Bayesian Point Estimation

Statistics · Module 6 — Bayesian Inference

Node 6.3

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. $\square$

MAP Estimator and Its Decision-Theoretic Justification

For a continuous parameter space, the MAP estimator is

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

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

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

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 $r(\pi, \hat{\theta}^{\mathrm{Bayes}}) < \infty$, then $\hat{\theta}^{\mathrm{Bayes}}$ is admissible.

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

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

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

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.

Statistics · Module 6 — Bayesian Inference · Node 6.3
Concepts: prior-posterior, bayesian-updating, cost-function
Prerequisites: 6.1, 3.1
Enables: 6.5

hbar.university — own.your.cognition