Lecture 2: Basic Bayesian calculus

Thibault Randrianarisoa

UTSC

January 15, 2026

Computer & Mathematical Sciences
UNIVERSITY OF TORONTO
SCARBOROUGH

- Frequentist vs. Bayesian

- Second part: Prior choice

Frequentist vs. Bayesian?

Frequentist approach: basic elements

Setup

- Data: $X_1, \ldots, X_n$ are viewed as random variables, generated i.i.d. from a distribution P_{θ_0}.
- Parameter: θ_0 is an *unknown but fixed* quantity (no probability distribution on θ_0).
- Randomness comes <u>only</u> from the sampling of the data.
- Probability is seen as the limit of the frequency of an event if I repeat an experiment indefinitely.

Main inferential tasks

- **Estimation:** construct an estimator $\hat{\theta}(X)$ with good long-run properties (bias, variance, risk, asymptotic normality).
- **Confidence sets:** build random sets $\mathcal{R}(X)$ such that $\mathbb{P}_\theta(\theta \in \mathcal{R}(X)) \approx 1 - \alpha$.
- **Hypothesis tests:** design tests $\varphi(X) \in \{0, 1\}$ with controlled type I error and good power.
- **Prediction:** predict a future observation X_{n+1} using $f(X_{n+1} \mid X_1, \ldots, X_n, \hat{\theta}_n)$.

Some drawbacks of the frequentist approach

1 Practical issues with small samples
  - Asymptotic theory may no longer be reliable for small n.
  - Comparison of estimators must use non-asymptotic criteria; many tools based on convergence in distribution (e.g. asymptotic confidence regions, test statistics) can become unusable.

2 Tension with the likelihood principle
  The likelihood principle says that all information about θ in an observation x is contained in the likelihood $L_\theta(X) = p_\theta(X)$.
  If two observations x_1, x_2 satisfy

  $$L_\theta(x_1) = c\, L_\theta(x_2) \quad \forall \theta,$$

  they should lead to the same inference.
  Frequentist procedures can violate this, because they may depend on other aspects beyond the likelihood.

Some drawbacks of the frequentist approach

❸ Maximum likelihood and prediction
- The MLE, often viewed as "most efficient", may fail to exist or be non-unique in some models.
- For prediction, the classical plug-in density

$$p_{\hat{\theta}_n}(X_{n+1} \mid X_1, \ldots, X_n) = \frac{p_{\hat{\theta}_n}(X_1, \ldots, X_n, X_{n+1})}{p_{\hat{\theta}_n}(X_1, \ldots, X_n)}$$

uses the data twice (to estimate θ and to condition), which can underestimate uncertainty (too narrow confidence intervals, overconfident forecasts).

Bayesian statistical framework

Statistical experiment

- We observe a random object X taking values in a measurable space $(E, \mathcal{E})$ (like $\mathbb{R}^p$).
- The distribution of X is assumed to belong to a parametric model

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where the parameter space satisfies $\Theta \subset \mathbb{R}^d$ for some fixed $d \geqslant 1$.

Bayesian point of view

- First step: equip the parameter space Θ with a probability measure Π, called the prior distribution.
- The parameter becomes a random variable

$$\theta \sim \Pi \quad \text{on } \Theta.$$

Prior, likelihood and joint law

Densities

We assume from now on that

- for every $\theta \in \Theta$, P_θ has a density $p_\theta(x)$ with respect to a sigma-finite measure μ on E:

$$dP_\theta(x) = p_\theta(x)\, d\mu(x);$$

- the prior Π has a density $\pi(\theta)$ with respect to a sigma-finite measure ν on Θ:

$$d\Pi(\theta) = \pi(\theta)\, d\nu(\theta).$$

Joint distribution of (X, θ)

We define the joint law $\mathcal{L}(\theta, X)$ by the density

$$(x, \theta) \mapsto \pi(\theta)\, p_\theta(x)$$

with respect to the product measure $\nu \otimes \mu$.

