Statistics, Inference & Information

A collection of definitions, theorems, and formulae

March 2, 2026

Contents

Part I Statistics and data analysis

1 Fundamentals

1.1 Numerical summaries of data

Given data $x_1, \ldots, x_n \in \mathbb{R}$,

- The sample mean is
$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$
- The (biased) sample variance is
$$\sigma^2_{\text{bias}} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$
- The (unbiased) sample variance is
$$\sigma^2_{\text{non-bias}} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$
- The sample median x_{median} is $x_{(\lceil n/2 \rceil)}$ where
$$x_{(1)} \leq \ldots \leq x_{(n)}$$

1.2 Empirical covariance of vectored data

Let $X \in \mathbb{R}^{n\times d}$ be a data matrix whose ith row $x_i^\top$ is an observation in $\mathbb{R}^d$. Define the sample mean

$$\bar{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathbb{R}^d, \qquad \text{and the centering operator } \; H \;=\; I_n - \frac{1}{n}\,11^\top,$$

therefore the centered data matrix is

$$\tilde{X} \;=\; HX \in \mathbb{R}^{n\times d}, \quad \text{whose rows are } (x_i - \bar{x})^\top.$$

If $x \in \mathbb{R}^d$ is a zero-mean random vector, then its population covariance is $\Sigma_X = \mathbb{E}[xx^\top]$ (if $x \in \mathbb{R}^d$ is not zero mean, this is simply the second moment).

From data iid data, the (biased) sample covariance is

$$S_{\text{bias}} \;=\; \frac{1}{n}\,\tilde{X}^\top\tilde{X} \in \mathbb{R}^{d\times d}, \qquad \text{which satisfies} \qquad \mathbb{E}[S_{\text{bias}}] = \frac{n-1}{n}\,\Sigma_X.$$

The unbiased version is

$$S_{\text{unbiased}} \;=\; \frac{1}{n-1}\,\tilde{X}^\top\tilde{X}, \qquad \text{which satisfies} \qquad \mathbb{E}[S] = \Sigma_X.$$

The scalar factor between S_{unbiased} and S_{bias} does not affect eigenvectors and only rescales eigenvalues. In what follows, any choice proportional to $\tilde{X}^\top\tilde{X}$ is acceptable; for definiteness we

use

$$S = \frac{1}{n}\tilde{X}^\top \tilde{X}.$$

1.3 Empirical pmf $\widehat{\mathbb{P}}$ and cdf $\widehat{\mathbb{F}}$

Given a collection of data $x_1, \ldots, x_n$, the empirical distribution assigns equal mass to each observation:

$$\begin{aligned}\widehat{\mathbb{P}}(\mathsf{x} = x) &:= \frac{\#\{i : x_i = x\}}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} \delta(x_i = x)\end{aligned}$$

Under this definition, we have that for any measurable set $\mathcal{A}$

$$\widehat{\mathbb{P}}(\mathcal{A}) = \frac{1}{n}\#\{i : x_i \in \mathcal{A}\}.$$

The empirical cdf $\widehat{\mathbb{F}}$ is the distribution function associated with $\widehat{\mathbb{P}}$, namely

$$\widehat{\mathbb{F}}(\alpha) := \frac{1}{n}\#\{i : x_i \leq \alpha\},$$

which is a right-continuous step function with jumps of size $1/n$ and values $\widehat{\mathbb{F}} \in [0, 1]$.

1.4 Law of Large Numbers

Let $\mathsf{x}_1, \mathsf{x}_2, \ldots$ be i.i.d. with $\mathbb{E}[\mathsf{x}_i] = \mu$, and define the sample mean

$$\bar{\mathsf{x}}_n := \frac{1}{n}\sum_{i=1}^{n} \mathsf{x}_i \to \mu$$

then the *Law of Large Numbers* states that

$$\bar{\mathsf{x}}_n \xrightarrow[n\to\infty]{\cdot} \mu$$

in probability (weak LLN) and almost surely (strong LLN). More generally, for any integrable function $g(\cdot)$,

$$\frac{1}{n}\sum_{i=1}^{n} g(\mathsf{x}_i) \xrightarrow[n\to\infty]{\cdot} \mathbb{E}[g(\mathsf{x})].$$

While this is not obvious, the LLN has an impact at the level of distributions, namely

$$\widehat{\mathbb{P}} \to \mathbb{P} \quad \text{as } n \to \infty,$$

and consequently $\widehat{\mathbb{F}} \to \mathbb{F}$. Here $\mathbb{P}$ is the true (possibly continuous) distribution, whereas $\widehat{\mathbb{P}}$ is discrete by construction.

1.5 Central Limit Theorem

Let $\mathsf{X}_1, \mathsf{X}_2, \ldots$ be i.i.d. with mean μ and variance $\sigma^2 < \infty$. Then

$$\sqrt{n}\,\frac{\bar{\mathsf{x}}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, 1),$$

The CLT says that at higher values of n, the random variable $\bar{x}_n$ behaves according to a standard normal distribution.

1.6 In what sense do we have $\widehat{\mathbb{P}} \to \mathbb{P}$?

In what sense does the empirical distribution $\widehat{\mathbb{P}}$ converge to the true distribution $\mathbb{P}$? Let $x_1, \ldots, x_n \sim \mathbb{P}$ i.i.d. with the associated CDF $\mathbb{F}(x) := \mathbb{P}(x \leq x)$. Define the empirical measure and empirical CDF

$$\widehat{\mathbb{P}}_n(x = x) := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i = x}, \qquad \widehat{\mathbb{F}}_n(x) := \widehat{\mathbb{P}}_n(x \in (-\infty, x]) = \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \leq x\},$$

To answer the question of convergence of distributions, first define the distance between distributions using the Kolmogorov–Smirnov statistic (goodness-of-fit distance):

$$\begin{aligned} D_n &= \|\widehat{\mathbb{F}}_n - \mathbb{F}\|_\infty \\ &= \sup_{x \in \mathbb{R}} |\widehat{\mathbb{F}}_n(x) - \mathbb{F}(x)| \end{aligned}$$

A standard notion of convergence is *uniform convergence of CDFs*, in particular, the Glivenko–Cantelli theorem says that

$$\|\widehat{\mathbb{F}}_n - \mathbb{F}\|_\infty \xrightarrow[n\to\infty]{a.s.} 0$$

If we want a measure of convergence for a finite sample, the DKW inequality (finite-sample uniform deviation bound) says for all $\varepsilon > 0$,

$$\mathbb{P}\Big(\|\widehat{\mathbb{F}}_n - \mathbb{F}\|_\infty > \varepsilon \Big) \leq 2e^{-2n\varepsilon^2}.$$

1.7 Empirical cross-entropy = negative log-likelihood

Let $\{(x_i, y_i)\}_{i=1}^n$ be data, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Let $p_\theta(y \mid x)$ be our parametric conditional model with parameter θ that we seek to optimize for θ. Let $\hat{p}_n(y \mid x)$ denote the empirical conditional distribution induced by the sample (i.e. it puts mass $1/n$ on each observed pair (x_i, y_i)).

Define the empirical cross-entropy between the empirical distribution $\hat{p}_n$ relative to p_θ by

$$\begin{aligned} H(\hat{p}_n, p^*_\theta) &:= \mathbb{E}_{(x,y)\sim\hat{p}_n}\big[-\log p^*_\theta(y \mid x)\big] \\ &= \sum_{i=1}^{n} \hat{p}_n(y_i \mid x_i)\big(-\log p_\theta(y_i \mid x_i)\big) \\ &= \sum_{i=1}^{n} \frac{1}{n}\big(-\log p_\theta(y_i \mid x_i)\big) \end{aligned}$$

Define the conditional log-likelihood by

$$\ell(\theta) := \sum_{i=1}^{n} \log p_\theta(y_i \mid x_i).$$

Then we can express the empirical cross-entropy as

$$H(\hat{p}_n, p_\theta) \;=\; -\frac{1}{n}\,\ell(\theta), \qquad \text{and therefore} \qquad \arg\min_\theta H(\hat{p}_n, p_\theta) \;=\; \arg\max_\theta \ell(\theta).$$

In the above setting, we can interpret $-\log p_\theta(y_i \mid x_i)$ as the surprise assigned by our model p_θ to the event $(y_i \mid x_i)$. Therefore, we can interpret the empirical cross entropy $H(\hat{p}_n, p^*_\theta)$ as the average surprisal of our model on the observed data, so minimizing it over θ makes the data "less surprising" under our model.

If there is a true conditional distribution $p^\star(y \mid x)$ and the expectation is taken under $p^\star$, then the cross entropy between $p^\star$ and p_θ can be written as

$$H(p^\star, p_\theta) := \mathbb{E}_{(x,y)\sim p^\star}\big[-\log p_\theta(y \mid x)\big] = H(p^\star) + \mathrm{KL}(p^\star \,\|\, p_\theta)\,,$$

where the entropy of the true model p^* and the KL divergence between our model p_θ and the true p^* are defined as

$$H(p^\star) := \mathbb{E}_{(x,y)\sim p^\star}\big[-\log p^\star(y \mid x)\big], \qquad \mathrm{KL}(p^\star \| p_\theta) := \mathbb{E}_{p^\star}\left[\log \frac{p^\star(y \mid x)}{p_\theta(y \mid x)}\right].$$

As a consequence of this definition, minimizing the cross entropy is equivalent to minimizing the KL between our model p_θ and the true p^*:

$$\min_\theta H(p^\star, p_\theta) \iff \min_\theta \mathrm{KL}(p^\star \| p_\theta)$$

and

$$H(p^\star, p_\theta) \geq H(p^\star)$$

with equality if and only if $\mathrm{KL}(p^\star \,\|\, p_\theta) = 0$, hence if and only if $p_\theta(y \mid x) = p^\star(y \mid x)$.

2 Function and parameter estimation using data

In this section, we describe how to compute a function of our ditribution (e.g. the second moment) using the plug-in method, then how to estimate the parameter using feature matching and maximum likelihood estimation.

2.1 Feature of a distribution

A feature is a functional of a distribution:

$$h = h(\mathbb{P}),$$

- expectation $h(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[x]$
- variance $h(\mathbb{P}) = \mathrm{Var}_{\mathbb{P}}(x)$
- quantile $h(\mathbb{P}) = \mathbb{F}^{-1}(0.75)$
- tail probability $h(\mathbb{P}) = \mathbb{P}(x \geq 1/2)$

2.2 Plug in estimators to compute functions $h(\mathbb{P})$

Suppose we have a random variable $x \sim \mathbb{P}$ (non-parametric) and a feature $h = h(\mathbb{P})$. The plug-in method allows us to compute an estimate of this function $\widehat{h}$ by replacing the true distribution $\mathbb{P}$ by the empirical distribution $\widehat{\mathbb{P}}$, so that the estimate is $\widehat{h} = h(\widehat{\mathbb{P}})$. In particular

$$h(\mathbb{P}) := \mathbb{E}[g(x)] \quad \rightsquigarrow \quad \widehat{h} = h(\widehat{\mathbb{P}}) = \sum_{i=1}^{n} \widehat{\mathbb{P}}(x = x_i) g(x_i).$$

For example,

$$h(\mathbb{P}) = \mathbb{E}[x^3] \quad \rightsquigarrow \quad \widehat{h} = h(\widehat{\mathbb{P}}) = \sum_{i=1}^{n} \widehat{\mathbb{P}}(x = x_i) x_i^3$$

2.3 Feature matching (method of moments) to esimate θ

If the model is parametric, $\{\mathbb{P}^\theta : \theta \in \Theta\}$, then the feature $h(\mathbb{P})$ becomes a function of the parameter θ underlying the distribution

$$h^\theta := h(\mathbb{P}^\theta).$$

Suppose we are given data $x_1, \ldots, x_n \sim \mathbb{P}^\theta$. Feature mathcing is a way to estimate the true parameter θ:

1. From data we compute the empirical feature

$$\widehat{h} := h(\widehat{\mathbb{P}}_n),$$

2. Find analytically h^θ as a function of θ
3. Then estimate the distribution parameter θ by solving

$$h^\theta = \widehat{h}.$$

   to obtain the value of θ which produces a distribution that has the same feature h^θ we observe in our data $\widehat{h}$.

A special case of feature matching is the method of moments, where we set up an system of independant equations to facilitate parameter estimation.

2.4 Method of moments on a Gaussian mixture

The method of moments is a way to do parameter estimation by setting up a system of equations involving the moments of the distribution. Consider the mixture model defined by

$$x \;=\; z\,x_1 \;+\; (1-z)\,x_2,$$

where $z \sim \mathrm{Bern}(p)$ and, conditionally on z, we draw

$$x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \qquad x_2 \sim \mathcal{N}(\mu_2, \sigma_2^2),$$

with z, x_1, x_2 independent. Equivalently, x is drawn from $\mathcal{N}(\mu_1, \sigma_1^2)$ with probability p and from $\mathcal{N}(\mu_2, \sigma_2^2)$ with probability $(1-p)$.
For any integer $k \geq 1$, conditioning on z gives the mixture-moment identity

$$\mathbb{E}\Big[x^k\Big] \;=\; p\,\mathbb{E}\Big[x_1^k\Big] \;+\; (1-p)\,\mathbb{E}\Big[x_2^k\Big].$$

If the parameters are unknown (e.g. $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p$), one can match multiple moments (e.g. $\mathbb{E}[x], \mathbb{E}[x^2], \ldots$) to obtain a system of equations in the unknown parameters.

2.5 Maximum likelihood

Assume a parametric model $x_1, \ldots, x_n \sim \mathbb{P}^\theta$ i.i.d. with density or pmf $p_\theta(\cdot)$. Given observed data $\vec{x} = (x_1, \ldots, x_n)$, define the likelihood

$$L(\theta; \vec{x}) := p_\theta(\vec{x}) = \prod_{i=1}^{n} p_\theta(x_i),$$

and the log-likelihood

$$\ell(\theta; \vec{x}) := \log L(\theta; \vec{x}) = \sum_{i=1}^{n} \log p_\theta(x_i).$$

The maximum likelihood estimator (MLE) is

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta \in \Theta} L(\theta; \vec{x}) \quad \Longleftrightarrow \quad \widehat{\theta}_{\mathrm{MLE}} \in \arg\max_{\theta \in \Theta} \ell(\theta; \vec{x}).$$

Interpretation: $L(\theta; \vec{x})$ is the probability assigned by the model $\mathbb{P}^\theta$ to the observed sample; MLE chooses the parameter that makes the observed data most likely. To actually compute the MLE, we can sometimes solve first-order conditions

$$\nabla_\theta \ell(\theta; \vec{x}) = 0$$

3 Analysis of estimators

In the previous section, we discussed two methods for parameter estimation: method of moments and maximum likelihood. In this section, we consider a general estimation algorithm and model it as a random variable in an effort to answer questions regarding the properties of the estimator (such as its mean and variance) and questions regarding the level of confidence we should have in the estimator.

3.1 Estimators as random variables under the sampling distribution

An *estimation algorithm* $g(\cdot) : \mathcal{X}^n \to \Theta$, is a deterministic function which maps data $x^n = (x_1, \ldots, x_n)$ to an estimate

$$\widehat{\theta} \;=\; g(x^n).$$

Here $\widehat{\theta}$ is a *number* because the input x^n is a realized dataset. To study the behavior of the estimator, we model the data as random:

$$x^n = (x_1, \ldots, x_n) \;\sim\; \mathbb{P}^\theta.$$

Since the data is now modeled as random, the estimator becomes a random variable too:

$$\widehat{\theta} \;=\; g(x^n).$$

The *sampling distribution* is the law of $\widehat{\theta}$ when $x^n \sim \mathbb{P}^\theta$. We may denote this distribution by $\mathbb{Q}^\theta$, i.e.

$$\widehat{\theta} \sim \mathbb{Q}^\theta \qquad \text{when} \qquad x^n \sim \mathbb{P}^\theta.$$

3.2 Explicit sampling distribution example (sum of Bernoulli → Binomial)

Under our assumption that the data $x \sim \mathbb{P}^\theta$, sometimes we can explicitly calculate the sampling distribution $\mathbb{Q}^\theta$ by analyzing the estimator $\widehat{\theta} \sim \mathbb{Q}^\theta$. For example, let $x_i \overset{\text{iid}}{\sim} \text{Bern}(\theta)$ and define

$$g(x^n) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Since the draws are i.i.d. Bernoulli, the sum is Binomial:

$$\sum_{i=1}^{n} x_i \;\sim\; \text{Bin}(n, \theta).$$

Hence the estimator

$$\widehat{\theta} \;=\; g(x^n) \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i$$

takes values in $\{0, \frac{1}{n}, \ldots, 1\}$, and its sampling distribution is

$$\mathbb{P}\left(\widehat{\theta} = \frac{k}{n}\right) \;=\; \binom{n}{k}\theta^k(1-\theta)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Equivalently, $n\widehat{\theta} \sim \text{Bin}(n, \theta)$.

