow with n.

In practice:

- If ϕ is unknown, we plug in a consistent estimate $\widehat{\phi}$ into both models.
- We reject $\mathcal{M}_i$ in favor of $\mathcal{M}_j$ if the observed ΔD exceeds $\chi^2_{p_j - p_i, 1-\alpha}$.
- It is important to stress that rejecting $\mathcal{M}_i$ does not imply that $\mathcal{M}_j$ fits the data well; it only says that $\mathcal{M}_j$ provides a significantly better fit than $\mathcal{M}_i$.

Example 9.13 (Linear regression $Y_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$)**.** In the normal linear model, one can verify that the deviance is proportional to the residual sum of squares:

$$\begin{aligned} D &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})^2 \\ &= \frac{1}{\sigma^2} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \frac{1}{\sigma^2} \big(\boldsymbol{Y}^\top \boldsymbol{Y} - 2\widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y} + \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}}\big) \\ &= \frac{1}{\sigma^2} \Big(\boldsymbol{Y}^\top \boldsymbol{Y} - \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}}\Big) \\ &= \frac{1}{\sigma^2} \Big(\boldsymbol{Y}^\top \boldsymbol{Y} - \widehat{\boldsymbol{Y}}^\top \widehat{\boldsymbol{Y}}\Big) = \frac{1}{\sigma^2} RSS, \end{aligned}$$

because $\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{Y}$ by the normal equations for least squares. Thus deviance-based tests for nested normal linear models reduce to the usual F-tests based on sums of squares.

For two nested linear models $\mathcal{M}_0$ and $\mathcal{M}_1$ with $q_0 < q_1$ parameters, if the null model $\mathcal{M}_0$ is true, then

$$\Delta D = D_0 - D_1 = \frac{1}{\sigma^2} \Big(\widehat{\boldsymbol{\beta}}_{(1)}^\top \boldsymbol{X}^\top \boldsymbol{Y} - \widehat{\boldsymbol{\beta}}_{(0)}^\top \boldsymbol{X}^\top \boldsymbol{Y}\Big) \sim \chi^2_{q_1 - q_0}.$$

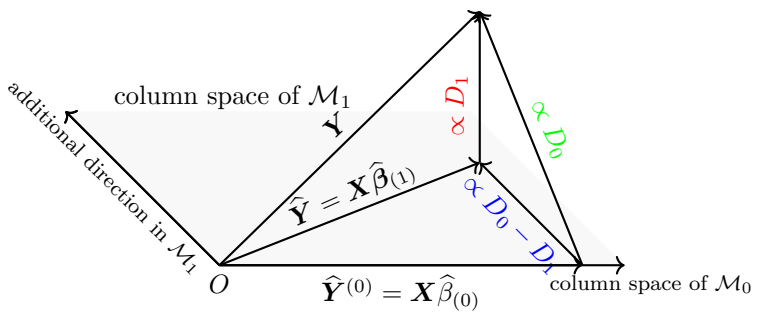
Because $\mathcal{M}_0 \subset \mathcal{M}_1$, the larger model $\mathcal{M}_1$ is also correct under H_0, and its deviance satisfies

$$D_1 = \frac{1}{\sigma^2} \Big(\boldsymbol{Y}^\top \boldsymbol{Y} - \widehat{\boldsymbol{\beta}}_{(1)}^\top \boldsymbol{X}^\top \boldsymbol{Y}\Big) \sim \chi^2_{n - q_1}.$$

Moreover, D_1 is independent of $D_0 - D_1$. Hence the ratio

$$\frac{D_0 - D_1}{q_1 - q_0} \Big/ \frac{D_1}{n - q_1} \sim F_{q_1 - q_0, n - q_1},$$

where *the unknown variance σ^2 cancels out.* This recovers the classical F-test for comparing nested linear regression models.



Example 9.14 (GLM with two groups of factors)**.** Suppose we consider a GLM with two sets of predictors, $\boldsymbol{x}_i \in \mathbb{R}^{p_1}$ and $\boldsymbol{z}_i \in \mathbb{R}^{p_2}$. Define the following models:

$$\begin{aligned} &\text{Model 1:} \quad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_1, \\ &\text{Model 2:} \quad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \boldsymbol{z}_i^\top \boldsymbol{\beta}_2. \end{aligned}$$

Then $\mathcal{M}_1 \subset \mathcal{M}_2$. More generally, we may consider the following sequence:

Predictor	Model	#parameters	Deviance	DF
Intercept only	$g(\mu) = \alpha_0$	1	D_0	$n-1$
Single factor $\boldsymbol{x}$	$g(\mu) = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}$	p_1	D_1	$n-p_1$
Single factor $\boldsymbol{z}$	$g(\mu) = \lambda_0 + \boldsymbol{z}^\top \boldsymbol{\lambda}$	p_2	D_2	$n-p_2$
Two factors	$g(\mu) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}_1 + \boldsymbol{z}^\top \boldsymbol{\beta}_2$	$p_3 = p_1 + p_2 - 1$	D_3	$n-p_3$

Goodness-of-fit tests based on deviance differences can then be constructed as follows:

Hypothesis	Effect to be detected	Test statistic	DF
$H_0 : \boldsymbol{\alpha} = 0$ vs. $H_1 : \boldsymbol{\alpha} \neq 0$	Effect of $\boldsymbol{x}$ ignoring $\boldsymbol{z}$	$D_0 - D_1$	$p_1 - 1$
$H_0 : \boldsymbol{\beta}_1 = 0$ vs. $H_1 : \boldsymbol{\beta}_1 \neq 0$	Effect of $\boldsymbol{x}$ with $\boldsymbol{z}$ in the model	$D_2 - D_3$	$p_1 - 1$
$H_0 : \boldsymbol{\lambda} = 0$ vs. $H_1 : \boldsymbol{\lambda} \neq 0$	Effect of $\boldsymbol{z}$ ignoring $\boldsymbol{x}$	$D_0 - D_2$	$p_2 - 1$
$H_0 : \boldsymbol{\beta}_2 = 0$ vs. $H_1 : \boldsymbol{\beta}_2 \neq 0$	Effect of $\boldsymbol{z}$ with $\boldsymbol{x}$ in the model	$D_1 - D_3$	$p_2 - 1$

These comparisons allow us to assess the marginal and conditional contributions of each predictor group to the overall fit.

Example 9.15 (Poisson regression for three-way contingency tables)**.** Finally, deviance-based nested model tests are widely used in log-linear Poisson regression for contingency tables. For example, in a three-way contingency table, one considers a hierarchy of log-linear models (independence, pairwise interactions, full three-way interaction, etc.), and the corresponding sequence of nested Poisson regression models can be visualized schematically. In such settings, deviance differences provide a natural way to test for the presence of specific interaction terms.

Reference Kevin Murphy, *Machine Learning: A Probabilistic Perspective.*
Annette Dobson, *An Introduction to Generalized Linear Models*, 2nd ed.
Germán Rodríguez, `https://data.princeton.edu/wws509`

10 Classification Methods

10.1 Introduction

Introduction We begin with a familiar supervised learning setup. Suppose we observe a sample

$$\{(\boldsymbol{X}_i, Y_i) : 1 \le i \le n\},$$

where $\boldsymbol{X}_i \in \mathbb{R}^p$ collects p predictors (or features), and Y_i is a class label taking values in a finite set $\mathcal{C}$ (for binary problems, $\mathcal{C} = \{0, 1\}$ or $\{-1, 1\}$). The goal of *classification* is to use the training sample to construct a rule that maps a future predictor vector $\boldsymbol{X}$ to a label in $\mathcal{C}$ as accurately as possible.

Classification rule

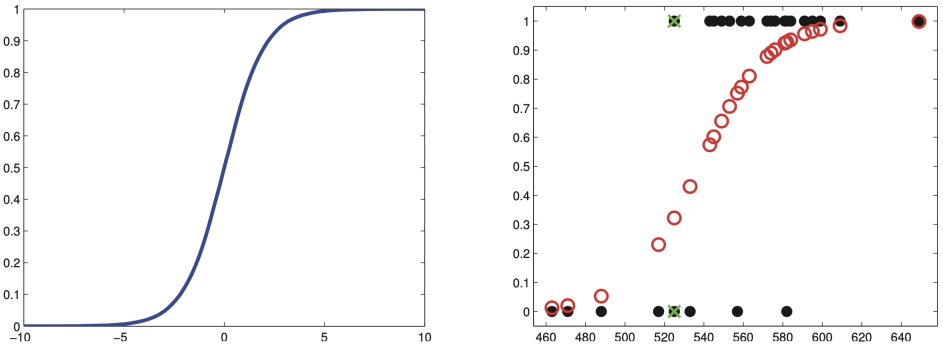
A classification rule is any mapping $\delta : \mathbb{R}^p \to \mathcal{C}$ that assigns a label $\delta(\boldsymbol{X})$ to each point $\boldsymbol{X}$.

Beyond this formal definition, a good classifier should be *interpretable* (we can understand why it predicts a certain label), *robust* (small perturbations of the data should not lead to wild changes), and *calibrated* when possible (its score relates sensibly to class probabilities). The methods in these notes strike different balances among these desiderata.