Posterior distribution and Bayes formula

Marginals and conditionals

From the joint density $\pi(\theta)p_\theta(x)$ we recover:

- the prior density of θ by integrating out x: $\forall \theta \in \Theta, \quad \int_E \pi(\theta)p_\theta(x)\,d\mu(x) = \pi(\theta)$
- the conditional law $X \mid \theta \sim P_\theta$ with density $p_\theta(x)$
- the marginal density of X with respect to μ: $\quad\quad$ ⚠ This is not $p_\theta(x)$

$$f(x) = \int_\Theta p_\theta(x)\,\pi(\theta)\,d\nu(\theta)$$

Posterior and Bayes formula

- The posterior distribution is the conditional law $\mathcal{L}(\theta \mid X)$, denoted $\Pi(\cdot \mid X)$.
- Under the density assumptions above, it admits a density w.r.t. ν (Bayes formula):

$$\forall \theta \in \Theta, \qquad \pi(\theta \mid X) = \frac{p_\theta(X)\,\pi(\theta)}{f(X)},$$

where $f(X) = \int_\Theta \pi(\theta')p_{\theta'}(X)\,d\nu(\theta')$ is the marginal likelihood.

Why Bayesian? De Finetti's theorem

Definition: Exchangeability

Random variables $X_1, \ldots, X_n$ are exchangeable if for any permutation σ, the laws of $(X_1, \ldots, X_n)$ and $(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ are identical.

De Finetti (1931): representation theorem

For any *exchangeable* sequence $(X_1, X_2, \ldots)$ of $\{0, 1\}$-valued random variables, there exists a unique probability density π on $[0, 1]$ such that, for every n and every $x_1, \ldots, x_n \in \{0, 1\}$,

$$P(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \pi(\theta) \, d\nu(\theta).$$

The joint law is a mixture of i.i.d. Bernoulli laws.

Why Bayesian? De Finetti's theorem

- Exchangeable binary data can always be represented as i.i.d. given a parameter θ with prior $\pi(\theta)$.

- The prior $\pi(\theta)$ is not an arbitrary trick: while we do not know what it is exactly, it always exists.

- De Finetti-type results extend to more general cases, giving a strong justification for Bayesian modeling.

Prior as information

A prior $\pi(\theta)$ is a probability measure/density that encodes uncertain information about the parameter θ before seeing the data.

The prior allows us to

- satisfy the likelihood principle: inferences depend on the likelihood $L_\theta(X)$ only
- represent all uncertainties about θ
- integrate external or expert knowledge a priori, instead of relying solely on the sample/observation X

Model.

$$X \mid \theta \sim \mathcal{N}(\theta, 1), \qquad \theta \sim \mathcal{N}(0, 1)$$

Densities (w.r.t. Lebesgue measure).

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right), \qquad \pi(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right)$$

Posterior for one observation $X = x$.

$$\pi(\theta \mid X = x) \propto \pi(\theta)\, p_\theta(x) \propto \exp\left(-\frac{1}{2}\left[\theta^2 + (x-\theta)^2\right]\right)$$

Complete the square:

$$\theta^2 + (x-\theta)^2 = 2\left(\theta - \frac{x}{2}\right)^2 + \frac{x^2}{2}$$

Hence, up to a normalising constant,

$$\pi(\theta \mid X = x) \propto \exp\left(-\left(\theta - \frac{x}{2}\right)^2\right) \qquad \text{or, equivalently} \qquad \boxed{\theta \mid X = x \sim \mathcal{N}\left(\frac{x}{2}, \frac{1}{2}\right)}$$

Now take $X_1, \ldots, X_n$ i.i.d. given θ:

$$X_i \mid \theta \sim \mathcal{N}(\theta, 1), \qquad \theta \sim \mathcal{N}(0, 1)$$

Likelihood

$$\prod_{i=1}^{n} p_\theta(x_i) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2\right)$$

