Module 6 — Bayesian Inference

Statistics

hbar.university

Priors and Posteriors

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.1

The prior-to-posterior update is the foundational operation of Bayesian inference. Every subsequent node in this module — conjugate families, point estimation, credible sets, predictive distributions, model comparison, and asymptotic reconciliation — is a derived consequence of the posterior distribution defined here. This node depends on the parametric model framework (Node 2.1), the likelihood function (Node 2.3), and the law of total probability (Node 1.7).

Formal Definition

Let $X = (X_1, \ldots, X_n)$ be a random sample with joint density or mass function $p(x \mid \theta)$ for $\theta \in \Theta \subseteq \mathbb{R}^k$.

Prior Distribution

A **prior distribution** $\pi(\theta)$ is a probability measure on $(\Theta, \mathcal{B}(\Theta))$ specified before observing X. It encodes the state of knowledge about θ prior to data collection. A prior is **proper** if $\int_\Theta \pi(\theta)\, d\theta = 1$, and **improper** if this integral diverges.

Posterior Distribution

Given observed data $X = x$, the **posterior distribution** is

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

where the **marginal likelihood** (or evidence) is

$$m(x) = \int_\Theta p(x \mid \theta)\, \pi(\theta)\, d\theta.$$

The posterior exists as a proper distribution whenever $0 < m(x) < \infty$.

Jeffreys Prior

The **Jeffreys prior** is defined as

$$\pi^J(\theta) \propto \sqrt{\det I(\theta)},$$

where $I(\theta)$ is the Fisher information matrix. This prior is invariant under reparametrization: if $\phi = g(\theta)$ is a smooth bijection, then $\pi^J(\phi) \propto \sqrt{\det I_\phi(\phi)}$, where I_ϕ is the Fisher information in the ϕ-parametrization.

Intuition

The posterior is a compromise between what we believed before seeing data (the prior) and what the data tell us (the likelihood). In regions where the likelihood is large, the posterior concentrates mass; in regions where the prior assigns little mass, the posterior is suppressed even if the likelihood is moderate.

The marginal likelihood $m(x)$ serves purely as a normalizing constant for the posterior but carries independent significance as the model's overall predictive score for the observed data (Node 6.6).

Improper priors are useful default specifications — they express maximal pre-data ignorance within a parametric family — but require verification that the resulting posterior is proper.

The Jeffreys prior is the canonical "objective" prior: it assigns prior mass in a way that respects the geometry of the statistical model, treating reparametrizations as coordinate changes on the same underlying manifold.

Structural Role

The prior-to-posterior update is the complete description of Bayesian learning. All Bayesian procedures — point estimates (Node 6.3), credible sets (Node 6.4), posterior predictive distributions (Node 6.5), and Bayes factors (Node 6.6) — are functionals of the posterior.

This node requires: parametric models (Node 2.1), the likelihood principle (Node 2.3), and marginalization via the law of total probability (Node 1.7).

This node enables: every subsequent node in Module 6. The posterior is also the bridge to asymptotic reconciliation with frequentist inference via the Bernstein–von Mises theorem (Node 6.7).

Mathematical Development

Derivation of Bayes' Theorem for Densities

Starting from the joint density of (X, θ):

$$p(x, \theta) = p(x \mid \theta)\, \pi(\theta) = \pi(\theta \mid x)\, m(x).$$

Solving for $\pi(\theta \mid x)$ yields Bayes' theorem. The key requirement is $m(x) > 0$, which holds whenever the data are compatible with at least one parameter value that receives prior mass.

Sequential Updating

For i.i.d. data, the posterior after n observations can be computed sequentially. Let $\pi_0(\theta) = \pi(\theta)$ and define

$$\pi_k(\theta) \propto p(x_k \mid \theta)\, \pi_{k-1}(\theta), \quad k = 1, \ldots, n.$$

Then $\pi_n(\theta) = \pi(\theta \mid x_1, \dots, x_n)$. The order of updating does not matter:

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

Sufficiency and the Posterior

If $T(X)$ is a sufficient statistic for θ, then by the factorization theorem $p(x \mid \theta) = g(T(x), \theta)\, h(x)$, so

$$\pi(\theta \mid x) = \frac{g(T(x), \theta)\, h(x)\, \pi(\theta)}{\int g(T(x), \theta)\, h(x)\, \pi(\theta)\, d\theta} = \frac{g(T(x), \theta)\, \pi(\theta)}{\int g(T(x), \theta)\, \pi(\theta)\, d\theta} = \pi(\theta \mid T(x)).$$

The posterior depends on the data only through the sufficient statistic.

Proper vs Improper Priors

An improper prior $\pi(\theta)$ with $\int_\Theta \pi(\theta)\, d\theta = \infty$ can still yield a proper posterior, provided $m(x) = \int p(x \mid \theta)\, \pi(\theta)\, d\theta < \infty$ for almost all x. For example, the flat prior $\pi(\mu) = 1$ on $\mathbb{R}$ for the normal mean yields proper posterior $\mu \mid x \sim \mathcal{N}(\bar{x}, \sigma^2/n)$.

Derivation of the Jeffreys Prior

Consider a one-parameter model with Fisher information $I(\theta) = -\mathbb{E}_\theta[\ell''(\theta)]$, where $\ell(\theta) = \log p(X \mid \theta)$. Under reparametrization $\phi = g(\theta)$, the Fisher information transforms as

$$I_\phi(\phi) = I(\theta) \left(\frac{d\theta}{d\phi} \right)^2.$$

The prior density transforms as $\pi_\phi(\phi) = \pi(\theta) \left| \frac{d\theta}{d\phi} \right|$. Setting $\pi(\theta) \propto \sqrt{I(\theta)}$ gives

$$\pi_\phi(\phi) \propto \sqrt{I(\theta)} \left| \frac{d\theta}{d\phi} \right| = \sqrt{I(\theta) \left(\frac{d\theta}{d\phi} \right)^2} = \sqrt{I_\phi(\phi)}.$$

Hence the Jeffreys prior is invariant under reparametrization.

Posterior Concentration

As $n \to \infty$, the posterior concentrates around the true parameter value θ_0. Informally, the log-posterior is

$$\log \pi(\theta \mid x) = \sum_{i=1}^{n} \log p(x_i \mid \theta) + \log \pi(\theta) - \log m(x).$$

The likelihood sum grows as $O(n)$ while $\log \pi(\theta)$ is $O(1)$, so the prior is overwhelmed. Under regularity conditions, this is made precise by the Bernstein–von Mises theorem (Node 6.7).

Worked Examples

Example 1: Normal Mean with Flat Prior

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. Use the improper flat prior $\pi(\mu) \propto 1$.

The likelihood is

$$p(x \mid \mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = \exp\Big(-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\Big) \cdot C(x),$$

where $C(x)$ does not depend on μ. Thus

$$\pi(\mu \mid x) \propto \exp\Big(-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\Big),$$

which is the kernel of $\mathcal{N}(\bar{x}, \sigma^2/n)$. The posterior is proper despite the improper prior.

Example 2: Jeffreys Prior for the Bernoulli Model

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0, 1)$. The Fisher information is

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

The Jeffreys prior is $\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, which is $\text{Beta}(1/2, 1/2)$. With $s = \sum x_i$ successes, the posterior is

$$\pi(\theta \mid x) = \text{Beta}\big(\tfrac{1}{2} + s,\ \tfrac{1}{2} + n - s\big).$$

The posterior mean is $(s + 1/2)/(n + 1)$, which shrinks the MLE s/n toward $1/2$ by an amount that vanishes as $n \to \infty$.

Cross-Links

- **Linear Algebra:** The Fisher information matrix $I(\theta)$ is a Riemannian metric on the parameter manifold; the Jeffreys prior is the volume element of this metric, connecting Bayesian inference to differential geometry on statistical manifolds.
- **AI:** Variational inference approximates the posterior $\pi(\theta \mid x)$ by optimizing over a tractable family when the marginal likelihood $m(x)$ is intractable. The prior-to-posterior update is the exact operation that variational and MCMC methods approximate.
- **Economics:** Bayesian agents update beliefs via Bayes' rule; rational expectations equilibria assume all agents compute posteriors from public signals. The prior encodes heterogeneous beliefs across agents.
- **Quantum Mechanics:** State update upon measurement (the Luders rule) is the non-commutative analogue of Bayes' theorem. The prior is a density matrix; the posterior is the post-measurement state.

Structural Compression

Invariant: $\pi(\theta \mid x) \propto p(x \mid \theta)\,\pi(\theta)$. The posterior is proportional to likelihood times prior.

Mechanism: Bayes' theorem normalizes the product $p(x \mid \theta)\,\pi(\theta)$ over Θ to produce a proper probability measure. The likelihood reweights prior mass according to data compatibility. The normalizing constant $m(x)$ ensures coherence.