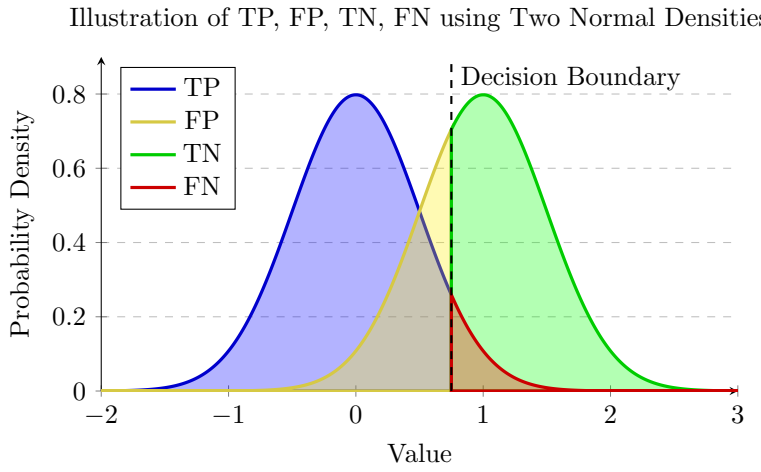
Logistic Regression as a Classifier A common entry point is logistic regression, which models the conditional probability of class "1" via a sigmoid transform of a linear score:

$$\mathbb{P}(Y = 1 \mid \boldsymbol{x}) = \mu(\boldsymbol{x}) = \frac{1}{1 + \exp(-\eta(\boldsymbol{x}))}, \qquad \eta(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}.$$

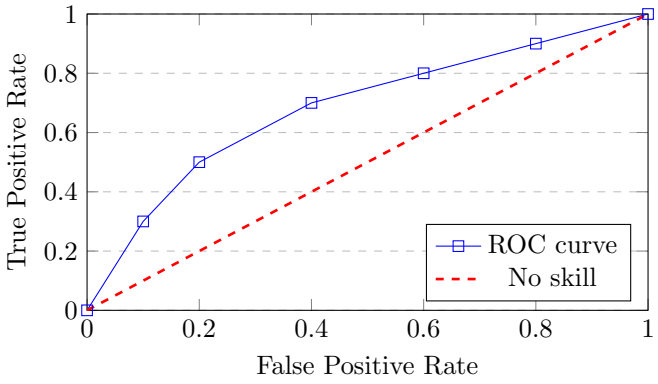
A natural (though not unique) decision rule is the 0.5 cutoff: predict $\widehat{Y} = 1$ if and only if $\mathbb{P}(Y = 1 \mid \boldsymbol{x}) > 0.5$. This threshold corresponds to equal misclassification costs; different costs lead to different cutoffs and therefore different points along the ROC curve.



Receiver operating characteristic (ROC) curve To appreciate the effect of moving the cutoff, imagine a one–dimensional score with two overlapping class-conditional densities. A vertical threshold partitions the axis into predicted negative/positive regions. Areas under the class-1 density to the right of the threshold are true positives, while areas under the class-0 density to the right are false positives.



ROC curve The ROC curve summarizes all cutoffs by plotting the *true positive rate* (TPR = TP/(TP + FN)) against the *false positive rate* (FPR = FP/(FP + TN)). A classifier that guesses at random lies on the diagonal; better classifiers bow towards the upper-left.



Bayes Rule The gold standard for classification under 0–1 loss is the Bayes rule.

Misclassification error

The expected 0–1 loss (risk) of a rule δ is

$$R(\delta) = \mathbb{P}(Y \neq \delta(\boldsymbol{X})) = \mathbb{E}_{\boldsymbol{X},Y}\big[\mathbb{1}\{Y \neq \delta(\boldsymbol{X})\}\big].$$

Bayes classifier

A Bayes rule predicts the most probable class given $\boldsymbol{X}$:

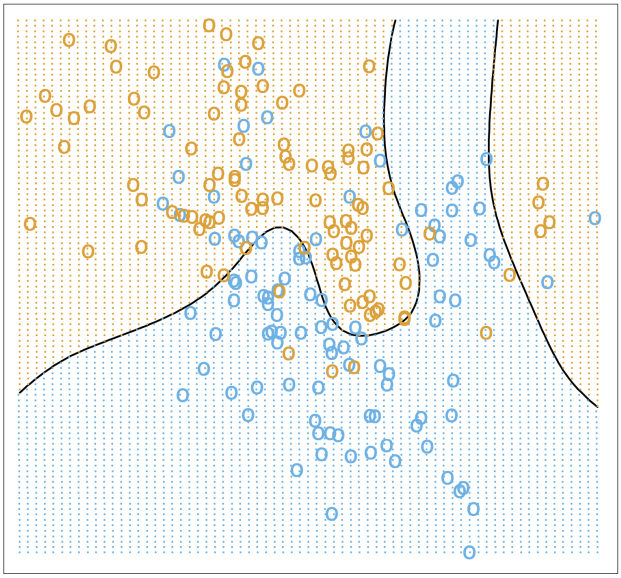
$$\delta^*(\boldsymbol{X}) = \arg\max_{c \in \mathcal{C}} \mathbb{P}(Y = c \mid \boldsymbol{X}).$$

The Bayes rule minimizes the misclassification error For the binary case $\mathcal{C} = \{0, 1\}$, write $\eta(\boldsymbol{X}) = \mathbb{P}(Y = 1 \mid \boldsymbol{X})$ and $\mathbb{P}(Y = 0 \mid \boldsymbol{X}) = 1 - \eta(\boldsymbol{X})$. Then

$$\begin{aligned}
R(\delta) &= \mathbb{E}_{\boldsymbol{X},Y}\big[\mathbb{1}\{Y \neq \delta(\boldsymbol{X})\}\big] = \mathbb{E}_{\boldsymbol{X}}\Big[\mathbb{E}_{Y|\boldsymbol{X}}\big[\mathbb{1}\{Y = 0, \delta(\boldsymbol{X}) = 1\} + \mathbb{1}\{Y = 1, \delta(\boldsymbol{X}) = 0\}\big]\Big] \\
&= \mathbb{E}_{\boldsymbol{X}}\big[(1 - \eta(\boldsymbol{X}))\mathbb{1}\{\delta(\boldsymbol{X}) = 1\} + \eta(\boldsymbol{X})\mathbb{1}\{\delta(\boldsymbol{X}) = 0\}\big] \\
&= \mathbb{E}_{\boldsymbol{X}}[\mathbb{P}(Y = 0 \mid \boldsymbol{X})\mathbb{1}(\delta(\boldsymbol{X}) = 1) + \mathbb{P}(Y = 1 \mid \boldsymbol{X})\mathbb{1}(\delta(\boldsymbol{X}) = 0)] \\
&= \mathbb{E}_{\boldsymbol{X}}\big[\eta(\boldsymbol{X}) + (1 - 2\eta(\boldsymbol{X}))\mathbb{1}\{\delta(\boldsymbol{X}) = 1\}\big].
\end{aligned}$$

For each fixed $\boldsymbol{X} = \boldsymbol{x}$, the integrand is minimized by predicting 1 whenever $1 - 2\eta(\boldsymbol{x}) < 0$, i.e., whenever $\eta(\boldsymbol{x}) > 1/2$. Therefore the pointwise minimizer—and hence the minimizer of the overall risk—is the Bayes rule $\delta^*(\boldsymbol{x}) = \mathbb{1}\{\eta(\boldsymbol{x}) > 1/2\}$.

Bayes Rule (visual intuition)



The optimal decision boundary occurs where the class posteriors are equal. In regions where the density of one class dominates, the risk of choosing that class is smallest.

Modeling the Bayes rule Directly applying δ^* requires $\mathbb{P}(Y = c \mid \boldsymbol{X})$, which is rarely available. The practical aim is to *approximate* δ^* as closely as possible from finite data, either by modeling the posteriors or by constructing discriminative rules whose behavior mimics the Bayes decision.

Bayes rule via Bayes formula Bayes' formula decomposes the posterior into class priors and class-conditional densities:

$$\mathbb{P}(Y = c_k \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{f(\boldsymbol{x})}, \text{ where}$$
$$\pi_k = \mathbb{P}(Y = c_k), \quad f_k(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid Y = c_k), \quad f(\boldsymbol{x}) = \sum_k \pi_k f_k(\boldsymbol{x}).$$

Thus, modeling posteriors is equivalent to modeling (π_k, f_k): priors can be estimated by $\widehat{\pi}_k = n_k/n$ and the f_k's by, for example, *kernel density estimators*, see Section 12.2 of Fan et al. (2020). This leads to model-based classifiers such as LDA/QDA.

10.2 LDA/QDA

Quadratic Discriminant Analysis A classical model-based approach posits multivariate normal class-conditional distributions:

$$\boldsymbol{X} \mid Y = c_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k).$$

Under this model, the log-posterior (up to an additive $\boldsymbol{x}$-independent constant) is

$$\delta_k^{\text{qda}}(\boldsymbol{x}) = \log \pi_k - \tfrac{1}{2} \log |\Sigma_k| - \tfrac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k),$$

and the QDA decision (Bayes rule) is $\arg\max_k \delta_k^{\text{qda}}(\boldsymbol{x})$.

- QDA boundaries are *quadratic* in $\boldsymbol{x}$;
- The term $(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)$ is the squared *Mahalanobis distance* to the class centroid.

Estimating QDA Plug-in estimates approximate the Bayes rule:

$$\widehat{\pi}_k = \frac{n_k}{n}, \qquad \widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{Y_i = c_k} \boldsymbol{X}_i, \qquad \widehat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{Y_i = c_k} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_k)^\top.$$

QDA

$$\arg\max_k \left\{ \log \widehat{\pi}_k - \tfrac{1}{2} \log |\widehat{\Sigma}_k| - \tfrac{1}{2} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k)^\top \widehat{\Sigma}_k^{-1} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_k) \right\}.$$

Linear Discriminant Analysis LDA assumes *homoscedasticity*:

$$\Sigma_k = \Sigma \quad \text{for all } k.$$

Terms independent of k cancel, leaving a linear score

$$\delta_k^{\text{lda}}(\boldsymbol{x}) = \log \pi_k + \boldsymbol{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \tfrac{1}{2} \boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k,$$

and the decision $\arg\max_k \delta_k^{\text{lda}}(\boldsymbol{x})$. Hence the name *linear* discriminant.