Posterior

$$\pi(\theta \mid x_1, \ldots, x_n) \propto \pi(\theta) \prod_{i=1}^{n} p_\theta(x_i) \propto \exp\left(-\frac{1}{2}\left[\theta^2 + \sum_{i=1}^{n} (x_i - \theta)^2\right]\right)$$

Using $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ and completing the square,

$$\pi(\theta \mid x_1, \ldots, x_n) \propto \exp\left(-\frac{n+1}{2}\left(\theta - \frac{n\bar{x}_n}{n+1}\right)^2\right) \text{ or, equivalently, } \boxed{\theta \mid X_1, \ldots, X_n \sim \mathcal{N}\left(\frac{n\bar{X}_n}{n+1}, \frac{1}{n+1}\right)}$$

What do we look at in the posterior?

- Posterior mean

$$m_X = \mathbb{E}[\theta \mid X] = \int_\Theta \theta \, d\pi(\theta \mid X).$$

- Posterior mode (MAP estimator)

$$\text{mode}(\theta \mid X) \in \arg\max_{\theta \in \Theta} \pi(\theta \mid X) = \arg\max_{\theta \in \Theta} \pi(\theta) p_\theta(X),$$

  where $\pi(\theta \mid X)$ is the posterior density.

- Posterior dispersion
  - For $\Theta \subset \mathbb{R}$:

$$v_X = \text{Var}(\theta \mid X) = \int_\Theta (\theta - m_X)^2 \, d\pi(\theta \mid X).$$

  - For $\Theta \subset \mathbb{R}^d$:

$$\Sigma_X = \int_\Theta (\theta - m_X)(\theta - m_X)^\mathsf{T} \, d\pi(\theta \mid X).$$

What do we look at in the posterior?

- **Posterior quantiles**
  Let $F_{\theta|X}$ be the cdf of $\pi(\,\cdot\mid X)$ and $F_{\theta|X}^{-1}$ its (generalised) inverse. For $p \in (0,1)$:

$$q_p(X) \;=\; F_{\theta|X}^{-1}(p)$$

  is the posterior p-quantile (for example $q_{1/2}(X)$ is the posterior median).

Penalized linear regression

Linear regression model.
We observe (x_i, y_i), $i = 1, \ldots, n$, and assume

$$y_i = x_i^\top \theta + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Penalized least squares. We choose $\hat{\theta}_n$ as a minimizer of

$$\sum_{i=1}^{n} (y_i - x_i^\top \theta)^2 + \text{pen}(\theta).$$

Typical choices:

- Ridge: $\text{pen}(\theta) = \lambda \|\theta\|_2^2$,
- Lasso: $\text{pen}(\theta) = \lambda \|\theta\|_1$.

Penalized linear regression: Bayesian view

Bayesian interpretation. Under the Gaussian noise model,

$$p_\theta(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - x_i^\top \theta)^2\right)$$

is the likelihood. If we choose a prior

$$\pi(\theta) \propto \exp\big(-\operatorname{pen}(\theta)\big),$$

then we also have

$$\hat{\theta}_n = \arg\max_\theta \pi(\theta \mid y_1, \ldots, y_n)$$

is a MAP estimator.

Penalty $\iff$ prior

- Ridge: $\operatorname{pen}(\theta) = \lambda\|\theta\|_2^2 \implies$ Gaussian prior $\pi(\theta) \propto \exp(-\lambda\|\theta\|_2^2)$.
- Lasso: $\operatorname{pen}(\theta) = \lambda\|\theta\|_1 \implies$ Laplace prior $\pi(\theta) \propto \exp(-\lambda\|\theta\|_1)$.

Take-home message: penalized linear regression is Bayesian estimation with an explicit prior on θ (MAP).

Why even non-Bayesians may like Bayesian methods

Even a true non-Bayesian may like Bayesian methods, because

- they are elegant;
- they allow us to incorporate prior information in a principled way;
- they may be easier to implement in complex models.

