# Lecture 4: Bayesian tests and Model selection

Thibault Randrianarisoa

UTSC

February 5, 2026

# Outline

- Bayesian tests
- Loss Functions and Risk
  - Frequentist Risk vs. Bayesian Risk
  - Posterior Risk
- Bayes Factors
- Model selection

# Bayesian Testing: Framework

In a model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, our goal is to test a property of the parameter $\theta$. We want to determine if $\theta$ belongs to a region $\Theta_0 \subset \Theta$ or $\Theta_1 \subset \Theta$, where $\Theta_0 \cap \Theta_1 = \varnothing$.

- **Frequentist vs. Bayesian Support:** Unlike the frequentist approach, we don't always assume $\Theta_0 \cup \Theta_1 = \Theta$. However, $\Theta_0 \cup \Theta_1$ will always correspond to the support of the prior distribution $\Pi$.

- **Hypotheses:**
  - $H_0 : \theta \in \Theta_0$ is the null hypothesis.
  - $H_1 : \theta \in \Theta_1$ is the alternative hypothesis.

- **Simple vs. Composite:** A hypothesis is simple if the region is a singleton (e.g., $\Theta_0 = \{\theta_0\}$); otherwise, it is composite.

# Definition of a Statistical Test

## Definition

A test is a measurable function $\varphi(X_1, \ldots, X_n)$ of the observations, taking values in $\{0, 1\}$.

- $\varphi(\mathbf{X}) = 1$ means we reject $H_0$.
- $\varphi(\mathbf{X}) = 0$ means we accept $H_0$.

- **Generalized Test:** Sometimes we allow $\varphi$ to take values in the interval $[0, 1]$, in which case we call it a generalized test.

# Frequentist Approach: Error Types

There are two types of possible errors when testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$:

1. **Type I Error:** Rejecting $H_0$ while $\theta \in \Theta_0$.

$$\theta \in \Theta_0 \mapsto P_\theta(\varphi(\mathbf{X}) = 1)$$

2. **Type II Error:** Accepting $H_0$ while $\theta \in \Theta_1$.

$$\theta \in \Theta_1 \mapsto P_\theta(\varphi(\mathbf{X}) = 0)$$

⚠ *Asymmetry: Usually, $H_0$ is the "default" or "base" hypothesis (like the presumption of innocence), while $H_1$ is only chosen if the data provides strong evidence (of culpability, for analogy).*

## Definition

The size of a test $\varphi$ is the quantity:

$$\sup_{\theta \in \Theta_0} \mathbf{E}_\theta \varphi(\mathbf{X}) = \sup_{\theta \in \Theta_0} P_\theta(\varphi(\mathbf{X}) = 1)$$

- A test $\varphi$ is of **level** $\alpha$ if its size is at most $\alpha$.
- The power function $\pi : \Theta \to [0, 1]$ is defined by:

$$\pi : \theta \mapsto \mathbf{E}_\theta[\varphi(\mathbf{X})]$$

- The frequentist strategy is to fix $\alpha$ and seek a test $\varphi$ of level $\alpha$ that maximizes the power (brings it closest to 1) over $\Theta_1$.

# Example: Gaussian Model

Consider $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n}, \theta \in \mathbb{R}\}$, with $\Theta_0 = \mathbb{R}_-$ and $\Theta_1 = \mathbb{R}_+^*$.

The test defined by:

$$\varphi(\mathbf{X}) = \mathbb{1}_{\{\sqrt{n}\bar{X}_n > q_{1-\alpha}\}}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\mathcal{N}(0,1)$, is a test of level $\alpha$.

We can also show that:

- This test is uniformly more powerful among all tests of level $\alpha$.
- For any other test $\tilde{\varphi}$ of level $\alpha$, its power function $\pi$ satisfies:

$$\forall \theta \in \Theta_1, \quad \pi(\theta) \geqslant \tilde{\pi}(\theta)$$

# The Bayesian Approach to Testing

The Bayesian approach consists of choosing a prior distribution $\pi$ with support $\Theta_0 \cup \Theta_1$.

- We have $\pi(\Theta_0 \cup \Theta_1) = 1$.
- Note that with this assumption, $\pi$ is not necessarily defined on the entire parameter space $\Theta$.