3.3 Variance and standard error of an estimator

Under $\mathbb{P}^\theta$, the variance of the estimator $\widehat{\theta} = g(x^n)$ is

$$\text{Var}_{\mathbb{P}^\theta}(\widehat{\theta}) \;=\; \mathbb{E}_{\mathbb{P}^\theta}\left[\left(\widehat{\theta} - \mathbb{E}_{\mathbb{P}^\theta}[\widehat{\theta}]\right)^2\right] \;=\; v^\theta$$

The *standard error* is

$$s^\theta \;=\; \sqrt{v^\theta}.$$

3.4 Bias and MSE, and what b^θ means

Define the *bias* (at parameter value θ) by

$$b^\theta \;=\; \mathbb{E}_{\mathbb{P}^\theta}[\hat{\theta}] - \theta.$$

Note the types:

$$\widehat{\theta} \text{ is random}, \qquad \mathbb{E}_{\mathbb{P}^\theta}[\widehat{\theta}] \text{ is deterministic}, \qquad b^\theta \text{ is deterministic}.$$

The *mean squared error* (MSE) is

$$\mathrm{MSE}^\theta \;=\; \mathbb{E}_{\mathbb{P}^\theta}\Big[\big(\widehat{\theta}-\theta\big)^2\Big].$$

Expanding $\widehat{\theta}-\theta = (\widehat{\theta}-\mathbb{E}[\widehat{\theta}]) + (\mathbb{E}[\widehat{\theta}]-\theta)$ yields the bias–variance decomposition

$$\mathrm{MSE}^\theta \;=\; \underbrace{\big(b^\theta\big)^2}_{\text{squared bias}} \;+\; \underbrace{v^\theta}_{\text{variance}}.$$

This highlights a common tradeoff: procedures with smaller variance may have larger bias (and vice versa), and MSE captures both effects.

3.5 Parametric bootstrap sampling and re-estimation

We've established that when we treat the data as random $\mathsf{x} \sim \mathbb{P}^\theta$ for some parametric model $\{\mathbb{P}^\theta : \theta \in \Theta\}$, the estimator $\widehat{\theta} = g(\mathsf{x}^n)$ is a random variable $\widehat{\theta} \sim \mathbb{Q}^\theta$. Our goal now is to approximate this sampling distribution of the estimator $\mathbb{Q}^\theta$.

Suppose we observe data $x^n = (x_1, \ldots, x_n)$ and compute an estimate via an algorithm $g(\cdot)$:

$$\hat{\theta} \;=\; g(x^n).$$

Using this estimate, we then form the estimated distribution $\mathbb{P}^{\hat{\theta}}$. The idea of bootstrap is to then sample from $\mathbb{P}^{\hat{\theta}}$, obtain a new parameter estimate, then repeat:

$$x^n \xrightarrow{\text{use } g(\cdot)} \hat{\theta} \xrightarrow{\text{plug in}} \mathbb{P}^{\hat{\theta}} \xrightarrow{\text{resample}} \mathsf{x}^{n,(b)} \xrightarrow{\text{use } g(\cdot)} \hat{\theta}^{(b)}.$$

In practice, we collect $\{\hat{\theta}^{(b)}\}_{b=1}^B$ and form a histogram or empirical CDF as an approximation of the sampling distribution $\mathbb{Q}^\theta$. From this estimate of the sampling distribution we can further estimate quantities such as variance, standard error, and confidence intervals.

3.6 Non-parametric bootstrap sampling and re-estimation

In non-parametric bootstrap, we do *not* assume a parametric model $\mathbb{P}^\theta$. To approximate the law of some function/estimator, we approximate the unknown data-generating law of $\mathsf{x} \sim \mathbb{P}$ by the empirical distribution

$$\widehat{\mathbb{P}}_n \;:=\; \frac{1}{n}\sum_{i=1}^n \delta_{x_i},$$

where δ_{x_i} is a point mass at x_i. Recall in the parametric setup we approximated $\mathsf{x} \sim \mathbb{P}^\theta$ using the distribution $\mathbb{P}^{\hat{\theta}}$.

Given observed data $x^n = (x_1, \ldots, x_n)$ and an estimator computed by an algorithm $g(\cdot)$,

$$\hat{\theta} \;=\; g(x^n),$$

the bootstrap procedure resamples *from the data itself* according to $\widehat{\mathbb{P}}_n$:

$$x^n \xrightarrow{\text{form } \widehat{\mathbb{P}}_n} \widehat{\mathbb{P}}_n \xrightarrow{\text{resample}} \mathsf{x}^{n,(b)} \xrightarrow{\text{use } g(\cdot)} \hat{\theta}^{(b)} = g(\mathsf{x}^{n,(b)}).$$

Concretely, we form the empirical distribution $\widehat{\mathbb{P}}_n$ once, then for each $b = 1, \ldots, B$:

1. Sample $x^{n,(b)} = (x_1^{(b)}, \ldots, x_n^{(b)})$ i.i.d. from $\widehat{\mathbb{P}}_n$ (equivalently: sample with replacement from $\{x_1, \ldots, x_n\}$).
2. Recompute the estimator $\hat{\theta}^{(b)} = g(x^{n,(b)})$.

We then use the empirical distribution of $\{\hat{\theta}^{(b)}\}_{b=1}^{B}$ (histogram or empirical CDF) as an approximation to the sampling distribution $\mathbb{Q}^\theta$ of $\widehat{\theta} = g(x^n)$, and use it to estimate variance, standard error, and confidence intervals of this estimator $\widehat{\theta} = g(x^n)$.

4 Frequentist confidence intervals

4.1 A confidence interval is defined as a random interval

Fix a parametric model $\{\mathbb{P}^\theta : \theta \in \Theta\}$ and data $x_i \sim \mathbb{P}^\theta$ for $i = 1, \ldots, n$. In the frequentist setting, the *parameter* θ is a fixed (unknown) constant and the *data* $x^n = (x_1, \ldots, x_n)$ is random. Therefore, any estimator

$$\widehat{\theta} \;=\; g(x^n)$$

is a random variable with sampling distribution $\widehat{\theta} \sim \mathbb{Q}^\theta$ (the law induced by $\mathbb{P}^\theta$ through g).
A $(1-\alpha)$ confidence interval is a *random interval*

$$[\,\widehat{\theta}^{\,-},\ \widehat{\theta}^{\,+}\,],$$

where the endpoints $\widehat{\theta}^{\,-}, \widehat{\theta}^{\,+}$ are random variables (functions of x^n), such that the *coverage* property holds:

$$\mathbb{P}^\theta\Big(\widehat{\theta}^{\,-} \le \theta \le \widehat{\theta}^{\,+}\Big) \;\ge\; 1-\alpha, \qquad \text{for all } \theta \in \Theta.$$

Interpretation: if we repeatedly sample a new dataset $x^n \sim \mathbb{P}^\theta$ (with the same fixed θ) and recompute the interval each time, then the random interval will contain the fixed truth θ at least a fraction $1-\alpha$ of the time. It is *not* correct to say "θ has probability $1-\alpha$ of lying in this fixed interval."

4.2 Designing a confidence interval of an estimator: CLT case

To construct a confidence interval, use the CLT to say that for large n,

$$\frac{\widehat{\theta} - \theta}{\mathrm{se}_n} \;\approx\; \mathcal{N}(0,1),$$

where se_n is the (possibly estimated) standard error. Then letting $z_{1-\alpha/2}$ be the $(1-\alpha/2)$ quantile of $\mathcal{N}(0,1)$, we get that

$$\mathbb{P}^\theta\Big(|\widehat{\theta} - \theta| \le z_{1-\alpha/2}\,\mathrm{se}_n\Big) \;\approx\; 1-\alpha,$$

which is equivalent to the (approximate) $(1-\alpha)$ CI

$$\Big[\ \widehat{\theta} - z_{1-\alpha/2}\,\mathrm{se}_n\ ,\ \widehat{\theta} + z_{1-\alpha/2}\,\mathrm{se}_n\ \Big].$$

For example, when $\alpha = 0.05$, $z_{0.975} \approx 1.96$.

4.3 Designing a confidence interval f an estimator: general case

More generally, if we can find constants a, b (possibly depending on n and α) such that

$$\mathbb{P}^{\theta}\Big(\theta - a \le \widehat{\theta} \le \theta + b\Big) \;=\; 1-\alpha,$$

then by rearranging the inequalities we obtain the equivalent statement

$$\mathbb{P}^{\theta}\Big(\widehat{\theta} - b \le \theta \le \widehat{\theta} + a\Big) \;=\; 1-\alpha,$$

so a valid $(1-\alpha)$ CI is

$$\Big[\ \widehat{\theta} - b\ ,\ \widehat{\theta} + a\ \Big].$$

Here again: θ is fixed; the randomness is only through $\widehat{\theta}$ (hence through the endpoints).

4.4 Using bootstrap to produce confidence intervals of an estimator

Designing confidence intervals as above requires knowing something about the distribution of the estimator $\widehat{\theta} \sim \mathbb{Q}^{\theta}$. Hence we use bootstrap to calculate an empirical distribution of $\widehat{\theta}$.

- *Parametric bootstrap:*

$$x^n \xrightarrow{g(\cdot)} \hat{\theta} = g(x^n) \quad \Longrightarrow \quad \text{sample } \mathsf{x}^{n,(b)} \sim \mathbb{P}^{\hat{\theta}} \quad \Longrightarrow \quad \hat{\theta}^{(b)} = g(\mathsf{x}^{n,(b)}).$$

  Then use the empirical quantiles of $\{\hat{\theta}^{(b)}\}_{b=1}^{B}$ to approximate quantiles of $\widehat{\theta}$.
- *Non-parametric bootstrap:* First define the empircal distribution $\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ then for iteration $b = 1, \ldots, B$

$$\text{resample } \mathsf{x}^{n,(b)} \sim \widehat{\mathbb{P}}_n \quad \Longrightarrow \quad \hat{\theta}^{(b)} = g(\mathsf{x}^{n,(b)}).$$

  e.g. $\mathsf{x}^{n,(b)} \sim \widehat{\mathbb{P}}_n$ means we resample with replacement from $\{x_i\}$.

The output of either case is an empirical distribution of $\widehat{\theta}$. We then use this to construct the *percentile interval*: if $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ are the empirical quantiles of $\{\hat{\theta}^{(b)}\}_{b=1}^{B}$, then take

$$\Big[\ \hat{q}_{\alpha/2}\ ,\ \hat{q}_{1-\alpha/2}\ \Big].$$

5 Principal component analysis

5.1 Form the empirical covariance matrix

Given data points $x_1, \ldots, x_n \in \mathbb{R}^d$, define the sample mean

$$\bar{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i$$

and form the (empirical) covariance matrix

$$S \;=\; \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^{\top} \in \mathbb{R}^{d\times d}$$

Then compute the eigenvalues/eigenvectors of S:

$$Sv_j \;=\; \lambda_j v_j, \qquad \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d,$$

where $v_1, \ldots, v_d$ are orthonormal. Since this S is symmetric and positive semidefinite, its SVD coincides with its eigen-decomposition, since for a symmetric PSD matrix like S, the singular values σ_j equal the eigenvalues λ_j. Write the SVD as

$$S \;=\; U\Sigma U^\top,$$

where $U \in \mathbb{R}^{d\times d}$ is orthonormal ($U^\top U = I$) and

$$\Sigma \;=\; \mathrm{diag}(\sigma_1, \ldots, \sigma_d), \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0.$$

Writing u_j for the jth column of U, we can equivalently write that

$$Su_j \;=\; \sigma_j u_j, \qquad j = 1, \ldots, d,$$

and

$$S \;=\; \sum_{j=1}^{d} \sigma_j \, u_j u_j^\top.$$

5.2 k-dimensional representation of x_i

The k-dimensional representation of a data point x_i is the vector of projections onto the top-k eigen-directions $v_1, \ldots, v_k$:

$$\big((x_i - \bar{x})^\top v_1,\; (x_i - \bar{x})^\top v_2,\; \ldots,\; (x_i - \bar{x})^\top v_k\big) \in \mathbb{R}^k$$

More broadly, think of PCA as mapping each point x_i to a coordinate vector

$$x_1 \mapsto (\alpha_{11}, \ldots, \alpha_{1k}), \quad x_2 \mapsto (\alpha_{21}, \ldots, \alpha_{2k}), \quad \ldots, \quad x_n \mapsto (\alpha_{n1}, \ldots, \alpha_{nk}),$$

where these coefficients

$$\alpha_{ij} \;=\; (x_i - \bar{x})^\top v_j, \qquad i \in [n],\; j \in [k].$$

Using these coefficients, a rank-k approximation of x_i is

$$x_i \;\approx\; \bar{x} \;+\; \alpha_{i1} v_1 \;+\; \cdots \;+\; \alpha_{ik} v_k.$$

5.3 Rank-r truncation of S and Frobenius error

Keeping the top k components gives the rank-k approximation

$$S_k \;=\; \sum_{\ell=1}^{r} \lambda_\ell v_\ell v_\ell^\top$$

then

$$S - S_k \;=\; \sum_{k=r+1}^{d} \lambda_k v_k v_k^\top.$$

Using $\|v_k v_k^\top\|_F^2 = 1$ and orthogonality across k, we obtain

$$\|S - S_k\|_F^2 \;=\; \sum_{\ell=k+1}^{d} \lambda_\ell^2, \qquad \|S\|_F^2 \;=\; \sum_{\ell=1}^{d} \lambda_\ell^2,$$

and hence the relative (squared) Frobenius error is

$$\frac{\|S - S_r\|_F^2}{\|S\|_F^2} \;=\; \frac{\sum_{k=r+1}^{d} \lambda_k^2}{\sum_{k=1}^{d} \lambda_k^2}.$$

6 Akaike model selection

6.1 Model fit vs number of parameters

Suppose we are comparing two parametric models, e.g. Model 1 has p_1 parameters and Model 2 has p_2 parameters, with $p_2 > p_1$. Increasing the number of parameters often increases in-sample fit, but we need a way to express the price of higher model complexity

Suppose we observe the training dataset $x^n = (x_1, \ldots, x_n)$, typically modeled as i.i.d. draws from an unknown true distribution $\mathbb{P}$. We consider a parametric model class

$$\mathcal{M} \;=\; \{\, p_\theta(\cdot) : \theta \in \Theta \subset \mathbb{R}^p \,\},$$

where p is the number of free parameters.

6.2 Fit the model by maximum likelihood

Assume the training dataset $x^n = (x_1, \ldots, x_n)$ is generated as i.i.d. draws from an unknown true distribution $\mathbb{P}$, i.e.

$$x^n \sim \mathbb{P}^{\otimes n}.$$

For a parametric model class $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$, the MLE is the (random) estimator

$$\hat{\theta}_{\mathrm{MLE}}(x^n) \;\in\; \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p_\theta(x_i).$$

Here $\hat{\theta}_{\mathrm{MLE}}$ is random because it is a function of the random sample x^n.

6.3 What AIC is trying to estimate (make the randomness explicit)

We care about predictive performance on new data. Let x' denote a fresh test draw. Given the fitted parameter $\hat{\theta}_{\mathrm{MLE}}(x^n)$, the test log score is the random variable

$$\log p_{\hat{\theta}_{\mathrm{MLE}}(x^n)}(x').$$

The target quantity AIC approximates is the expected test log score, with expectations taken over both sources of randomness:

$$\mathbb{E}_{x^n \sim \mathbb{P}^{\otimes n},\; x' \sim \mathbb{P}} \Big[\log p_{\hat{\theta}_{\mathrm{MLE}}(x^n)}(x') \Big].$$

6.4 Key intuition: in-sample fit is optimistically biased

Because $\hat{\theta}_{\mathrm{MLE}}(x^n)$ is chosen to maximize the training log-likelihood, the in-sample score is typically larger than the out-of-sample score. Akaike's asymptotic bias correction states

$$\mathbb{E}_{x^n \sim \mathbb{P}^{\otimes n},\; x' \sim \mathbb{P}} \Big[\log p_{\hat{\theta}_{\mathrm{MLE}}(x^n)}(x') \Big] \;=\; \mathbb{E}_{x^n \sim \mathbb{P}^{\otimes n}} \left[\frac{1}{n} \sum_{i=1}^{n} \log p_{\hat{\theta}_{\mathrm{MLE}}(x^n)}(x_i) \right] \;-\; \frac{p}{n} \;+\; o\Big(\frac{1}{n}\Big).$$

Interpretation: the expected training log score overestimates the expected test log score by about p/n.

6.5 Definition of AIC and how to use it in practice

Multiplying the preceding expansion by $-2n$ yields a criterion equivalent (up to asymptotically negligible terms) to maximizing expected test log score:

$$\mathrm{AIC} \;=\; -2\sum_{i=1}^{n} \log p_{\hat{\theta}_{\mathrm{MLE}}(x^n)}(x_i) \;+\; 2p.$$

How to use AIC to compare models:

- For each candidate model class $\mathcal{M}_m = \{p_\theta : \theta \in \Theta_m \subset \mathbb{R}^{p_m}\}$:
  1. Fit by MLE on the observed training data x^n to obtain $\hat{\theta}_{m,\mathrm{MLE}}(x^n)$.
  2. Compute
  $$\mathrm{AIC}_m \;=\; -2\sum_{i=1}^{n} \log p_{\hat{\theta}_{m,\mathrm{MLE}}(x^n)}(x_i) \;+\; 2p_m.$$
- Choose the model with the smallest AIC_m (best estimated out-of-sample performance).