Cross-System Echo: The posterior update is the universal belief revision rule: it appears as Bayesian filtering in signal processing, as the E-step kernel in EM algorithms, as density matrix conditioning in quantum information, and as belief propagation in graphical models.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$. Derive the posterior under the Jeffreys prior $\pi(\lambda) \propto 1/\lambda$. Verify the posterior is proper for $n \geq 1$.
2. Prove that the posterior distribution depends on the data only through a sufficient statistic by applying the factorization theorem.
3. Show that the Jeffreys prior for the normal variance σ^2 (with known mean) is $\pi(\sigma^2) \propto 1/\sigma^2$. Derive the resulting posterior and identify its distributional family.
4. Compute the posterior for θ in the model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, \theta)$ under the prior $\pi(\theta) \propto \theta^{-1}$. Show that the MLE and the posterior mode differ.
5. Demonstrate sequential updating: let X_1, X_2, X_3 be i.i.d. Bernoulli(θ) with $\pi(\theta) = \text{Beta}(2, 2)$. Compute the posterior after each observation for the sequence $(1, 0, 1)$ and verify the final posterior matches the batch computation.
6. Prove that if $\pi(\theta)$ is a proper prior and $p(x \mid \theta)$ is a bounded likelihood, then the posterior $\pi(\theta \mid x)$ is proper for all x.
7. Argue, structurally, why the prior is asymptotically irrelevant in regular parametric models: identify the rate at which the likelihood dominates the prior, and explain why this argument fails in nonparametric settings where Θ is infinite-dimensional.

Conjugate Families

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.2

Conjugate priors provide the rare setting in which Bayesian updating is analytically tractable, yielding closed-form posteriors that reveal the exact mechanism by which data modify beliefs. This node depends on the prior-to-posterior update (Node 6.1) and the exponential family structure (Node 2.6). It enables tractable computation of Bayesian point estimates (Node 6.3) and credible sets (Node 6.4), and supplies the canonical examples used throughout Module 6.

Formal Definition

Let $\{p(x \mid \theta) : \theta \in \Theta\}$ be a parametric likelihood family and let $\mathcal{F} = \{\pi(\theta \mid \eta) : \eta \in H\}$ be a family of prior distributions indexed by hyperparameters $\eta \in H$.

Conjugate prior family. The family $\mathcal{F}$ is **conjugate** to the likelihood $p(x \mid \theta)$ if for every $\eta \in H$ and every observation x,

$$\pi(\theta \mid \eta) \in \mathcal{F} \implies \pi(\theta \mid x) \in \mathcal{F}.$$

That is, the posterior belongs to the same parametric family as the prior, differing only in the value of the hyperparameter.

Exponential Family Conjugacy

For a likelihood in the canonical exponential family,

$$p(x \mid \theta) = h(x) \exp\big\{\eta(\theta)^\top T(x) - A(\theta)\big\},$$

the conjugate prior takes the form

$$\pi(\theta \mid \lambda_0, n_0) \propto \exp\big\{\lambda_0^\top \eta(\theta) - n_0\, A(\theta)\big\},$$

where (λ_0, n_0) are hyperparameters. After n i.i.d. observations, the posterior hyperparameters update as

$$\lambda_0 \mapsto \lambda_n = \lambda_0 + \sum_{i=1}^{n} T(x_i), \qquad n_0 \mapsto n_n = n_0 + n.$$

Intuition

Conjugacy means the prior and posterior speak the same parametric language. The data do not change the functional form of our belief — they only shift the hyperparameters. The hyperparameter n_0 acts as a prior pseudo-sample size: it quantifies the strength of prior information in units commensurate with data. The ratio λ_0/n_0 is a prior guess at the sufficient statistic per observation. As data accumulate, the posterior hyperparameters drift toward data-driven values and the prior pseudo-sample size becomes negligible relative to the actual sample size.

Conjugacy is the exception rather than the rule. Most interesting models — mixture models, hierarchical models, regression with non-conjugate priors — require computational approximation (MCMC, variational inference). But conjugate families remain the essential test bed for understanding Bayesian mechanics.

Structural Role

Conjugate priors make the posterior analytically tractable by closing the family under Bayesian updating. The hyperparameter update equations reveal the precise mechanism by which data accumulate: additive updates to the natural parameter and sample-size components.

This node requires the exponential family structure (Node 2.6) and the prior-to-posterior update (Node 6.1). It enables closed-form computation in Bayesian point estimation (Node 6.3), credible sets (Node 6.4), and posterior predictive distributions (Node 6.5). When conjugacy is unavailable, computation requires MCMC or variational methods (Module 9).

Mathematical Development

Proof of Beta-Binomial Conjugacy

Let $X_1, \ldots, X_n \mid \theta \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ and $\pi(\theta) = \text{Beta}(\alpha_0, \beta_0)$.

The prior density is

$$\pi(\theta) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\,\Gamma(\beta_0)}\,\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}, \quad \theta \in (0,1).$$

The likelihood for $s = \sum_{i=1}^{n} x_i$ successes in n trials is

$$p(x \mid \theta) = \theta^s (1-\theta)^{n-s}.$$

By Bayes' theorem:

$$\pi(\theta \mid x) \propto \theta^s(1-\theta)^{n-s} \cdot \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1} = \theta^{\alpha_0 + s - 1}(1-\theta)^{\beta_0 + n - s - 1}.$$

This is the kernel of $\text{Beta}(\alpha_0 + s,\ \beta_0 + n - s)$. Since the kernel uniquely determines the distribution, the posterior is Beta, proving conjugacy. □

Normal-Normal Conjugacy (Known Variance)

Let $X_i \mid \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known, and $\pi(\mu) = \mathcal{N}(\mu_0, \tau_0^2)$.

Define the prior precision $\kappa_0 = 1/\tau_0^2$ and data precision $\kappa_{\text{data}} = n/\sigma^2$. The posterior is

$$\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2),$$

where

$$\tau_n^2 = \frac{1}{\kappa_0 + \kappa_{\text{data}}}, \qquad \mu_n = \tau_n^2(\kappa_0 \mu_0 + \kappa_{\text{data}} \bar{x})\,.$$

The posterior mean is a precision-weighted average of the prior mean and sample mean. As $n \to \infty$, $\mu_n \to \bar{x}$ and $\tau_n^2 \to \sigma^2/n$.

Gamma-Poisson Conjugacy

Let $X_i \mid \lambda \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ and $\pi(\lambda) = \text{Gamma}(\alpha_0, \beta_0)$ with density $\pi(\lambda) \propto \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda}$.

The likelihood is $p(x \mid \lambda) \propto \lambda^{\sum x_i} e^{-n\lambda}$. Thus

$$\pi(\lambda \mid x) \propto \lambda^{\alpha_0 + \sum x_i - 1} e^{-(\beta_0 + n)\lambda},$$

which is $\text{Gamma}(\alpha_0 + \sum x_i,\ \beta_0 + n)$. The posterior mean is $(\alpha_0 + \sum x_i)/(\beta_0 + n)$.

Dirichlet-Multinomial Conjugacy

Let $X \mid \theta \sim \text{Multinomial}(n, \theta_1, \ldots, \theta_K)$ with $\pi(\theta) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$. The posterior is

$$\theta \mid x \sim \text{Dirichlet}(\alpha_1 + x_1, \ldots, \alpha_K + x_K).$$

Each hyperparameter α_k acts as a prior pseudo-count for category k. The posterior mean for component k is $(\alpha_k + x_k)/(\sum_j \alpha_j + n)$.

Conjugate Table Summary

Likelihood	Conjugate Prior	Posterior	Update Rule
Bernoulli(θ)	Beta(α_0, β_0)	Beta($\alpha_0 + s, \beta_0 + n - s$)	add successes, failures
$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(\mu_0, \tau_0^2)$	$\mathcal{N}(\mu_n, \tau_n^2)$	precision-weighted mean
Poisson(λ)	Gamma(α_0, β_0)	Gamma($\alpha_0 + \sum x_i, \beta_0 + n$)	add counts, observations
Multinomial(n, θ)	Dirichlet(α)	Dirichlet($\alpha + x$)	add category counts
$\mathcal{N}(\mu, \sigma^2)$ (both unknown)	NIG($\mu_0, \kappa_0, \alpha_0, \beta_0$)	NIG($\mu_n, \kappa_n, \alpha_n, \beta_n$)	see standard references

Interpretation of Hyperparameters

In every conjugate pair, the hyperparameters decompose into a **pseudo-sample size** n_0 and a **pseudo-sufficient statistic** λ_0. For Beta-Binomial: $n_0 = \alpha_0 + \beta_0$ (prior trials), $\lambda_0/n_0 = \alpha_0/(\alpha_0 + \beta_0)$ (prior success rate). For Gamma-Poisson: $n_0 = \beta_0$ (prior observations), $\lambda_0/n_0 = \alpha_0/\beta_0$ (prior rate).

Worked Examples

Example 1: Beta-Binomial with Clinical Trial Data

A drug trial tests 20 patients; 14 respond. The prior on the response rate is $\theta \sim \text{Beta}(2, 2)$, encoding a mild belief toward 1/2.

Posterior: $\text{Beta}(2 + 14, 2 + 6) = \text{Beta}(16, 8)$.

Posterior mean: $16/24 = 2/3 \approx 0.667$.

Prior mean: $2/4 = 0.5$. MLE: $14/20 = 0.7$.

The posterior mean lies between the prior mean and the MLE, closer to the MLE because the data (pseudo-sample size 20) dominate the prior (pseudo-sample size 4).

Posterior variance: $(16 \cdot 8)/(24^2 \cdot 25) = 128/14400 \approx 0.0089$. Posterior standard deviation: ≈ 0.094.

Example 2: Normal-Normal with Sensor Calibration

A sensor measures a voltage with known noise $\sigma^2 = 4$. Prior belief: $\mu \sim \mathcal{N}(10, 1)$, i.e., $\mu_0 = 10, \tau_0^2 = 1, \kappa_0 = 1$.