Next step: We will use the posterior probabilities of $\Theta_0$ and $\Theta_1$ to make a decision.

## Bayesian Decision Theory for Testing

We extend the framework from Lecture 3 to testing problems. Here, a loss function $L$ is a map:

$$L : \Theta \times \{0, 1\} \to \mathbb{R}_+$$

where $\varphi \in \{0, 1\}$ represents our decision (test).

### Definition

The 0-1 loss function (or balanced loss) is defined by:

$$L(\theta, \varphi) = \begin{cases} 1 & \text{if } (\theta \in \Theta_0 \text{ and } \varphi = 1) \text{ or } (\theta \in \Theta_1 \text{ and } \varphi = 0) \\ 0 & \text{otherwise} \end{cases}$$

- This loss penalizes both types of errors equally with a cost of 1.
- We speak of a Bayes test rather than an estimator because the minimization (in the definition) occurs over the set of measurable tests $\varphi : \mathcal{X} \to \{0, 1\}$.

## Proposition

A Bayes test $\varphi^*$ for the 0-1 loss function is given by:

$$\varphi^*(\mathbf{X}) = \mathbb{1}_{\pi(\Theta_0|\mathbf{X}) \leqslant \pi(\Theta_1|\mathbf{X})} = \mathbb{1}_{\pi(\Theta_0|\mathbf{X}) \leqslant \frac{1}{2}}$$

## Proposition

A Bayes test $\varphi^*$ for the 0-1 loss function is given by:

$$\varphi^*(\mathbf{X}) = \mathbb{1}_{\pi(\Theta_0|\mathbf{X}) \leqslant \pi(\Theta_1|\mathbf{X})} = \mathbb{1}_{\pi(\Theta_0|\mathbf{X}) \leqslant \frac{1}{2}}$$

**Proof Sketch:** A Bayes test (or *estimator*) minimizes the posterior risk $\int L(\theta, \varphi) d\pi(\theta \mid \mathbf{X})$.

$$\begin{aligned} \int L(\theta, \varphi) d\pi(\theta \mid \mathbf{X}) &= \int (\mathbb{1}_{\theta \in \Theta_0, \varphi = 1} + \mathbb{1}_{\theta \in \Theta_1, \varphi = 0}) d\pi(\theta \mid \mathbf{X}) \\ &= \pi(\Theta_0 \mid \mathbf{X}) \mathbb{1}_{\varphi = 1} + \pi(\Theta_1 \mid \mathbf{X}) \mathbb{1}_{\varphi = 0} \end{aligned}$$

This is minimized when we choose the hypothesis with the highest posterior probability.

# Reminders: Total Variation Distance

## Definition

The total variation distance between two probability measures $P$ and $Q$ is:

$$d_{\mathrm{TV}}(P, Q) = \sup_A |P(A) - Q(A)|$$

## Proposition

If $P$ and $Q$ have densities $p$ and $q$ with respect to a measure $\mu$, then:

- $d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \int_E |p(x) - q(x)| d\mu(x)$
- The affinity $\mathcal{A}(P, Q) = 1 - d_{\mathrm{TV}}(P, Q)$ can be written as:

$$\mathcal{A}(P, Q) = \int_E (p \wedge q)(x) d\mu(x)$$

The affinity represents the "overlap" between two distributions.

## Two Simple Hypotheses of Equal Weight

Suppose $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. Let $p_0, p_1$ be the respective densities. If the prior is $\pi = \frac{1}{2}\delta_{\theta_0} + \frac{1}{2}\delta_{\theta_1}$, the Bayes test is:

### Bayes Risk for Simple Hypotheses

The Bayes risk $\mathbf{R}_B(\pi)$ is related to the Total Variation distance:

$$\mathbf{R}_B(\pi) = \frac{1}{2}(1 - d_{\mathrm{TV}}(P_0, P_1))$$

Intuition: The closer the distributions (smaller $d_{\mathrm{TV}}$), the higher the minimum risk.