7 Stein's paradox and the James–Stein estimator

7.1 Model setup and MLE benchmark

We observe a single random vector $\mathsf{z} \in \mathbb{R}^N$ drawn as

$$\mathsf{z} \sim \mathcal{N}(\mu, I_N),$$

where the mean vector $\mu \in \mathbb{R}^N$ is unknown and treated as a fixed (non-random) parameter. Equivalently, for each coordinate,

$$\mathsf{z}_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1), \qquad i \in [N],$$

and the coordinates are independent under $\mathbb{P}_\mu$. We evaluate an estimator $\hat{\mu}(\cdot)$ under squared-error loss

$$L(\mu, \hat{\mu}(\mathsf{z})) \;=\; \|\hat{\mu}(\mathsf{z}) - \mu\|_2^2,$$

and its risk at parameter value μ is

$$R(\mu, \hat{\mu}) \;=\; \mathbb{E}_{\mathsf{z}\sim\mathcal{N}(\mu, I_N)}\Big[\|\hat{\mu}(\mathsf{z}) - \mu\|_2^2\Big].$$

The likelihood factorizes across coordinates,

$$p_{\mathsf{z}|\mu}(z \mid \mu) \;=\; \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\!\Big(-\frac{1}{2}(z_i - \mu_i)^2\Big),$$

so maximizing coordinatewise yields the MLE

$$\hat{\mu}^{\mathrm{MLE}}(z) \;=\; z.$$

7.2 Stein's paradox: the MLE is not optimal for $N \geq 3$

For $N \geq 3$, the estimator $\hat{\mu}^{\mathrm{MLE}}(\mathsf{z}) = \mathsf{z}$ is *inadmissible* under squared-error loss: there exists another estimator $\hat{\mu}^{\star}$ such that

$$R(\mu, \hat{\mu}^{\star}) \;<\; R(\mu, \hat{\mu}^{\mathrm{MLE}}) \qquad \text{for all } \mu \in \mathbb{R}^N.$$

What makes it feel paradoxical is the tension between:

- *Decoupled sampling model:* $p_{\mathsf{z}|\mu}(z \mid \mu)$ factorizes across coordinates, suggesting each μ_i can be estimated separately by z_i.
- *Coupled objective:* the loss $\|\hat{\mu} - \mu\|_2^2 = \sum_{i=1}^N (\hat{\mu}_i - \mu_i)^2$ aggregates error across all coordinates, so it is meaningful to trade a small amount of bias for a larger reduction in total variance.

7.3 James–Stein shrinkage estimator dominates MLE

A concrete estimator that dominates the MLE for $N \geq 3$ is the James–Stein (JS) estimator

$$\hat{\mu}^{\mathrm{JS}}(\mathsf{z}) \;=\; \left(1 - \frac{N-2}{\|\mathsf{z}\|_2^2}\right) \mathsf{z}.$$

which *shrinks* the entire vector toward 0 by a data-dependent factor.

7.4 The Bayesian shrinkage approach to hierarchical gaussians

An empirical-Bayes perspective explains why shrinkage can help, and also motivates JS-like rules. First, assume a hierarchical Gaussian model:

$$\mu_i \sim \mathcal{N}(M, A), \qquad \mathsf{z}_i \mid \mu_i \sim \mathcal{N}(\mu_i, 1), \qquad i \in [N],$$

independently across i given (M, A). Here the randomness is:

- μ is random under the prior $\mathbb{P}_{M,A}$,
- z is random under the conditional law $\mathbb{P}_{\mathsf{z}|\mu}$,
- and after observing z, the posterior law is $\mathbb{P}_{\mu|z;M,A}$.

Bayes estimator: For squared-error loss, the Bayes estimator is the posterior mean, coordinate-wise:

$$\hat{\mu}_i^{\mathrm{Bayes}}(z) \;=\; \mathbb{E}_{\mu_i \sim \mathbb{P}_{\mu_i|z_i=z_i;M,A}}[\mu_i \mid z_i;\; A, M].$$

By Bayes' rule

$$p_{\mu|z}(\mu \mid z) \;\propto\; p_{z|\mu}(z \mid \mu)\, p_{\mu}(\mu).$$

where from above we have

$$p_{z|\mu}(z \mid \mu) \;\propto\; \exp\!\left(-\frac{1}{2}(z-\mu)^2\right), \qquad p_{\mu}(\mu) \;\propto\; \exp\!\left(-\frac{1}{2A}(\mu - M)^2\right).$$

and after computing this collecting terms we get that

$$\mu_i \mid z_i \sim \mathcal{N}\!\left(\frac{A}{A+1} z_i + \frac{1}{A+1} M,\; \frac{A}{A+1}\right),$$

so

$$\hat{\mu}_i^{\mathrm{Bayes}}(z) \;=\; \frac{A}{A+1}\, z_i \;+\; \frac{1}{A+1}\, M, \qquad i \in [N].$$

Thus the Bayes rule *shrinks* each z_i toward the common mean M.

Empirical Bayes: If (M, A) are unknown, one approach is to estimate them from the marginal

distribution of z_i under the hierarchical model.

$$\hat{M}(z) \;=\; \frac{1}{N}\sum_{i=1}^{N} z_i, \qquad \hat{A}(z) \;=\; \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \hat{M}(z))^2.$$

Plugging into the Bayes posterior mean yields an empirical-Bayes shrinkage rule

$$\hat{\mu}_i^{\mathrm{EB}}(z) \;=\; \frac{\hat{A}(z)}{\hat{A}(z)+1}\, z_i \;+\; \frac{1}{\hat{A}(z)+1}\, \hat{M}(z), \qquad i \in [N],$$

or equivalently,

$$\hat{\mu}_i^{\mathrm{EB}}(z) \;=\; \hat{M}(z) \;+\; \frac{\hat{A}(z)}{\hat{A}(z)+1}\,(z_i - \hat{M}(z)).$$

So empirical Bayes says: *pool information across coordinates to learn a common location $\hat{M}(z)$ and then shrink each coordinate toward it.*

7.5 Connection between Bayesian shrinkage and James–Stein

Both James–Stein and empirical-Bayes estimators share the same structural idea:

$$\hat{\mu}(z) \;=\; (\text{shrinkage center}) \;+\; (\text{shrinkage factor}) \cdot (z - \text{shrinkage center}).$$

James–Stein uses center 0 (or, in variants, $\hat{M}(z)1$) and a data-dependent shrinkage factor that depends on $\|z\|_2^2$. The empirical-Bayes rule uses center $\hat{M}(z)$ and a shrinkage factor determined by an estimated signal variance $\hat{A}(z)$.

The key shared message is the same as Stein's paradox: when $N \geq 3$, appropriately shrinking a multivariate mean estimate can reduce total risk

$$\mathbb{E}_{z\sim\mathcal{N}(\mu,I_N)}\left[\sum_{i=1}^{N}(\hat{\mu}_i(z) - \mu_i)^2\right]$$

relative to the naive coordinatewise estimator $\hat{\mu}(z) = z$.

8 Linear regression and confiednce intervals

8.1 Setup and notation

We observe training data as pairs

$$(x^{(i)}, y^{(i)}), \qquad i = 1, \ldots, n,$$

where $x^{(i)}$ are treated as deterministic inputs (features) and $y^{(i)}$ are responses. Define the empirical quantities

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)},$$

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \bar{x})^2.$$

8.2 Linear regression as prediction (ERM viewpoint)

We consider the linear predictor

$$\widehat{Y}(x) = \beta_0 + \beta_1 x.$$

and seek to find it by minimizing empirical squared error:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x^{(i)} - y^{(i)}\right)^2.$$

The solution is

$$\hat{\beta}_1 = \frac{1}{n s_X^2} \sum_{i=1}^{n} (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Practical recipe:

1. Compute $\bar{x}, \bar{y}$.
2. Compute s_X^2.
3. Compute $\hat{\beta}_1$ using the covariance formula above.
4. Compute $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.
5. Predict via $\widehat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$.

8.3 Goodness of fit: R^2

We now seek to answer how well our model fits the data. First define the Residual Sum of Squares and Total Sum of Squares as

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y^{(i)} - \widehat{Y}(x^{(i)})\right)^2,$$

$$\mathrm{TSS} = \sum_{i=1}^{n} (y^{(i)} - \bar{y})^2.$$

The coefficient of determination is

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}.$$

In simple linear regression,

$$R^2 = \mathrm{corr}^2(X, Y),$$

the squared empirical correlation.

8.4 Linear regression as inference

Assume the true model

$$Y^{(i)} = \beta_0 + \beta_1 x^{(i)} + \epsilon_i,$$

where the noise satisfies

$$\mathbb{E}[\epsilon_i] = 0, \qquad \mathrm{Var}(\epsilon_i) = \sigma^2,$$

and the noise ϵ_i are uncorrelated. Then our estimates are random variables

$$\hat{\beta}_1 = \beta_1 + W_1, \qquad \hat{\beta}_0 = \beta_0 + W_0,$$

where

$$\mathrm{Var}(W_1) = \frac{\sigma^2}{n s_X^2}, \qquad \mathrm{Var}(W_0) = \frac{\sigma^2(s_X^2 + \bar{x}^2)}{n s_X^2}.$$

However, our original estimators $\hat{\beta}_0, \hat{\beta}_1$ are unbiased:

$$\mathbb{E}[\hat{\beta}_0] = \beta_0, \qquad \mathbb{E}[\hat{\beta}_1] = \beta_1.$$

If ϵ_i are Gaussian, then $(\hat{\beta}_0, \hat{\beta}_1)$ are jointly Gaussian. Otherwise, they are approximately Gaussian for large n.

8.5 Estimating the noise variance

The variance σ^2 of our noise ϵ_i is unknown and must be estimated in order to construct confidence intervals for out points. As a first step, define the residuals

$$\hat{Z}^{(i)} = y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x^{(i)}.$$

The unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y^{(i)} - \hat{\beta}_0 - \hat{\beta}_1 x^{(i)} \right)^2.$$

The denominator $n-2$ accounts for estimating two parameters.

8.6 Confidence intervals for $\hat{\beta}_0, \hat{\beta}_1$

Since our estimators $(\hat{\beta}_0, \hat{\beta}_1)$ are jointly Gaussian due to the noise in our data model, we can form a confidence interval for them. Approximate 95% confidence intervals:

$$\hat{\beta}_1 \pm 1.96 \sqrt{\frac{\hat{\sigma}^2}{n s_X^2}},$$

$$\hat{\beta}_0 \pm 1.96 \sqrt{\frac{\hat{\sigma}^2 (s_X^2 + \bar{x}^2)}{n s_X^2}}.$$

Practical method:

1. Fit $\hat{\beta}_0, \hat{\beta}_1$.
2. Compute residuals and $\hat{\sigma}^2$.
3. Plug $\hat{\sigma}^2$ into the variance formulas.
4. Form the normal-based interval above.

8.7 Hypothesis testing for slope

To test

$$H : \beta_1 = 0 \quad \text{vs} \quad K : \beta_1 \neq 0,$$

compute the test statistic

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / (n s_X^2)}}.$$

For moderate/large n, compare $|T|$ to 1.96 for a 5% test (or use the t_{n-2} distribution for exact finite-sample inference under Gaussian noise).

8.8 Prediction error at a new point

For a new input x_{new}, prediction is

$$\widehat{Y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}.$$

The conditional mean squared prediction error is

$$\mathbb{E}\left[\left(\widehat{Y}_{\text{new}} - Y_{\text{new}}\right)^2 \,\middle|\, X = x_{\text{new}}\right] = \sigma^2 + \frac{\sigma^2}{n s_X^2}\left(s_X^2 + (x_{\text{new}} - \bar{x})^2\right).$$

or symbolically,

$$\text{prediction error} = \underbrace{\sigma^2}_{\text{irreducible noise}} + \underbrace{\text{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}\right)}_{\text{estimation variance}}$$

Part II Inference and information

1 Preliminaries and introduction

1.1 Notation for common objects

- Random quantities are written as x
- Deterministic quantities, such as sample realizations, are x
- Randomized functions are written as $\mathsf{f}(\cdot)$.
- Deterministic functions are written $f(\cdot)$
- Consequently, $f(\mathsf{x})$, $\mathsf{f}(x)$, and $\mathsf{f}(\mathsf{x})$ are random variables.
- Sets and events are written as $\mathcal{X}$. Sets that have structure are called ordinal, and sets without structure are called categorical. For an alphabet $\mathcal{X}$ and a positive integer N, we write
$$\mathcal{X}^N \;=\; \underbrace{\mathcal{X} \times \cdots \times \mathcal{X}}_{N\text{-fold}}$$
for the alphabet of N-tuples with entries in $\mathcal{X}$. Equivalently,
$$(x_1, x_2, \ldots, x_N) \in \mathcal{X}^N \quad \Longleftrightarrow \quad x_i \in \mathcal{X} \;\text{ for } i = 1, 2, \ldots, N.$$

1.2 PMF notation

For a discrete random variable $\mathsf{x} \in \mathcal{X}$, we denote its PMF by $p_{\mathsf{x}}(\cdot)$ and write (the slightly abused notation)
$$p_{\mathsf{x}}(x) = \mathbb{P}(\mathsf{x} = x),$$
where $\mathbb{P}$ denotes the underlying probability law. For an event $\mathcal{E} \subset \mathcal{X}$, we write
$$\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathsf{x} \in \mathcal{E}).$$
When needed to avoid confusion, we explicitly write the distruction with respect to which the probability function corresponds. For example, if x has PMF q, we may write
$$\mathbb{P}_q(\mathsf{x} \in \mathcal{E}) = \sum_{x \in \mathcal{E}} q(x).$$
Sometimes, when convenient and unambiguous, we also write $\mathbb{Q}(\mathcal{E})$ or $\mathbb{Q}(\mathsf{x} \in \mathcal{E})$ for the law corresponding to q.

1.3 Expectation and conditioning operator

We use $\mathbb{E}[\cdot]$ to denote expectation. When needed to avoid confusion, we make explicit the underlying probability law by adding a subscript. For example, if x has density q, then
$$\mathbb{E}_q[f(\mathsf{x})] \;=\; \int f(x)\, q(x)\, dx.$$
For a conditional distribution $p_{\mathsf{x}|y}(\cdot \mid y)$, we write
$$\mathbb{E}_{p_{\mathsf{x}|y}(\cdot|y)}[f(\mathsf{x})] \;=\; \mathbb{E}[f(\mathsf{x}) \mid y = y] \;=\; \int f(x)\, p_{\mathsf{x}|y}(x \mid y)\, dx.$$

1.4 Moment generating functions and characteristic functions

When well-defined, the moment-generating function of x is

$$M_x(s) \;=\; \mathbb{E}[e^{sx}]\,,$$

and the characteristic function is

$$M_x(j\omega) \;=\; \mathbb{E}\Big[e^{j\omega x}\Big]\,.$$

1.5 Sequence and subsequence notation

For a sequence $(x_1, x_2, \ldots)$, we use script notation to denote subsequences. In particular, for indices $j \geq i$,

$$x_i^j \;=\; (x_i, x_{i+1}, \ldots, x_j).$$

As a further shorthand, for $n \geq 1$ we often write

$$x^n \;=\; x_1^n.$$

1.6 Big-O notation

For functions $f(\cdot)$ and $g(\cdot)$ defined on a suitable alphabet $\mathcal{U}$, we write

$$g(u) = O(f(u)) \quad \text{as } u \to a$$

to mean that there exist finite constants $c > 0$ and $\epsilon > 0$ such that

$$|g(u)| \leq c\,|f(u)| \quad \text{for all } u \text{ with } |u - a| < \epsilon.$$

1.7 Little-o notation

For functions $f(\cdot)$ and $g(\cdot)$ defined on a suitable alphabet $\mathcal{U}$, we write

$$g(u) = o(f(u)) \quad \text{as } u \to a$$

to mean that for every constant $c > 0$, there exists $\epsilon > 0$ such that

$$|g(u)| \leq c\,|f(u)| \quad \text{for all } u \text{ with } |u - a| < \epsilon.$$

1.8 Gaussian tail (Q-function)

The tail probability of a standard normal random variable is denoted by the Q-function,

$$Q(x) \;\coloneqq\; \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt.$$

1.9 Gaussian random vectors

We write $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Lambda})$ to denote that $\mathbf{x}$ is a Gaussian random vector with mean

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

and covariance

$$\mathbb{E}\Big[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\Big] = \boldsymbol{\Lambda}.$$

When $\boldsymbol{\Lambda}$ is nonsingular, $\mathbf{x}$ has density

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Lambda}) \;\coloneqq\; \frac{1}{|2\pi\boldsymbol{\Lambda}|^{1/2}}\exp\!\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

2 Bayesian hypothysis testing

2.1 The typical model of inference

We typically think of inference as the following process

$$\textit{phenomena} \longrightarrow x \longrightarrow \boxed{\text{model}} \longrightarrow y \longrightarrow \boxed{f(\cdot)} \longrightarrow \hat{x}$$

- We care about x, which is determined from some phenomena. We do not observe this phenomena or its latent representation x
- However, we do see data y, which is x through some model that we assume. We often model this data y as being random.
- $\widehat{x}$ is the estimate of x, generated by applying our design $f(\cdot)$ over data y to obtain our prediction of the x that generated y. We will set this method $f(\cdot)$ based on desired performance criteria.