Estimating LDA The shared covariance is estimated by the pooled sample covariance

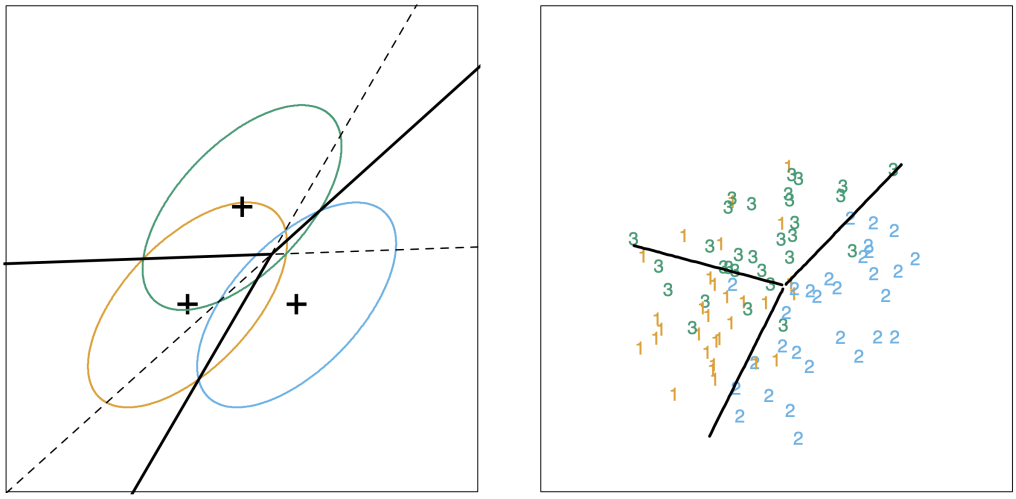
$$\widehat{\Sigma} = \frac{1}{\sum_{k=1}^K (n_k - 1)} \sum_{k=1}^K (n_k - 1) \widehat{\Sigma}_k,$$

leading to

LDA

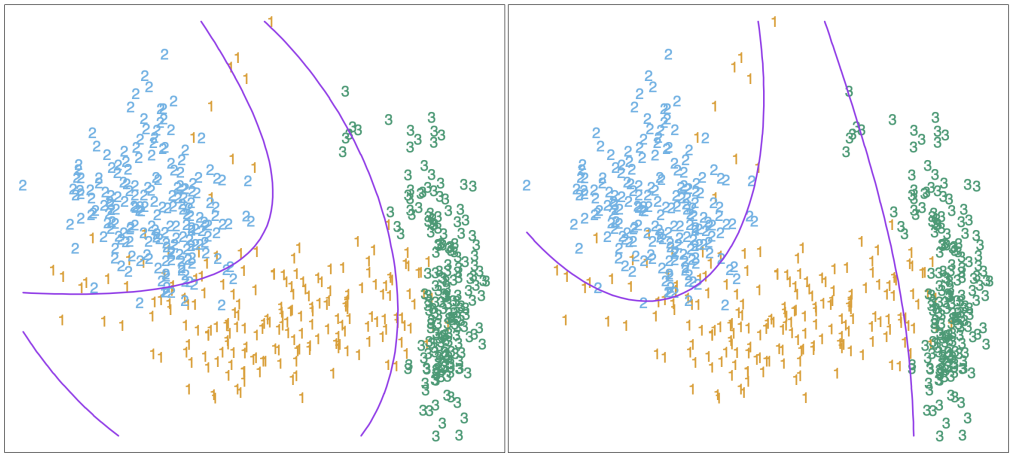
$$\arg\max_k \left\{ \log \widehat{\pi}_k + \boldsymbol{x}^\top \widehat{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_k - \tfrac{1}{2} \widehat{\boldsymbol{\mu}}_k^\top \widehat{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_k \right\}.$$

LDA is also known as *Fisher's discriminant analysis*. It is often more sample-efficient than purely discriminative methods when the Gaussian assumption is reasonable.



LDA

Regularized Discriminant Analysis When p is moderate-to-large relative to n, covariance estimates are noisy and $\widehat{\Sigma}^{-1}$ may be unstable (or undefined if $p > n$). *RDA* stabilizes estimation by shrinking towards a spherical structure:



(Left) LDA with quadratic boundary vs. (right) QDA. In many low-dimensional problems the two can be quite similar; QDA gains flexibility at the cost of estimating a covariance for each class.

RDA shrinkage

$$\widehat{\Sigma}^{\mathrm{rda}}(\gamma) = \gamma\widehat{\Sigma} + (1-\gamma)\frac{\operatorname{tr}(\widehat{\Sigma})}{p} I, \quad \widehat{\Sigma}_k^{\mathrm{rda}}(\alpha) = \alpha\widehat{\Sigma}_k + (1-\alpha)\widehat{\Sigma}^{\mathrm{rda}}(\gamma), \quad 0 \le \alpha, \gamma \le 1.$$

The tuning parameters (α, γ) are typically chosen by cross-validation.

Connection to Logistic Regression In the binary case $\mathcal{C} = \{0, 1\}$, the LDA log-odds are

$$\log \frac{\mathbb{P}(Y=1 \mid \boldsymbol{x})}{\mathbb{P}(Y=0 \mid \boldsymbol{x})} = \log \frac{\pi_1}{\pi_0} + \boldsymbol{x}^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x},$$

where $\boldsymbol{\beta} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$. Thus LDA implies a *particular* logistic model, and models $\mathbb{P}(\boldsymbol{X} \mid Y)$ explicitly as Gaussian. Logistic regression, by contrast, does not model $\mathbb{P}(\boldsymbol{X} \mid Y)$ and can be more robust under non-Gaussian features; LDA is more efficient when its modeling assumptions hold.

10.3 k-NN

The Nearest Neighbor Classifier Nearest neighbors are a localized approach: nearby points in predictor space tend to share labels. With a distance metric d (e.g. ℓ_q norms),

$$d(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|_q = \Big(\sum_{j=1}^{p} |x_j - x_j'|^q \Big)^{1/q},$$

we recover Manhattan ($q = 1$), Euclidean ($q = 2$), and Hamming ($q = 0$) distances.

k nearest neighbors

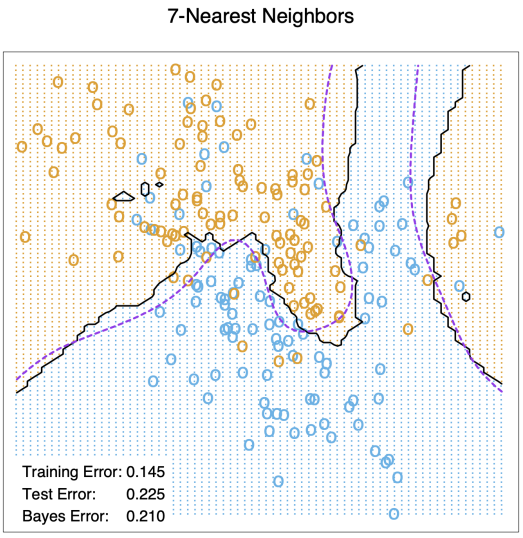
For any query $\boldsymbol{x}$, let $\mathcal{N}_k(\boldsymbol{x})$ be the set of k training points closest to $\boldsymbol{x}$ under d.

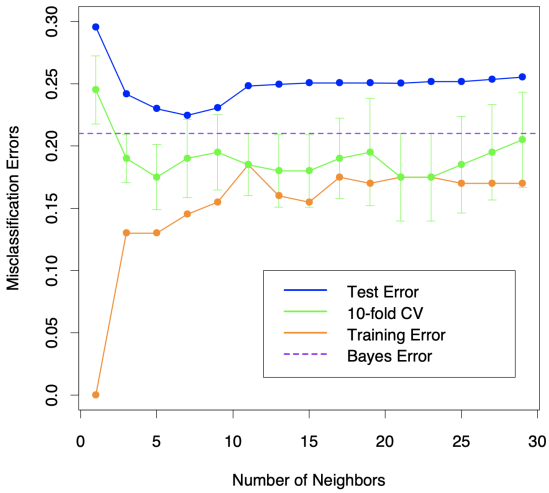
k-NN classifier

Predict by a majority vote among the neighbors:

$$\arg\max_{c_j \in \mathcal{C}} \sum_{i \in \mathcal{N}_k(\boldsymbol{x})} \mathbb{1}(Y_i = c_j).$$

Example 10.1 (k-NN approximating the Bayes rule)**.** As $k \to \infty$ slowly with n, k-NN averages local labels and can approach Bayes risk in benign settings.





Figures from ESL, Chapter 13.

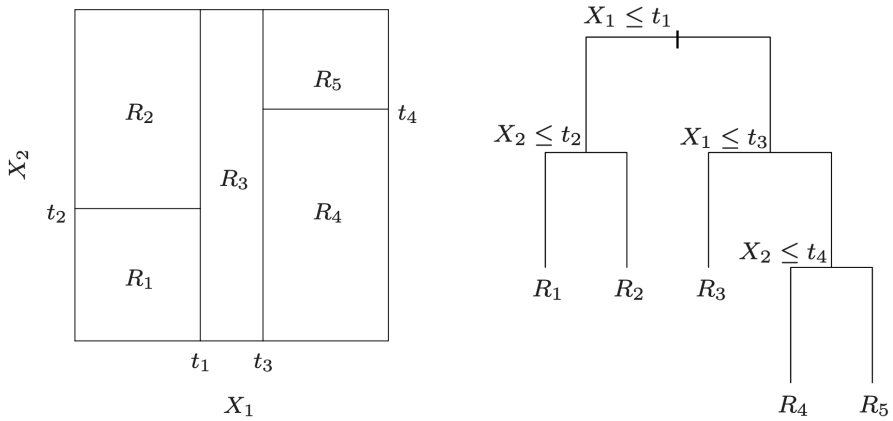
Remarks—k-NN **Advantages:** conceptually simple, no explicit training, and flexible decision boundaries. **Disadvantages:** slow queries (must search the training set), sensitivity to irrelevant features, and *curse of dimensionality*—in high p, "nearest" points may still be far away, degrading locality.

10.4 Tree-based Methods and Ensemble

We now turn to recursive partitioning and ensemble strategies for nonlinear classification boundaries.

10.4.1 Classification Trees

Classification Trees Tree-structured classifiers recursively partition the predictor space into regions (leaves) so that points within a leaf are predominantly of a single class. Similar to k-NN, the idea of classification tree is that predictors in the same leaf should be more likely to share the same label.



From partitions to predictions Given leaves $\{\mathcal{R}_1, \ldots, \mathcal{R}_S\}$, a query $\boldsymbol{x}$ falls in $\mathcal{R}(\boldsymbol{x})$ and the prediction is the leaf-wise majority:

$$\widehat{y} = \arg\max_{c_k \in \mathcal{C}} \sum_{\boldsymbol{X}_i \in \mathcal{R}(\boldsymbol{x})} \mathbb{1}(Y_i = c_k).$$

Trees may allow multiway splits, but binary splits are generally preferred: they are simpler, avoid fragmenting data too quickly, and can emulate multiway splits through sequences of binaries.

CART Classification and Regression Trees (CART) perform *binary splits* to *one variable at a time*:

Decision stump

For a node $\mathcal{R}$, a candidate split uses a feature j and threshold t:

$$\mathcal{R}_1(j,t) = \{\boldsymbol{X} \in \mathcal{R} : X_j \le t\}, \quad \mathcal{R}_2(j,t) = \{\boldsymbol{X} \in \mathcal{R} : X_j > t\}.$$

CART grows a tree greedily by choosing the split that most reduces node's *expected impurity*.