After $n = 5$ readings with $\bar{x} = 11.2$: data precision $\kappa_{\text{data}} = 5/4 = 1.25$.

Posterior precision: $\kappa_0 + \kappa_{\text{data}} = 1 + 1.25 = 2.25$.

Posterior variance: $\tau_n^2 = 1/2.25 \approx 0.444$.

Posterior mean: $\mu_n = (1/2.25)(1 \cdot 10 + 1.25 \cdot 11.2) = (1/2.25)(10 + 14) = 24/2.25 \approx 10.667$.

The posterior shifts from the prior mean 10 toward the sample mean 11.2, with the shift proportional to the relative data precision.

Cross-Links

- **Linear Algebra:** The precision-weighted update in Normal-Normal conjugacy is a special case of the Woodbury matrix identity applied to precision matrices. In the multivariate case, the posterior precision matrix is the sum of the prior precision and the data precision: $\Sigma_n^{-1} = \Sigma_0^{-1} + n\Sigma^{-1}$.
- **AI:** Conjugate priors underlie Bayesian online learning: each mini-batch update is a hyperparameter shift. Latent Dirichlet Allocation uses Dirichlet-Multinomial conjugacy as its core inferential engine.
- **Economics:** Beta-Bernoulli conjugacy models Bayesian learning by agents who observe binary signals (buy/sell, default/no-default) and update beliefs about an unknown probability.
- **Quantum Mechanics:** The Gaussian state update under homodyne measurement in quantum optics is formally identical to Normal-Normal conjugacy, with the Wigner function playing the role of the prior density.

Structural Compression

Invariant: The posterior hyperparameters are additive in the sufficient statistic of the data: $\eta_n = \eta_0 + T_n$, where T_n is the vector of accumulated sufficient statistics.

Mechanism: Exponential family structure ensures the product of likelihood and conjugate prior remains in the same exponential family. The natural parameter of the posterior is a linear function of the natural parameter of the prior and the sufficient statistic of the data.

Cross-System Echo: Conjugate updating is the Bayesian instance of Kalman filtering (linear-Gaussian conjugacy is the Kalman filter in one dimension). In information geometry, conjugate priors lie on the same exponential family manifold as the likelihood, and the posterior update is a geodesic step on this manifold.

Exercises

1. Prove Gamma-Poisson conjugacy: show that if $X_i \mid \lambda \sim \text{Poisson}(\lambda)$ and $\pi(\lambda) = \text{Gamma}(\alpha_0, \beta_0)$, then the posterior is $\text{Gamma}(\alpha_0 + \sum x_i, \beta_0 + n)$.
2. For the Normal-Normal model with known variance, show that the posterior mean can be written as $\mu_n = w \cdot \mu_0 + (1 - w) \cdot \bar{x}$ where $w = \tau_0^{-2}/(\tau_0^{-2} + n\sigma^{-2})$. Interpret w as a function of n and verify $w \to 0$ as $n \to \infty$.
3. Derive the conjugate prior for the exponential distribution $p(x \mid \lambda) = \lambda e^{-\lambda x}$. Identify the conjugate family, state the posterior hyperparameter update, and compute the posterior mean.

4. In the Dirichlet-Multinomial model with $K = 3$ and prior Dirichlet$(1, 1, 1)$, compute the posterior after observing counts $(x_1, x_2, x_3) = (10, 5, 3)$. Find the posterior mean and the posterior mode of each component.

5. Show that for the Beta-Binomial model, the prior predictive probability of s successes in n trials is the Beta-Binomial distribution: $p(s) = \binom{n}{s} \frac{B(\alpha_0+s, \beta_0+n-s)}{B(\alpha_0, \beta_0)}$.

6. Prove that the Normal-Normal posterior variance τ_n^2 is always smaller than both the prior variance τ_0^2 and the sampling variance σ^2/n. When are the two bounds approximately equal?

7. Argue, structurally, why conjugacy is restricted to exponential families: identify the algebraic property of the likelihood that allows the prior-posterior product to remain in the same family, and explain why non-exponential families (e.g., Cauchy) do not admit conjugate priors of finite parametric dimension.

Bayesian Point Estimation

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.3

Bayesian point estimation selects a single summary of the posterior distribution by minimizing posterior expected loss. This node connects the posterior (Node 6.1) to the decision-theoretic framework (Node 3.1), replacing the frequentist risk with posterior risk. It provides the estimators — posterior mean, MAP, posterior median — that serve as inputs to prediction (Node 6.5) and model comparison (Node 6.6).

Formal Definition

Loss Function and Posterior Risk

A **loss function** is a measurable map $L : \Theta \times \mathcal{A} \to [0, \infty)$, where $\mathcal{A}$ is the action space. The **posterior expected loss** (posterior risk) of action a given data x is

$$\rho(a, x) = \int_{\Theta} L(\theta, a)\, \pi(\theta \mid x)\, d\theta.$$

Bayes Estimator

The **Bayes estimator** under loss L is

$$\hat{\theta}^{\text{Bayes}}(x) = \arg\min_{a \in \mathcal{A}} \ \rho(a, x) = \arg\min_{a \in \mathcal{A}} \int_{\Theta} L(\theta, a)\, \pi(\theta \mid x)\, d\theta.$$

Bayes Risk

The **Bayes risk** of an estimator $\hat{\theta}$ under prior π is the prior expectation of the frequentist risk:

$$r(\pi, \hat{\theta}) = \int_{\Theta} R(\theta, \hat{\theta})\, \pi(\theta)\, d\theta = \int_{\Theta} \int_{\mathcal{X}} L(\theta, \hat{\theta}(x))\, p(x \mid \theta)\, dx\, \pi(\theta)\, d\theta.$$

The Bayes estimator minimizes the Bayes risk over all estimators.

Intuition

The Bayesian framework converts estimation into an optimization problem: given the posterior, choose the single value that incurs the smallest expected loss. Different loss functions extract different features of the posterior.

Squared error loss penalizes large deviations quadratically and yields the posterior mean — the center of mass. Absolute error loss penalizes proportionally and yields the posterior median — the balance point. Zero-one loss (for discrete parameters) or a limiting peaked loss yields the posterior mode — the single most probable value.

The Bayes estimator is always admissible (cannot be dominated in risk by any other estimator at every parameter value), provided the prior assigns positive mass to all open sets. This is a deep connection between Bayesian and frequentist optimality.

Structural Role

Bayesian point estimation reframes optimality from the frequentist minimax or unbiasedness criteria to posterior loss minimization. The choice of loss function determines which posterior functional is optimal.

This node requires: the posterior distribution (Node 6.1) and the decision-theoretic framework (Node 3.1). It enables: posterior predictive distributions (Node 6.5) via plug-in vs integrated prediction, and connects to credible sets (Node 6.4) through the relationship between the MAP estimator and the HPD region.

Mathematical Development

Proof: Posterior Mean Minimizes Squared Error Loss

Theorem. Under squared error loss $L(\theta, a) = (\theta - a)^2$, the Bayes estimator is the posterior mean $\hat{\theta}^{\text{Bayes}}(x) = \mathbb{E}[\theta \mid x]$.

Proof. The posterior expected loss is

$$\rho(a, x) = \int_{\Theta} (\theta - a)^2 \, \pi(\theta \mid x) \, d\theta.$$

Expand the square:

$$\rho(a, x) = \int_{\Theta} \theta^2 \, \pi(\theta \mid x) \, d\theta - 2a \int_{\Theta} \theta \, \pi(\theta \mid x) \, d\theta + a^2.$$

Differentiate with respect to a:

$$\frac{\partial}{\partial a} \rho(a, x) = -2 \, \mathbb{E}[\theta \mid x] + 2a.$$

Setting this to zero gives $a^* = \mathbb{E}[\theta \mid x]$. The second derivative is $2 > 0$, confirming a minimum. □

Proof: Posterior Median Minimizes Absolute Error Loss

Theorem. Under absolute error loss $L(\theta, a) = |\theta - a|$, the Bayes estimator is the posterior median.

Proof. Write

$$\rho(a, x) = \int_{-\infty}^{a} (a - \theta) \, \pi(\theta \mid x) \, d\theta + \int_{a}^{\infty} (\theta - a) \, \pi(\theta \mid x) \, d\theta.$$

Differentiate under the integral sign (using Leibniz's rule):

$$\frac{\partial}{\partial a} \rho(a, x) = \int_{-\infty}^{a} \pi(\theta \mid x) \, d\theta - \int_{a}^{\infty} \pi(\theta \mid x) \, d\theta = 2F_{\pi}(a \mid x) - 1,$$

where $F_{\pi}(\cdot \mid x)$ is the posterior CDF. Setting this to zero gives $F_{\pi}(a^* \mid x) = 1/2$, i.e., a^* is the posterior median. □

MAP Estimator and Its Decision-Theoretic Justification

For a continuous parameter space, the MAP estimator is

$$\hat{\theta}^{\text{MAP}}(x) = \arg\max_{\theta \in \Theta} \ \pi(\theta \mid x) = \arg\max_{\theta \in \Theta} \ \big[\log p(x \mid \theta) + \log \pi(\theta)\big].$$

The MAP is the Bayes estimator under the limiting loss $L_\varepsilon(\theta, a) = \mathbf{1}(|\theta - a| > \varepsilon)$ as $\varepsilon \to 0$, provided the posterior is unimodal and continuous.