2.2 Bayesian hypothesis testing

We consider $M \geq 2$ competing hypotheses

$$\mathcal{H} = \{H_0, H_1, \ldots, H_{M-1}\},$$

where each hypothesis specifies a different probabilistic model for the observed data. The observed data is a random vector y and the unknown hypothesis is a random variable $\mathsf{H} \in \mathcal{H}$. A Bayesian specification consists of:

- prior (a priori) probabilities over hypotheses:

$$p_{\mathsf{H}}(H_m) = \mathbb{P}_{\mathsf{H}}(\mathsf{H} = H_m), \qquad m = 0, 1, \ldots, M-1;$$

- conditional distributions (likelihood models) for the data under each hypothesis:

$$p_{\mathsf{y}|\mathsf{H}}(\,\cdot \mid H_m), \qquad m = 0, 1, \ldots, M-1.$$

After observing $\mathsf{y} = y$, our knowledge of the true hypothesis is summarized by the posterior (a posteriori) probabilities

$$p_{\mathsf{H}|\mathsf{y}}(H_m \mid y) = \frac{p_{\mathsf{y}|\mathsf{H}}(y \mid H_m)\, p_{\mathsf{H}}(H_m)}{\sum_{m'} p_{\mathsf{y}|\mathsf{H}}(y \mid H_{m'})\, p_{\mathsf{H}}(H_{m'})}.$$

for each $m = 0, 1, \ldots, M-1$. The prior p_{H} is the prior belief, and the posterior $p_{\mathsf{H}|\mathsf{y}}(\cdot \mid y)$ is the revised belief after observing the data y. In many applications one subsequently makes a decision (i.e., selects a hypothesis) based on the posterior.

2.3 Binary hypothesis testing

In the binary case, the hypothesis set is

$$\mathcal{H} = \{H_0, H_1\},$$

and our Bayesian model has two components: prior probabilities on H and likelihood models for observation y.

- *Prior on each hypothesis in $\mathcal{H}$.* Define the priors

$$P_0 := \mathbb{P}_{\mathsf{H}}(\mathsf{H} = H_0), \qquad P_1 := \mathbb{P}_{\mathsf{H}}(\mathsf{H} = H_1) = 1 - P_0.$$

- *Likelihoods.* The conditional distributions of y under each hypothesis in $\mathcal{H}$ are

$$p_{\mathsf{y}|\mathsf{H}}(y \mid H_0), \qquad p_{\mathsf{y}|\mathsf{H}}(y \mid H_1).$$

2.4 The *solution* to hypothysis testing

The solution to a hypothesis test is specified in terms of a decision rule. For now, we focus on deterministic decision rules that decide over hypotheses. In particular, a binary decision rule is a function $\widehat{H}(\cdot)$ that maps each observation $y \in \mathcal{Y}$ to one of the two hypotheses:

$$\widehat{H} : \mathcal{Y} \to \mathcal{H}, \qquad \mathcal{H} = \{H_0, H_1\}.$$

Choosing $\widehat{H}(\cdot)$ is equivalent to partitioning the observation space $\mathcal{Y}$ into two disjoint decision regions corresponding to the two possible decisions.

$$\mathcal{Y}_0 = \{y \in \mathcal{Y} : \widehat{H}(y) = H_0\},$$
$$\mathcal{Y}_1 = \{y \in \mathcal{Y} : \widehat{H}(y) = H_1\}.$$

2.5 Cost functions measure the performance of $\widehat{H}(y) \in \mathcal{H}$

We evaluate the quality of a decision rule via an expected cost. Let $C(H_j, H_i) = C_{ij}$ denote the cost of deciding $\widehat{H}(\cdot) = H_i$ when the true hypothesis is $H = H_j$. Equivalently, for a generic decision rule $f(\cdot) : \mathcal{Y} \to \mathcal{H}$, the incurred loss is $C(\mathsf{H}, f(\mathsf{y}))$, and we seek to minimize the the expected cost, known as Bayes risk

$$\varphi(f) := \mathbb{E}_{(\mathsf{H},\mathsf{y})}[C(\mathsf{H}, f(\mathsf{y}))]\,,$$

In this setting, the Bayes optimal decision rule is

$$\begin{aligned}\widehat{H}(\cdot) &= \arg\min_{f(\cdot)} \mathbb{E}_{(\mathsf{H},\mathsf{y})}[C(\mathsf{H}, f(\mathsf{y})).] \\ &= \arg\min_{f(\cdot)} \varphi(f)\end{aligned}$$

A common symmetric choice is the 0–1 loss function,

$$C_{ij} = 1 - 1_{i=j},$$

i.e.,

$$C_{00} = C_{11} = 0, \qquad C_{01} = C_{10} = 1,$$

Note that this $0-1$ loss function corresponds to minimizing the probability of classification error $\mathbb{P}_{\text{error}}$

2.6 Finding the Bayes optimum decision rule using likelihood ratio

We now need to find

$$\widehat{H}(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}_{(\boldsymbol{H},\boldsymbol{y})}[C(\boldsymbol{H}, f(\boldsymbol{y})).]$$

Fix an arbitrary (deterministic) decision rule $f(\cdot) : \mathcal{Y} \to \mathcal{H}$, where $\mathcal{H} = \{H_0, H_1\}$. And recall we call the Bayes risk

$$\varphi(f) \;:=\; \mathbb{E}[C(\boldsymbol{H}, f(\boldsymbol{y}))]\,.$$

By iterated expectation,

$$\varphi(f) = \mathbb{E}_{\boldsymbol{y}}\Big[\mathbb{E}_{\boldsymbol{H}}[C(\boldsymbol{H}, f(\boldsymbol{y})) \,|\, \boldsymbol{y} = y]\Big] = \int \tilde{\varphi}(f(y), y)\, p_{\boldsymbol{y}}(y)\, dy,$$

where we define the conditional (posterior) risk

$$\tilde{\varphi}(f(y), y) \;:=\; \mathbb{E}_{\boldsymbol{H}}[C(\boldsymbol{H}, f(\boldsymbol{y})) \,|\, \boldsymbol{y} = y]\,.$$

Since $p_{\boldsymbol{y}}(y) \geq 0$, we minimize φ by minimizing $\tilde{\varphi}(f(y), y)$ pointwise in y. Thus, for each y we choose $f(y) \in \{H_0, H_1\}$ to minimize $\tilde{\varphi}(f(y), y)$. Writing $C(H_j, H_i) = C_{ji}$, we have

$$\begin{aligned}
\tilde{\varphi}(H_0, y) &= C_{00}\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_0 \mid y) + C_{01}\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_1 \mid y) \\
\tilde{\varphi}(H_1, y) &= C_{10}\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_0 \mid y) + C_{11}\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_1 \mid y).
\end{aligned}$$

Therefore, the optimal assignment is simply to choose the hypothysis that corresponds with the lower expected conditional risk

$$C_{00}\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_0 \mid y) + C_{01}\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_1 \mid y) \underset{\widehat{H}(y)=H_0}{\overset{\widehat{H}(y)=H_1}{\gtrless}} C_{10}\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_0 \mid y) + C_{11}\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_1 \mid y).$$

Rearranging,

$$(C_{01} - C_{11})\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_1 \mid y) \underset{\widehat{H}(y)=H_0}{\overset{\widehat{H}(y)=H_1}{\gtrless}} (C_{10} - C_{00})\, p_{\boldsymbol{H}|\boldsymbol{y}}(H_0 \mid y),$$

and equivalently (assuming $C_{01} - C_{11} > 0$ and $C_{10} - C_{00} > 0$),

$$\frac{p_{\boldsymbol{H}|\boldsymbol{y}}(H_1 \mid y)}{p_{\boldsymbol{H}|\boldsymbol{y}}(H_0 \mid y)} \underset{\widehat{H}(y)=H_0}{\overset{\widehat{H}(y)=H_1}{\gtrless}} \frac{C_{10} - C_{00}}{C_{01} - C_{11}}.$$

Using Bayes' rule with priors $P_m := \mathbb{P}(\boldsymbol{H} = H_m)$ and likelihoods $p_{\boldsymbol{y}|\boldsymbol{H}}(y \mid H_m)$,

$$p_{\boldsymbol{H}|\boldsymbol{y}}(H_m \mid y) = \frac{p_{\boldsymbol{y}|\boldsymbol{H}}(y \mid H_m)\, P_m}{p_{\boldsymbol{y}}(y)}, \qquad m \in \{0, 1\}.$$

Substituting into the posterior-odds test gives the *likelihood ratio test*:

$$\boxed{\mathcal{L}(\mathbf{y}) = \frac{p_{\boldsymbol{y}|\boldsymbol{H}}(y \mid H_1)}{p_{\boldsymbol{y}|\boldsymbol{H}}(y \mid H_0)} \underset{\widehat{H}(y)=H_0}{\overset{\widehat{H}(y)=H_1}{\gtrless}} \frac{C_{10} - C_{00}}{C_{01} - C_{11}} \cdot \frac{P_0}{P_1} = \eta.}$$

Notice that the threshold is completely determined *a priori* probabilities and costs, while the likelihood is only a function of the data and observation model.

2.7 MAP rule and probability of error under the 0–1 loss

Assume the 0–1 loss

$$C_{ij} = 1 - 1_{i=j}, \qquad C_{00} = C_{11} = 0, \quad C_{01} = C_{10} = 1$$

For a decision rule $f(\cdot)$, the Bayes risk equals the probability of error:

$$\begin{aligned}\varphi(f) &= \mathbb{E}[C(H, f(y))] \\ &= \mathbb{E}[\mathbb{E}[C(H, f(y)) \mid y]] \\ &= \mathbb{P}(f(y) = H_0,\ H = H_1) + \mathbb{P}(f(y) = H_1,\ H = H_0) \\ &= \mathbb{P}_{\text{error}}\end{aligned}$$

Moreover, the Bayes optimal decision rule for this setting of the 0–1 loss is the function that chooses the hypothesis with the larger posterior probability:

$$\widehat{H}(y) \in \arg \max_{H \in \{H_0, H_1\}} p_{H|y}(H \mid y).$$

This is the *maximum a posteriori* (MAP) decision rule. Note that this uses the likelihood of the hypothysis under the data, which we often rewrite using Bayes rule.

3 Non-Bayesian hypothesis testing

We look at hypothesis testing when either priors or costs are unavailable.

3.1 Error probabilities for a binary decision rule

Consider a binary hypothesis test with $H \in \{H_0, H_1\}$ and a decision rule $\widehat{H}(y) \in \{H_0, H_1\}$ based on observation y. The performance of any decision rule $\widehat{H}(\cdot)$ is fully specified by the pair

$$P_{\mathrm{F}} \coloneqq \mathbb{P}\Big(\widehat{H}(y) = H_1 \,\Big|\, H = H_0\Big), \qquad P_{\mathrm{D}} \coloneqq \mathbb{P}\Big(\widehat{H}(y) = H_1 \,\Big|\, H = H_1\Big).$$

which correspond to false alarm probability and true detection probability. Any deterministic decision rule is equivalent to specifying a partition of the observation space $\mathcal{Y}$ into two decision regions

$$\mathcal{Y}_0 \coloneqq \{y \in \mathcal{Y} : \widehat{H}(y) = H_0\}, \qquad \mathcal{Y}_1 \coloneqq \{y \in \mathcal{Y} : \widehat{H}(y) = H_1\},$$

Let $p_{\mathsf{y}|H}(\cdot \mid H_i)$ denote the conditional density of y under hypothesis H_i. Then

$$P_{\mathrm{F}} \;=\; \mathbb{P}(\mathsf{y} \in \mathcal{Y}_1 \mid H = H_0) \;=\; \int_{\mathcal{Y}_1} p_{\mathsf{y}|H}(y \mid H_0)\, dy,$$

and

$$P_{\mathrm{D}} \;=\; \mathbb{P}(\mathsf{y} \in \mathcal{Y}_1 \mid H = H_1) \;=\; \int_{\mathcal{Y}_1} p_{\mathsf{y}|H}(y \mid H_1)\, dy.$$

Typically P_{D} and P_{F} are competing objectives: enlarging the acceptance region $\mathcal{Y}_1$ tends to increase both P_{D} and P_{F}, while shrinking $\mathcal{Y}_1$ tends to decrease both P_{D} and P_{F}.

3.2 Likelihood ratio test family indexed by η

Consider the family of likelihood ratio test (LRT) decision rules indexed by a threshold $\eta \in \mathbb{R}$, Equivalently,

$$\mathcal{L}(y) \;\begin{cases} \geq \eta, & \widehat{H}(y) = H_1, \\ < \eta, & \widehat{H}(y) = H_0. \end{cases}$$

where

$$\mathcal{L}(y) \coloneqq \frac{p_{\mathsf{y}|H}(y \mid H_1)}{p_{\mathsf{y}|H}(y \mid H_0)}.$$

3.3 Operating characteristic traced by varying η

Each threshold η uniquely specifies a decision rule, and hence an operating point $(P_{\mathrm{F}}(\eta), P_{\mathrm{D}}(\eta))$ in the $(P_{\mathrm{F}}, P_{\mathrm{D}})$ plane:

$$P_{\mathrm{F}}(\eta) \;=\; \mathbb{P}(\mathcal{L}(\mathsf{y}) \geq \eta \,|\, H = H_0), \qquad P_{\mathrm{D}}(\eta) \;=\; \mathbb{P}(\mathcal{L}(\mathsf{y}) \geq \eta \,|\, H = H_1).$$

As η is varied from 0 to ∞, the corresponding collection of operating points

$$\{(P_{\mathrm{F}}(\eta), P_{\mathrm{D}}(\eta)) : \eta \in (0, \infty)\}$$

is traced out in the $(P_{\mathrm{F}}, P_{\mathrm{D}})$ plane. This collection is called the operating characteristic of the LRT (OC-LRT). We note that OC-LRT is non-decreasing, since for $\eta_2 > \eta_1$ we have that P_{D} is non-increasing and P_{D} is non-increasing.

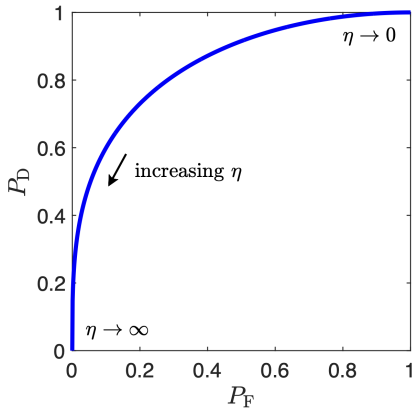

Figure 1: Operating characteristic of the likelihood ratio test.

3.4 Bayes risk as a linear objective in (P_F, P_D)

Assume prior probabilities

$$P_0 = \mathbb{P}(H = H_0), \qquad P_1 = \mathbb{P}(H = H_1)$$

and costs C_{ij} of deciding H_i when the true $H = H_j$. For a deterministic rule $\hat{H}(\cdot)$, the Bayes risk is

$$\begin{aligned}\varphi(\hat{H}) &= \mathbb{E}_{y,H}\Big[C(H, \hat{H}(y))\Big] \\ &\qquad \text{and using law of iterated expectation} \\ &= \sum_{i,j\in\{0,1\}} C_{ij}\, \mathbb{P}\big(\hat{H}(y) = H_i \,\big|\, H = H_j\big)\, P_j,\end{aligned}$$

where probabilities are taken under the corresponding conditional law $y \sim p_{y|H}(\cdot \mid H_j)$.
Write the two key error probabilities:

$$P_F = \mathbb{P}\big(\hat{H}(y) = H_1 \mid H = H_0\big), \qquad P_D = \mathbb{P}\big(\hat{H}(y) = H_1 \mid H = H_1\big).$$

Then

$$\mathbb{P}\big(\hat{H}(y) = H_0 \mid H = H_0\big) = 1 - P_F, \qquad \mathbb{P}\big(\hat{H}(y) = H_0 \mid H = H_1\big) = 1 - P_D.$$

Substitute into $\varphi(\hat{H})$ and collect terms:

$$\begin{aligned}\varphi(\hat{H}) &= C_{00}(1 - P_F)P_0 + C_{10}P_F P_0 + C_{01}(1 - P_D)P_1 + C_{11}P_D P_1 \\ &= \underbrace{(C_{10} - C_{00})P_0}_{\alpha}\, P_F \;-\; \underbrace{(C_{01} - C_{11})P_1}_{\beta}\, P_D \;+\; \underbrace{(C_{00}P_0 + C_{01}P_1)}_{\gamma}.\end{aligned}$$

Hence the Bayes risk can be written as

$$\varphi(\hat{H}) = \alpha P_F - \beta P_D + \gamma,$$

for

$$\alpha = (C_{10} - C_{00})P_0, \quad \beta = (C_{01} - C_{11})P_1, \quad \gamma = C_{00}P_0 + C_{01}P_1.$$

Thus, among LRTs (or more generally among any tests), the Bayesian choice is the one that

minimizes a linear combination of P_F and P_D.

3.5 Equi-Bayes-risk lines and the Bayesian operating point

Fix a constant c and consider the set of tests with the same Bayes risk $\varphi(\hat{H}) = c$. From

$$\alpha P_F - \beta P_D + \gamma = c$$

we can solve for P_D in terms of P_F:

$$P_D = \frac{\alpha}{\beta} P_F \; + \; \frac{\gamma - c}{\beta}.$$

So the level sets of Bayes risk are straight lines in the (P_F, P_D) plane, all with slope

$$\eta_B \; \coloneqq \; \frac{\alpha}{\beta} = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}.$$

The Bayesian operating point on the OC/ROC curve is obtained by finding the smallest value of c such that an equi-risk line intersects the OC/ROC curve. Geometrically, the optimal point is where an equi-risk line is just tangent to (or first touches) the achievable curve.