Impurity functions Let p_k be the proportion of class k inside node $\mathcal{R}$,

$$p_k = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{X}_i \in \mathcal{R}} \mathbb{1}(Y_i = c_k).$$

Two common impurities are

$$\text{Gini: } \mathrm{GI}(\mathcal{R}) = \sum_k p_k(1-p_k), \qquad \text{Cross-entropy: } \mathrm{CE}(\mathcal{R}) = -\sum_k p_k \log p_k.$$

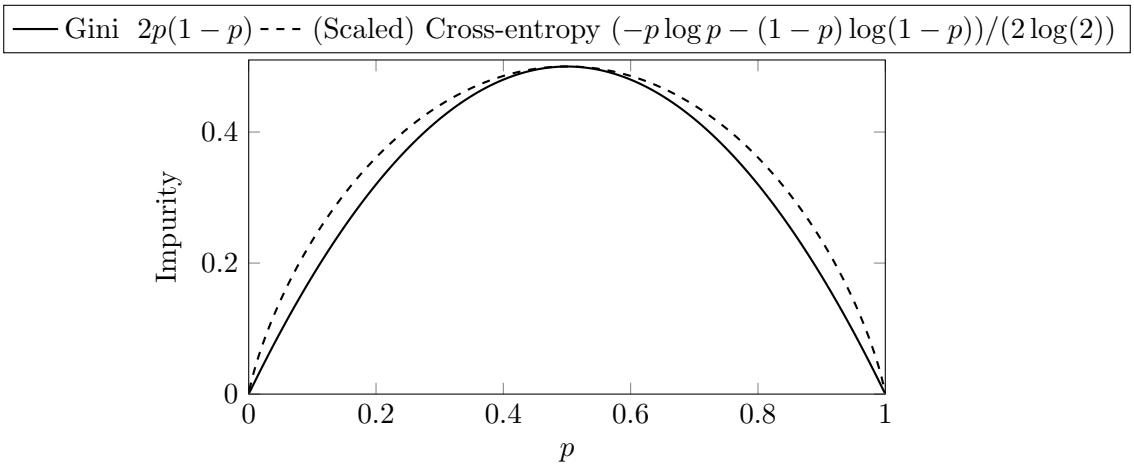


Figure 13: Impurity functions in the binary case ($p = p_1$). Both are maximized at $p = 0.5$ and minimized at $p = 0, 1$.

Remark 10.1. In CART, both the Gini index and the cross-entropy serve as impurity measures for evaluating candidate splits. For a node $\mathcal{R}$ with class proportions $(p_k)_k$, the Gini index $\mathrm{GI}(\mathcal{R}) = \sum_k p_k(1-p_k)$ and the cross-entropy $\mathrm{CE}(\mathcal{R}) = -\sum_k p_k \log p_k$ are both

- minimized when the node is pure; and
- maximized when the classes are evenly mixed ($p_i = 1/K$).

However, for very small p_k, we have

$$\mathrm{GI}(\mathcal{R}) \approx p_k \quad \text{whereas} \quad \mathrm{CE}(\mathcal{R}) \approx -p_k \log p_k,$$

so cross-entropy penalizes small minority probabilities more strongly. As a result, cross-entropy is often more sensitive to imbalanced nodes, while the Gini index tends to provide a slightly more conservative measure of impurity.

Optimal split

CART grows a tree greedily by choosing the split that most reduces node's *expected impurity*:

$$(j,t)^{\text{opt}} = \arg\min_{j,t} \left[\frac{|\mathcal{R}_1(j,t)|}{|\mathcal{R}|} F(\mathcal{R}_1(j,t)) + \frac{|\mathcal{R}_2(j,t)|}{|\mathcal{R}|} F(\mathcal{R}_2(j,t)) \right],$$

where F is GI or CE and

$$\mathcal{R}_1(j,t) = \{\boldsymbol{X} \in \mathcal{R} : X_j \le t\}, \quad \mathcal{R}_2(j,t) = \{\boldsymbol{X} \in \mathcal{R} : X_j > t\}.$$

Splitting until every leaf is pure yields a *fully grown* tree, which typically overfits.

Stopping and pruning Depth controls the bias–variance tradeoff. (Depth = longest path from root to a leaf.)

- If a tree is too shallow, it underfits (high bias).
- If a tree is too deep, it overfits (high variance).

CART therefore prunes a fully grown tree using a leaf-penalty $\alpha > 0$: select the subtree minimizing (empirical impurity + $\alpha\times$ #leaves). Cross-validation chooses α.

Regression Trees For data set $\{(\boldsymbol{X}_i, Y_i), 1 \le i \le n\}$, a regression problem asks for a function f such that $f(\boldsymbol{X}_i) \approx Y_i$. Tree methodology extends seamlessly to regression. Consider a piece-wise constant regression model:

$$f(\boldsymbol{x}) = \sum_{k=1}^{S} \beta_k \mathbb{1}(\boldsymbol{x} \in \mathcal{R}_k),$$

with leaf labels β_k equal to within-leaf sample means. Greedy splitting minimizes *squared error* within children:

$$(j,t)^{\text{opt}} = \arg\min_{j,t} \left[\frac{|\mathcal{R}_1(j,t)|}{|\mathcal{R}|} F(\mathcal{R}_1(j,t)) + \frac{|\mathcal{R}_2(j,t)|}{|\mathcal{R}|} F(\mathcal{R}_2(j,t)) \right],$$

where

$$F(\mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{X}_i \in \mathcal{R}} (Y_i - \widehat{\beta})^2, \quad \text{and} \quad \widehat{\beta} = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{X}_i \in \mathcal{R}} Y_i$$

and

$$\mathcal{R}_1(j,t) = \{\boldsymbol{X} \in \mathcal{R} : X_j \le t\}, \quad \mathcal{R}_2(j,t) = \{\boldsymbol{X} \in \mathcal{R} : X_j > t\}.$$

The prediction is

$$\widehat{f}(\boldsymbol{x}) = \sum_{k=1}^{S} \widehat{\beta}_k \mathbb{1}(\boldsymbol{x} \in \mathcal{R}_k).$$

Tree-based methods do not posit a fixed parametric form for the prediction (such as linearity); instead it approximates f by a data–adaptive, piecewise-constant function defined on a recursive partition of the predictor space. The effective number of parameters (one label/mean per terminal node, plus the split points) is not fixed in advance and can grow with the sample size as the tree becomes more complex. In this sense, CART is viewed as a nonparametric method.

- CART is highly interpretable: the prediction for a given input can be traced through a sequence of human-readable decision rules.
- Although CART is computationally efficient, it relies on a greedy search procedure and thus comes with limited theoretical guarantees of global optimality.
- Classification trees are notoriously unstable: small perturbations in the training data can lead to substantially different fitted trees.
- This instability can be mitigated by *combining* multiple trees to form a final classifier; such techniques are known collectively as *ensemble learning*.

10.4.2 Bagging

Bootstrap **agg**regat**ing** (Bagging) reduces variance by averaging many models trained on *bootstrap resamples*.

Bootstrap resamples

Given a dataset $\mathcal{D} = \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, a *bootstrap resample* is obtained by drawing n observations *with replacement* from $\mathcal{D}$. Formally, we sample indices $I_1, \dots, I_n$ independently and uniformly from $\{1, \dots, n\}$, and form the resampled dataset

$$\mathcal{D}^* = \{(\boldsymbol{X}_{I_j}, Y_{I_j})\}_{j=1}^n.$$

Bagging as a meta-algorithm Bagging is a *meta-algorithm* built on top of a given base learning procedure. Let g denote a base learner that maps a dataset to a fitted predictor. Given the training data

$$\mathcal{D} = \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n,$$

bagging proceeds as follows.

Bagging

1. Generate B bootstrap resamples

$$\mathcal{D}^{*(b)} = \{(\boldsymbol{X}_{I_j^{(b)}}, Y_{I_j^{(b)}})\}_{j=1}^n, \qquad b = 1, \dots, B,$$

   by sampling indices $I_1^{(b)}, \dots, I_n^{(b)}$ i.i.d. from $\{1, \dots, n\}$ with replacement.

2. Fit the base learner on each resample to obtain predictors

$$\widehat{f}^{(b)} = g\big(\mathcal{D}^{*(b)}\big), \qquad b = 1, \dots, B.$$

3. Aggregate predictions:

   - For classification with label set $\mathcal{C}$, use majority vote at a query point $\boldsymbol{x}$:

$$\widehat{y}^{\,\mathrm{bag}}(\boldsymbol{x}) = \arg\max_{c_k \in \mathcal{C}} \sum_{b=1}^{B} \mathbb{1}\big(\widehat{f}^{(b)}(\boldsymbol{x}) = c_k\big).$$

   - For regression, average the fitted values:

$$\widehat{f}^{\,\mathrm{bag}}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{f}^{(b)}(\boldsymbol{x}).$$

Bagging is most effective for base learners that are *unstable*, in the sense that small perturbations in the training data can cause large changes in the fitted predictor, with CART being a prime example. To see this in regression trees, let $\widehat{f}$ denote the predictor obtained by applying the base learner to the original sample, and $\widehat{f}^{*(b)}$ the predictors trained on bootstrap resamples. In an idealized setting where the $\widehat{f}^{*(b)}(\boldsymbol{x})$ are independent with variance $\mathrm{Var}^*(\widehat{f}^*(\boldsymbol{x}))$, the variance of their average decreases as

$$\mathrm{Var}^*\left(\frac{1}{B}\sum_{b=1}^{B} \widehat{f}^{*(b)}(\boldsymbol{x})\right) = \frac{1}{B}\,\mathrm{Var}^*(\widehat{f}^*(\boldsymbol{x})).$$

In practice the bootstrap predictors are correlated, so the reduction is less dramatic, but averaging still smooths out idiosyncratic fluctuations that arise from individual resamples. The bias of the procedure typically changes little, so bagging can substantially reduce mean squared error by trading a negligible increase in bias for a meaningful reduction in variance.

10.4.3 Random Forests

Bagging already improves on a single CART tree by averaging over many bootstrap trees, thereby reducing variance and stabilizing predictions. However, bagging does not fully address a key limitation: the individual trees in the ensemble are often *highly correlated*. If there is a very strong predictor (or a small set of dominant predictors), most bootstrap trees will repeatedly split on the same variables near the top of the tree. In the variance–reduction picture, averaging many highly correlated predictors does not reduce variance as effectively as averaging many decorrelated ones.