Connection to Regularized Maximum Likelihood

Under a Gaussian prior $\pi(\theta) = \mathcal{N}(0, \sigma_\pi^2 I)$, the MAP estimator solves

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} \ \left[\sum_{i=1}^{n} \log p(x_i \mid \theta) - \frac{\|\theta\|^2}{2\sigma_\pi^2}\right],$$

which is **ridge regression** (L2-regularized MLE) with penalty $\lambda = 1/(2\sigma_\pi^2)$.

Under a Laplace prior $\pi(\theta_j) \propto \exp(-|\theta_j|/b)$, the MAP estimator becomes the **LASSO** (L1-regularized MLE).

Admissibility of Bayes Estimators

Theorem. If $\pi(\theta) > 0$ for all $\theta \in \Theta$ and the Bayes risk $r(\pi, \hat{\theta}^{\text{Bayes}}) < \infty$, then $\hat{\theta}^{\text{Bayes}}$ is admissible.

Proof sketch. Suppose $\hat{\theta}'$ dominates $\hat{\theta}^{\text{Bayes}}$: $R(\theta, \hat{\theta}') \le R(\theta, \hat{\theta}^{\text{Bayes}})$ for all θ, with strict inequality on a set of positive prior measure. Then $r(\pi, \hat{\theta}') < r(\pi, \hat{\theta}^{\text{Bayes}})$, contradicting the minimality of the Bayes risk. □

Posterior Mean as Shrinkage Estimator

In the Normal-Normal conjugate setting with prior $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$ and data $X_i \sim \mathcal{N}(\mu, \sigma^2)$, the posterior mean is

$$\hat{\mu}^{\text{Bayes}} = w\,\mu_0 + (1 - w)\,\bar{X}, \quad w = \frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n}.$$

This shrinks the MLE $\bar{X}$ toward the prior mean μ_0. The shrinkage factor w decreases with n and with prior variance τ_0^2, vanishing as either grows.

Worked Examples

Example 1: Posterior Mean for Beta-Binomial

Let $X_1, \ldots, X_{20} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ with $s = 12$ successes. Prior: $\theta \sim \text{Beta}(3, 3)$.

Posterior: $\text{Beta}(15, 11)$.

Posterior mean (Bayes estimator under squared error loss):

$$\hat{\theta}^{\text{Bayes}} = \frac{15}{15 + 11} = \frac{15}{26} \approx 0.577.$$

MLE: $12/20 = 0.6$. Prior mean: $3/6 = 0.5$.

Posterior mode (MAP estimator): $(15 - 1)/(15 + 11 - 2) = 14/24 \approx 0.583$.

Posterior median: approximately 0.577 (close to the mean since Beta(15,11) is nearly symmetric).

Example 2: Posterior Mean under Normal-Normal with Asymmetric Information

Prior: $\mu \sim \mathcal{N}(0, 100)$ (vague). Data: $n = 10$, $\bar{x} = 3.5$, $\sigma^2 = 4$.

Prior precision: $\kappa_0 = 1/100 = 0.01$. Data precision: $\kappa_{\text{data}} = 10/4 = 2.5$.

Posterior mean:

$$\hat{\mu}^{\text{Bayes}} = \frac{0.01 \cdot 0 + 2.5 \cdot 3.5}{0.01 + 2.5} = \frac{8.75}{2.51} \approx 3.486.$$

The vague prior barely influences the estimate: shrinkage weight $w = 0.01/2.51 \approx 0.004$.

Posterior variance: $1/2.51 \approx 0.398$. Compare to sampling variance $\sigma^2/n = 0.4$. The prior contributes almost no precision.

Cross-Links

- **Linear Algebra:** The posterior mean in multivariate normal models is a solution to a linear system involving precision matrices: $\hat{\mu}^{\text{Bayes}} = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})$. This is a Tikhonov-regularized least squares problem.
- **AI:** MAP estimation under structured priors produces regularized learning: L2 prior gives weight decay, L1 prior gives sparsity, and the full posterior mean gives Bayesian model averaging. Deep learning approximations to the posterior mean include MC dropout and deep ensembles.
- **Economics:** In rational expectations models, agents form point estimates of unknown parameters (productivity shocks, inflation rates) by computing posterior means under squared error loss, producing optimal forecasts given their information sets.
- **Quantum Mechanics:** Quantum state estimation uses Bayesian point estimators to reconstruct density matrices from measurement data. The posterior mean over density matrices (under the Hilbert-Schmidt metric) is the Bayes estimator under Frobenius loss.

Structural Compression

Invariant: The Bayes estimator minimizes posterior expected loss. The specific functional of the posterior — mean, median, or mode — is uniquely determined by the loss geometry.

Mechanism: Differentiate the posterior expected loss with respect to the action. The first-order condition selects the posterior summary matching the loss function. Squared error yields the expectation, absolute error yields the median, zero-one loss yields the mode.

Cross-System Echo: MAP estimation under Gaussian prior is ridge regression; under Laplace prior, it is the LASSO. The Bayes estimator is the admissible estimator corresponding to the prior π, bridging Bayesian and frequentist admissibility theory.

Exercises

1. Prove that the posterior mean minimizes posterior expected loss under weighted squared error loss $L(\theta, a) = w(\theta)(\theta - a)^2$, and find the resulting estimator in terms of the posterior.
2. For the Gamma-Poisson model with prior $\text{Gamma}(\alpha_0, \beta_0)$ and data $\sum x_i = 45$, $n = 10$, compute the posterior mean, posterior mode, and posterior median. Compare all three to the MLE.

3. Show that the MAP estimator is not invariant under reparametrization: if $\hat{\theta}^{\text{MAP}}$ is the MAP for θ, then $g(\hat{\theta}^{\text{MAP}})$ is generally not the MAP for $\phi = g(\theta)$. Give an explicit example.

4. Prove that the Bayes estimator under squared error loss has smaller Bayes risk than the MLE, for the Beta-Binomial model with prior $\text{Beta}(\alpha, \alpha)$. Under what conditions is the improvement largest?

5. Derive the posterior mean for the multivariate normal model: $X_i \sim \mathcal{N}_k(\mu, \Sigma)$, prior $\mu \sim \mathcal{N}_k(\mu_0, \Sigma_0)$, with Σ known. Express the answer in terms of precision matrices.

6. In the James-Stein setting ($k \geq 3$, $X \sim \mathcal{N}_k(\theta, I)$), show that the posterior mean under the prior $\theta \sim \mathcal{N}_k(0, \tau^2 I)$ produces a shrinkage estimator of the form $(1 - c/\|X\|^2)X$ for appropriate c, and connect this to the inadmissibility of the MLE.

7. Argue, structurally, why the choice of loss function cannot be resolved within the Bayesian framework itself: explain what additional information or judgment is required, and how the decision-theoretic framework (Node 3.1) addresses this gap.

Credible Sets

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.4

Credible sets are the Bayesian analogue of confidence sets, providing posterior probability statements about parameter regions. This node depends on the posterior distribution (Node 6.1) and the frequentist confidence set framework (Node 4.1), which it reinterprets from the Bayesian perspective. It enables posterior predictive model checking (Node 6.5) and connects to asymptotic frequentist-Bayesian reconciliation via the Bernstein–von Mises theorem (Node 6.7).

Formal Definition

Let $\pi(\theta \mid x)$ be the posterior distribution on $\Theta \subseteq \mathbb{R}^k$.

Credible Set

A set $C(x) \subseteq \Theta$ is a $(1-\alpha)$**-credible set** if

$$\int_{C(x)} \pi(\theta \mid x)\, d\theta \geq 1 - \alpha.$$

The probability is with respect to the posterior. The parameter θ is the random quantity; the data x are fixed.

Highest Posterior Density (HPD) Region

The **HPD region** at level $1-\alpha$ is

$$C^{\text{HPD}}(x) = \big\{\theta \in \Theta : \pi(\theta \mid x) \geq k_\alpha\big\},$$

where k_α is the largest constant such that

$$\int_{C^{\text{HPD}}(x)} \pi(\theta \mid x)\, d\theta = 1 - \alpha.$$

Equal-Tailed Credible Interval

For scalar θ, the **equal-tailed** $(1-\alpha)$**-credible interval** is

$$C^{\text{ET}}(x) = \big[Q_{\alpha/2},\ Q_{1-\alpha/2}\big],$$

where $Q_p = Q_p(\pi(\cdot \mid x))$ denotes the p-quantile of the posterior distribution.

Intuition

A credible set directly answers: given the observed data and the model, what region contains θ with posterior probability at least $1-\alpha$?

The HPD region collects the most probable parameter values first, so it is the shortest interval (smallest volume region) achieving the desired posterior coverage. For symmetric unimodal posteriors, the HPD and equal-tailed intervals coincide. For skewed posteriors, the HPD region is shorter but asymmetric, while the equal-tailed interval is symmetric in probability but longer.

The credible set interpretation is conceptually simpler than the frequentist confidence set: it conditions on the observed data and makes a probability statement about the parameter. The price is dependence on the prior.

Structural Role

Credible sets are the canonical Bayesian uncertainty quantification tool for parameters. They require the posterior (Node 6.1) and parallel the frequentist confidence sets (Node 4.1).

The HPD region connects to Bayesian point estimation (Node 6.3): the MAP estimator is always contained in the HPD region, and as $\alpha \to 1$ the HPD region shrinks to the MAP.