## Two Simple Hypotheses of Equal Weight

Suppose $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. Let $p_0, p_1$ be the respective densities. If the prior is $\pi = \frac{1}{2}\delta_{\theta_0} + \frac{1}{2}\delta_{\theta_1}$, the Bayes test is:

$$\varphi^*(\mathbf{X}) = \mathbb{1}_{p_0(\mathbf{X}) \leqslant p_1(\mathbf{X})}$$

### Bayes Risk for Simple Hypotheses

The Bayes risk $\mathbf{R}_B(\pi)$ is related to the Total Variation distance:

$$\mathbf{R}_B(\pi) = \frac{1}{2}(1 - d_{\mathrm{TV}}(P_0, P_1))$$

Intuition: The closer the distributions (smaller $d_{\mathrm{TV}}$), the higher the minimum risk.

## Weighted Loss Functions

To penalize the two types of errors differently, we can use a weighted loss function.

### Definition

A loss function is weighted if:

$$L(\theta, \varphi) = \begin{cases} a_0 & \text{if } \theta \in \Theta_0, \varphi = 1 \\ a_1 & \text{if } \theta \in \Theta_1, \varphi = 0 \\ 0 & \text{sinon} \end{cases}$$

with $a_0, a_1 \in \mathbb{R}_+$.

### Proposition

The Bayes test for the weighted loss is:

## Weighted Loss Functions

To penalize the two types of errors differently, we can use a weighted loss function.

### Definition

A loss function is weighted if:

$$L(\theta, \varphi) = \begin{cases} a_0 & \text{if } \theta \in \Theta_0, \varphi = 1 \\ a_1 & \text{if } \theta \in \Theta_1, \varphi = 0 \\ 0 & \text{sinon} \end{cases}$$

with $a_0, a_1 \in \mathbb{R}_+$.

### Proposition

The Bayes test for the weighted loss is:

$$\varphi^*(\mathbf{X}) = \mathbb{1}_{a_0 \pi(\Theta_0|\mathbf{X}) \leqslant a_1 \pi(\Theta_1|\mathbf{X})} = \mathbb{1}_{\pi(\Theta_0|\mathbf{X}) \leqslant \frac{a_1}{a_0 + a_1}}$$

# Example: Gaussian Model (Simple Hypotheses)

Test $H_0 : \{\theta = 0\}$ vs $H_1 : \{\theta = 1\}$ with prior $\pi = \pi_0\delta_0 + \pi_1\delta_1$. The posterior probability for $H_0$ is:

$$\pi(\{0\} \mid \mathbf{X}) =$$

The weighted Bayes test takes the form:

Note: If $a_0 = a_1$ and $\pi_0 = \pi_1 = 1/2$, the threshold is $1/2$, making the hypotheses symmetric.

# Example: Gaussian Model (Simple Hypotheses)

Test $H_0 : \{\theta = 0\}$ vs $H_1 : \{\theta = 1\}$ with prior $\pi = \pi_0\delta_0 + \pi_1\delta_1$. The posterior probability for $H_0$ is:

$$\pi(\{0\} \mid \mathbf{X}) = \frac{\pi_0 \prod e^{-\frac{1}{2}(X_i-0)^2}}{\pi_0 \prod e^{-\frac{1}{2}(X_i-0)^2} + \pi_1 \prod e^{-\frac{1}{2}(X_i-1)^2}} = \frac{\pi_0}{\pi_0 + \pi_1 e^{n\bar{X}_n - n/2}}$$

The weighted Bayes test takes the form:

$$\varphi^*(\mathbf{X}) = \mathbb{1}_{\{\bar{X}_n \geqslant \frac{1}{2} + \frac{1}{n}\log(\frac{a_0\pi_0}{a_1\pi_1})\}}$$

Note: If $a_0 = a_1$ and $\pi_0 = \pi_1 = 1/2$, the threshold is $1/2$, making the hypotheses symmetric.

# Example: Gaussian Model (Composite Hypotheses)

Test $H_0 : \{\theta \leqslant 0\}$ vs $H_1 : \{\theta > 0\}$. Let the prior be $\pi = \mathcal{N}(\mu, 1)$.

The posterior distribution is $\mathcal{N}(\hat{\theta}(\mathbf{X}), \frac{1}{n+1})$ where $\hat{\theta}(\mathbf{X}) = \frac{\mu + n\bar{X}_n}{n+1}$.