A true non-Bayesian will still want to understand the performance of Bayesian procedures in a non-Bayesian framework: frequentist Bayesian theory (see Lecture 7)

Frequentist Bayesian theory. Assume the data X are generated under a fixed "true'" parameter θ_0 and consider the posterior $\Pi(\theta \in \cdot \mid X)$ as a random probability measure on the parameter space. We would like $\Pi(\theta \in \cdot \mid X)$ to put most of its mass near θ_0 for "most" samples X.

Asymptotic setting. For a growing sample $X^{(n)}$ where the information increases as $n \to \infty$, we want the posterior $\Pi(\theta \in \cdot \mid X^{(n)})$ to contract around θ_0 fast.

Prior choice

Why talk about priors?

- The prior Π encodes information we have about the parameter before seeing the data (expert opinion, physical constraints, etc.).

- Different priors can lead to very different posterior distributions $\pi(\,\cdot \mid X)$, especially with small samples.

- In many applications the available prior information is vague: several priors are compatible with it, so the choice is often partly arbitrary.

Criteria for choosing a prior

There are many possible criteria for selecting π.

- **Practical / computational:** choose priors that make posterior calculations simple, e.g. conjugate priors.
- **Invariance and objective rules:** priors such as Jeffreys prior are motivated by invariance or information arguments.
- **Empirical Bayes:** estimate hyperparameters of the prior from the data.
- **Hierarchical modelling:** use several levels of priors to represent different sources of variability or uncertainty.
- **Physical or qualitative information:** prior support reflects constraints on the parameter (positivity, being in a given interval, order restrictions, etc.).

These ideas will guide the different approaches to prior construction described in the following.

Subjectivist and objective viewpoints

Two Bayesian mindsets.

- Subjectivist: the prior represents genuine prior beliefs, informed by past experience and expert knowledge.

- Objective: the prior is not derived from personal beliefs, but constructed in order to "let the data speak" as much as possible (non informative priors, reference priors, empirical Bayes, . . .).

Remarks:

- Prior information is rarely precise enough to determine a unique prior; several priors may be compatible with the same background information ⇒ the choice is often partly arbitrary.

- There is no single universally correct prior, and the choice of prior has an impact on the inference.

- Ambiguity is not specific to Bayes: frequentists also choose among many estimators (MLE, penalized MLE, . . .).

Objective ("non-informative") priors as regularization

- In many statistical learning methods, a prior can be viewed as a regularization term on the likelihood: it penalizes complex models and helps prevent overfitting.
- However, we often do not want to privilege any particular parametrization of θ.

Example

A variable X with Weibull law can be parametrized in different ways:

$$f(x \mid \eta, \beta) = \frac{\beta}{\eta^\beta} x^{\beta-1} \exp\big(-(x/\eta)^\beta\big)\mathbf{1}_{x \geqslant 0},$$

or, equivalently,

$$f(x \mid \mu, \beta) = \mu\beta x^{\beta-1} \exp\big(-\mu x^\beta\big)\mathbf{1}_{x \geqslant 0}.$$

The prior information we might have about X should not depend on whether we use (η, β) or (μ, β).

- Objective priors aim to encode only minimal information, in a way that is as invariant to reparametrization as possible.

Exercise

Let $\theta \in [1, 2]$ be the parameter of a model $X \sim p_\theta$. Assume we do not know anything else about X or about θ.

- We decide to use the prior $\theta \sim \mathcal{U}[1, 2]$.
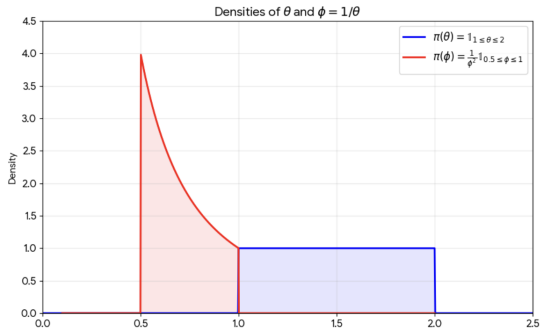- Now reparametrize the model in terms of

$$\phi = 1/\theta \in [1/2, 1],$$

so that $X \sim q_\phi$, where $q_\phi = p_\theta$.