The Bernstein–von Mises theorem (Node 6.7) establishes that in regular parametric models, $(1-\alpha)$-credible sets have asymptotic frequentist coverage $1-\alpha$, reconciling the two frameworks.

Mathematical Development

Proof: HPD is the Shortest Credible Interval

Theorem. Among all $(1-\alpha)$-credible sets in $\mathbb{R}$, the HPD region has the smallest Lebesgue measure.

Proof. Let C be any $(1-\alpha)$-credible set with $\int_C \pi(\theta \mid x)\, d\theta \geq 1-\alpha$. Let C^{HPD} be the HPD region with threshold k_α.

Consider the symmetric difference. For any $\theta_1 \in C^{\text{HPD}} \setminus C$ and $\theta_2 \in C \setminus C^{\text{HPD}}$, we have $\pi(\theta_1 \mid x) \geq k_\alpha > \pi(\theta_2 \mid x)$.

The posterior mass of $C \setminus C^{\text{HPD}}$ must be at least that of $C^{\text{HPD}} \setminus C$ (since both C and C^{HPD} achieve coverage $1-\alpha$). But every point in $C \setminus C^{\text{HPD}}$ has posterior density strictly less than every point in $C^{\text{HPD}} \setminus C$. Therefore

$$\lambda(C \setminus C^{\text{HPD}}) \geq \frac{\int_{C^{\text{HPD}} \setminus C} \pi(\theta \mid x)\, d\theta}{k'_\alpha} \geq \lambda(C^{\text{HPD}} \setminus C),$$

where $k'_\alpha \leq k_\alpha$ bounds the density on $C \setminus C^{\text{HPD}}$ from above. Thus $\lambda(C) = \lambda(C \cap C^{\text{HPD}}) + \lambda(C \setminus C^{\text{HPD}}) \geq \lambda(C^{\text{HPD}})$. $\square$

Formal Distinction from Confidence Sets

A frequentist $(1-\alpha)$-confidence set $C_{\text{freq}}(X)$ satisfies

$$\mathbb{P}_\theta\big(\theta \in C_{\text{freq}}(X)\big) \geq 1-\alpha \quad \text{for all } \theta \in \Theta.$$

Here θ is fixed and X varies. The probability is over repetitions of the experiment.

A Bayesian $(1-\alpha)$-credible set $C(x)$ satisfies

$$\pi\big(\theta \in C(x) \mid X = x\big) \geq 1 - \alpha.$$

Here x is fixed and θ varies under the posterior. The probability is conditional on the observed data.

These are fundamentally different probability statements. A credible set with good posterior coverage need not have good frequentist coverage, and vice versa.

When Credible Sets Have Frequentist Coverage

Under the Bernstein–von Mises theorem (Node 6.7), for regular parametric models with fixed dimension:

$$\pi(\theta \mid X_1, \ldots, X_n) \approx \mathcal{N}\left(\hat{\theta}_n,\ \frac{I(\theta_0)^{-1}}{n}\right)$$

in total variation. The HPD credible region based on this posterior is asymptotically

$$C_n^{\mathrm{HPD}} \approx \big\{\theta : n(\theta - \hat{\theta}_n)^\top I(\theta_0)(\theta - \hat{\theta}_n) \leq \chi^2_{k,1-\alpha}\big\},$$

which coincides with the Wald confidence ellipsoid. Therefore $\mathbb{P}_{\theta_0}(\theta_0 \in C_n^{\mathrm{HPD}}) \to 1 - \alpha$.

HPD for Common Posterior Families

Normal posterior: If $\theta \mid x \sim \mathcal{N}(\mu_n, \sigma_n^2)$, the HPD interval is symmetric about μ_n:

$$C^{\mathrm{HPD}} = \big[\mu_n - z_{1-\alpha/2}\,\sigma_n,\ \mu_n + z_{1-\alpha/2}\,\sigma_n\big].$$

This coincides with the equal-tailed interval by symmetry.

Beta posterior: If $\theta \mid x \sim \mathrm{Beta}(\alpha_n, \beta_n)$ with $\alpha_n \neq \beta_n$, the HPD interval is shorter than the equal-tailed interval. It must be computed numerically by finding k_α such that the density level set has posterior mass $1 - \alpha$.

Gamma posterior: Similarly asymmetric; the HPD interval shifts toward smaller values relative to the equal-tailed interval due to right skewness.

Multivariate HPD Regions

For $\theta \in \mathbb{R}^k$, the HPD region is a level set of the joint posterior density. For a multivariate normal posterior $\theta \mid x \sim \mathcal{N}_k(\mu_n, \Sigma_n)$, the HPD region is the ellipsoid

$$C^{\mathrm{HPD}} = \big\{\theta : (\theta - \mu_n)^\top \Sigma_n^{-1}(\theta - \mu_n) \leq \chi^2_{k,1-\alpha}\big\}.$$

Worked Examples

Example 1: HPD Interval for a Beta Posterior

Data: $n = 30$ Bernoulli trials, $s = 21$ successes. Prior: $\mathrm{Beta}(1, 1)$ (uniform).

Posterior: $\mathrm{Beta}(22, 10)$. Posterior mean: $22/32 = 0.6875$. Posterior mode: $21/30 = 0.7$.

Equal-tailed 95% interval: Using Beta quantiles, $[Q_{0.025}, Q_{0.975}] \approx [0.518, 0.832]$. Length: 0.314.

HPD 95% interval: The density level set $\{\theta : f(\theta) \geq k\}$ must be found numerically. For $\mathrm{Beta}(22, 10)$: $C^{\mathrm{HPD}} \approx [0.530, 0.840]$. Length: 0.310.

The HPD interval is shifted slightly toward the mode and is shorter than the equal-tailed interval due to the right skewness of Beta$(22, 10)$.

Example 2: Normal Posterior Credible Interval vs Confidence Interval

Data: $n = 25$, $\bar{x} = 5.2$, $\sigma^2 = 9$ (known). Prior: $\mu \sim \mathcal{N}(4, 4)$.

Posterior: $\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2)$ with

$$\tau_n^2 = \frac{1}{1/4 + 25/9} = \frac{1}{0.25 + 2.778} = \frac{1}{3.028} \approx 0.330,$$

$$\mu_n = 0.330 \cdot (0.25 \cdot 4 + 2.778 \cdot 5.2) = 0.330 \cdot (1 + 14.444) = 0.330 \cdot 15.444 \approx 5.097.$$

95% credible interval: $5.097 \pm 1.96\sqrt{0.330} = 5.097 \pm 1.125 = [3.972, 6.222]$.

95% confidence interval (frequentist, no prior): $5.2 \pm 1.96 \cdot 3/5 = 5.2 \pm 1.176 = [4.024, 6.376]$.

The credible interval is shorter (width 2.25 vs 2.35) and shifted toward the prior mean of 4. As $n \to \infty$, the two intervals converge.

Cross-Links

- **Linear Algebra:** Multivariate HPD regions are ellipsoids defined by the posterior precision matrix. The eigenvectors of the precision matrix give the principal axes of the credible ellipsoid, connecting to PCA-like decompositions of posterior uncertainty.
- **AI:** Bayesian neural networks produce posterior distributions over weights; credible sets for predictions quantify epistemic uncertainty. Approximate HPD regions are computed via MCMC samples or variational posteriors.
- **Economics:** Bayesian credible intervals for structural parameters (elasticities, policy multipliers) directly quantify parameter uncertainty conditional on observed data, which is the natural object for policy decisions.
- **Quantum Mechanics:** Credible regions for quantum states (density matrices) arise in quantum state tomography. The HPD region over the space of density matrices provides the smallest set of states consistent with measurement data at a given posterior probability level.

Structural Compression

Invariant: $\pi(\theta \in C(x) \mid X = x) \geq 1 - \alpha$. The posterior mass of the credible set is at least $1 - \alpha$.

Mechanism: Integrate the posterior density over the candidate region $C(x)$. The HPD region selects the level set of the posterior density, which minimizes volume subject to the coverage constraint. This is a constrained optimization: minimize Lebesgue measure subject to posterior mass $\geq 1 - \alpha$.

Cross-System Echo: Credible sets are the Bayesian analogue of confidence sets. The Bernstein–von Mises theorem (Node 6.7) establishes their asymptotic equivalence in regular models, making credible sets the bridge between Bayesian and frequentist uncertainty quantification.

Exercises

1. For the posterior $\theta \mid x \sim \text{Gamma}(10, 2)$, compute the equal-tailed 95% credible interval and explain why the HPD interval will be shorter.
2. Prove that for a unimodal symmetric posterior, the HPD interval and the equal-tailed interval coincide.
3. Let $\theta \mid x \sim \text{Beta}(5, 15)$. Compute the posterior mean, posterior mode, and the equal-tailed 90% credible interval. Without computing it exactly, argue whether the HPD interval will shift left or right relative to the equal-tailed interval.
4. Construct an example where a 95% credible interval has frequentist coverage far from 95% for a specific θ_0. Specify the model, prior, and true parameter.
5. For the multivariate normal posterior $\theta \mid x \sim \mathcal{N}_2(\mu_n, \Sigma_n)$ with $\mu_n = (3, 1)^\top$ and $\Sigma_n = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, derive the equation of the 95% HPD ellipse and find its semi-axes.
6. Prove that the MAP estimator is always contained in the HPD region $C^{\text{HPD}}(x)$ for every $\alpha \in (0, 1)$, and that the HPD region shrinks to $\{\hat{\theta}^{\text{MAP}}\}$ as $\alpha \to 1$.
7. Argue, structurally, why the distinction between credible sets and confidence sets is sharpest in small samples with informative priors, and vanishes asymptotically under the conditions of the Bernstein–von Mises theorem: identify the role of the prior, the likelihood dominance rate, and the resulting posterior shape.

