Bayes Factors

Statistics · Module 6 — Bayesian Inference

Node 6.6

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.6

Bayes factors are the Bayesian instrument for comparing models or hypotheses. They quantify the relative evidence the data provide for one model over another, on the scale of marginal likelihoods. This node depends on the posterior and marginal likelihood (Node 6.1), Bayesian estimation (Node 6.3), and the likelihood function (Node 2.3). It connects to the Neyman-Pearson framework (Node 5.2) as a special case and enables the Bernstein–von Mises reconciliation (Node 6.7) by establishing the role of the marginal likelihood in model evaluation.

Formal Definition

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two statistical models with parameter spaces Θ_0 and Θ_1, likelihoods $p_0(x \mid \theta_0)$ and $p_1(x \mid \theta_1)$, and priors $\pi_0(\theta_0)$ and $\pi_1(\theta_1)$.

Marginal Likelihood

The **marginal likelihood** under model $\mathcal{M}_j$ is

$$m_j(x) = \int_{\Theta_j} p_j(x \mid \theta_j)\, \pi_j(\theta_j)\, d\theta_j.$$

Bayes Factor

The **Bayes factor** in favor of $\mathcal{M}_1$ against $\mathcal{M}_0$ is

$$\mathrm{BF}_{10}(x) = \frac{m_1(x)}{m_0(x)}.$$

Posterior Model Odds

Given prior model probabilities $\pi(\mathcal{M}_j)$, the posterior odds satisfy

$$\frac{\pi(\mathcal{M}_1 \mid x)}{\pi(\mathcal{M}_0 \mid x)} = \mathrm{BF}_{10}(x) \cdot \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_0)}.$$

The Bayes factor transforms prior odds into posterior odds. It is the data's contribution to model comparison, separated from the prior model weights.

Intuition

The Bayes factor is a likelihood ratio averaged over parameter uncertainty. Unlike the frequentist likelihood ratio, which evaluates each model at its best-fitting parameter value, the Bayes factor evaluates each model at its average performance over the prior. This builds in an automatic complexity penalty: a model with a diffuse prior over many parameter values will spread its marginal likelihood thinly, while a model that concentrates prior mass near the data-generating region will score well.

The Bayes factor does not require choosing a significance level, does not depend on the sampling distribution of a test statistic, and can provide evidence in favor of the null hypothesis — not merely a failure to reject.

Structural Role

The Bayes factor is the coherent Bayesian update factor for model probabilities. It requires marginal likelihoods (Node 6.1) and connects to the likelihood ratio test (Node 5.2) as a special case for simple hypotheses.

It depends on prior specification over parameters within each model. Improper priors within models render the Bayes factor undefined (the marginal likelihood is defined only up to an arbitrary constant), so model comparison requires proper priors or careful calibration.

This node enables: the Bernstein–von Mises theorem (Node 6.7), which provides asymptotic approximations to the marginal likelihood and hence to Bayes factors via the BIC.

Mathematical Development

Relation to Neyman-Pearson Likelihood Ratio

For simple hypotheses $H_0 : P = P_0$, $H_1 : P = P_1$ (no free parameters):

$$\mathrm{BF}_{10}(x) = \frac{p_1(x)}{p_0(x)},$$

which is exactly the likelihood ratio of the Neyman-Pearson lemma (Node 5.2). The Bayes factor generalizes the likelihood ratio to composite hypotheses by integrating over parameters.

Jeffreys' Interpretive Scale

Harold Jeffreys proposed the following scale: $\log_{10} \mathrm{BF}_{10} \in (0, 1/2]$ is "barely worth mentioning," $(1/2, 1]$ is "substantial," $(1, 3/2]$ is "strong," $(3/2, 2]$ is "very strong," and > 2 is "decisive." This is a guideline, not a formal decision rule.

Savage-Dickey Density Ratio

For nested models where $\mathcal{M}_0$ is $\mathcal{M}_1$ with the restriction $\psi = \psi_0$ (a subvector of the parameter), the Bayes factor simplifies to

$$\mathrm{BF}_{01}(x) = \frac{\pi_1(\psi_0 \mid x)}{\pi_1(\psi_0)},$$

where $\pi_1(\psi_0 \mid x)$ is the marginal posterior density of ψ under $\mathcal{M}_1$ evaluated at ψ_0, and $\pi_1(\psi_0)$ is the corresponding prior density. This avoids computing the marginal likelihoods directly.