Question. Can we also choose a *uniform* prior

$$\phi \sim \mathcal{U}[1/2, 1] \ ?$$

We do not have the same prior if we put a uniform distribution on θ or ϕ

Densities of θ and $\phi = 1/\theta$

$\pi(\theta) = 1_{1 \le \theta \le 2}$

$\pi(\phi) = \frac{1}{\phi^2} 1_{0.5 \le \phi \le 1}$

We used the change-of-variable formula $\pi_\phi(\phi) = \pi_\theta(h(\phi)) \left| \frac{dh}{d\phi} \right|$ for $h(\phi) = 1/\phi$.

Improper and weakly informative priors

- Objectively, we often only have very weak information such as *"the likelihood of a potential dataset should have this form"*.

- General construction rules can also lead to priors $\pi(\theta)$ that are not probability measures, in the sense that

$$\int_{\Theta} \pi(\theta)\, \mathrm{d}\theta = \infty.$$

  These are called improper priors.

- In the literature they are sometimes called *non-informative* priors, but strictly speaking *no* prior is completely information-free. A better description is *weakly informative*.

⚠ Such priors are useful only if the resulting posterior is a proper probability distribution (integrable and normalizable).

Definition

Suppose we use an improper prior π on θ and assume that, for the observed data X,

$$\int_{\Theta} p_{\theta}(X) \, \mathrm{d}\pi(\theta) < \infty \quad \text{almost surely.}$$

Then the corresponding posterior distribution $\pi[\cdot \mid X]$ is a probability measure with density given by

$$\theta \longmapsto \pi(\theta \mid X) = \frac{p_{\theta}(X) \, \pi(\theta)}{\displaystyle\int_{\Theta} p_{\theta}(X) \, \pi(\theta) \, \mathrm{d}\nu(\theta)}.$$

Invariance principle

If we move from θ to $\eta = g(\theta)$ by a bijection g, the amount of prior information should not change:

$$\pi^*(\eta) = \left| \det \frac{\partial \eta}{\partial \theta} \right| \pi(g^{-1}(\eta))$$

should encode the same beliefs as $\pi(\theta)$.

To construct such a prior, Jeffreys proposes to use the Fisher information $I(\theta)$, which measures how informative the model P_θ is about θ.

Fisher information

Consider a regular parametric model $\{P_\theta,\ \theta \in \Theta\}$ on X with density $p_\theta(x)$ and log-likelihood

$$\ell_\theta(X) = \log p_\theta(X).$$

Score

$$\ell'_\theta(X) = \frac{\partial}{\partial \theta}\, \ell_\theta(X) = \frac{p'_\theta(X)}{p_\theta(X)}.$$

Fisher information at θ

$$I(\theta) = \mathbb{E}_\theta\big[\ell'_\theta(X)^2\big].$$

For an i.i.d. sample $X^{(n)} = (X_1, \ldots, X_n)$ from P_θ, the information adds up:

$$I_n(\theta) = n\, I(\theta).$$

Large $I(\theta)$ means the likelihood is very peaked around θ, so the data dominate the prior there.

Definition: Jeffreys prior, 1D

For $\Theta \subset \mathbb{R}$, if $I(\theta)$ exists, the Jeffreys prior is

$$\pi(\theta) = \sqrt{I(\theta)}.$$

- This construction uses only the model $p_\theta(x)$.
- Regions where the model is very informative ($I(\theta)$ large) receive more prior mass, so that the prior has less influence on the posterior.

Examples

- Bernoulli model $\mathcal{B}(\theta)$, $\theta \in (0,1)$: $I(\theta) = \frac{1}{\theta(1-\theta)}$, hence

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2},$$

  i.e. a $\mathrm{Beta}(1/2, 1/2)$ prior.