Random Forests (RF) build on bagging by deliberately injecting additional randomness into the tree-growing process to reduce this correlation. At each split, instead of searching over *all* predictors for the best split, a random forest restricts the search to a small, randomly chosen subset of features. This "feature subsampling" encourages different trees to explore different splitting structures, even when some predictors are very strong, thus increasing the diversity of the ensemble. The resulting trees are still low-bias (they are grown deeply, as in standard CART), but their predictions are less aligned with each other.

Random forests

Given the training data $\mathcal{D} = \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, a random forest is constructed as follows.

1. Generate B bootstrap resamples

$$\mathcal{D}^{*(b)} = \{(\boldsymbol{X}_{I_j^{(b)}}, Y_{I_j^{(b)}})\}_{j=1}^n, \qquad b = 1, \dots, B,$$

   by sampling indices $I_1^{(b)}, \dots, I_n^{(b)}$ i.i.d. from $\{1, \dots, n\}$ with replacement.

2. For each bootstrap resample $\mathcal{D}^{*(b)}$, grow a randomized CART tree $\widehat{f}^{(b)}$ by:
   (a) at each splitting step, randomly select a subset of m predictors;
   (b) among these m predictors, find the best split for a chosen impurity measure;
   (c) repeat recursively until the tree is fully grown (or until a stopping rule is met).

3. For a query point $\boldsymbol{x}$, let $\widehat{f}^{(b)}(\boldsymbol{x})$ be the prediction of the bth tree,

$$\widehat{y}^{\mathrm{RF}}(\boldsymbol{x}) = \arg\max_{c_k \in \mathcal{C}} \sum_{b=1}^{B} \mathbb{1}\big(\widehat{f}^{(b)}(\boldsymbol{x}) = c_k\big).$$

From a bias–variance viewpoint, random forests aim to preserve the low bias of deep trees while further lowering variance relative to plain bagging, by averaging many *less correlated* trees. Empirically, this often translates into better predictive performance and more robust behavior, especially in high-dimensional problems where many predictors are noisy or redundant.

10.4.4 Boosting

Boosting

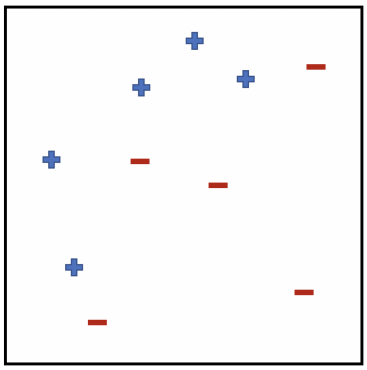
Boosting asks whether a highly accurate *strong* learner can be constructed by aggregating many *weak* learners (such as decision stumps or shallow trees) whose individual performance is only slightly better than random guessing.

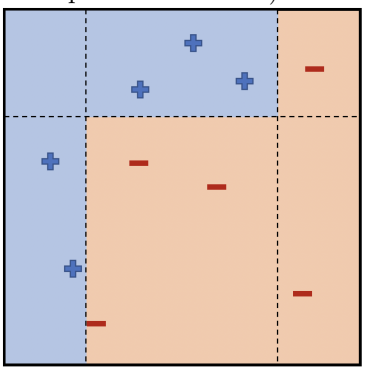
Adaptive **Boost**ing (Freund & Schapire, 1996) implements this idea by fitting weak learners sequentially to *reweighted versions of the training data*. Operationally, AdaBoost proceeds as follows:

- Base learners are fitted *sequentially* on reweighted data:
  - the weights of misclassified observations are inflated,
  - the weights of correctly classified observations are shrunk,
  - the guiding idea is to encourage future learners to focus on examples that previous learners have misclassified.

- Each base learner is assigned a weight (or "credibility") that depends on its classification error.
- Final predictions are obtained by a weighted majority vote over all fitted base learners.

Example 10.2 (Decision stumps and AdaBoost).





AdaBoost

1. Initialize weights $w_i = 1/n$ and recode $Y_i \in \{-1, 1\}$.
2. For $m = 1, \dots, M$:
   a. Fit classifier $C_m(\boldsymbol{x})$ to minimize *weighted* error $\sum_i w_i \mathbb{1}\{Y_i \neq C_m(\boldsymbol{X}_i)\}$.
   b. Compute error $\mathrm{err}^{(m)} = \frac{\sum_i w_i \mathbb{1}\{Y_i \neq C_m(\boldsymbol{X}_i)\}}{\sum_i w_i}$.
   c. Set learner weight $\alpha_m = \log \frac{1-\mathrm{err}^{(m)}}{\mathrm{err}^{(m)}}$.
   d. Update $w_i \leftarrow w_i \exp\big(\alpha_m \mathbb{1}\{Y_i \neq C_m(\boldsymbol{X}_i)\}\big)$.
3. Final prediction: $\widehat{Y} = \mathrm{sign}\left(\sum_{m=1}^{M} \alpha_m C_m(\boldsymbol{x})\right)$.

Loss-function view of AdaBoost For $Y \in \{-1, 1\}$ and score $f(\boldsymbol{x})$ with rule $\mathrm{sign}(f(\boldsymbol{x}))$, consider the *expected exponential loss*

$$L(f) = \mathbb{E}\big[\exp(-Y f(\boldsymbol{X}))\big] = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{P}(Y = 1 \mid \boldsymbol{X}) e^{-f(\boldsymbol{X})} + \mathbb{P}(Y = -1 \mid \boldsymbol{X}) e^{f(\boldsymbol{X})}\right].$$

The minimizer is

$$f^*(\boldsymbol{x}) = \tfrac{1}{2} \log\left(\frac{\mathbb{P}(Y = 1 \mid \boldsymbol{x})}{\mathbb{P}(Y = -1 \mid \boldsymbol{x})}\right),$$

whose sign is the Bayes rule.

Proof.

$$\begin{aligned} L(f) &= \mathbb{E}\big[\exp(-Y f(\boldsymbol{X}))\big] \\ &= \mathbb{E}_{\boldsymbol{X}}\Big[\underbrace{\eta(\boldsymbol{X}) e^{-f(\boldsymbol{X})} + \big(1 - \eta(\boldsymbol{X})\big) e^{f(\boldsymbol{X})}}_{\ell(f(\boldsymbol{X});\, \eta(\boldsymbol{X}))}\Big], \quad \text{where } \eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid \boldsymbol{x}). \end{aligned}$$

For each fixed $\boldsymbol{x}$, minimizing $L(f)$ is equivalent to minimizing

$$\ell(u;\eta) = \eta e^{-u} + (1-\eta)e^{u} \quad \text{over } u \in \mathbb{R}.$$

Differentiating and setting to zero,

$$\frac{\partial \ell}{\partial u} = -\eta e^{-u} + (1-\eta)e^{u} = 0 \implies (1-\eta)e^{u} = \eta e^{-u} \implies (1-\eta)e^{2u} = \eta \implies u^{*} = \tfrac{1}{2}\log\left(\frac{\eta}{1-\eta}\right),$$

so

$$f^{*}(\boldsymbol{x}) = \tfrac{1}{2}\log\left(\frac{\mathbb{P}(Y=1 \mid \boldsymbol{x})}{\mathbb{P}(Y=-1 \mid \boldsymbol{x})}\right). \qquad \square$$

Loss function interpretation of AdaBoost Minimizing $L(f)$ over all functions f is essentially as hard as constructing the Bayes rule itself. In practice, we therefore approximate the minimizer by restricting attention to a simpler class of functions and using an iterative, greedy procedure.

First, we constrain f to lie in a finite-dimensional linear span of base classifiers:

$$\mathcal{F}_M = \left\{ f(\boldsymbol{x}) : f(\boldsymbol{x}) = \sum_{m=1}^{M} \beta_m C_m(\boldsymbol{x}) \right\},$$

where each $C_m(\boldsymbol{x})$ is a classifier (a "weak learner") taking values in $\{-1, 1\}$ and the coefficients $\beta_m > 0$. Within this class, we approximate the minimizer of the (empirical) exponential loss by forward stagewise (greedy) fitting.

We start from the constant function $f^{(0)}(\boldsymbol{x}) \equiv 0$. At iteration m, given the current function $f^{(m-1)}$, we update

$$f^{(m)}(\boldsymbol{x}) = f^{(m-1)}(\boldsymbol{x}) + \alpha^{(m)} C_m(\boldsymbol{x}),$$

where $(\alpha^{(m)}, C_m)$ is chosen to minimize the empirical exponential loss:

$$\begin{aligned}(\alpha^{(m)}, C_m) &= \arg\min_{\alpha, C} \frac{1}{n}\sum_{i=1}^{n} \exp\big\{-Y_i\big[f^{(m-1)}(\boldsymbol{X}_i) + \alpha C(\boldsymbol{X}_i)\big]\big\} \\ &= \arg\min_{\alpha, C} \sum_{i=1}^{n} w_i^{(m)} \exp\big\{-Y_i \alpha C(\boldsymbol{X}_i)\big\},\end{aligned}$$

with

$$w_i^{(m)} = \frac{1}{n}\exp\big\{-Y_i f^{(m-1)}(\boldsymbol{X}_i)\big\}$$

interpreted as the current observation weights.

(AdaBoost Step 2.a) For fixed α, the optimal classifier C_m minimizes a weighted misclassification error. Indeed,

$$\begin{aligned}\sum_{i=1}^{n} w_i^{(m)} \exp\big\{-Y_i \alpha C(\boldsymbol{X}_i)\big\} &= \sum_{i=1}^{n} w_i^{(m)} \Big[e^{-\alpha}\mathbb{1}\big(Y_i = C(\boldsymbol{X}_i)\big) + e^{\alpha}\mathbb{1}\big(Y_i \neq C(\boldsymbol{X}_i)\big)\Big] \\ &= e^{-\alpha}\sum_{i=1}^{n} w_i^{(m)} + \big(e^{\alpha} - e^{-\alpha}\big)\sum_{i=1}^{n} w_i^{(m)}\mathbb{1}\big(Y_i \neq C(\boldsymbol{X}_i)\big),\end{aligned}$$

so for any fixed α the term $e^{-\alpha}\sum_i w_i^{(m)}$ is constant in C, and the minimizer is

$$C_m = \arg\min_{C} \sum_{i=1}^{n} w_i^{(m)}\mathbb{1}\big(Y_i \neq C(\boldsymbol{X}_i)\big),$$

i.e., a weak learner trained to minimize the weighted classification error.