Derivation. Write $\theta_1 = (\psi, \lambda)$ where λ is the nuisance parameter. Under $\mathcal{M}_0$, $\psi = \psi_0$ is fixed and

$$m_0(x) = \int p(x \mid \psi_0, \lambda)\, \pi_0(\lambda)\, d\lambda.$$

Under $\mathcal{M}_1$, the marginal posterior of ψ evaluated at ψ_0 is

$$\pi_1(\psi_0 \mid x) = \frac{\int p(x \mid \psi_0, \lambda)\, \pi_1(\lambda \mid \psi_0)\, d\lambda \cdot \pi_1(\psi_0)}{m_1(x)}.$$

If the conditional prior $\pi_1(\lambda \mid \psi_0) = \pi_0(\lambda)$ (the priors on nuisance parameters agree), then

$$\pi_1(\psi_0 \mid x) = \frac{m_0(x)\, \pi_1(\psi_0)}{m_1(x)},$$

so $\mathrm{BF}_{01} = m_0/m_1 = \pi_1(\psi_0 \mid x)/\pi_1(\psi_0)$. $\square$

BIC Approximation to the Bayes Factor

The Schwarz (BIC) approximation to the log marginal likelihood is

$$\log m_j(x) \approx \ell_j(\hat{\theta}_j) - \frac{d_j}{2} \log n,$$

where $\ell_j(\hat{\theta}_j)$ is the maximized log-likelihood, $d_j = \dim(\Theta_j)$, and n is the sample size.

Derivation. Apply the Laplace approximation to the marginal likelihood integral. The log-likelihood near the MLE is

$$\ell(\theta) \approx \ell(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^\top \hat{J}(\theta - \hat{\theta}),$$

where $\hat{J} = -\nabla^2 \ell(\hat{\theta})$. The marginal likelihood becomes

$$m(x) \approx p(x \mid \hat{\theta})\, \pi(\hat{\theta})\, (2\pi)^{d/2}\, |\hat{J}|^{-1/2}.$$

Taking logarithms: $\log m(x) \approx \ell(\hat{\theta}) + \log \pi(\hat{\theta}) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\hat{J}|$. Since $\hat{J} \approx n\, I(\theta_0)$, $\log|\hat{J}| \approx d \log n + \log|I(\theta_0)|$. Dropping $O(1)$ terms (which include the prior and the Fisher information determinant) leaves

$$\log m(x) \approx \ell(\hat{\theta}) - \frac{d}{2} \log n + O(1).$$

The BIC approximation to the log Bayes factor is therefore

$$\log \mathrm{BF}_{10} \approx \left[\ell_1(\hat{\theta}_1) - \ell_0(\hat{\theta}_0)\right] - \frac{d_1 - d_0}{2} \log n.$$

Lindley's Paradox

Consider testing $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ with $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$ and prior $\mu \sim \mathcal{N}(0, \tau^2)$ under H_1.

The Bayes factor in favor of H_0 is

$$\mathrm{BF}_{01} = \sqrt{1 + n\tau^2}\, \exp\left(-\frac{n\bar{x}^2}{2} \cdot \frac{n\tau^2}{1 + n\tau^2}\right).$$

Fix $\bar{x} = z/\sqrt{n}$ (so that the z-statistic is fixed at some value that would reject H_0 in a frequentist test). As $n \to \infty$ with z fixed, $\mathrm{BF}_{01} \to \infty$: the Bayes factor increasingly favors H_0, even though the frequentist test rejects. This is **Lindley's paradox**: at any fixed significance level, there exist sample sizes for which the data reject H_0 but the Bayes factor favors H_0.

The resolution: the frequentist test and the Bayes factor answer different questions. The p-value measures tail probability; the Bayes factor measures average likelihood.

Worked Examples

Example 1: Two Poisson Rates

$\mathcal{M}_0$: $X_1, \dots, X_n \sim \mathrm{Poisson}(3)$ (simple).

$\mathcal{M}_1$: $X_i \sim \mathrm{Poisson}(\lambda)$, $\lambda \sim \mathrm{Gamma}(3, 1)$.

Data: $n = 20$, $\sum x_i = 72$, $\bar{x} = 3.6$.

$m_0 = \prod_{i=1}^{20} \frac{3^{x_i} e^{-3}}{x_i!} = \frac{3^{72} e^{-60}}{\prod x_i!}$.

$m_1 = \int_0^\infty \frac{\lambda^{72} e^{-20\lambda}}{\prod x_i!} \cdot \frac{\lambda^2 e^{-\lambda}}{\Gamma(3)}\, d\lambda = \frac{1}{\prod x_i! \cdot \Gamma(3)} \int_0^\infty \lambda^{74} e^{-21\lambda}\, d\lambda = \frac{\Gamma(75)}{\prod x_i! \cdot \Gamma(3) \cdot 21^{75}}$.

The Bayes factor $\mathrm{BF}_{10} = m_1/m_0$. Taking the log ratio:

$$\log \mathrm{BF}_{10} = \log \Gamma(75) - \log \Gamma(3) - 75 \log 21 - 72 \log 3 + 60.$$

Numerical evaluation gives $\log \mathrm{BF}_{10} \approx -1.8$, so $\mathrm{BF}_{10} \approx 0.17$. The data moderately favor $\mathcal{M}_0$ (the simple Poisson(3) model) over the Gamma-Poisson model, since the observed rate 3.6 is close to 3.