Posterior Predictive Distribution

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.5

The posterior predictive distribution is the Bayesian answer to prediction: it integrates parameter uncertainty into the distribution of future observations. This node depends on the posterior (Node 6.1), conjugate families for tractable computation (Node 6.2), and marginalization (Node 1.7). It enables model comparison (Node 6.6) through the marginal likelihood, which is the prior predictive evaluated at the observed data, and provides the foundation for posterior predictive model checking.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p(x \mid \theta)$ with posterior $\pi(\theta \mid x_{1:n})$. Let $\tilde{X}$ denote a future observation from the same model.

Posterior Predictive Distribution

The **posterior predictive distribution** of $\tilde{X}$ given observed data $x_{1:n}$ is

$$p(\tilde{x} \mid x_{1:n}) = \int_\Theta p(\tilde{x} \mid \theta)\, \pi(\theta \mid x_{1:n})\, d\theta.$$

Prior Predictive Distribution

Before data are observed, the **prior predictive distribution** is

$$p(\tilde{x}) = \int_\Theta p(\tilde{x} \mid \theta)\, \pi(\theta)\, d\theta.$$

Marginal Likelihood

The marginal likelihood of the observed data is the prior predictive evaluated at $x_{1:n}$:

$$m(x_{1:n}) = \int_\Theta p(x_{1:n} \mid \theta)\, \pi(\theta)\, d\theta = p(x_{1:n}).$$

This quantity appears in the denominator of Bayes' theorem (Node 6.1) and in Bayes factors (Node 6.6).

Intuition

The posterior predictive averages the sampling model over all parameter values, weighted by their posterior plausibility. It does not commit to a single estimate of θ; instead, it propagates the full posterior uncertainty into the prediction.

Compared to plug-in prediction $p(\tilde{x} \mid \hat{\theta})$, which conditions on a single estimate, the posterior predictive is typically more dispersed. This extra spread reflects parameter uncertainty, which the plug-in approach ignores. The posterior predictive is the honest assessment of predictive uncertainty given the model and data.

The prior predictive distribution has a dual role: it describes the marginal behavior of the data before any conditioning, and its evaluation at the observed data yields the marginal likelihood — the key ingredient in Bayesian model comparison.

Structural Role

The posterior predictive distribution requires the posterior (Node 6.1) and marginalization (Node 1.7). Conjugate families (Node 6.2) provide closed-form solutions in standard cases.

This node enables: Bayes factors (Node 6.6) through the marginal likelihood, and posterior predictive checks for model adequacy. It connects to the decision-theoretic framework: under squared error loss, the optimal point prediction of $\tilde{X}$ is $\mathbb{E}[\tilde{X} \mid x_{1:n}]$, computed from the posterior predictive.

Mathematical Development

Derivation as Marginalization

The joint distribution of $(\tilde{X}, \theta)$ given data is

$$p(\tilde{x}, \theta \mid x_{1:n}) = p(\tilde{x} \mid \theta)\, \pi(\theta \mid x_{1:n}),$$

using the conditional independence $\tilde{X} \perp X_{1:n} \mid \theta$. Marginalizing over θ:

$$p(\tilde{x} \mid x_{1:n}) = \int_\Theta p(\tilde{x}, \theta \mid x_{1:n})\, d\theta = \int_\Theta p(\tilde{x} \mid \theta)\, \pi(\theta \mid x_{1:n})\, d\theta.$$

Posterior Predictive Moments

The mean of the posterior predictive is

$$\mathbb{E}[\tilde{X} \mid x_{1:n}] = \int \tilde{x}\, p(\tilde{x} \mid x_{1:n})\, d\tilde{x} = \mathbb{E}_{\theta \mid x}\big[\mathbb{E}[\tilde{X} \mid \theta]\big] = \mathbb{E}[\mu(\theta) \mid x_{1:n}],$$

where $\mu(\theta) = \mathbb{E}[\tilde{X} \mid \theta]$. By the law of total variance:

$$\mathrm{Var}(\tilde{X} \mid x_{1:n}) = \mathbb{E}_{\theta \mid x}\big[\mathrm{Var}(\tilde{X} \mid \theta)\big] + \mathrm{Var}_{\theta \mid x}\big[\mathbb{E}[\tilde{X} \mid \theta]\big].$$

The first term is the average sampling variance (aleatoric uncertainty). The second term is the posterior variance of the conditional mean (epistemic uncertainty). The posterior predictive variance is always at least as large as the plug-in predictive variance $\mathrm{Var}(\tilde{X} \mid \hat{\theta})$.

Conjugate Posterior Predictives

Beta-Binomial: Let $X_i \mid \theta \sim \mathrm{Bernoulli}(\theta)$, $\pi(\theta) = \mathrm{Beta}(\alpha_0, \beta_0)$. After s successes in n trials, the posterior is $\mathrm{Beta}(\alpha_n, \beta_n)$ with $\alpha_n = \alpha_0 + s$, $\beta_n = \beta_0 + n - s$. The posterior predictive for a single new trial is

$$p(\tilde{X} = 1 \mid x_{1:n}) = \mathbb{E}[\theta \mid x_{1:n}] = \frac{\alpha_n}{\alpha_n + \beta_n}.$$

For m future trials, the number of successes $\tilde{S}$ follows a Beta-Binomial distribution:

$$p(\tilde{S} = k \mid x_{1:n}) = \binom{m}{k} \frac{B(\alpha_n + k,\ \beta_n + m - k)}{B(\alpha_n, \beta_n)}, \quad k = 0, 1, \ldots, m.$$

Normal-Normal: Let $X_i \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2, and posterior $\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2)$. The posterior predictive is

$$\tilde{X} \mid x_{1:n} \sim \mathcal{N}(\mu_n,\ \sigma^2 + \tau_n^2).$$

The predictive variance $\sigma^2 + \tau_n^2$ exceeds the sampling variance σ^2 by the posterior uncertainty τ_n^2. As $n \to \infty$, $\tau_n^2 \to 0$ and the predictive distribution approaches the plug-in distribution.

Gamma-Poisson: Let $X_i \mid \lambda \sim \text{Poisson}(\lambda)$, posterior $\lambda \mid x \sim \text{Gamma}(\alpha_n, \beta_n)$. The posterior predictive for a single new observation is Negative Binomial:

$$\tilde{X} \mid x_{1:n} \sim \text{NegBin}\left(\alpha_n,\ \frac{\beta_n}{\beta_n + 1}\right).$$

Posterior Predictive Checks

A **posterior predictive check** assesses model adequacy by comparing observed data summaries to their distribution under the posterior predictive. Define a test statistic $T(x)$. Generate replicated data $x^{\text{rep}} \sim p(x^{\text{rep}} \mid x_{1:n})$ by:

1. Draw $\theta^{(j)} \sim \pi(\theta \mid x_{1:n})$.
2. Draw $x^{\text{rep},(j)} \sim p(x \mid \theta^{(j)})$.

The **posterior predictive p-value** is

$$p_B = \Pr\big(T(x^{\text{rep}}) \geq T(x_{\text{obs}}) \mid x_{1:n}\big).$$

Values near 0 or 1 indicate model inadequacy: the observed summary is extreme relative to what the fitted model predicts.

Marginal Likelihood via Sequential Prediction

The marginal likelihood decomposes as a product of one-step-ahead predictive densities:

$$m(x_{1:n}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1}),$$

where $p(x_1 \mid x_0) = p(x_1)$ is the prior predictive. This decomposition is the **prequential** form of the marginal likelihood and connects to information-theoretic model evaluation.

Worked Examples

Example 1: Normal Posterior Predictive

Data: $n = 16$, $\bar{x} = 50$, $\sigma^2 = 64$ (known). Prior: $\mu \sim \mathcal{N}(45, 16)$.

Posterior: $\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2)$.

$$\tau_n^2 = \frac{1}{1/16 + 16/64} = \frac{1}{0.0625 + 0.25} = \frac{1}{0.3125} = 3.2.$$

$$\mu_n = 3.2 \cdot (0.0625 \cdot 45 + 0.25 \cdot 50) = 3.2 \cdot (2.8125 + 12.5) = 3.2 \cdot 15.3125 = 49.0.$$

Posterior predictive: $\tilde{X} \mid x \sim \mathcal{N}(49.0,\ 64 + 3.2) = \mathcal{N}(49.0,\ 67.2)$.

Predictive standard deviation: $\sqrt{67.2} \approx 8.20$.

A 95% posterior predictive interval for a new observation: $49.0 \pm 1.96 \cdot 8.20 = [32.93,\ 65.07]$.

Compare to plug-in prediction at MLE $\hat{\mu} = 50$: $\mathcal{N}(50, 64)$, giving interval $[34.32, 65.68]$. The Bayesian interval is centered slightly differently and accounts for parameter uncertainty.

Example 2: Beta-Binomial Posterior Predictive

Data: $n = 50$, $s = 30$. Prior: Beta$(1, 1)$.