The posterior probability of $H_0$ is:

$$\pi(\Theta_0 \mid \mathbf{X}) = P\left(\hat{\theta}(\mathbf{X}) + \frac{1}{\sqrt{n+1}}\mathcal{N}(0,1) \leqslant 0 \mid \mathbf{X}\right) = \Phi(-\sqrt{n+1}\hat{\theta}(\mathbf{X}))$$

where $\Phi$ is the standard normal CDF. The Bayes test $\varphi^*$ compares this probability to the threshold $\frac{a_1}{a_0 + a_1}$.

# Risk Function for Weighted Loss

The risk function for the weighted loss $L_{a_0,a_1}$ associated with a test $\varphi$ is:

$$\begin{aligned}\theta \longmapsto \mathbf{R}(\theta,\varphi) &= \mathbf{E}_\theta[L_{a_0,a_1}(\theta,\varphi(\mathbf{X}))] \\ &= \mathbf{E}_\theta[a_0 \mathbb{1}_{\varphi(\mathbf{X})=1}\mathbb{1}_{\theta\in\Theta_0} + a_1 \mathbb{1}_{\varphi(\mathbf{X})=0}\mathbb{1}_{\theta\in\Theta_1}] \\ &= a_0 P_\theta(\varphi(\mathbf{X})=1)\mathbb{1}_{\theta\in\Theta_0} + a_1 P_\theta(\varphi(\mathbf{X})=0)\mathbb{1}_{\theta\in\Theta_1}\end{aligned}$$

The Bayes risk of the test $\varphi$ is then the integral against the prior $\pi$:

$$\mathbf{R}_B(\pi,\varphi) = a_0 \int_{\Theta_0} P_\theta(\varphi(\mathbf{X})=1)d\pi(\theta) + a_1 \int_{\Theta_1} P_\theta(\varphi(\mathbf{X})=0)d\pi(\theta)$$

## Interpretation of the Bayes Test

By definition, the Bayes test minimizes the global Bayes risk $\mathbf{R}_B(\pi, \varphi)$.

- **Averaging Errors:** Type I and Type II errors are averaged according to the prior distribution $\pi$.
- **Weighting:** The constants $a_0$ and $a_1$ introduce additional weighting to penalize one type of error more heavily than the other if necessary.
- **Optimality:** This provides a systematic way to balance the trade-off between the two error types, which are not symmetric in practical applications.

The Bayes test naturally incorporates both our prior beliefs and the specific costs of making wrong decisions.

# Definition of the Bayes Factor

The Bayes Factor provides a way to evaluate the evidence provided by the data in favor of one hypothesis over another, independent of the prior probabilities of the hypotheses themselves.

## Definition

The Bayes factor $B^{\pi}_{0/1}(\mathbf{X})$ is defined as the ratio of posterior odds to prior odds:

$$B^{\pi}_{0/1}(\mathbf{X}) = \frac{\pi(\Theta_0 \mid \mathbf{X})/\pi(\Theta_1 \mid \mathbf{X})}{\pi(\Theta_0)/\pi(\Theta_1)} = \frac{\pi(\Theta_0 \mid \mathbf{X})\pi(\Theta_1)}{\pi(\Theta_1 \mid \mathbf{X})\pi(\Theta_0)}$$

- We similarly define $B^{\pi}_{1/0}(\mathbf{X}) = \left(B^{\pi}_{0/1}(\mathbf{X})\right)^{-1}$.
- It is interpreted as a Bayesian likelihood ratio.

# Marginal Likelihood Interpretation

Let $f(\mathbf{X}) = \int p_\theta(\mathbf{X}) d\pi(\theta)$ be the marginal density (marginal likelihood) of $\mathbf{X}$. Let $\pi_0$ and $\pi_1$ be the restrictions of the prior $\pi$ to $\Theta_0$ and $\Theta_1$:

$$\pi_0 = \frac{\pi[\cdot \cap \Theta_0]}{\pi[\Theta_0]}, \quad \pi_1 = \frac{\pi[\cdot \cap \Theta_1]}{\pi[\Theta_1]}$$

The marginal likelihood of $\mathbf{X}$ under $\pi_i$ is $\int_{\Theta_i} p_\theta(\mathbf{X}) d\pi_i(\theta)$. We can show:

$$B^\pi_{0/1}(\mathbf{X}) =$$

Thus, the Bayes factor is the ratio of marginal likelihoods under $H_0$ and $H_1$.