3.6 Connection to the Bayesian LRT threshold

The slope parameter

$$\eta_B = \frac{\alpha}{\beta} = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}$$

is exactly the likelihood-ratio threshold that appears in the optimal Bayesian LRT. Thus:

- choosing priors and costs determines η_B,
- choosing η_B selects a single LRT $\hat{H}_{\eta_B}(\cdot)$,
- and that LRT corresponds to a single operating point $(P_F(\eta_B), P_D(\eta_B))$ on the OC/ROC curve.

3.7 Neyman–Pearson criterion: maximize detection subject to a false-alarm constraint

When it is difficult or unnatural to assign costs, a common alternative is the Neyman–Pearson formulation:

$$\max_{\hat{H}(\cdot)} \; P_D \qquad \text{subject to} \qquad P_F \leq \alpha,$$

where $\alpha \in (0, 1)$ is the maximum allowable false-alarm probability.

3.8 Neyman–Pearson lemma (deterministic rules)

Assume the likelihood ratio $\mathcal{L}(y)$ is a purely continuous random variable under each hypothesis (i.e., $\mathbb{P}_{y \sim p_{y|H}(\cdot|H_j)}(\mathcal{L}(y) = \eta) = 0$ for all η and $j \in \{0, 1\}$). Then an optimal solution to the Neyman–Pearson problem among deterministic rules is an LRT:

$$\hat{H}(y) = H_1 \, 1\{\mathcal{L}(y) \geq \eta\} \; + \; H_0 \, 1\{\mathcal{L}(y) < \eta\},$$

where the threshold η is chosen so that the constraint is met with equality:

$$P_F = \mathbb{P}_{y \sim p_{y|H}(\cdot|H_0)}(\mathcal{L}(y) \geq \eta) = \alpha.$$

(When $\mathcal{L}(y)$ has point masses, exact equality may require randomization on the boundary $\{\mathcal{L}(y) = \eta\}$.)

3.9 Varying η traces out achievable operating characteristic curve of LRTs

Varying η traces out the achievable operating characteristic curve of LRTs: each η produces a pair $(P_F(\eta), P_D(\eta))$. Under the Neyman–Pearson criterion, the operating point is found by selecting the η such that $P_F(\eta) = \alpha$ and reading off the corresponding $P_D(\eta)$.

4 Performance limits of hypothesis testing

4.1 The likelihood ratio and operating characteristic may be discontinuous

Consider binary-valued data $y \in \mathcal{Y} = \{0, 1\}$ with conditional PMFs

$$p_{y|H}(y \mid H_0) = \begin{cases} \frac{1}{2}, & y = 0, \\ \frac{1}{2}, & y = 1, \end{cases} \qquad p_{y|H}(y \mid H_1) = \begin{cases} \frac{2}{3}, & y = 0, \\ \frac{1}{3}, & y = 1. \end{cases}$$

The likelihood ratio is discontinuous

$$\mathcal{L}(y) \;=\; \frac{p_{y|H}(y \mid H_1)}{p_{y|H}(y \mid H_0)} = \begin{cases} \frac{4}{3}, & y = 0, \\ \frac{2}{3}, & y = 1. \end{cases}$$

The likelihood ratio test (LRT) with threshold η decides H_1 iff $\mathcal{L}(y) \geq \eta$. The false-alarm and detection probabilities are

$$P_F(\eta) \;=\; \mathbb{P}(\mathcal{L}(y) \geq \eta \mid H = H_0) = \begin{cases} 0, & \frac{4}{3} < \eta, \\ \frac{1}{2}, & \frac{2}{3} < \eta \leq \frac{4}{3}, \\ 1, & \eta \leq \frac{2}{3}, \end{cases}$$

$$P_D(\eta) \;=\; \mathbb{P}(\mathcal{L}(y) \geq \eta \mid H = H_1) = \begin{cases} 0, & \frac{4}{3} < \eta, \\ \frac{2}{3}, & \frac{2}{3} < \eta \leq \frac{4}{3}, \\ 1, & \eta \leq \frac{2}{3}. \end{cases}$$

Hence the operating characteristic (ROC) consists of exactly three achievable (P_F, P_D) points:

$$(0, 0), \quad \left(\frac{1}{2}, \frac{2}{3}\right), \quad (1, 1).$$

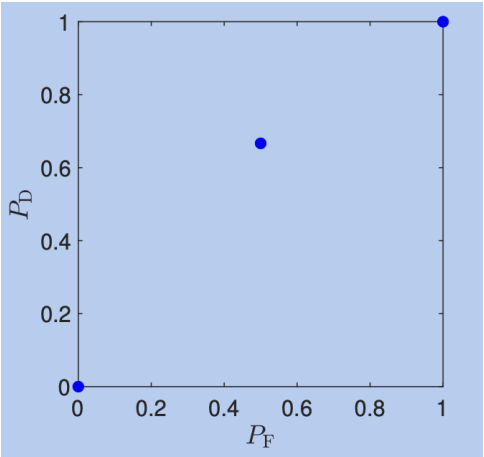

Figure 2: Operating characteristic may be discontinuous in η

It is important to emphasize that discontinuities in the operating characteristic $\{(P_F(\eta), P_D(\eta)) : \eta \geq 0\}$ arise due to discrete components in the distribution of the likelihood ratio $\mathcal{L}(y)$ *under* the constituent hypotheses (note definition of $P_F(\eta)$, $P_D(\eta)$ to see this). In particular, whether the OC is discontinuous is unrelated to whether there are discontinuities in the likelihood ratio $\mathcal{L}(y)$ as a function of the observed data y.

4.2 NP value $P_D^{\mathrm{NP}}(\alpha)$ is discontinuous for discrete $\mathcal{L}(\cdot)$

Consider the Neyman–Pearson problem

$$P_D^{\mathrm{NP}}(\alpha) \;:=\; \max_{\hat{H}(\cdot)} \; P_D \quad \text{subject to} \quad P_F \leq \alpha,$$

when we restrict $\hat{H}(\cdot)$ to be a (deterministic) likelihood-ratio test (LRT) with some threshold η. If the likelihood ratio $\mathcal{L}(y)$ has discrete components, then as η varies the achievable operating points $(P_F(\eta), P_D(\eta))$ form a discrete set (the OC-LRT). As α increases continuously, the selected OC-LRT point stays fixed until α crosses the next achievable P_F level, at which time the optimizer jumps to a new point with a larger P_D. This implies that $P_D^{\mathrm{NP}}(\alpha)$ (under the restriction to deterministic LRTs) is a step function of α.

4.3 Randomizing between two LRT thresholds yields convex combinations on the OC

Under the Neyman–Pearson criterion, we can improve achievable operating points by *randomization*. Fix two LRTs $\hat{H}'(\cdot)$ and $\hat{H}''(\cdot)$ corresponding to thresholds η' and η'', with $\eta'' > \eta'$. Let $u \in \{\eta', \eta''\}$ be independent of y with

$$\mathbb{P}(u = \eta') = p, \qquad \mathbb{P}(u = \eta'') = 1 - p, \qquad p \in [0, 1].$$

Define the *randomized test*

$$\hat{H}(y) = \begin{cases} \hat{H}'(y), & \text{if } u = \eta', \\ \hat{H}''(y), & \text{if } u = \eta''. \end{cases}$$

Then its detection and false-alarm probabilities are the convex combinations

$$P_D \;=\; p\, P_D(\eta') + (1-p)\, P_D(\eta''), \qquad P_F \;=\; p\, P_F(\eta') + (1-p)\, P_F(\eta'').$$

Equivalently, the operating point (P_F, P_D) lies on the line segment joining $(P_F(\eta''), P_D(\eta''))$ and $(P_F(\eta'), P_D(\eta'))$ and as p increases from 0 to 1, the operating point moves from the η''-point

to the η'-point.

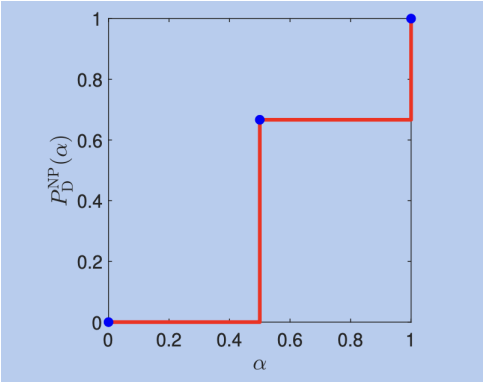

Figure 3: Oringal LRT-OC

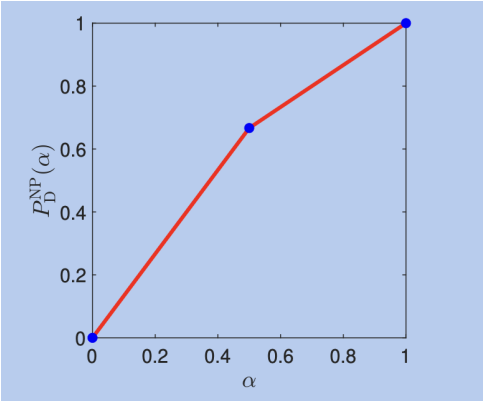

Figure 4: Randomization leasds to convex combination

4.4 Claim: randomization cannot beat the Bayes optimal LRT

In binary Bayesian hypothesis testing, allowing randomized decision rules does *not* reduce the minimum achievable Bayes risk: an optimal deterministic rule already attains the Bayes-optimal risk among *all* (possibly randomized) tests.

A randomized test is specified by a function $q(\cdot)$ with $q(y) \in [0,1]$, where

$$q(y) \;\coloneqq\; \mathbb{P}(\hat{H} = H_1 \mid \mathsf{y} = y), \qquad \mathbb{P}(\hat{H} = H_0 \mid \mathsf{y} = y) = 1 - q(y).$$

The Bayes risk of the randomized rule $q(\cdot)$ is

$$\varphi(q) \;\coloneqq\; \mathbb{E}\Big[C(H, \hat{H})\Big].$$

By iterated expectation,

$$\varphi(q) = \mathbb{E}\Big[\,\mathbb{E}\Big[C(H,\hat{H})\,\Big|\,\mathsf{y}\Big]\Big] = \mathbb{E}[\tilde{\varphi}(\mathsf{y})]\,, \qquad \tilde{\varphi}(y) \;\coloneqq\; \mathbb{E}\Big[C(H,\hat{H})\,\Big|\,\mathsf{y} = y\Big]\,.$$

Therefore, to minimize $\varphi(q)$ it suffices to minimize $\tilde{\varphi}(y)$ *pointwise* for each fixed y.

Fix y. Under the posterior on H given $\mathsf{y} = y$, define the two posterior expected costs

$$r_0(y) \;\coloneqq\; \mathbb{E}[C(H, H_0) \mid \mathsf{y} = y]\,, \qquad r_1(y) \;\coloneqq\; \mathbb{E}[C(H, H_1) \mid \mathsf{y} = y]\,.$$

If we randomize at y with probability $q(y)$ of choosing H_1, then

$$\tilde{\varphi}(y) = (1 - q(y))\, r_0(y) + q(y)\, r_1(y),$$

which is a convex combination of $r_0(y)$ and $r_1(y)$. Hence the minimum over $q(y) \in [0,1]$ is achieved at an endpoint:

$$\min_{q(y)\in[0,1]} \tilde{\varphi}(y) = \min\{r_0(y), r_1(y)\},$$

so there exists an optimal rule that is *deterministic* at each y:

$$\hat{H}(y) \in \arg\min_{h\in\{H_0,H_1\}} \mathbb{E}[C(H,h) \mid \mathsf{y} = y]\,.$$

4.5 False positive and detection probabilities under randomized rules

A randomized binary decision rule is specified by

$$q(y) \;\coloneqq\; p_{\hat{H}|\mathsf{y}}(H_1 \mid y) \in [0,1],$$

therefore

$$P_F(q) \;=\; \mathbb{P}(\hat{H}(\mathsf{y}) = H_1 \mid H = H_0) \;=\; \int q(y)\, p_{\mathsf{y}|H}(y \mid H_0)\, dy \;=\; \mathbb{E}[q(\mathsf{y}) \mid H = H_0],$$

$$P_D(q) \;=\; \mathbb{P}(\hat{H}(\mathsf{y}) = H_1 \mid H = H_1) \;=\; \int q(y)\, p_{\mathsf{y}|H}(y \mid H_1)\, dy \;=\; \mathbb{E}[q(\mathsf{y}) \mid H = H_1].$$

4.6 Neyman–Pearson lemma, full version (with randomization)

Neyman–Pearson lemma: Given distributions of data under hypothyses

$$p_{\mathsf{y}|H}(\cdot \mid H_0) \quad \text{and} \quad p_{\mathsf{y}|H}(\cdot \mid H_1)$$

and any $\alpha \in [0,1]$, there exist threshold η and $p \in [0,1]$ such that the randomized rule

$$q_\star(y) = \begin{cases} 0, & \mathcal{L}(y) < \eta, \\ p, & \mathcal{L}(y) = \eta, \\ 1, & \mathcal{L}(y) > \eta, \end{cases}$$

satisfies

$$P_F(q_\star) = \alpha, \qquad P_D(q_\star) \;\geq\; P_D(q) \;\text{ for any } q(\cdot) \text{ with } P_F(q) \leq \alpha.$$

Earlier, when we restricted attention to *deterministic* LRTs, the set of achievable (P_F, P_D) points can be discrete when $\mathcal{L}(\mathsf{y})$ has atoms. Then for many values of α, no deterministic LRT achieves $P_F = \alpha$; the optimizer picks the largest achievable $P_F < \alpha$, making the NP value $P_D^{\mathrm{NP}}(\alpha)$ a step function of α (hence discontinuous).

The full NP lemma resolves this by allowing *randomization on the boundary* $\{\mathcal{L}(y) = \eta\}$. By choosing the tie-breaking probability p, we can tune $P_F(q_\star)$ continuously to hit *any* target level $\alpha \in [0,1]$ exactly.

4.7 The Neyman–Pearson function $\zeta_{\mathrm{NP}}(\alpha)$

By the theorem above, the best performance curves are achievable *without* restricting the form of the test: optimizing over all tests is equivalent to optimizing over (possibly randomized)

likelihood-ratio tests.

We define the *Neyman–Pearson function* of the model as the mapping $\zeta_{\mathrm{NP}}(\cdot) : [0,1] \to [0,1]$ of false-alarm constraint levels to detection probability of the optimal randomized decision rule at false-alarm level $\alpha \in [0,1]$,

$$\zeta_{\mathrm{NP}}(\alpha) \;\coloneqq\; P_D(q_\star), \qquad \text{with } q_\star \text{ chosen so that } P_F(q_\star) = \alpha.$$

Equivalently, $\zeta_{\mathrm{NP}}(\alpha)$ is the maximum achievable detection probability among all (possibly randomized) decision rules with false-alarm probability exactly α.

4.8 Efficient frontier and the Neyman–Pearson function

Given the Neyman–Pearson function

$$\zeta_{\mathrm{NP}}(\alpha) \;\coloneqq\; \max_{q(\cdot)} \; P_D(q) \quad \text{subject to} \quad P_F(q) \le \alpha,$$

where $q(y) = p_{\hat{H}|\mathsf{y}}(H_1 \mid y)$ parametrizes (possibly randomized) tests. Then the *efficient frontier* of achievable operating points is

$$\mathcal{F}_{p_{\mathsf{y}|H}} \;\coloneqq\; \{(P_F, P_D) :\; P_D = \zeta_{\mathrm{NP}}(P_F)\},$$

because by optimality no test can achieve $P_D > \zeta_{\mathrm{NP}}(P_F)$ at a given P_F.

4.9 ζ_{NP} Property I on endpoints

- $\zeta_{\mathrm{NP}}(1) = 1$, i.e., $(1,1) \in \mathcal{F}_{p_{\mathsf{y}|H}}$ (declare H_1 always).
- $\zeta_{\mathrm{NP}}(P_F) \ge P_F$, hence every frontier point satisfies $P_D \ge P_F$.

4.10 ζ_{NP} Property II on concavity

The function $\zeta_{\mathrm{NP}}(\cdot)$ is concave. Indeed, if two tests achieve operating points $(P_F^{(1)}, P_D^{(1)})$ and $(P_F^{(2)}, P_D^{(2)})$, then randomizing between them with probability $\lambda \in [0,1]$ yields

$$P_F = \lambda P_F^{(1)} + (1-\lambda) P_F^{(2)}, \qquad P_D = \lambda P_D^{(1)} + (1-\lambda) P_D^{(2)},$$

so the achievable region is convex, and the upper boundary $P_D = \zeta_{\mathrm{NP}}(P_F)$ is therefore concave.

4.11 ζ_{NP} Property III: slope interpretation

Let η_0 be an LRT threshold such that the likelihood ratio has no point mass at η_0 under either hypothesis:

$$\mathbb{P}(\mathcal{L}(\mathsf{y}) = \eta_0 \mid H = H_0) = \mathbb{P}(\mathcal{L}(\mathsf{y}) = \eta_0 \mid H = H_1) = 0.$$

Then the corresponding deterministic LRT operating point lies on the frontier:

$$P_D(\eta_0) = \zeta_{\mathrm{NP}}(P_F(\eta_0)),$$

and the tail probabilities

$$P_F(\eta) = \mathbb{P}(\mathcal{L}(\mathsf{y}) \ge \eta \mid H = H_0), \qquad P_D(\eta) = \mathbb{P}(\mathcal{L}(\mathsf{y}) \ge \eta \mid H = H_1)$$

are continuous in η at $\eta = \eta_0$. At such a "smooth" operating point (no atom at η_0), the Neyman–Pearson function is differentiable at $P_F(\eta_0)$ and its slope equals the corresponding

LRT threshold:

$$\dot{\zeta}_{\mathrm{NP}}(P_F(\eta_0)) \;=\; \eta_0.$$

Equivalently, the supporting line to the concave frontier at that point has slope η_0, linking the local tradeoff $\frac{dP_D}{dP_F}$ to the LRT threshold.

