Conjugate Families

Statistics · Module 6 — Bayesian Inference

Node 6.2

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. $\square$

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 $\text{Dirichlet}(1, 1, 1)$, 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.

Statistics · Module 6 — Bayesian Inference · Node 6.2
Concepts: prior-posterior, bayesian-updating, exponential-families
Prerequisites: 6.1, 2.6
Enables: 6.3, 6.4

hbar.university — own.your.cognition