# Marginal Likelihood Interpretation

Let $f(\mathbf{X}) = \int p_\theta(\mathbf{X}) d\pi(\theta)$ be the marginal density (marginal likelihood) of $\mathbf{X}$. Let $\pi_0$ and $\pi_1$ be the restrictions of the prior $\pi$ to $\Theta_0$ and $\Theta_1$:

$$\pi_0 = \frac{\pi[\cdot \cap \Theta_0]}{\pi[\Theta_0]}, \quad \pi_1 = \frac{\pi[\cdot \cap \Theta_1]}{\pi[\Theta_1]}$$

The marginal likelihood of $\mathbf{X}$ under $\pi_i$ is $\int_{\Theta_i} p_\theta(\mathbf{X}) d\pi_i(\theta)$. We can show:

$$B^\pi_{0/1}(\mathbf{X}) = \frac{\int_{\Theta_0} p_\theta(\mathbf{X}) d\pi_0(\theta)}{\int_{\Theta_1} p_\theta(\mathbf{X}) d\pi_1(\theta)}$$

Thus, the Bayes factor is the ratio of marginal likelihoods under $H_0$ and $H_1$.

## Case: Simple Hypotheses

If $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, the prior is $\pi = \tilde{\pi}_0 \delta_{\theta_0} + \tilde{\pi}_1 \delta_{\theta_1}$.

In this specific case, the Bayes factor simplifies significantly:

$$B^{\pi}_{0/1}(\mathbf{X}) = \frac{p_{\theta_0}(\mathbf{X})}{p_{\theta_1}(\mathbf{X})}$$

- This is exactly the classical likelihood ratio.
- Note that in this simple case, the Bayes factor does not depend on the prior weights $\pi_0, \pi_1$.

# Point Null vs. Composite Hypothesis

Consider $H_0 : \{\theta = \theta_0\}$ vs $H_1 : \{\theta \in \Theta_1\}$. If we use a continuous prior $Q$ on $\{\theta_0\} \cup \Theta_1$, then $\pi(\Theta_0) = 0$, and we would always reject $H_0$.

**Solution:** We must assign a prior mass to the singleton $\theta_0$:

$$\pi = \pi_0 \delta_{\theta_0} + (1 - \pi_0) Q$$

where $dQ(\theta) = q(\theta) d\theta$ is a density on $\Theta_1$.

The posterior probability of $H_0$ becomes:

# Point Null vs. Composite Hypothesis

Consider $H_0 : \{\theta = \theta_0\}$ vs $H_1 : \{\theta \in \Theta_1\}$. If we use a continuous prior $Q$ on $\{\theta_0\} \cup \Theta_1$, then $\pi(\Theta_0) = 0$, and we would always reject $H_0$.

**Solution:** We must assign a prior mass to the singleton $\theta_0$:

$$\pi = \pi_0 \delta_{\theta_0} + (1 - \pi_0) Q$$

where $dQ(\theta) = q(\theta) d\theta$ is a density on $\Theta_1$.

The posterior probability of $H_0$ becomes:

$$\pi(\{\theta_0\} \mid \mathbf{X}) = \frac{\pi_0 p_{\theta_0}(\mathbf{X})}{\pi_0 p_{\theta_0}(\mathbf{X}) + (1 - \pi_0) \int_{\Theta_1} p_\theta(\mathbf{X}) q(\theta) d\theta}$$

## Example: Gaussian Point Null

Test $H_0 : \{\theta = 0\}$ vs $H_1 : \{\theta \neq 0\}$ in $\mathcal{N}(\theta, 1)^{\otimes n}$. Prior: $\pi = \pi_0 \delta_0 + (1 - \pi_0)\mathcal{N}(0, 1)$.

The marginal likelihood under $H_1$ is computed via integration:

$$\int p_\theta(\mathbf{X}) q(\theta) d\theta$$

where $p_0(\mathbf{X})$ is the likelihood under $\theta = 0$.