4.12

4.13 OVERVIEW: Bayesian vs. Neyman–Pearson formulations for binary testing

We observe data $y \in \mathcal{Y}$ and must decide between H_0 and H_1. A deterministic decision rule is a function $f(\cdot) : y \mapsto \{H_0, H_1\}$. On the other hand a randomized rule is specified by

$$r(y) \;\coloneqq\; \mathbb{P}(\hat{H} = H_1 \mid \mathsf{y} = y) \in [0,1], \qquad \mathbb{P}(\hat{H} = H_0 \mid \mathsf{y} = y) = 1 - r(y).$$

- *Bayesian formulation (prior + costs):* Model the hypothesis as a random variable $\mathsf{H} \in \{H_0, H_1\}$ with prior $\pi_1 = \mathbb{P}(\mathsf{H} = \mathcal{H}_1)$ (and $\pi_0 = 1 - \pi_1$), and specify costs $C(\mathcal{H}_j, H_i)$ for deciding $\hat{H} = H_i$ when $\mathsf{H} = H_j$. Choose a rule to minimize Bayes risk:

  $$\hat{H}_{\mathrm{B}}(\cdot) \in \arg\min_{f(\cdot)} \; \varphi_{\mathrm{B}}(f), \qquad \varphi_{\mathrm{B}}(f) \;=\; \mathbb{E}[C(\mathsf{H}, f(\mathsf{y}))]\,.$$

  Allowing randomization does not improve the minimum attainable Bayes risk; a Bayes optimal rule can be taken deterministic and is an LRT with a threshold determined by priors and costs.
- *Neyman–Pearson formulation (no prior or costs, set error constraint):* Do *not* assume a prior or costs. Fix a false-alarm tolerance $\alpha \in (0,1)$ and choose a rule to maximize detection subject to the constraint:

  $$\hat{H}_{\mathrm{NP}}(\cdot) \in \arg \max_{f(\cdot):\, P_{\mathrm{F}}(f) \le \alpha} \; P_{\mathrm{D}}(f).$$

  For a randomized rule $r(\cdot)$, the operating probabilities are linear in r:

  $$P_{\mathrm{F}}(r) \;=\; \mathbb{E}[r(\mathsf{y}) \mid \mathsf{H} = H_0]\,, \qquad P_{\mathrm{D}}(r) \;=\; \mathbb{E}[r(\mathsf{y}) \mid \mathsf{H} = H_1]\,.$$

  Here randomization *can* matter: it may be needed to hit the constraint $P_{\mathrm{F}} = \alpha$ exactly (especially with discrete components)

5 Minimax

5.1 Minimax formulation (costs known, prior unknown)

Sometimes we can meaningfully specify costs $C_{ij} = C(H_j, H_i)$ for $i, j \in \{0,1\}$, but we do *not* want to commit to a prior $p = \mathbb{P}(H = H_1)$. The minimax approach models this as a zero sum game in which:

- We choose a possibly randomized decision rule $r(\cdot) : y \mapsto [0,1]$, where

  $$r(y) = \mathbb{P}(\hat{H} = H_1 \mid \mathsf{y} = y)$$

- "Nature" chooses the most adverse prior $p \in [0,1]$.

For a fixed prior p and rule r, the Bayes risk can be written as

$$\varphi(p,r) \;\coloneqq\; \mathbb{E}\Big[C(H,\hat{H})\Big] \;=\; (1-p)\,\mathbb{E}\Big[C(H,\hat{H}) \mid H=H_0\Big]\,, +p\,\mathbb{E}\Big[C(H,\hat{H}) \mid H=H_1\Big]\,,$$

and let us define the conditional risks as

$$\varphi_0(r) \;\coloneqq\; \mathbb{E}\Big[C(H,\hat{H}) \mid H=H_0\Big]\,, \qquad \varphi_1(r) \;\coloneqq\; \mathbb{E}\Big[C(H,\hat{H}) \mid H=H_1\Big]\,.$$

The minimax decision rule is then

$$r_M(\cdot) \in \arg\min_{r:\mathcal{Y}\to[0,1]} \;\varphi_M(r), \qquad \varphi_M(r) \;\coloneqq\; \max_{p\in[0,1]} \varphi(p,r).$$

A decision rule r is called an *equalizer rule* if $\varphi_0(r) = \varphi_1(r)$, i.e., the risk $\varphi(p,r)$ does not depend on the prior p. In a minimax setting, equalizer rules are natural candidates because they "hedge" against the worst-case prior.

5.2 Matched Bayes LRT under prior p (0–1 costs)

For 0–1 costs and prior $p = \mathbb{P}(H = H_1)$, the Bayes decision rule is an LRT with threshold

$$\eta(p) \;= \frac{C_{10}-C_{00}}{C_{01}-C_{11}} \cdot \frac{P_0}{P_1} = \;\frac{1-p}{p}.$$

Let $P_F(\eta(\cdot))$ and $P_D(\eta(\cdot))$ denote the false-alarm and detection probabilities achieved by the LRT with threshold $\eta(\cdot)$. The resulting probability of error is

$$P_e(p) \;=\; (1-p)\,P_F(\eta(p)) \;+\; p\Big(1-P_D(\eta(p))\Big).$$

5.3 Mismatched Bayes rule: design for q, operate under p

Suppose the true prior is still p, but we use the Bayes LRT designed for a different prior q. Then the rule uses threshold

$$\eta(q) \;=\; \frac{1-q}{q},$$

and its (mismatched) probability of error is

$$P_e(p;q) \;=\; (1-p)\,P_F(\eta(q)) \;+\; p\Big(1-P_D(\eta(q))\Big).$$

For each fixed q, $P_e(p;q)$ is an affine (linear) function of p.

5.4 Minimax choice within the LRT family and equalizer condition

In a minimax setting where nature chooses p after we choose the (LRT) design parameter q, we want to minimize

$$\max_{p\in[0,1]} P_e(p;q)$$

which is the worst that nature can choose p to be given our choice of q. Since $P_e(p;q)$ is affine in p, the worst case occurs at an endpoint:

$$\max_{p\in[0,1]} P_e(p;q) = \max\Big\{P_e(0;q),\, P_e(1;q)\Big\} = \max\Big\{P_F(\eta(q)),\, 1-P_D(\eta(q))\Big\}.$$

Therefore, our best choice of q equalizes these two endpoint errors: pick q such that

$$P_F(\eta(q)) \;=\; 1 - P_D(\eta(q)).$$

5.5 Randomized LRT with general costs: Bayes vs Neyman-Pearson

With general costs $C_{ij} = C(H_j, H_i)$ and design prior $p \in [0,1]$, a randomized LRT can be written as

$$r_B(y; p, \lambda) \;\coloneqq\; \begin{cases} 1, & \mathcal{L}(y) > \eta(p), \\ \lambda, & \mathcal{L}(y) = \eta(p), \\ 0, & \mathcal{L}(y) < \eta(p), \end{cases} \qquad \eta(p) \;\coloneqq\; \frac{1-p}{p} \cdot \frac{C_{10} - C_{00}}{C_{01} - C_{11}}, \qquad \lambda \in [0,1].$$

where p is the prior and $\lambda \in [0,1]$ is the randomization parameter. Recall that randomization cannot improve aBayes decision rule, but nor does it incur additional cost. At the same time, recall from our development of Neyman-Pearson hypothesis testing that, in general, the associated false-alarm and detection probabilities

$$P_F^B(p,\lambda) = \mathbb{E}[r_B(y;p,\lambda) \mid H = H_0] = \mathbb{P}(\mathcal{L}(y) > \eta(p) \mid H = H_0) + \lambda\, \mathbb{P}(\mathcal{L}(y) = \eta(p) \mid H = H_0),$$

$$P_D^B(p,\lambda) = \mathbb{E}[r_B(y;p,\lambda) \mid H = H_1] = \mathbb{P}(\mathcal{L}(y) > \eta(p) \mid H = H_1) + \lambda\, \mathbb{P}(\mathcal{L}(y) = \eta(p) \mid H = H_1).$$

do depend on the randomization λ, and that every point on the efficient frontier is achievable by a randomized test of this form.

5.6 Notation for mismatched Bayes risk in randomized LRT

We define the *mismatched Bayes risk* as the risk under true prior p when using the Bayes rule designed for prior q:

$$\varphi_B(p,q,\lambda) \;\coloneqq\; \varphi\big(p,\; r_B(\cdot; q, \lambda)\big).$$

The *matched* Bayes risk is

$$\varphi_B^\star(p) \;\coloneqq\; \varphi_B(p,p,\lambda),$$

and for the matched case the value does not depend on λ (randomization only affects tie-breaking on a set where both decisions have equal conditional Bayes cost).

5.7 Claim 5.1: key properties and brief proof sketches

Fix a randomized Bayes LRT $r_B(\cdot; q, \lambda)$ designed for prior $q \in [0,1]$ (and tie-randomization $\lambda \in [0,1]$). Let

$$\varphi_B(p,q,\lambda) \;\coloneqq\; \varphi\big(p,\, r_B(\cdot;q,\lambda)\big), \qquad \varphi_B^\star(p) \;\coloneqq\; \min_r\; \varphi(p,r).$$

Define the conditional Bayes risks of the rule $r_B(\cdot; q, \lambda)$ by

$$\varphi_B^0(q,\lambda) \;\coloneqq\; \mathbb{E}\Big[C(H,\hat{H}) \mid H = H_0\Big], \qquad \varphi_B^1(q,\lambda) \;\coloneqq\; \mathbb{E}\Big[C(H,\hat{H}) \mid H = H_1\Big],$$

so that, in terms of $P_F^B(q,\lambda)$ and $P_D^B(q,\lambda)$,

$$\varphi_B^0(q,\lambda) = C_{00}(1 - P_F^B(q,\lambda)) + C_{10} P_F^B(q,\lambda), \qquad \varphi_B^1(q,\lambda) = C_{01}(1 - P_D^B(q,\lambda)) + C_{11} P_D^B(q,\lambda).$$

P1. *$\varphi_B(\cdot, q, \lambda)$ is affine in p.*

Proof sketch. Use the law of total expectation under prior p:

$$\varphi_B(p,q,\lambda) = (1-p)\,\varphi_B^0(q,\lambda) + p\,\varphi_B^1(q,\lambda) = \varphi_B^0(q,\lambda) + p\big(\varphi_B^1(q,\lambda) - \varphi_B^0(q,\lambda)\big),$$

which is linear in $p \in [0,1]$.

P2. *Lower bound:* $\varphi_B(p,q,\lambda) \geq \varphi_B^\star(p)$*, with equality when* $q = p$.

Proof sketch. By definition, $\varphi_B^\star(p)$ is the minimum Bayes risk over *all* decision rules under prior p. Since $r_B(\cdot; q, \lambda)$ is just one feasible rule,

$$\varphi_B(p,q,\lambda) = \varphi\big(p, r_B(\cdot;q,\lambda)\big) \;\geq\; \min_r \varphi(p,r) \;=\; \varphi_B^\star(p).$$

If $q = p$, then $r_B(\cdot; p, \lambda)$ is Bayes-optimal for that prior, so it achieves the minimum and equality holds.

P3. $\varphi_B^\star(\cdot)$ *is concave and continuous on* $[0,1]$.

Proof sketch (concavity). Fix $p_1, p_2 \in [0,1]$ and $\theta \in [0,1]$, and set $p = \theta p_1 + (1-\theta)p_2$. For any fixed rule r, the Bayes risk is affine in p:

$$\varphi(p,r) = \theta\,\varphi(p_1,r) + (1-\theta)\,\varphi(p_2,r).$$

Taking minima over r and using $\min_r(\theta a_r + (1-\theta)b_r) \geq \theta \min_r a_r + (1-\theta)\min_r b_r$ gives

$$\varphi_B^\star(p) \geq \theta\,\varphi_B^\star(p_1) + (1-\theta)\,\varphi_B^\star(p_2),$$

so $\varphi_B^\star$ is concave.

Proof sketch (continuity). A concave function on an interval is continuous on the interior $(0,1)$. Continuity at the endpoints follows by bounding $\varphi_B^\star(p)$ between endpoint risks and using that $\varphi(p,r)$ is affine in p for each fixed r, hence cannot jump at 0 or 1.

P4. *Endpoint values:* $\varphi_B^\star(0) = C_{00}$ *and* $\varphi_B^\star(1) = C_{11}$.

Proof sketch. If $p = 0$, then $H = H_0$ almost surely, so the Bayes risk reduces to the conditional risk under H_0. The decision rule that always declares $\hat{H} = H_0$ achieves risk C_{00}, hence $\varphi_B^\star(0) \leq C_{00}$. Conversely, no rule can achieve risk below C_{00} when $H = H_0$ is certain, because C_{00} is precisely the cost incurred when deciding H_0 under H_0. Therefore $\varphi_B^\star(0) = C_{00}$.

Similarly, if $p = 1$, then $H = H_1$ almost surely. The rule that always declares $\hat{H} = H_1$ achieves risk C_{11}, and no rule can do better when H_1 is certain. Thus $\varphi_B^\star(1) = C_{11}$.

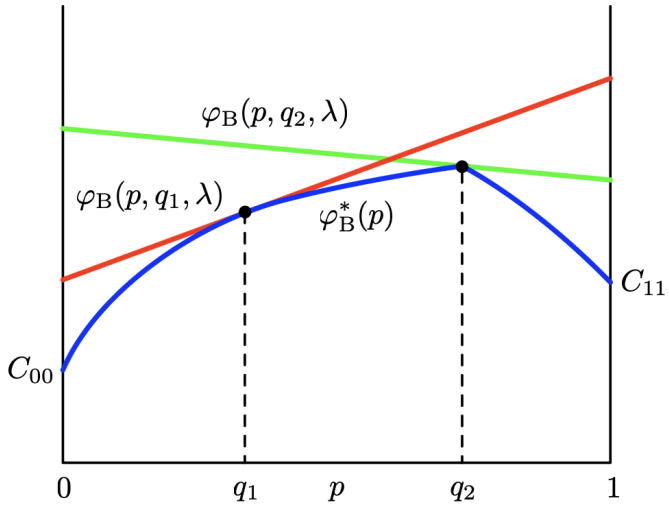

Figure 5.3: Mismatched and matched Bayes risk relationships. The blue curve indicates the Bayes risk as a function of the actual prior p for a Bayes decision rule that is correctly matched to p. The red and green lines indicate the mismatch Bayes risk when the Bayes decision rule is designed for prior q_1 and q_2, respectively (but the actual prior is p).

5.8 Setup: what is given vs. what we choose

Given:

- Two hypotheses $H \in \{H_0, H_1\}$ and data models $p_{y|H}(\cdot \mid H_i)$ for $i \in \{0, 1\}$.
- A cost matrix $C_{ij} = C(H_j, H_i)$ for $i, j \in \{0, 1\}$, where $C(H_j, H_i)$ is the cost of deciding $\hat{H} = H_i$ when the true hypothesis is $H = H_j$.

We choose: a (possibly randomized) decision rule $r : \mathcal{Y} \to [0, 1]$ with

$$r(y) \;\coloneqq\; \mathbb{P}(\hat{H} = H_1 \mid y = y), \qquad \mathbb{P}(\hat{H} = H_0 \mid y = y) = 1 - r(y).$$

Nature chooses (adversarially): an unknown prior

$$p \in [0, 1], \qquad p \;\coloneqq\; \mathbb{P}(H = H_1), \quad (1 - p) = \mathbb{P}(H = H_0).$$

For any rule r, define its conditional risks

$$\varphi_0(r) \;\coloneqq\; \mathbb{E}\Big[C(H, \hat{H}) \mid H = H_0\Big], \qquad \varphi_1(r) \;\coloneqq\; \mathbb{E}\Big[C(H, \hat{H}) \mid H = H_1\Big],$$

and the Bayes risk under prior p as

$$\varphi(p, r) \;\coloneqq\; (1 - p)\, \varphi_0(r) + p\, \varphi_1(r).$$

The *minimax* problem is

$$\min_{r:\mathcal{Y}\to[0,1]} \; \max_{p\in[0,1]} \; \varphi(p, r)$$

where we use the notation $r_M(\cdot) = r_\star(\cdot)$ such that

$$r_\star \in \arg \min_{r:\mathcal{Y}\to[0,1]} \max_{p\in[0,1]} \varphi(p, r),$$

and r_M is simply a name for such an optimizer.