- Normal model $X \mid \theta \sim \mathcal{N}(\theta, 1)$: $I(\theta) = 1$, so $\pi(\theta) \propto 1$ (improper flat prior).

Jeffreys prior in higher dimensions

For $\theta \in \Theta \subset \mathbb{R}^d$, the Fisher information matrix is

$$I_{ij}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(X \mid \theta)\right].$$

Definition: Jeffreys prior, d-dimensional

If $I(\theta)$ exists, define

$$\pi(\theta) \propto \sqrt{\det I(\theta)}.$$

Invariance property Let $\eta = g(\theta)$ be any smooth bijective reparametrization. If $\pi_\theta(\theta) \propto \sqrt{\det I(\theta)}$, then the induced density on η satisfies

$$\pi_\eta(\eta) \propto \sqrt{\det I(\eta)}.$$

Hence Jeffreys prior automatically respects the invariance principle.

Proof:

Jeffreys prior: exercises

Exercise 1 (Exponential model). Let $X \mid \theta \sim \mathcal{E}(\theta)$ with rate $\theta > 0$.

- Compute the Fisher information $I(\theta)$.
- Deduce the Jeffreys prior $\pi(\theta) \propto \sqrt{I(\theta)}$.

Exercise 2 (Weibull model). Let X follow a Weibull law with two common parametrizations

$$p(x \mid \eta, \beta) = \frac{\beta}{\eta} c \left(\frac{x}{\eta}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\eta}\right)^{\beta}\right] 1_{\{x \geq 0\}},$$

$$p(x \mid \mu, \beta) = \beta \mu x^{\beta-1} \exp(-\mu x^{\beta}) \, 1_{\{x \geq 0\}}.$$

- Compute the Jeffreys prior in each parametrization.
- Check that the two expressions are coherent by using the change-of-variables formula.

Conjugate priors: idea

Goal. Choose a prior family that is stable under Bayesian updating.

Definition (conjugate family)

Let $\mathcal{P} = \{P_\theta, \ \theta \in \Theta\}$ be a statistical model and $\mathcal{F}$ a family of prior distributions on Θ. We say that $\mathcal{F}$ is *conjugate* for $\mathcal{P}$ if, for every $\pi \in \mathcal{F}$, the posterior law $\pi[\cdot \mid X]$ also belongs to $\mathcal{F}$.

Why it is useful.

- Posterior has the same functional form as the prior; only hyperparameters change, not structural form.
- Closed forms for posterior mean, variance, credible sets, predictions, etc.
- Easy to simulate from the posterior if we know how to simulate from the prior.

Exponential family and natural conjugate priors

Consider a k-dimensional exponential family in natural form

$$p_\theta(x) = h(x) \exp\{\theta \cdot T(x) - \psi(\theta)\}, \qquad \theta \in \Theta \subset \mathbb{R}^k.$$

A standard **natural conjugate prior** for θ is

$$\pi(\theta \mid a, b) \propto \exp\{\theta \cdot a - b\,\psi(\theta)\}, \qquad a \in \mathbb{R}^k, \ b > 0.$$

Given one observation x, Bayes rule gives the posterior

$$\pi(\theta \mid a, b, x) \propto \exp\{\theta \cdot (a + T(x)) - (b+1)\psi(\theta)\},$$