Example 2: Savage-Dickey for a Normal Mean

Test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. Under H_1: $X_i \sim \mathcal{N}(\mu, 1)$, $\mu \sim \mathcal{N}(0, \tau^2)$ with $\tau^2 = 10$.

Data: $n = 25$, $\bar{x} = 0.8$.

Posterior under H_1: $\mu \mid x \sim \mathcal{N}(\mu_n, \sigma_n^2)$ with $\sigma_n^2 = (1/10 + 25)^{-1} = 1/25.1 \approx 0.0398$, $\mu_n = 0.0398 \cdot (0 + 25 \cdot 0.8) = 0.796$.

Prior density at 0: $\pi_1(0) = (2\pi \cdot 10)^{-1/2} = 0.0399/\sqrt{2\pi} \approx 0.1262$.

Posterior density at 0: $\pi_1(0 \mid x) = (2\pi \cdot 0.0398)^{-1/2} \exp(-0.796^2/(2 \cdot 0.0398)) = 3.989 \cdot \exp(-7.96) \approx 3.989 \cdot 0.000349 \approx 0.00139$.

Savage-Dickey: $\mathrm{BF}_{01} = 0.00139/0.1262 \approx 0.011$.

Thus $\mathrm{BF}_{10} \approx 91$: strong evidence against H_0.

Cross-Links

- **Linear Algebra:** The Laplace approximation to the marginal likelihood involves the determinant of the observed information matrix $|\hat{J}|$, connecting model comparison to the spectral properties of the Hessian of the log-likelihood.
- **AI:** Bayesian model selection via Bayes factors underlies model comparison in Gaussian process regression (choosing kernel hyperparameters) and Bayesian neural architecture search. The marginal likelihood is the "type-II likelihood" maximized in empirical Bayes.
- **Economics:** Bayesian model comparison using Bayes factors is standard in macroeconomics for comparing DSGE models. The marginal likelihood penalizes overfit, addressing the same concern as information criteria.
- **Quantum Mechanics:** Quantum hypothesis testing (discriminating between quantum states) uses a quantum analogue of the Bayes factor, where the marginal likelihoods are replaced by traces of the state operators against measurement outcomes.

Structural Compression

Invariant: $\mathrm{BF}_{10}(x) = m_1(x)/m_0(x)$. The ratio of marginal likelihoods is the coherent Bayesian update factor for model probabilities.

Mechanism: Each marginal likelihood averages the data likelihood over prior uncertainty in the parameters. The ratio compares these averages, automatically penalizing model complexity through prior dispersion: a model that spreads prior mass over regions of low likelihood is penalized.

Cross-System Echo: Bayes factors generalize the Neyman-Pearson likelihood ratio to composite hypotheses. The BIC approximation $2 \log \mathrm{BF} \approx 2\Delta\ell - \Delta d \cdot \log n$ connects Bayesian model comparison to penalized likelihood criteria (AIC, BIC). In coding theory, the marginal likelihood is the inverse of the minimum description length under the prior.

Exercises

1. Derive the Bayes factor for testing $H_0 : \theta = 1/2$ vs $H_1 : \theta \neq 1/2$ in the Binomial model with prior $\mathrm{Beta}(1, 1)$ under H_1. Compute numerically for $n = 100$, $s = 55$.
2. Prove the Savage-Dickey density ratio for a one-dimensional parameter of interest ψ, starting from the marginal likelihood ratio and the definition of the marginal posterior.
3. Derive the BIC approximation to $\log m(x)$ via the Laplace method, keeping all terms through $O(1)$ before dropping to the $O(\log n)$ approximation. Identify precisely which terms are absorbed into the $O(1)$ remainder.
4. Demonstrate Lindley's paradox numerically: for the normal mean test with $\tau^2 = 1$ and a fixed z-statistic of 2.5, compute BF_{01} for $n = 10, 100, 1000, 10000$ and verify the Bayes factor increasingly favors H_0.
5. Show that the Bayes factor is symmetric: $\mathrm{BF}_{01} = 1/\mathrm{BF}_{10}$, and that this property fails for p-values (a p-value against H_0 is not the complement of a p-value against H_1).
6. For two nested normal linear regression models with d_0 and $d_1 > d_0$ predictors, express the BIC-approximated Bayes factor in terms of the residual sums of squares and the sample size. Show that BIC penalizes additional parameters more heavily than AIC when $n \geq 8$.
7. Argue, structurally, why improper priors make the Bayes factor undefined: identify where the arbitrary normalizing constant enters and explain why the prior predictive interpretation of the marginal likelihood fails when the prior is not a probability measure.

Statistics · Module 6 — Bayesian Inference · Node 6.6
Concepts: model-selection, likelihood, prior-posterior
Prerequisites: 6.1, 6.3, 2.3
Enables: 6.7

hbar.university — own.your.cognition