5.9 Theorem 5.1 (characterization of the minimax rule)

Theorem. There exist $(p_\star, \lambda_\star) \in [0,1]^2$ such that a minimax-optimal rule is a *randomized Bayes likelihood-ratio test* (LRT)

$$r_M(\cdot) = r_B(\cdot; p_\star, \lambda_\star),$$

where $r_B(\cdot; p, \lambda)$ denotes the Bayes LRT designed for prior p (with tie-randomization parameter λ on the threshold set). Moreover, $(p_\star, r_\star)$ form a saddle point, meaning that

$$\min_r \max_{p\in[0,1]} \varphi(p, r) = \max_{p\in[0,1]} \min_r \varphi(p, r) = \varphi(p_\star, r_\star).$$

The prior $p_\star$ is called the *least-favorable (minimax) prior.*

5.10 How to *find* $(p_\star, \lambda_\star)$ (what you compute)

Define the *matched Bayes risk* (Bayes-optimal value if the prior were p) by

$$\varphi_B^\star(p) \;\coloneqq\; \min_{r:\mathcal{Y}\to[0,1]} \varphi(p, r).$$

where we recall that $\varphi(p, r)$ is the Bayes risk under prior p is

$$\varphi(p, r) \;\coloneqq\; (1-p)\,\varphi_0(r) + p\,\varphi_1(r).$$

Also define the Neyman–Pearson efficient frontier

$$\zeta_{\mathrm{NP}}(P_F) \;\coloneqq\; \sup\Big\{P_D(r) :\; r : \mathcal{Y} \to [0,1],\; P_F(r) \le P_F\Big\},$$

where, for any randomized rule r,

$$P_F(r) \;\coloneqq\; \mathbb{E}[r(\mathsf{y}) \mid H = H_0]\,, \qquad P_D(r) \;\coloneqq\; \mathbb{E}[r(\mathsf{y}) \mid H = H_1]\,.$$

From the costs, form the *equalizer line* (a linear function of P_F)

$$g_M(P_F) \;\coloneqq\; \frac{C_{01} - C_{00}}{C_{01} - C_{11}} \;-\; \frac{C_{10} - C_{00}}{C_{01} - C_{11}}\, P_F.$$

- *Interior case (solve an intersection):* if there exists $P_F^\star \in [0,1]$ such that

  $$\zeta_{\mathrm{NP}}(P_F^\star) = g_M(P_F^\star),$$

  then one can choose $(p_\star, \lambda_\star)$ so that the Bayes LRT $r_B(\cdot; p_\star, \lambda_\star)$ attains the operating point

  $$P_F\big(r_B(\cdot; p_\star, \lambda_\star)\big) = P_F^\star, \qquad P_D\big(r_B(\cdot; p_\star, \lambda_\star)\big) = \zeta_{\mathrm{NP}}(P_F^\star).$$

  Geometrically: the minimax operating point is where the NP frontier intersects the equalizer line.
- *Endpoint case (no intersection):* if $\zeta_{\mathrm{NP}}(P_F) > g_M(P_F)$ for all $P_F \in [0,1]$, then $p_\star = 0$. If on the other hand $\zeta_{\mathrm{NP}}(P_F) < g_M(P_F)$ for all $P_F \in [0,1]$, then $p_\star = 1$.

5.11 Two exam-useful corollaries (quick checks)

- *Least-favorable prior maximizes matched Bayes risk:* if $r_B(\cdot; p_\star, \lambda_\star)$ is minimax, then

  $$p_\star \in \arg\max_{p\in[0,1]} \varphi_B^\star(p).$$

- *Equalizer condition (interior solution):* if the least-favorable prior satisfies $p_\star \in (0,1)$,

then the corresponding minimax Bayes LRT equalizes conditional risks:

$$\varphi_0(r_B(\cdot; p_\star, \lambda_\star)) = \varphi_1(r_B(\cdot; p_\star, \lambda_\star)).$$

(If $p_\star \in \{0, 1\}$, the minimax solution is at an endpoint and tie-randomization is immaterial.)

5.12 How to plot and find the intersection in the (P_F, P_D) plane

1. *Trace the NP frontier* For each threshold η of the likelihood ratio $\mathcal{L}(y) = \frac{p_{y|H}(y|H_1)}{p_{y|H}(y|H_0)}$ (and tie-randomization λ on $\{\mathcal{L}(y) = \eta\}$), form the randomized LRT

$$r_\eta(y) = \begin{cases} 1, & \mathcal{L}(y) > \eta, \\ \lambda, & \mathcal{L}(y) = \eta, \\ 0, & \mathcal{L}(y) < \eta, \end{cases}$$

and compute

$$P_F(\eta, \lambda) = \mathbb{E}[r_\eta(y) \mid H = H_0], \qquad P_D(\eta, \lambda) = \mathbb{E}[r_\eta(y) \mid H = H_1].$$

Plot the resulting points $(P_F(\eta, \lambda), P_D(\eta, \lambda))$; as η varies you trace the ROC frontier (randomization fills in gaps if there are discrete components).

2. *Plot the equalizer line.* On the same axes, plot

$$P_D = g_M(P_F) = \frac{C_{01} - C_{00}}{C_{01} - C_{11}} - \frac{C_{10} - C_{00}}{C_{01} - C_{11}} P_F, \qquad P_F \in [0, 1].$$

3. *Read off the minimax operating point.* If the line intersects the ROC frontier, the intersection point $(P_F^\star, P_D^\star)$ is the minimax operating point:

$$P_D^\star = \zeta_{\mathrm{NP}}(P_F^\star) = g_M(P_F^\star).$$

Choose (η, λ) that achieves $(P_F^\star, P_D^\star)$; that LRT is minimax.

4. *If there is no intersection.* If $\zeta_{\mathrm{NP}}(P_F) > g_M(P_F)$ for all P_F, the minimax prior is $p_\star = 0$ (nature concentrates on H_0); if $\zeta_{\mathrm{NP}}(P_F) < g_M(P_F)$ for all P_F, then $p_\star = 1$.

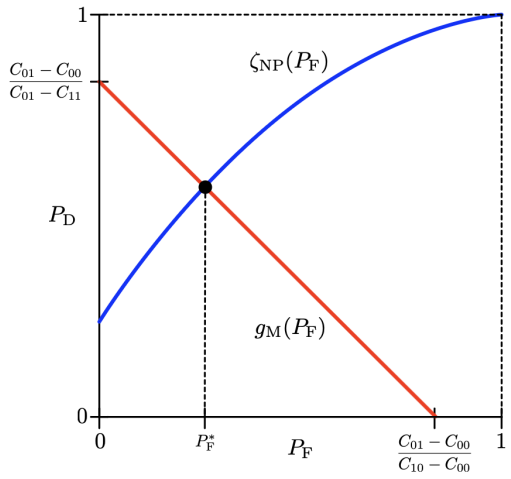

Figure 5: The intersection $(P_F^\star, P_D^\star)$ is the minimax operating point: $P_D^\star = \zeta_{\mathrm{NP}}(P_F^\star) = g_M(P_F^\star)$.

5.13 From an intersection point $P_F^\star$ to $(p_\star, \lambda_\star)$

Assume we have identified an intersection point $P_F^\star \in [0,1]$ such that

$$\zeta_{\mathrm{NP}}(P_F^\star) = g_M(P_F^\star).$$

By definition of the NP frontier, there exists a (possibly randomized) likelihood-ratio test (LRT) with some threshold $\eta_\star$ and tie-randomization $\lambda_\star \in [0,1]$ whose operating point equals

$$P_F(\eta_\star, \lambda_\star) = P_F^\star, \qquad P_D(\eta_\star, \lambda_\star) = \zeta_{\mathrm{NP}}(P_F^\star).$$

Concretely, take $\eta_\star$ so that

$$P_F(\eta_\star, 0) \;\le\; P_F^\star \;\le\; P_F(\eta_\star, 1),$$

where $P_F(\eta, 0) = \mathbb{P}(\mathcal{L}(y) > \eta \mid H = H_0)$ and $P_F(\eta, 1) = \mathbb{P}(\mathcal{L}(y) \ge \eta \mid H = H_0)$, and then set

$$\lambda_\star \;=\; \frac{P_F^\star - \mathbb{P}(\mathcal{L}(y) > \eta_\star \mid H = H_0)}{\mathbb{P}(\mathcal{L}(y) = \eta_\star \mid H = H_0)} \;\in [0,1],$$

(with the convention that if $\mathbb{P}(\mathcal{L}(y) = \eta_\star \mid H = H_0) = 0$, then no randomization is needed and any $\lambda_\star$ works).

It remains to choose $p_\star$ so that the Bayes LRT threshold induced by $p_\star$ equals $\eta_\star$. Recall the Bayes LRT threshold for prior p and costs C_{ij} is

$$\eta(p) \;=\; \frac{1-p}{p} \cdot \frac{C_{10} - C_{00}}{C_{01} - C_{11}}.$$

Thus we choose $p_\star$ by solving $\eta(p_\star) = \eta_\star$, i.e.,

$$\frac{1-p_\star}{p_\star} = \eta_\star \cdot \frac{C_{01} - C_{11}}{C_{10} - C_{00}}, \qquad \Longrightarrow \qquad p_\star = \frac{1}{1 + \eta_\star \frac{C_{01}-C_{11}}{C_{10}-C_{00}}}.$$

With this choice, the Bayes LRT $r_B(\cdot; p_\star, \lambda_\star)$ uses threshold $\eta_\star$ and (by the construction of $\lambda_\star$) attains the desired operating point

$$P_F(r_B(\cdot; p_\star, \lambda_\star)) = P_F^\star, \qquad P_D(r_B(\cdot; p_\star, \lambda_\star)) = \zeta_{\mathrm{NP}}(P_F^\star).$$

6 Bayesian parameter estimation

6.1 Bayesian estimation: setup and posterior

In Bayesian estimation, the unknown parameter is modeled as a random variable $\mathsf{x} \in \mathcal{X}$, and the observation is a random variable $\mathsf{y} \in \mathcal{Y}$. The Bayesian model is specified by:

- a prior distribution $p_{\mathsf{x}}(\cdot)$ on the unknown parameter x
- a likelihood (data model) $p_{\mathsf{y}|\mathsf{x}}(\cdot \mid x)$ describing how the data y depends on the parameter x

After observing $\mathsf{y} = y$, we update our belief about x via Bayes' rule:

$$p_{\mathsf{x}|\mathsf{y}}(x \mid y) = \frac{p_{\mathsf{y}|\mathsf{x}}(y \mid x)\, p_{\mathsf{x}}(x)}{\int_{\mathcal{X}} p_{\mathsf{y}|\mathsf{x}}(y \mid x')\, p_{\mathsf{x}}(x')\, dx'}.$$

The posterior $p_{\mathsf{x}|\mathsf{y}}(\cdot \mid y)$ summarizes all information about x contained in the prior and the observed data.

6.2 Estimators and Bayes risk

An *estimator* is a deterministic function $\hat{x}(\cdot) : \mathcal{Y} \to \mathcal{X}$; given observation y, it outputs the estimate $\hat{x}(y)$. To choose an estimator, specify a loss function $C(x, \hat{x})$ (cost of estimating x by $\hat{x}$), and minimize the expected loss (Bayes risk):

$$\hat{x}(\cdot) \in \arg\min_{f(\cdot)} \ \mathbb{E}\big[C(\mathsf{x}, f(\mathsf{y}))\big]\,,$$

where the expectation is with respect to the joint law $p_{\mathsf{x},\mathsf{y}}(x, y) = p_{\mathsf{x}}(x)\, p_{\mathsf{y}|\mathsf{x}}(y \mid x)$. Equivalently (pointwise form), the Bayes-optimal estimator minimizes posterior expected loss for each realized y:

$$\hat{x}(y) \in \arg\min_{\hat{x} \in \mathcal{X}} \ \mathbb{E}[C(\mathsf{x}, \hat{x}) \mid \mathsf{y} = y]\,.$$

6.3 Pointwise Bayes optimality

Let $\mathsf{x} \in \mathcal{X}$ be the unknown parameter with prior $p_{\mathsf{x}}(\cdot)$, and let $\mathsf{y} \in \mathcal{Y}$ be the observation with likelihood $p_{\mathsf{y}|\mathsf{x}}(\cdot \mid x)$. An estimator is a deterministic function $\hat{x}(\cdot) : \mathcal{Y} \to \mathcal{X}$, and given a loss function $C(x, \hat{x})$ the Bayes risk is

$$R(\hat{x}) \ \coloneqq \ \mathbb{E}\big[C(\mathsf{x}, \hat{x}(\mathsf{y}))\big]\,.$$

Using $p_{\mathsf{x},\mathsf{y}}(x, y) = p_{\mathsf{x}|\mathsf{y}}(x \mid y)\, p_{\mathsf{y}}(y)$,

$$R(\hat{x}) = \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} C(x, \hat{x}(y))\, p_{\mathsf{x}|\mathsf{y}}(x \mid y)\, dx \right) p_{\mathsf{y}}(y)\, dy.$$

Since $p_{\mathsf{y}}(y) \geq 0$, minimizing $R(\hat{x})$ is equivalent to minimizing the inner (posterior) expected loss for each fixed y:

$$\hat{x}(y) \in \arg\min_{a \in \mathcal{X}} \int_{\mathcal{X}} C(x, a)\, p_{\mathsf{x}|\mathsf{y}}(x \mid y)\, dx,$$

with the integral replaced by a sum if $\mathcal{X}$ is discrete.

6.4 Three canonical Bayes estimators

These follow by plugging different losses into the pointwise rule above (assume scalar x for MAE/MAP; the MSE form holds for vectors).

- *MAE (using absolute error loss):* If $C(x,a) = |x-a|$, then the Bayes estimator is any posterior median:
$$\hat{x}_{\mathrm{MAE}}(y) \in \mathrm{median}\Big(p_{x|y}(\cdot \mid y)\Big).$$
- *MAP (using MUC loss):* For $\epsilon > 0$, define the minimum uniform cost (MUC) loss
$$C_\epsilon(x,a) = \begin{cases} 1, & |x-a| > \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$
As $\epsilon \to 0$, minimizing the posterior expected C_ϵ selects a posterior mode, i.e., the MAP estimator:
$$\hat{x}_{\mathrm{MAP}}(y) \in \arg\max_{a \in \mathcal{X}} p_{x|y}(a \mid y).$$
- *BLS (using MSE loss):* If $C(x,a) = \|x-a\|^2$, then the Bayes least-squares (BLS), also called the MMSE estimator, is the posterior mean:
$$\hat{x}_{\mathrm{BLS}}(y) = \mathbb{E}[x \mid y = y].$$

6.5 Bias–variance tradeoff

Let $x \in \mathbb{R}^N$ be the parameter and let y be the observation. For an estimator $\hat{x}(y)$, define the estimation error
$$e \;\coloneqq\; \hat{x}(y) - x.$$
Now define the mean error as
$$b \;\coloneqq\; \mathbb{E}[e],$$
and the error covariance as
$$\Lambda_e \;\coloneqq\; \mathbb{E}\Big[(e-b)(e-b)^\top\Big].$$
Then the error second moment (error correlation) decomposes as
$$\mathbb{E}[ee^\top] = \Lambda_e + bb^\top,$$
since the cross terms vanish under $\mathbb{E}[e] = b$. Taking the trace gives the mean-squared error:
$$\mathrm{MSE}(\hat{x}) \;\coloneqq\; \mathbb{E}\Big[\|\hat{x}(y) - x\|^2\Big] = \mathrm{tr}\Big(\mathbb{E}[ee^\top]\Big) = \mathrm{tr}(\Lambda_e) + \|b\|^2.$$
Thus MSE splits into a *variance term* $\mathrm{tr}(\Lambda_e)$ and a *bias term* $\|b\|^2$.

6.6 BLS is unbiased

Let $\hat{x}_{\mathrm{BLS}}(y) \coloneqq \mathbb{E}[x \mid y]$ and define the error $e \coloneqq \hat{x}_{\mathrm{BLS}}(y) - x$. The bias is
$$b_{\mathrm{BLS}} \;\coloneqq\; \mathbb{E}[e] = \mathbb{E}[\mathbb{E}[x \mid y] - x] = \mathbb{E}[\mathbb{E}[x \mid y]] - \mathbb{E}[x] = \mathbb{E}[x] - \mathbb{E}[x] = 0,$$
where we used the law of iterated expectation $\mathbb{E}[\mathbb{E}[x \mid y]] = \mathbb{E}[x]$. Hence the BLS estimator is unbiased:
$$\mathbb{E}[\hat{x}_{\mathrm{BLS}}(y) - x] = 0.$$

6.7 Error covariance equals expected posterior covariance

Let (x, y) be jointly distributed. Define the (random) BLS/MMSE estimator
$$\widehat{x}(y) \;\coloneqq\; \mathbb{E}_{p_{x|y}}[x \mid y],$$

i.e. $\widehat{x}(\cdot)$ is a deterministic function, and $\widehat{x}(y)$ is random only through y. For each fixed y, define the posterior covariance matrix

$$\Lambda_{x|y}(y) \;\coloneqq\; \mathbb{E}_{p_{x|y}(\cdot|y)}\Big[(\widehat{x}(y)-x)(\widehat{x}(y)-x)^\top\Big].$$

Define the estimation error random vector

$$e \;\coloneqq\; \widehat{x}(y)-x.$$

Then the (unconditional) BLS error covariance is

$$\Lambda_{\mathrm{BLS}} \;\coloneqq\; \mathbb{E}_{p_{x,y}}[ee^\top] = \mathbb{E}_{p_{x,y}}\Big[(\widehat{x}(y)-x)(\widehat{x}(y)-x)^\top\Big].$$

Applying iterated expectation (conditioning on y) gives

$$\Lambda_{\mathrm{BLS}} = \mathbb{E}_{p_y}\Big[\mathbb{E}_{p_{x|y}(\cdot|y)}\Big[(\widehat{x}(y)-x)(\widehat{x}(y)-x)^\top\Big]\Big] = \mathbb{E}_{p_y}[\Lambda_{x|y}(y)].$$

7 Linear least squares*

8 NonBayesian parameter estimation

8.1 Nonrandom parameter estimation and parameterized observation model

In many estimation problems it is unnatural to place a prior on the unknown quantity of interest. In such settings we treat the parameter as *deterministic but unknown*. We write

$$x \in \mathcal{X} \quad \text{(unknown, deterministic parameter)}, \qquad y \in \mathcal{Y} \quad \text{(random observation)}.$$

Since the parameter is fixed, the observation model is *not* a conditional law $p_{y|x}(\cdot \mid \cdot)$; instead, it is a *family of distributions* on $\mathcal{Y}$ parameterized by x, written

$$p_y(y;x), \qquad \mathcal{P} \;\coloneqq\; \{\, p_y(\cdot;x) \in \mathcal{P}^{\mathcal{Y}} \;:\; x \in \mathcal{X} \,\}.$$

8.2 Why the naive Bayesian MSE criterion fails without a prior

An estimator is a function $\widehat{x} : \mathcal{Y} \to \mathcal{X}$, and $\widehat{x}(y)$ is random only through y. Recall that in Bayesian parameter estimation we found the estimator to minimize Bayes risk. A tempting but incorrect adaptation of the Bayesian framework is

$$\widehat{x}(\cdot) \;\overset{?}{=}\; \arg\min_{f:\mathcal{Y}\to\mathcal{X}} \; \mathbb{E}_{p_y(\cdot;x)}\Big[(x-f(y))^2\Big].$$

Here the expectation is *only over* $y \sim p_y(\cdot;x)$ (since x is not random). But then the minimizer is

$$\widehat{x}(\cdot) = f(y) = x,$$

which is useless because it depends on the very unknown parameter we are trying to estimate.