(AdaBoost Step 2.b and 2.c) Given C_m as above, the optimal step size $\alpha^{(m)}$ that minimizes the one-dimensional function

$$\alpha \mapsto \sum_{i=1}^{n} w_i^{(m)} \exp\big\{-Y_i \alpha C_m(\boldsymbol{X}_i)\big\}$$

turns out to be $\alpha^{(m)} = \alpha_m/2 = \log\frac{1-\mathrm{err}^{(m)}}{\mathrm{err}^{(m)}}/2$ and $\mathrm{err}^{(m)} = \frac{\sum_i w_i \mathbb{1}\{Y_i \neq C_m(\boldsymbol{X}_i)\}}{\sum_i w_i}$.

Since $Y_i C_m(\boldsymbol{X}_i) = 1$ for correct and -1 for incorrect classifications, the objective is

$$\sum_{i=1}^{n} w_i^{(m)} e^{-Y_i \alpha C_m(\boldsymbol{X}_i)} = e^{-\alpha} \sum_{Y_i = C_m(\boldsymbol{X}_i)} w_i^{(m)} + e^{\alpha} \sum_{Y_i \neq C_m(\boldsymbol{X}_i)} w_i^{(m)} =: e^{-\alpha} W_c + e^{\alpha} W_m,$$

so setting $\frac{\mathrm{d}}{\mathrm{d}\alpha}(e^{-\alpha} W_c + e^{\alpha} W_m) = 0$ gives

$$-e^{-\alpha} W_c + e^{\alpha} W_m = 0 \implies e^{2\alpha} = \frac{W_c}{W_m} \implies \alpha^{(m)} = \frac{1}{2} \log \frac{W_c}{W_m} = \frac{1}{2} \log \frac{1 - \mathrm{err}^{(m)}}{\mathrm{err}^{(m)}},$$

where $\mathrm{err}^{(m)} = W_m / (W_c + W_m)$.

(AdaBoost Step 2.d) Finally, plugging the updated function

$$f^{(m)}(\boldsymbol{x}) = f^{(m-1)}(\boldsymbol{x}) + \frac{\alpha_m}{2} C_m(\boldsymbol{x})$$

into the weight definition

$$w_i^{(m+1)} = \frac{1}{n} \exp\{-Y_i f^{(m)}(\boldsymbol{X}_i)\},$$

we have

$$w_i^{(m+1)} = \frac{1}{n} \exp\{-Y_i f^{(m)}(\boldsymbol{X}_i)\} = \frac{1}{n} \exp\Big\{-Y_i \big(f^{(m-1)}(\boldsymbol{X}_i) + \tfrac{\alpha_m}{2} C_m(\boldsymbol{X}_i)\big)\Big\},$$

so

$$\begin{aligned} w_i^{(m+1)} &= \frac{1}{n} \exp\{-Y_i f^{(m-1)}(\boldsymbol{X}_i)\} \exp\Big\{-Y_i \tfrac{\alpha_m}{2} C_m(\boldsymbol{X}_i)\Big\} \\ &= w_i^{(m)} \cdot \exp\Big\{-Y_i \tfrac{\alpha_m}{2} C_m(\boldsymbol{X}_i)\Big\}. \end{aligned}$$

Since $Y_i, C_m(\boldsymbol{X}_i) \in \{-1, 1\}$, we have $Y_i C_m(\boldsymbol{X}_i) = 1 - 2\mathbb{1}\big(Y_i \neq C_m(\boldsymbol{X}_i)\big)$. Hence,

$$\exp\Big\{-Y_i \tfrac{\alpha_m}{2} C_m(\boldsymbol{X}_i)\Big\} = e^{-\alpha_m/2} \exp\big(\alpha_m \mathbb{1}(Y_i \neq C_m(\boldsymbol{X}_i))\big),$$

hence, up to the common factor $e^{-\alpha_m/2}$ (removed by renormalization),

$$w_i^{(m+1)} = w_i^{(m)} \cdot \exp\big(\alpha_m \mathbb{1}(Y_i \neq C_m(\boldsymbol{X}_i))\big),$$

which increases the weights of misclassified observations and decreases those of correctly classified ones.

Performance guarantee

Theorem 10.1 (Freund and Schapire, 1997)**.** *If weak learner C_m has accuracy $1 - \mathrm{err}^{(m)} = \frac{1}{2} + \gamma_m$ with $\gamma_m > 0$, then the training error of AdaBoost after M rounds is at most $\prod_{m=1}^{M} \sqrt{1 - 4\gamma_m^2} \leq \exp\big(-2 \sum_{m=1}^{M} \gamma_m^2\big)$.*

Remarks—Bagging, Random Forest, Boosting Bagging/RF are parallel (each tree independent), unlike AdaBoost (sequential). Bagging and RF mainly reduce variance; boosting often reduces bias (especially with simple weak learners) at the cost of some variance increase. All are *ensemble* methods.

10.5 SVM

This subsection develops margin-based classifiers, first in separable settings and then with regularization and kernels.

10.5.1 Separable Case

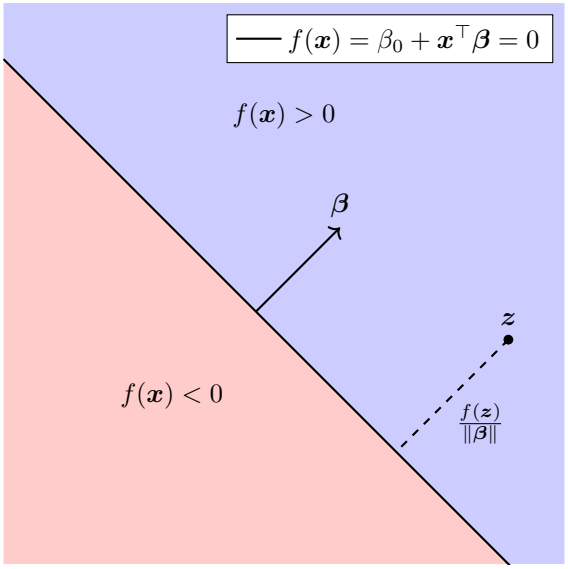
Hyperplanes in $\mathbb{R}^p$ A hyperplane is the set of $\boldsymbol{x} \in \mathbb{R}^p$ satisfying $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta} = 0$.

- $f(\boldsymbol{x}) > 0 \Leftrightarrow \boldsymbol{x}$ is on one side of the hyperplane pointed by $\boldsymbol{\beta}$.
- $f(\boldsymbol{x}) < 0 \Leftrightarrow \boldsymbol{x}$ is on the other side of the hyperplane.

- For any $\boldsymbol{z} \in \mathbb{R}^p$, the signed distance of $\boldsymbol{z}$ to the hyperplane is

$$(\langle \boldsymbol{z}, \boldsymbol{\beta} \rangle + \beta_0)/\|\boldsymbol{\beta}\|_2) = f(\boldsymbol{z})/\|\boldsymbol{\beta}\|_2.$$

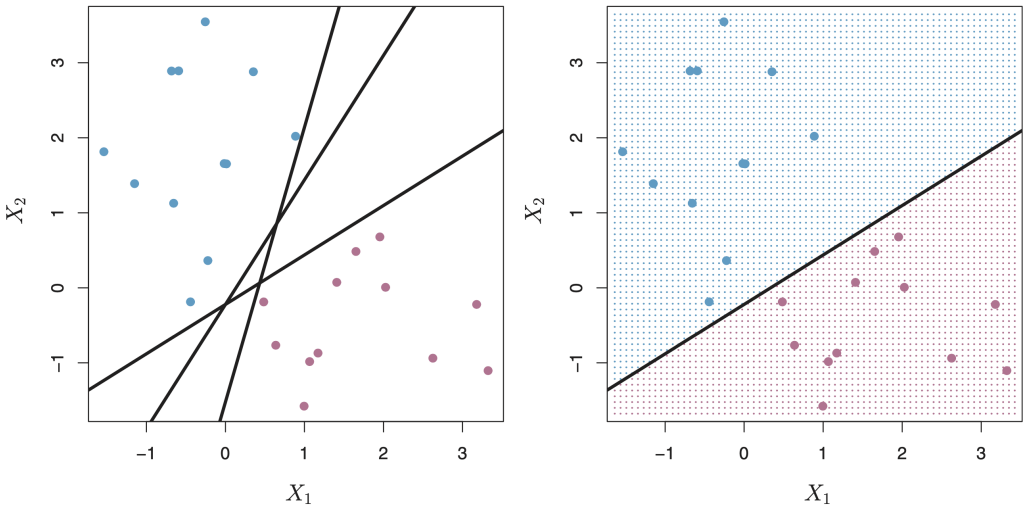
  $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^p$.
- Hence, if $\|\boldsymbol{\beta}\|_2 = 1$, then $f(\boldsymbol{z})$ is the signed distance of $\boldsymbol{z}$ to the hyperplane defined by $f(\boldsymbol{x}) = 0$.



Separating hyperplanes For linearly separable data with labels $Y_i \in \{-1, 1\}$, a *separating hyperplane* correctly classifies all points:

$$Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) \geq 0 \quad \forall i.$$

There are typically infinitely many such hyperplanes.



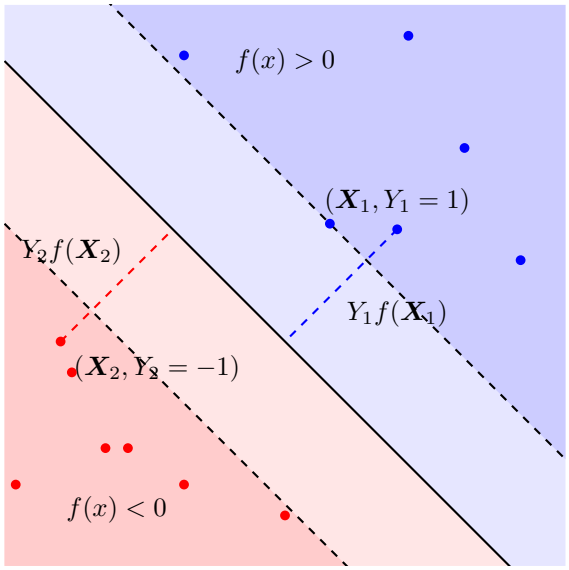
Support Vector Machine (SVM) SVM chooses among them the one with the largest *margin*—the smallest distance from any training point to the hyperplane—yielding a more robust decision boundary.

