Posterior Predictive Distribution

Statistics · Module 6 — Bayesian Inference

Node 6.5

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

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

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

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

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 $X_i \mid \theta \sim \text{Bernoulli}(\theta)$, $\pi(\theta) = \text{Beta}(\alpha_0, \beta_0)$. After s successes in n trials, the posterior is $\text{Beta}(\alpha_n, \beta_n)$ 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.

Statistics · Module 6 — Bayesian Inference · Node 6.5
Concepts: prior-posterior, bayesian-updating, conditional-probability
Prerequisites: 6.1, 6.2, 1.7
Enables: 6.6

hbar.university — own.your.cognition