8.3 Valid estimators are statistics (must not depend on the parameter)

The failure above shows that in nonrandom parameter estimation we must restrict attention to estimators $\widehat{x}(\cdot)$ that do not depend on the true parameter x

A valid estimator has the form $\widehat{x}(\cdot) = f(y)$ where $f(\cdot)$ does not depend on x.

8.4 Bias, error covariance, and MSE for nonrandom parameters

As in the case of random parameters, two important quantities that impact the mean-square error performance of an estimator for nonrandom parameters are the bias and error covariance. However, there are some significant distinctions between these quantities in the nonrandom case, which we emphasize.

Fix an unknown deterministic parameter $x \in \mathcal{X}$, and let $y \in \mathcal{Y}$ be the random observation with law $p_y(\cdot; x)$. Let a (valid) estimator be a deterministic function $\widehat{x} : \mathcal{Y} \to \mathcal{X}$, so that the random estimate is $\widehat{x}(y)$ (random only through y).

Define the (random) estimation error

$$e(y;\ x) \ \coloneqq\ \widehat{x}(y) - x.$$

The bias function is

$$b_{\widehat{x}}(x) \ \coloneqq\ \mathbb{E}_{p_y(\cdot;x)}[e(y;\ x)] = \mathbb{E}_{p_y(\cdot;x)}[\widehat{x}(y)] - x,$$

and the error covariance matrix is

$$\Lambda_e(x) \ \coloneqq\ \mathrm{Cov}_{p_y(\cdot;x)}(e(y;\ x), e(y;\ x)) = \mathbb{E}_{p_y(\cdot;x)}\Big[(e - b_{\widehat{x}}(x))\,(e - b_{\widehat{x}}(x))^{\top}\Big].$$

Both $b_{\widehat{x}}(x)$ and $\Lambda_e(x)$ generally depend on x. The mean-square error (MSE) is

$$\mathrm{MSE}(x) \ \coloneqq\ \mathbb{E}_{p_y(\cdot;x)}[e^{\top} e] = \mathrm{tr}\Big(\mathbb{E}_{p_y(\cdot;x)}[ee^{\top}]\Big),$$

with the bias–covariance decomposition

$$\mathbb{E}_{p_y(\cdot;x)}[ee^{\top}] = b_{\widehat{x}}(x)\, b_{\widehat{x}}(x)^{\top} + \Lambda_e(x).$$

8.5 Unbiasedness of an estimator

An estimator $\widehat{x}(\cdot)$ is *unbiased* if

$$b_{\widehat{x}}(x) = 0, \qquad \forall x \in \mathcal{X}.$$

8.6 $\Lambda_{\mathbf{e}}(x) = \Lambda_{\widehat{x}}(x)$

In the nonrandom-parameter setting, the error covariance equals the covariance of the estimator itself. Fix a deterministic $x \in \mathcal{X}$ and let $y \sim p_y(\cdot; x)$. Let $\widehat{x} : \mathcal{Y} \to \mathcal{X}$ be an estimator, and recall the random error

$$e \ \coloneqq\ \widehat{x}(y) - x, \qquad b_{\widehat{x}}(x) \ \coloneqq\ \mathbb{E}_{p_y(\cdot;x)}[e] = \mathbb{E}_{p_y(\cdot;x)}[\widehat{x}(y)] - x.$$

Then

$$\begin{aligned}\Lambda_{e}(x) &\coloneqq \mathbb{E}_{p_y(\cdot;x)}\Big[\big(e - b_{\widehat{x}}(x)\big)\big(e - b_{\widehat{x}}(x)\big)^{\top}\Big] \\ &= \mathbb{E}_{p_y(\cdot;x)}\Big[\Big(\big(\widehat{x}(y) - x\big) - \big(\mathbb{E}_{p_y(\cdot;x)}[\widehat{x}(y)] - x\big)\Big)\Big(\big(\widehat{x}(y) - x\big) - \big(\mathbb{E}_{p_y(\cdot;x)}[\widehat{x}(y)] - x\big)\Big)^{\top}\Big] \\ &= \mathbb{E}_{p_y(\cdot;x)}\Big[\big(\widehat{x}(y) - \mathbb{E}_{p_y(\cdot;x)}[\widehat{x}(y)]\big)\big(\widehat{x}(y) - \mathbb{E}_{p_y(\cdot;x)}[\widehat{x}(y)]\big)^{\top}\Big] \\ &=: \Lambda_{\widehat{x}}(x).\end{aligned}$$

8.7 Minimum-variance unbiased estimator

Let $\mathcal{A}$ denote the class of valid unbiased estimators:

$$\mathcal{A} \coloneqq \Big\{\widehat{x} : \mathcal{Y} \to \mathbb{R} \;:\; \widehat{x}(\cdot) \text{ does not depend on } x,\; b_{\widehat{x}}(x) = 0 \;\forall x \in \mathcal{X}\Big\}.$$

If there exists $\widehat{x}_\star(\cdot) \in \mathcal{A}$ such that for every $\widehat{x}(\cdot) \in \mathcal{A}$,

$$\Lambda_{\widehat{x}}(x) \succeq \Lambda_{\widehat{x}_\star}(x), \qquad \forall x \in \mathcal{X},$$

then $\widehat{x}_\star(\cdot)$ is the *minimum-variance unbiased (MVU)* estimator, denoted $\widehat{x}_{\mathrm{MVU}}(\cdot)$.

8.8 Why MVU estimators may fail to exist and why we use lower bounds

The MVU estimator $\widehat{x}_{\mathrm{MVU}}(\cdot)$ (if it exists) is an element of the class $\mathcal{A}$ of valid unbiased estimators that has the smallest variance for *every* $x \in \mathcal{X}$. There are two basic obstructions:

- $\mathcal{A}$ may be empty (there may be no valid unbiased estimators).
- Even if $\mathcal{A} \neq \emptyset$, there may be no single estimator whose variance is uniformly smallest over all $x \in \mathcal{X}$ (different unbiased estimators can dominate at different parameter values).

Moreover, even when an MVU estimator exists, it can be difficult to identify or compute. A common approach is therefore to derive a universal *lower bound* on the variance $\lambda_{\widehat{x}}(x)$ (or on the covariance in the vector case) that applies to *all* unbiased estimators; such a bound can be used both to rule out overly-optimistic performance claims and, when achieved, to certify optimality. This is why Cramér-Rao is userful.

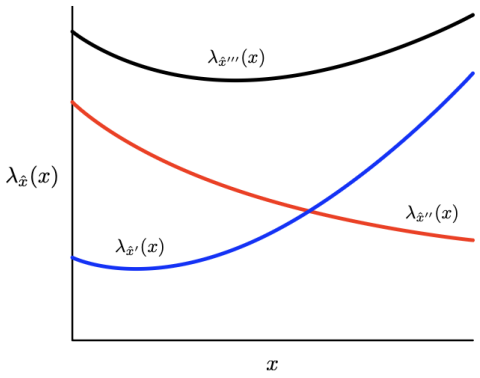

Figure 6: The variances of unbiased estimators $\widehat{x}'(\cdot)$, $\widehat{x}''(\cdot)$, $\widehat{x}'''(\cdot)$. No estimator achieves lower variances over all possible parameters $x \in \mathcal{X}$.

8.9 Preliminary of Cramér-Rao: score function and Fisher information

Fix $x \in \mathcal{X}$ and let $\mathsf{y} \sim p_{\mathsf{y}}(\cdot; x)$ with $p_{\mathsf{y}}(y; x) > 0$ and differentiable in x for each $y \in \mathcal{Y}$. The *score* is the random function of y,

$$S(\mathsf{y}; x) \;\coloneqq\; \frac{\partial}{\partial x} \ln p_{\mathsf{y}}(\mathsf{y}; x).$$

The *Fisher information in* y *about* x is

$$J_{\mathsf{y}}(x) \;\coloneqq\; \mathbb{E}_{p_{\mathsf{y}}(\cdot;x)}\big[S(\mathsf{y}; x)^2\big].$$

8.10 Cramér–Rao bound

Theorem (Cramér–Rao bound). Assume $p_{\mathsf{y}}(y; x)$ is positive and differentiable in $x \in \mathcal{X} \subset \mathbb{R}$ and satisfies the regularity condition

$$\mathbb{E}_{p_{\mathsf{y}}(\cdot;x)}\big[S(\mathsf{y}; x)\big] = 0, \qquad \forall x \in \mathcal{X}.$$

Let $\widehat{x} : \mathcal{Y} \to \mathbb{R}$ be a valid unbiased estimator, i.e.

$$\mathbb{E}_{p_{\mathsf{y}}(\cdot;x)}\big[\widehat{x}(\mathsf{y})\big] = x, \qquad \forall x \in \mathcal{X},$$

and suppose the further regularity condition

$$\mathbb{E}_{p_{\mathsf{y}}(\cdot;x)}\big[\widehat{x}(\mathsf{y})\, S(\mathsf{y}; x)\big] = 1, \qquad \forall x \in \mathcal{X},$$

holds. Define the variance of the estimator $\widehat{x}(\mathsf{y})$ as

$$\lambda_{\widehat{x}}(x) \;\coloneqq\; \mathrm{Var}_{p_{\mathsf{y}}(\cdot;x)}\big(\widehat{x}(\mathsf{y})\big).$$

then the Cramér Rao bound says that

$$\lambda_{\widehat{x}}(x) \;\geq\; \frac{1}{J_{\mathsf{y}}(x)}, \qquad \forall x \in \mathcal{X},$$

so we have a lower bound on the variance of the estimator. A few notes

- The Fisher information cannot be computed in all problems, i.e., the regularity condition may not be satisfied, in which case no Cram´er-Rao bound exists. For example, $p_{\mathsf{y}}(y; x)$ is not positive for all $x \in \mathcal{X}$.
- Any estimator that satisfies the Cramér-Rao bound with equality must be a MVU estimator.

8.11 Definition of an efficienct estimator using the Cramér–Rao bound

Fix a deterministic $x \in \mathcal{X} \subset \mathbb{R}$ and let $\mathsf{y} \sim p_{\mathsf{y}}(\cdot; x)$. Assume the conditions of the scalar Cramér–Rao theorem hold, and let

$$S(\mathsf{y}; x) \;\coloneqq\; \frac{\partial}{\partial x} \ln p_{\mathsf{y}}(\mathsf{y}; x), \qquad J_{\mathsf{y}}(x) \;\coloneqq\; \mathbb{E}_{p_{\mathsf{y}}(\cdot;x)}\big[S(\mathsf{y}; x)^2\big].$$

We say that a valid unbiased estimator $\widehat{x}(\cdot)$ is *efficient* if it attains the Cramér–Rao bound with equality for all $x \in \mathcal{X}$, i.e.

$$\lambda_{\widehat{x}}(x) \;=\; \frac{1}{J_{\mathsf{y}}(x)}, \qquad \forall x \in \mathcal{X}.$$

8.12 Characterization of efficient estimators

Corollary of Cramér Rao: A valid unbiased estimator $\widehat{x}(\cdot)$ is efficient if and only if it can be written as

$$\widehat{x}(\mathsf{y}) \;=\; x + \frac{1}{J_{\mathsf{y}}(x)}\, S(\mathsf{y}; x),$$

where the right-hand side, viewed as a function of y, must be independent of x (so that $\widehat{x}(\cdot)$ is a valid estimator).

8.13 Maximum likelihood estimator and and connections to effeciency

Fix a model family

$$\mathcal{P} \;\coloneqq\; \{\, p_{\mathsf{y}}(\cdot; x) \in \mathcal{P}^{\mathcal{Y}} :\; x \in \mathcal{X} \,\},$$

where $x \in \mathcal{X}$ is an unknown deterministic parameter and $\mathsf{y} \sim p_{\mathsf{y}}(\cdot; x)$ is the random observation.

Definition (Maximum likelihood estimator): The maximum likelihood (ML) estimate based on observation $y \in \mathcal{Y}$ is

$$\widehat{x}_{\mathrm{ML}}(y) \;\coloneqq\; \arg\max_{x \in \mathcal{X}} p_{\mathsf{y}}(y; x),$$

It is often convenient to define the (random) likelihood function of x given y by

$$L_{\mathsf{y}}(x) \;\coloneqq\; p_{\mathsf{y}}(\mathsf{y}; x),$$

and, for a fixed realized observation y, its deterministic version

$$L_y(x) \;\coloneqq\; p_{\mathsf{y}}(y; x),$$

which emphasizes the dependence on x for a fixed y.

Connection to efficiency

Corollary: Suppose, for the model family above, that an efficient estimator $\widehat{x}_{\mathrm{eff}}(\cdot)$ and an ML estimator $\widehat{x}_{\mathrm{ML}}(\cdot)$ both exist (and are unique). Then they coincide:

$$\widehat{x}_{\mathrm{eff}}(\cdot) = \widehat{x}_{\mathrm{ML}}(\cdot).$$

For convenience, we write

$$\lambda_{\mathrm{ML}}(x) \;\coloneqq\; \lambda_{\widehat{x}_{\mathrm{ML}}}(x) = \mathrm{Var}_{p_{\mathsf{y}}(\cdot;x)}(\widehat{x}_{\mathrm{ML}}(\mathsf{y})).$$

This does *not* mean ML estimators are always efficient: if an efficient estimator does not exist, the ML estimator need not be unbiased or have good finite-sample variance. Nonetheless, ML estimators often enjoy favorable *asymptotic* properties when many independent observations are available.

8.14 ML estimate for the linear-Gaussian model and least squares connection

Let $\mathbf{x} \in \mathcal{X} = \mathbb{R}^K$ denote an unknown deterministic parameter vector, and let $\mathbf{y} \in \mathbb{R}^N$ be the random observation vector. A *linear-Gaussian model* is

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \qquad \mathbf{w} \sim \mathcal{N}(0, \mathbf{\Lambda}_{\mathbf{w}}),$$

where $\mathbf{H} \in \mathbb{R}^{N \times K}$ is known and $\mathbf{\Lambda}_{\mathbf{w}} \in \mathbb{R}^{N \times N}$ is a known noise covariance matrix. Assume $\mathbf{\Lambda}_{\mathbf{w}} \succ 0$ and $\mathbf{H}$ has full column rank. Then the maximum likelihood estimator $\widehat{\mathbf{x}}_{\mathrm{ML}} : \mathbb{R}^N \to \mathbb{R}^K$ is

$$\widehat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = \left(\mathbf{H}^\top \mathbf{\Lambda}_{\mathbf{w}}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^\top \mathbf{\Lambda}_{\mathbf{w}}^{-1} \mathbf{y},$$

and it is MVU.

Gauss–Markov theorem as weighted least squares: For the same model, the ML estimate is also the solution to a weighted least-squares problem:

$$\widehat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^K} \varphi_{\mathrm{WLS}}(\mathbf{x}),$$

where

$$\varphi_{\mathrm{WLS}}(\mathbf{x}) \coloneqq (\mathbf{y}-\mathbf{H}\mathbf{x})^{\top}\boldsymbol{\Lambda}_{\mathbf{w}}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) = \left\|\boldsymbol{\Lambda}_{\mathbf{w}}^{-1/2}(\mathbf{y}-\mathbf{H}\mathbf{x})\right\|_2^2.$$

Additionally, the ordinary least squares can be obtained as a special case: If $\boldsymbol{\Lambda}_{\mathbf{w}} \propto \mathbf{I}$, then $\widehat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y})$ solves the ordinary least-squares problem

$$\widehat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathbb{R}^K} \|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2.$$

9 Exponential families