Margin

The margin is defined as the smallest distance from the training data to the hyperplane.

- Hyperplanes with larger margin are more robust.
- The Support Vector Machine (SVM) finds the hyperplane that maximizes the margin.

- The seperating hyperplane such that the minimum distance of any training point to the hyperplane is the largest.



Margin maximization Assuming $\|\boldsymbol{\beta}\|_2 = 1$, the unsigned distance of $(\boldsymbol{X}_i, Y_i)$ to the boundary is $Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta})$. The hard-margin SVM is

Hard-margin SVM

$$\max_{\beta_0, \boldsymbol{\beta},\ \|\boldsymbol{\beta}\|_2 = 1} C \quad \text{s.t.} \quad Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) \geq C, \quad \forall i.$$

For any feasible solution $\beta_0, \boldsymbol{\beta}$, perform a change of variable $\gamma_0 = \beta_0 / C, \boldsymbol{\gamma} = \boldsymbol{\beta}/C$, then $C\|\boldsymbol{\gamma}\|_2 = \|\boldsymbol{\beta}\|_2 = 1 \Rightarrow$ maximizing C is equivalent to minimizing $\|\boldsymbol{\gamma}\|_2$.

Equivalent convex program

SVM is equivalent to (with an abuse of notation $\gamma_0 = \beta_0, \boldsymbol{\gamma} = \boldsymbol{\beta}$

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) \geq 1, \quad \forall i.$$

KKT conditions and support vectors The Lagrangian is

$$\mathcal{L}(\alpha, \beta_0, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 - \sum_{i=1}^n \alpha_i \left[Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) - 1\right],$$

with optimality Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned}
\text{(stationarity)} \quad & \boldsymbol{\beta} = \sum_{i=1}^n \alpha_i Y_i \boldsymbol{X}_i, \quad 0 = \sum_{i=1}^n \alpha_i Y_i, \\
\text{(primal/dual feasibility)} \quad & Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) \geq 1, \quad \alpha_i \geq 0, \quad \forall i, \\
\text{(complementary slackness)} \quad & \alpha_i \left[Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) - 1\right] = 0, \quad \forall i.
\end{aligned}$$

Support Vectors

Points with $\alpha_i > 0$ lie exactly on the margin and are the *support vectors*:

$$\alpha_i \left[Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) - 1 \right] = 0 \Rightarrow \left[Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) - 1 \right] = 0.$$

The optimal solution $\boldsymbol{\beta}$ depends only on support vectors:

$$\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i Y_i \boldsymbol{X}_i.$$

The classifier is given by the sign of a linear function

$$\widehat{y} = \text{sign}\left(\widehat{\beta}_0 + \sum_{i=1}^{n} \widehat{\alpha}_i Y_i \boldsymbol{x}^\top \boldsymbol{X}_i \right).$$

Two practical issues remain: real data are rarely separable, and linear boundaries may be too rigid.

10.5.2 Non-Separable Case

Soft margin SVM Allow margin violations via slack variables:

$$\max_{\beta_0, \boldsymbol{\beta},\ \|\boldsymbol{\beta}\|_2 = 1} C \quad \text{s.t.} \quad Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) \geq C(1 - \xi_i), \quad \xi_i \geq 0, \quad \sum_i \xi_i \leq B,$$

where ξ_i are the *slack variables* and B is a *tuning parameter*.

The slack variable ξ_i encodes the position of observation i relative to the margin and the separating hyperplane:

- $\xi_i = 0$ if the point lies on the correct side of the margin;
- $\xi_i > 0$ if the point lies inside the margin or on the wrong side of it;
- $\xi_i > 1$ if the point lies on the wrong side of the separating hyperplane.

The parameter B plays the role of a "budget" for margin violations:

- If $B = 0$, no violations are allowed; a separating classifier exists only if the data are linearly separable.
- The larger B is, the more violations of the margin are permitted.
- In particular, no more than B observations can lie on the wrong side of the hyperplane.
- As B increases, the margin parameter C (the inverse of the regularization strength) effectively becomes larger, allowing a wider margin with more slack.

The SVM classifier is then given by the sign of a linear function

$$\widehat{y} = \text{sign}(\widehat{\beta}_0 + \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}).$$

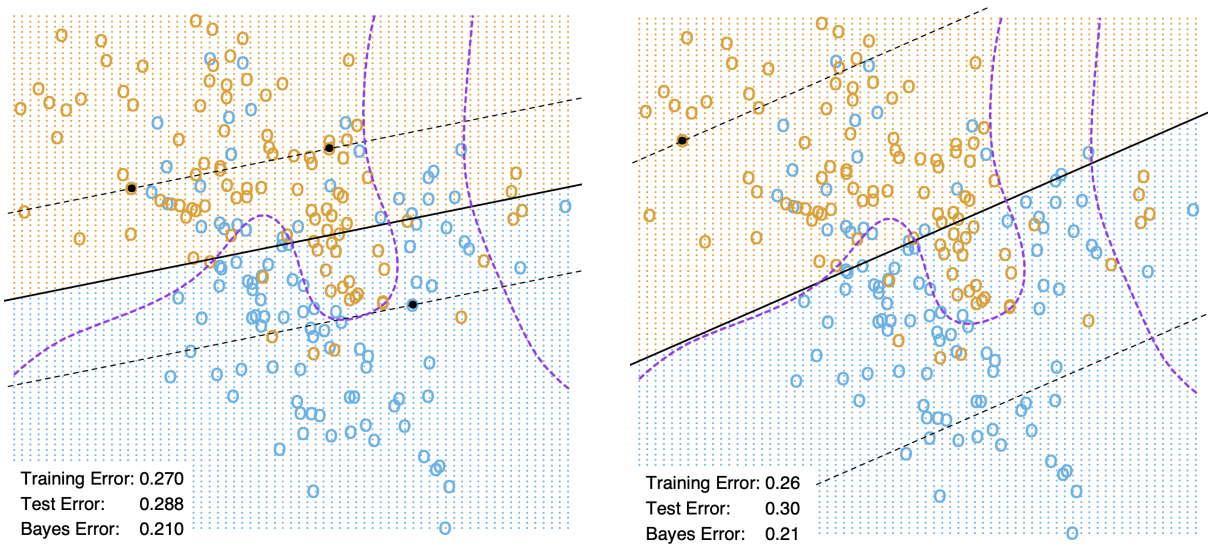
Equivalently, using Lagrangian multipliers, the convex *primal* problem is

Soft-margin SVM (primal)

$$\min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} \ \tfrac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \lambda \sum_i \xi_i \quad \text{s.t.} \quad Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

Here λ trades off margin width and violations. Large B (small λ) tolerates more violations, increases the margin, and often improves robustness by lowering variance.

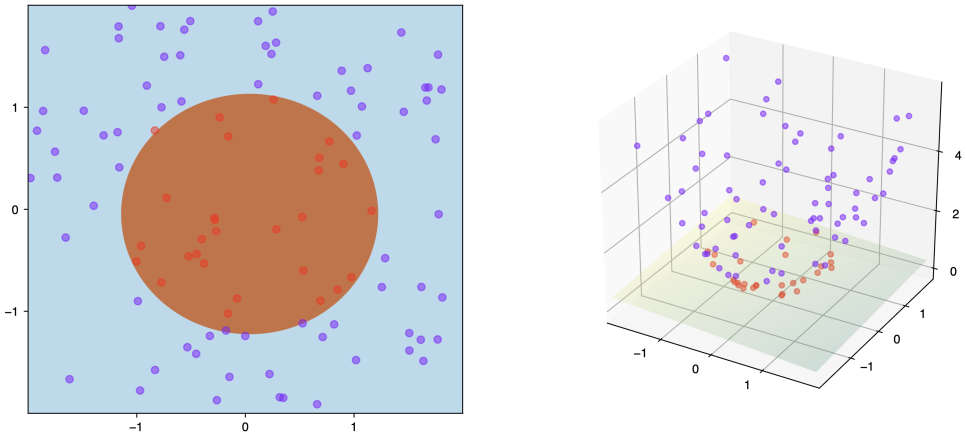
Who becomes a support vector? With slack, all points on or *inside* the margin (including misclassified ones) become support vectors. As the allowed budget of violations grows, the number of support vectors increases; the model becomes smoother but potentially more biased.



10.5.3 Kernel Method

Nonlinear boundaries via features As in polynomial regression, we can enlarge the feature space (e.g., $(X_1, X_2) \mapsto (X_1, X_2, X_1^2, X_2^2, X_1X_2, \dots)$) and fit a linear separator there, corresponding to a nonlinear boundary in the original space.

Example 10.3 (Quadratic boundary)**.**



A separating hyperplane in the enlarged space $(X_1, X_2, X_1^2, X_2^2, X_1X_2)$ has equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 = 0.$$

The curse of explicit features If $\boldsymbol{X} \in \mathbb{R}^m$ and we include all monomials up to degree d, the feature dimension is $\binom{m+d-1}{d}$. For $m = 100$ and $d = 6$ this is ≈ 1.6 billion—computationally prohibitive.

Key observations (the kernel trick) Recall the Karush-Kuhn-Tucker (KKT) conditions in the separable case

$$\begin{aligned}
&\boldsymbol{\beta} = \sum_{i=1}^{n} \alpha_i Y_i \boldsymbol{X}_i, \quad 0 = \sum_{i=1}^{n} \alpha_i Y_i, \\
&Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) \geq 1, \quad \alpha_i \geq 0, \quad \forall i, \\
&\alpha_i \left[Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) - 1 \right] = 0, \quad \forall i.
\end{aligned}$$

Hence, we observe that

$$\boldsymbol{\beta} = \sum_i \alpha_i Y_i \boldsymbol{X}_i \quad \Rightarrow \quad Y_i\big(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}\big) = Y_i \left(\beta_0 + \sum_j \alpha_j Y_j \boldsymbol{X}_i^\top \boldsymbol{X}_j\right).$$

Only inner products $\boldsymbol{X}_i^\top \boldsymbol{X}_j$ are needed for training, and $\boldsymbol{x}^\top \boldsymbol{X}_i$ for prediction:

$$\widehat{y} = \operatorname{sign}\left(\widehat{\beta}_0 + \sum_i \widehat{\alpha}_i Y_i \boldsymbol{x}^\top \boldsymbol{X}_i\right).$$

Thus we can replace inner products by a *kernel* $K(\boldsymbol{x}, \boldsymbol{z})$ without ever forming explicit features.