The resulting Bayes factor is:

$$B^\pi_{0/1}(\mathbf{X})$$

## Example: Gaussian Point Null

Test $H_0 : \{\theta = 0\}$ vs $H_1 : \{\theta \neq 0\}$ in $\mathcal{N}(\theta, 1)^{\otimes n}$. Prior: $\pi = \pi_0 \delta_0 + (1 - \pi_0)\mathcal{N}(0, 1)$.

The marginal likelihood under $H_1$ is computed via integration:

$$\int p_\theta(\mathbf{X}) q(\theta) d\theta = \frac{p_0(\mathbf{X}) \exp\left(\frac{(n\bar{X}_n)^2}{2(n+1)}\right)}{\sqrt{n+1}}$$

where $p_0(\mathbf{X})$ is the likelihood under $\theta = 0$.

The resulting Bayes factor is:

$$B^\pi_{0/1}(\mathbf{X}) = \sqrt{n+1} \exp\left(-\frac{n^2}{2(n+1)} \bar{X}_n^2\right)$$

# Decision Rule with Bayes Factor

Using the weighted Bayes test from before, we reject $H_0$ if:

$$B^{\pi}_{0/1}(\mathbf{X}) \leqslant \frac{a_1(1-\pi_0)}{a_0\pi_0}$$

- In the balanced case ($a_0 = a_1$ and $\pi_0 = 1/2$), this simplifies to:

$$B^{\pi}_{0/1}(\mathbf{X}) \leqslant 1$$

- This means we reject $H_0$ if the data is more likely under the alternative model than under the null model.

# Beyond Simple Testing: Model Comparison

We have seen how to test $H_0$ vs $H_1$. Often, we view this as comparing two different models $\mathcal{M}_0$ and $\mathcal{M}_1$.

- **Frequentist Problem:** The classical likelihood ratio $p_{\theta_0}(\mathbf{X})/p_{\theta_1}(\mathbf{X})$ typically increases with model complexity (of $\theta_0$).
- **Overfitting:** A more complex model (more parameters) will always provide a better fit to the data, even if it does not represent the true underlying process.
- **Bayesian Solution:** The marginal likelihood (ML) provides a natural mechanism to penalize overly complex models.

# The Marginal Likelihood: Natural Regularization

For a model $\mathcal{M}$ with parameters $\theta$ and prior $\pi(\theta \mid \mathcal{M})$, the marginal likelihood is:

$$m(\mathbf{X} \mid \mathcal{M}) = \int_{\Theta} f(\mathbf{X} \mid \theta, \mathcal{M}) \pi(\theta \mid \mathcal{M}) d\theta$$

- **The Averaging Argument:** Unlike the maximum likelihood which only cares about the single "best" $\theta$, the marginal likelihood averages the performance over the entire parameter space $\Theta$ or model $\mathcal{M}$.
- **Complexity vs. Evidence:** Even if a complex model contains a few "very good" parameters that fit the data well, it also includes a vast number of "bad" parameters that do not.
- **Regularization Effect:** As $\mathcal{M}$ becomes more complex, the prior mass prior $\pi(\theta \mid \mathcal{M})$ becomes more spread.
- **The Penalty:** Averaging a small region of high-fit parameters with a large region of low-fit parameters naturally regularizes the maximization problem.

The ML only increases if the improvement in fit outweighs the "dilution" caused by adding more bad models into the average.

# Summary: Bayesian Model Selection

1. **No Manual Tuning:** In frequentist statistics, many methods have been developed to penalize the likelihood with respect to increasing complexity of the model. However, these methods are not well-motivated. In the Bayesian framework, ML has a natural built-in penalty for more complex models.

2. **Evidence:** The Bayes factor $B_{0/1} = \frac{m(\mathbf{X}|\mathcal{M}_0)}{m(\mathbf{X}|\mathcal{M}_1)}$ is simply the ratio of these marginal likelihoods.

3. **Priors Matter:** Because the ML depends directly on the prior $\pi(\theta \mid \mathcal{M})$, model selection is much more sensitive to the choice of prior than point estimation is.