Posterior: Beta$(31, 21)$. Posterior mean: $31/52 \approx 0.596$.

Posterior predictive probability of success on the next trial: $31/52 \approx 0.596$.

For $m = 10$ future trials, the posterior predictive distribution of the number of successes is Beta-Binomial$(10, 31, 21)$. The predictive mean is $10 \cdot 31/52 \approx 5.96$. The predictive variance is

$$\mathrm{Var}(\tilde{S}) = m \cdot \frac{\alpha_n \beta_n (\alpha_n + \beta_n + m)}{(\alpha_n + \beta_n)^2 (\alpha_n + \beta_n + 1)} = 10 \cdot \frac{31 \cdot 21 \cdot 62}{52^2 \cdot 53} \approx 2.81.$$

Compare to the binomial variance at the plug-in $\hat{\theta} = 0.6$: $10 \cdot 0.6 \cdot 0.4 = 2.4$. The posterior predictive variance is larger, reflecting parameter uncertainty.

Cross-Links

- **Linear Algebra:** In Gaussian linear models, the posterior predictive at a new input $\tilde{x}$ is a quadratic form involving the posterior covariance of the coefficient vector, connecting matrix algebra of precision updates to predictive uncertainty.
- **AI:** Gaussian process regression computes the posterior predictive distribution at new inputs in closed form — it is the canonical example of an exact posterior predictive under a nonparametric prior. Bayesian neural networks approximate the posterior predictive via MC dropout or ensembles.
- **Economics:** The posterior predictive distribution is the natural object for Bayesian forecasting: given observed economic data, it provides the full distribution of future GDP, inflation, or asset returns, incorporating parameter uncertainty from the structural model.
- **Quantum Mechanics:** In quantum Bayesian inference (QBism), the predictive distribution for a future measurement outcome given past measurement results is obtained by marginalizing over the posterior distribution of the quantum state, exactly paralleling the posterior predictive.

Structural Compression

Invariant: $p(\tilde{x} \mid x_{1:n}) = \int p(\tilde{x} \mid \theta)\, \pi(\theta \mid x_{1:n})\, d\theta$. Prediction marginalizes over posterior uncertainty in the parameter.

Mechanism: The posterior reweights the parameter space according to data compatibility. The predictive integrates the sampling model against this reweighted measure, producing a distribution over new observations that is strictly more dispersed than the plug-in predictive, with the excess variance equal to the posterior variance of the conditional mean.

Cross-System Echo: The posterior predictive is the statistical instance of Bayesian model averaging. In ensemble methods, the averaged prediction across models implements the same marginalization. The marginal likelihood decomposition into sequential one-step predictives connects to online learning and prequential analysis.

Exercises

1. Derive the posterior predictive distribution for the Gamma-Poisson model: show that the posterior predictive for a single new observation is Negative Binomial, and identify its parameters in terms of the posterior hyperparameters.
2. For the Normal-Normal model, prove that the posterior predictive variance equals $\sigma^2 + \tau_n^2$ and decompose this into aleatoric and epistemic components. Show that the epistemic component vanishes as $n \to \infty$.
3. Compute the marginal likelihood $m(x_{1:n})$ for the Beta-Binomial model in closed form using the Beta function. Verify that it equals the product of sequential one-step predictive densities.
4. Design a posterior predictive check for the Poisson model: specify a test statistic $T(x)$ that would detect overdispersion, and describe the procedure for computing the posterior predictive p-value via simulation.
5. Show that the posterior predictive mean $\mathbb{E}[\tilde{X} \mid x_{1:n}]$ equals the posterior mean of $\mathbb{E}[\tilde{X} \mid \theta]$. Under what conditions does the posterior predictive median equal the posterior median of θ?
6. For the Normal-Normal model with prior $\mu \sim \mathcal{N}(0, \tau_0^2)$ and known σ^2, compute the marginal likelihood $m(x_{1:n})$ in closed form and show it is proportional to $\exp\left(-\frac{n\bar{x}^2}{2(\sigma^2 + n\tau_0^2)}\right)$.
7. Argue, structurally, why the posterior predictive distribution is the unique coherent predictive distribution given the model and data: identify the assumptions (conditional independence, coherent updating) that force the marginalization over the posterior, and explain why plug-in prediction violates coherence.

Bayes Factors

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

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

Lindley's Paradox

Consider testing $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ with $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ 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

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

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

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 $\text{BF}_{10} = m_1/m_0$. Taking the log ratio:

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

Numerical evaluation gives $\log \text{BF}_{10} \approx -1.8$, so $\text{BF}_{10} \approx 0.17$. 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: $\text{BF}_{01} = 0.00139/0.1262 \approx 0.011$.

Thus $\text{BF}_{10} \approx 91$: 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.

Bernstein–von Mises Theorem

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.7

The Bernstein–von Mises theorem is the fundamental asymptotic bridge between the Bayesian and frequentist frameworks. It establishes that in regular parametric models, the posterior distribution is asymptotically Gaussian, centered at the MLE, with variance equal to the inverse Fisher information — regardless of the prior. This node depends on the posterior (Node 6.1), the asymptotic theory of the MLE (Node 3.3), and the Fisher information (Node 7.2). It is the terminal node of Module 6: it closes the loop by showing that Bayesian and frequentist inference converge in large samples.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$ where $\theta_0 \in \Theta \subseteq \mathbb{R}^k$ is the true parameter value. Let $\hat{\theta}_n$ denote the MLE and $I(\theta_0)$ the Fisher information matrix at θ_0.

Regularity Conditions

1. **Identifiability:** $\theta \neq \theta' \implies p_\theta \neq p_{\theta'}$.
2. **Smoothness:** The log-likelihood $\ell(\theta) = \log p(x \mid \theta)$ is twice continuously differentiable in a neighborhood of θ_0.
3. **Fisher information:** $I(\theta_0) = -\mathbb{E}_{\theta_0}[\nabla^2 \ell(\theta_0)]$ is finite and positive definite.
4. **Prior positivity:** π is continuous and positive at θ_0: $\pi(\theta_0) > 0$.
5. **Consistency:** There exists a consistent sequence of estimators (e.g., the MLE $\hat{\theta}_n \to \theta_0$ in probability).
6. **Local asymptotic normality (LAN):** The log-likelihood ratio admits the expansion $\ell_n(\theta_0 + h/\sqrt{n}) - \ell_n(\theta_0) = h^\top \Delta_n - \frac{1}{2} h^\top I(\theta_0) h + o_P(1)$, where $\Delta_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$.

Theorem Statement

Under the regularity conditions above:

$$\left\| \pi\Big(\sqrt{n}(\theta - \hat{\theta}_n) \in \cdot \;\Big|\; X_1, \ldots, X_n\Big) - \mathcal{N}\big(0,\ I(\theta_0)^{-1}\big) \right\|_{\text{TV}} \xrightarrow{P^n_{\theta_0}} 0.$$

Equivalently, the posterior distribution of θ given the data satisfies

$$\pi(\theta \mid X_1, \ldots, X_n) \approx \mathcal{N}\left(\hat{\theta}_n,\ \frac{I(\theta_0)^{-1}}{n}\right)$$

in the sense of total variation convergence under the true data-generating distribution.

Intuition

As data accumulate, the likelihood dominates the prior. The log-posterior is approximately

$$\log \pi(\theta \mid x) \approx \ell_n(\theta) + \log \pi(\theta) - \text{const},$$

where $\ell_n(\theta) = \sum_{i=1}^{n} \log p(x_i \mid \theta)$ grows as $O(n)$ while $\log \pi(\theta)$ remains $O(1)$. A Taylor expansion of ℓ_n around the MLE produces a quadratic in θ, which is the kernel of a Gaussian.

The theorem says the prior washes out: any two investigators with different priors (both positive and continuous at θ_0) will have posteriors that converge to the same Gaussian as $n \to \infty$. This is the Bayesian analogue of the frequentist result that the MLE is asymptotically normal.

Structural Role

The Bernstein–von Mises theorem is the asymptotic reconciliation of Bayesian and frequentist inference. It depends on the posterior (Node 6.1), asymptotic MLE theory (Node 3.3), and Fisher information (Node 7.2).

It has consequences for every other node in Module 6: - **Node 6.3:** The posterior mean, MAP, and posterior median all converge to the MLE at rate $O_p(n^{-1/2})$. - **Node 6.4:** Credible sets have asymptotic frequentist coverage. - **Node 6.6:** The BIC approximation to the Bayes factor is derived from the same Laplace expansion.

The theorem fails in nonparametric models, misspecified models, and high-dimensional settings, delineating the boundary of Bayesian-frequentist agreement.

Mathematical Development

Proof Sketch

The argument proceeds in four steps.

Step 1: Posterior concentration. By the consistency of the MLE and the positivity of the prior at θ_0, the posterior mass outside any ball $B(\theta_0, \varepsilon)$ converges to zero in probability:

$$\pi\big(\|\theta - \theta_0\| > \varepsilon \mid X_{1:n}\big) \xrightarrow{P} 0.$$

This follows from the exponential growth of the likelihood ratio at the true value relative to distant parameter values.

Step 2: Log-posterior expansion. Expand the log-likelihood around the MLE:

$$\ell_n(\theta) = \ell_n(\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^\top \hat{J}_n (\theta - \hat{\theta}_n) + R_n(\theta),$$

where $\hat{J}_n = -\nabla^2 \ell_n(\hat{\theta}_n)$ is the observed information and $R_n(\theta)$ is a remainder. By the law of large numbers, $\hat{J}_n / n \to I(\theta_0)$ in probability.