Kernelized SVM

Choose a positive semidefinite kernel $K(\cdot, \cdot)$ (an inner product in some feature space). Then

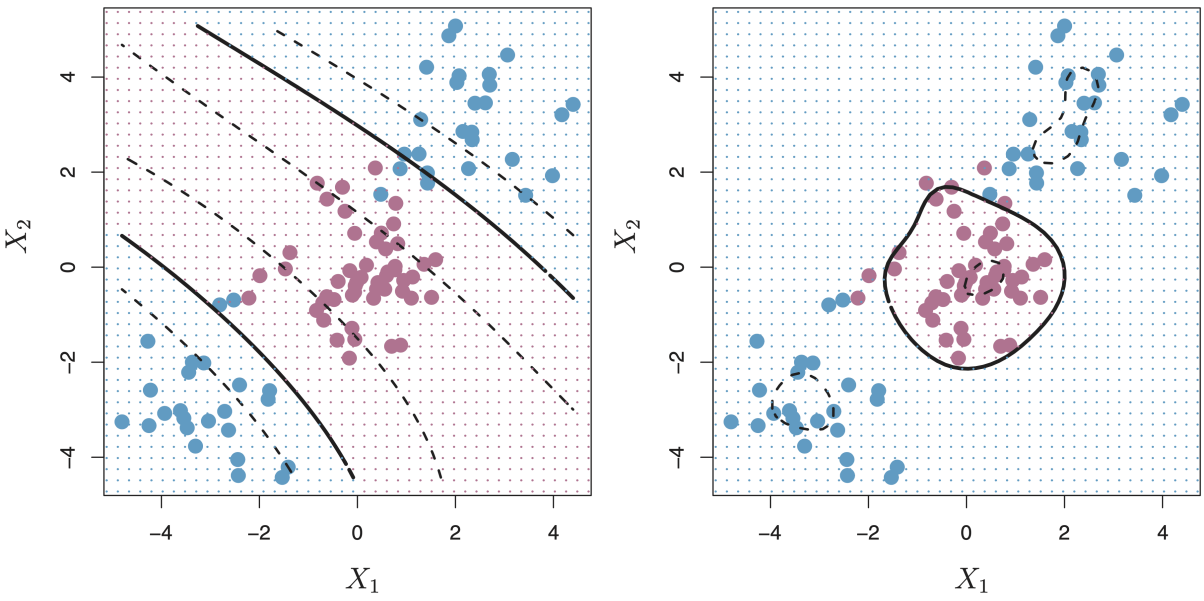
$$\widehat{y} = \operatorname{sign}\left(\widehat{\beta}_0 + \sum_{i=1}^{n} \widehat{\alpha}_i Y_i K(\boldsymbol{x}, \boldsymbol{X}_i)\right),$$

implicitly operating in a (possibly infinite-dimensional) space spanned by $\{\boldsymbol{x} \mapsto K(\boldsymbol{x}, \boldsymbol{X}_i)\}$.

Example 10.4 (Common kernels).

- Linear: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_j$.
- Polynomial of degree $\le d$: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i^\top \boldsymbol{x}_j)^d$ (all monomials up to degree d).
- Polynomial of degree exactly d: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j)^d$.
- Gaussian RBF: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$, with strong *locality* (far points have negligible influence) and an implicit infinite-dimensional feature map.

Example 10.5 (Polynomial kernel (degree 3) vs. Gaussian RBF).



10.5.4 Loss and Penalty Formula

"Loss + Penalty" viewpoint Many estimators solve

$$\min_f \sum_i L\big(Y_i, f(\boldsymbol{X}_i)\big) + \lambda P(f).$$

Ridge regression uses squared loss with $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$; lasso uses $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$.

SVM as Loss + Penalty Starting from the soft-margin constraints,

$$\begin{aligned} \max_{\beta_0,\boldsymbol{\beta},\|\beta\|_2=1} \quad & C, \\ \text{s.t.} \quad & Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta}) \geq C(1-\xi_i), \quad \forall i, \\ & \xi_i \geq 0, \sum_i \xi_i \leq B. \end{aligned}$$

Let $\gamma_0 = \beta_0/C, \boldsymbol{\gamma} = \boldsymbol{\beta}/C$, then $C\|\boldsymbol{\gamma}\|_2 = 1$. The constraints are then

$$\xi_i \geq 1 - Y_i(\gamma_0 + \boldsymbol{X}_i^\top \boldsymbol{\gamma}).$$

Combine this with the non-negative constraints of ξ_i, we have

$$\xi_i \geq \left[1 - Y_i(\gamma_0 + \boldsymbol{X}_i^\top \boldsymbol{\gamma})\right]_+.$$

Here $x_+ = \max\{0, x\}$ is the positive part of x.

The optimal slacks satisfy

$$\xi_i = \left[1 - Y_i(\gamma_0 + \boldsymbol{X}_i^\top \boldsymbol{\gamma})\right]_+,$$

giving the equivalent constrained problem

$$\min_{\gamma_0,\boldsymbol{\gamma}} \ \tfrac{1}{2}\|\boldsymbol{\gamma}\|_2^2 \quad \text{s.t.} \quad \sum_i \left[1 - Y_i(\gamma_0 + \boldsymbol{X}_i^\top \boldsymbol{\gamma})\right]_+ \leq B.$$

Introducing a Lagrange multiplier yields the unconstrained form

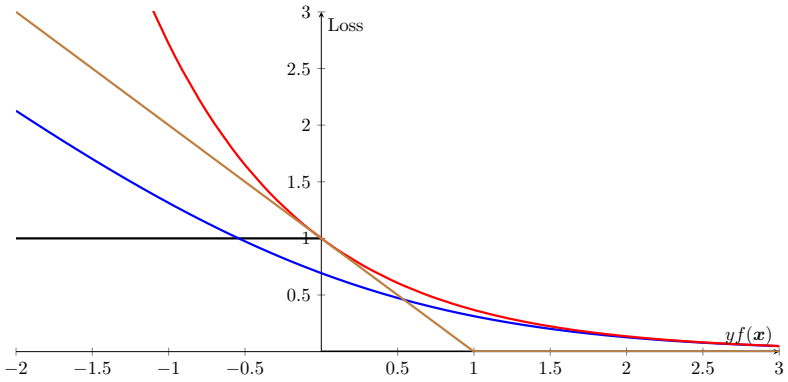
$$\min_{\beta_0,\boldsymbol{\beta}} \ \frac{1}{n}\sum_i \underbrace{\left[1 - Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta})\right]_+}_{\text{hinge loss}} + \lambda\|\boldsymbol{\beta}\|_2^2.$$

- $(1-t)_+$ is called the hinge loss.
- $\frac{1}{n}\sum_i \left[1 - Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta})\right]_+$ is the empirical hinge loss.
- $\lambda\|\boldsymbol{\beta}\|_2^2$ is the l_2 penalty.
- There is a one-to-one correspondence between B and λ.
- Notice that the loss is expressed in terms of the margin $Y_i f(\boldsymbol{X}_i)$. This is a general phenomenon in statistical learning.

Other losses (for comparison) For $Y \in \{-1, 1\}$:

- Logistic loss (binomial deviance): $\sum_i \log\left(1 + e^{-Y_i \boldsymbol{X}_i^\top \boldsymbol{\beta}}\right)$; with ℓ_2 or ℓ_1 penalty gives regularized logistic regression.
- Exponential loss (AdaBoost): $\frac{1}{n}\sum_i e^{-Y_i f(\boldsymbol{X}_i)}$, often with an ℓ_1 penalty on the expansion coefficients.

Comparing losses



The 0–1 loss is ideal but discontinuous; convex surrogates like hinge, logistic, and exponential balance optimization tractability with statistical properties.

Sparse Support Vector Machine High-dimensional settings benefit from sparse predictors for both accuracy and interpretability. Replacing the ℓ_2 penalty by ℓ_1 encourages sparsity:

ℓ_1-SVM

$$\min_{\beta_0,\boldsymbol{\beta}} \frac{1}{n}\sum_i \left[1 - Y_i(\beta_0 + \boldsymbol{X}_i^\top \boldsymbol{\beta})\right]_+ \;+\; \lambda\|\boldsymbol{\beta}\|_1.$$

10.5.5 Multi-class SVM

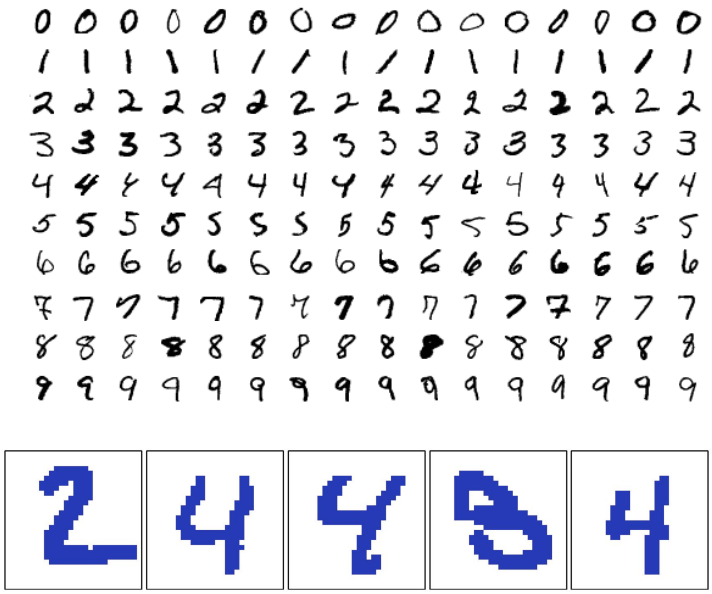
One-Versus-One (OvO) For $K > 2$ classes, fit SVMs to each of the $\binom{K}{2}$ pairs. At prediction time, each binary classifier casts one vote; assign the class with the most votes.

One-Versus-All (OvA) Alternatively, fit K SVMs, each distinguishing class k versus the rest (labels $+1$ vs. -1). For a new $\boldsymbol{x}$, compute

$$f^k(\boldsymbol{x}) = \widehat{\beta}_0^k + \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}^k$$

and assign the class with the largest $f^k(\boldsymbol{x})$.

Example 10.6 (Handwriting recognition).



Reading materials

- *The Elements of Statistical Learning*, Sections 4.3–4.5, 9.2, 10.1–10.6, 12.2–12.3, 13.3.