Step 3: Prior contribution. On the concentration set $\|\theta - \hat{\theta}_n\| = O(n^{-1/2})$, the prior satisfies $\log \pi(\theta) = \log \pi(\hat{\theta}_n) + O(n^{-1/2})$. Since $\log \pi$ is $O(1)$ and the quadratic term is $O(n)$, the prior contributes a negligible additive correction.

Step 4: Gaussian identification. Combining Steps 2 and 3:

$$\log \pi(\theta \mid x) = -\frac{1}{2}(\theta - \hat{\theta}_n)^\top \hat{J}_n (\theta - \hat{\theta}_n) + o_P(1)$$

on the set where $\sqrt{n}(\theta - \hat{\theta}_n)$ is bounded. This is the log-density of $\mathcal{N}(\hat{\theta}_n, \hat{J}_n^{-1})$. Since $\hat{J}_n / n \to I(\theta_0)$, the posterior is asymptotically $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$ in total variation. $\square$

Consequence: Posterior Mean and MLE Agreement

Since the posterior is approximately $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$, its mean is

$$\mathbb{E}[\theta \mid X_{1:n}] = \hat{\theta}_n + O_P(n^{-1}).$$

The posterior mean and the MLE agree to first order. Their difference is $O_P(n^{-1})$, a second-order correction driven by the prior and the curvature of the likelihood.

Consequence: Frequentist Coverage of Credible Sets

Let $C_n(x)$ be any $(1-\alpha)$-credible set derived from the posterior. By the total variation convergence:

$$\pi(\theta \in C_n \mid x) \to 1-\alpha \implies \Phi\left(\frac{\sqrt{n}\, I(\theta_0)^{1/2}(\theta_0 - \hat{\theta}_n)}{\text{boundary}}\right) \to 1-\alpha.$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$ under P_{θ_0}, the event $\theta_0 \in C_n(X)$ has probability converging to $1-\alpha$:

$$P_{\theta_0}\big(\theta_0 \in C_n(X)\big) \to 1-\alpha.$$

Asymptotically, Bayesian credible sets are frequentist confidence sets.

Consequence: HPD and Wald Interval Agreement

The HPD credible set based on the limiting Gaussian posterior is the ellipsoid

$$\big\{\theta : n(\theta - \hat{\theta}_n)^\top I(\theta_0)(\theta - \hat{\theta}_n) \le \chi^2_{k,1-\alpha}\big\},$$

which coincides with the Wald confidence ellipsoid based on the MLE.

Failure Modes

Nonparametric models. When Θ is infinite-dimensional (e.g., density estimation), the posterior can concentrate at rate slower than $1/\sqrt{n}$, and the Gaussian approximation fails. The Diaconis-Freedman theorem provides examples where the posterior concentrates on the wrong value.

Misspecified models. If the true distribution $P_0 \notin \{P_\theta : \theta \in \Theta\}$, the posterior concentrates around the KL-minimizer $\theta^* = \arg\min_\theta \mathrm{KL}(P_0 \| P_\theta)$, but the variance is no longer $I(\theta^*)^{-1}/n$ — a sandwich-type correction is needed.

High-dimensional models. When $k = k_n \to \infty$ with n, the prior contribution does not vanish relative to the likelihood, and the Bernstein–von Mises conclusion can fail even for parametric models.

Boundary parameters. If θ_0 lies on the boundary of Θ, the posterior may not be asymptotically normal (it may concentrate on a half-space).

Worked Examples

Example 1: Normal Mean

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\mu \sim \mathcal{N}(0, \tau^2)$. The exact posterior is $\mu \mid x \sim \mathcal{N}(\mu_n, \sigma_n^2)$ with

$$\sigma_n^2 = \frac{1}{1/\tau^2 + n} = \frac{\tau^2}{1 + n\tau^2}, \qquad \mu_n = \sigma_n^2 \cdot n\bar{x} = \frac{n\tau^2}{1 + n\tau^2}\bar{x}.$$

As $n \to \infty$: $\sigma_n^2 \to 1/n$ and $\mu_n \to \bar{x} = \hat{\mu}_n$. The Fisher information is $I(\mu) = 1$, so the BvM limit is $\mathcal{N}(\hat{\mu}_n, 1/n)$. The exact posterior converges to this limit, confirming the theorem. The convergence rate of the prior correction is $O(1/n)$: the term $1/\tau^2$ in the posterior precision becomes negligible relative to n.

Example 2: Bernoulli Model and Coverage Verification

Let $X_i \sim \text{Bernoulli}(\theta_0)$ with $\theta_0 = 0.3$. Prior: Beta$(2, 2)$. For large n, the posterior is approximately

$$\theta \mid x \dot{\sim} \mathcal{N}\left(\hat{\theta}_n, \frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}\right),$$

since $I(\theta) = 1/[\theta(1 - \theta)]$.

For $n = 500$: if $s = 155$ successes, then $\hat{\theta}_n = 0.31$. The asymptotic posterior variance is $0.31 \cdot 0.69/500 \approx 0.000428$. The 95% HPD interval is approximately $0.31 \pm 1.96\sqrt{0.000428} = 0.31 \pm 0.041 = [0.269, 0.351]$.

The exact posterior is Beta$(157, 347)$ with mean $157/504 \approx 0.3115$ and variance $157 \cdot 347/(504^2 \cdot 505) \approx 0.000429$. The Gaussian approximation is excellent: the prior contributes only 4 pseudo-observations out of 504 total.

The frequentist coverage: $\theta_0 = 0.3 \in [0.269, 0.351]$, and repeated sampling confirms coverage near 95%.

Cross-Links

- **Linear Algebra:** The quadratic expansion of the log-likelihood around the MLE is governed by the Hessian matrix $\hat{J}_n$. The Bernstein–von Mises theorem requires this matrix to converge (after scaling by $1/n$) to the positive definite Fisher information matrix, a spectral condition.
- **AI:** The Bernstein–von Mises theorem motivates Laplace approximations for Bayesian deep learning: approximate the posterior by a Gaussian centered at the MAP estimate with covariance given by the inverse Hessian. The theorem's failure in high dimensions explains why this approximation is unreliable for overparameterized networks.
- **Economics:** The theorem justifies the common econometric practice of interpreting Bayesian credible intervals as confidence intervals in large samples. It also explains why prior choice matters less in large macro panels than in small cross-sections.
- **Quantum Mechanics:** Quantum local asymptotic normality (quantum LAN) extends the Bernstein–von Mises theorem to quantum statistical models, showing that the posterior over quantum states converges to a Gaussian shift experiment in the limit of many copies of the state.

Structural Compression

Invariant: In regular parametric models, the large-sample posterior is asymptotically $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$, independent of the prior.

Mechanism: A second-order Taylor expansion of the log-likelihood around the MLE produces an $O(n)$ quadratic that dominates the $O(1)$ prior contribution. The resulting Gaussian shape is determined entirely by the Fisher information geometry. The prior survives only in the $O(n^{-1})$ correction to the posterior mean.

Cross-System Echo: The Bernstein–von Mises theorem is the statistical manifestation of the Laplace approximation to integrals dominated by a sharp peak. In physics, the saddle-point approximation to partition

functions is the same phenomenon. In AI, the Laplace approximation to the neural network posterior and the correspondence between MAP and MLE in the large-data regime are direct applications of this asymptotic equivalence.

Exercises

1. Verify the Bernstein–von Mises conclusion for the Poisson model: let $X_i \sim \text{Poisson}(\lambda_0)$ with prior $\text{Gamma}(\alpha, \beta)$. Show that the exact posterior is Gamma and that it converges in total variation to $\mathcal{N}(\hat{\lambda}_n, \hat{\lambda}_n/n)$ as $n \to \infty$.
2. Show that the Bernstein–von Mises conclusion implies the posterior variance satisfies $\text{Var}(\theta \mid X) \to I(\theta_0)^{-1}/n$ in probability. Derive the rate of convergence of the difference.
3. Construct an explicit example where the Bernstein–von Mises theorem fails: take a uniform prior on a misspecified model where the true density is not in the parametric family, and show the posterior concentrates at the KL-minimizer but with incorrect variance.
4. For the normal location model with known variance and prior $\mathcal{N}(\mu_0, \tau_0^2)$, compute the exact total variation distance between the posterior and the BvM Gaussian limit as a function of n, and verify it is $O(n^{-1})$.
5. Prove that the second-order correction to the posterior mean is $\mathbb{E}[\theta \mid x] - \hat{\theta}_n = O_P(n^{-1})$ by expanding the posterior mean in powers of $1/n$ using the exact posterior for a conjugate exponential family.
6. Explain why the Bernstein–von Mises theorem fails in the model $X_i \sim \text{Uniform}(0, \theta)$: identify which regularity condition is violated, and describe the actual asymptotic behavior of the posterior.
7. Argue, structurally, why the Bernstein–von Mises theorem establishes a hierarchy between Bayesian and frequentist inference: in the regular parametric setting, the Bayesian approach subsumes the frequentist one (credible sets are asymptotically confidence sets, posterior mean agrees with MLE), but the theorem's failure modes in nonparametric, misspecified, and high-dimensional settings show where this subsumption breaks down and the two frameworks genuinely diverge.