so the posterior is again in the same family, with updated hyperparameters

$$(a, b) \longrightarrow (a + T(x), \ b + 1).$$

For a sample $x_1, \ldots, x_n$ the update is

$$(a, b) \longrightarrow \left(a + \sum_{i=1}^n T(x_i), \ b + n\right).$$

Natural conjugate priors for some common models

$f(x \mid \theta)$	$\pi(\theta)$	$\pi(\theta \mid x)$
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}\big(\varrho(\sigma^2\mu + \tau^2 x),\ \varrho\sigma^2\tau^2\big), \quad \varrho^{-1} = \sigma^2 + \tau^2$
$\mathrm{Poisson}(\theta)$	$\mathrm{Gamma}(\alpha, \beta)$	$\mathrm{Gamma}(\alpha + x, \beta + 1)$
$\mathrm{Gamma}(\nu, \theta)$	$\mathrm{Gamma}(\alpha, \beta)$	$\mathrm{Gamma}(\alpha + \nu, \beta + x)$
$\mathrm{Binomial}(n, \theta)$	$\mathrm{Beta}(\alpha, \beta)$	$\mathrm{Beta}(\alpha + x, \beta + n - x)$
$\mathrm{NegBin}(m, \theta)$	$\mathrm{Beta}(\alpha, \beta)$	$\mathrm{Beta}(\alpha + m, \beta + x)$
$\mathrm{Multinomial}_k(\theta_1, \ldots, \theta_k)$	$\mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_k)$	$\mathrm{Dirichlet}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
$\mathcal{N}(\mu, 1/\theta)$	$\mathrm{Gamma}(\alpha, \beta)$	$\mathrm{Gamma}\big(\alpha + \tfrac{1}{2},\ \beta + \tfrac{(\mu-x)^2}{2}\big)$
$X_1, \ldots, X_n \mid \theta \sim \mathrm{Unif}(0, \theta)$	$\mathrm{Pareto}(\alpha, r)$	$\mathrm{Pareto}(\alpha + n,\ r_X), \quad r_X = \max\{r, X_1, \ldots, X_n\}$

Motivation.

- In many problems we need a prior on a parameter θ, but we are not sure how to choose it.
- We introduce a hyperparameter γ that controls a family of priors

$$\theta \mid \gamma \sim \pi(\theta \mid \gamma).$$

- Then we put a second–level prior on γ:

$$\gamma \sim \pi(\gamma).$$

Joint model.

$$X, \theta, \gamma \sim p_\theta(X)\, \pi(\theta \mid \gamma)\, \pi(\gamma).$$

Advantages.

- Provides a flexible framework for modeling families of priors.
- Allows us to encode partial prior information and share information across related parameters (random effects, panel data, etc.).
- Hyperparameters γ play the role of an *index* for a whole family $\{\pi(\cdot \mid \gamma)\}_\gamma$.

Hierarchical Bayes vs empirical Bayes

Hierarchical Bayes.

- We treat γ as an unknown random quantity:

$$\theta \mid \gamma \sim \pi(\theta \mid \gamma), \qquad \gamma \sim \eta(\gamma).$$

- Posterior inference is based on

$$\pi(\theta, \gamma \mid x) \propto p_\theta(X) \, \pi(\theta \mid \gamma) \, \pi(\gamma).$$

- Fully Bayesian: uncertainty on γ is propagated into the posterior of θ.

Empirical Bayes.

- We choose a parametric family of priors $\{\pi_\gamma(\theta)\}_{\gamma \in \Gamma}$ (e.g. Normal, Gamma, Beta).
- Use the data to estimate γ (for example by marginal likelihood):

$$f_\gamma(X) = \int p_\theta(X) \, \pi_\gamma(\theta) \, d\theta, \quad \hat{\gamma} = \arg\max_\gamma f_\gamma(X).$$

- Then treat $\pi_{\hat{\gamma}}(\theta)$ as the prior and perform standard Bayes.

Empirical Bayes: examples

Gaussian model.

- Data: $X_1, \ldots, X_n \mid \theta \sim \mathcal{N}(\theta, 1)$ i.i.d.
- Prior family: $\theta \sim \mathcal{N}(\mu, 1)$, with hyperparameter μ.
- Marginal likelihood for one observation:

$$f_\mu(X_1) = \int \mathcal{N}(X_1 \mid \theta, 1)\, \mathcal{N}(\theta \mid \mu, 1)\, d\theta = \mathcal{N}(X_1 \mid \mu, 2).$$

- Maximizing $f_\mu(X_1)$ gives $\hat{\mu} = X_1$; for n observations, $\hat{\mu} = \bar{X}_n$.
- Empirical Bayes prior: $\theta \sim \mathcal{N}(\bar{X}_n, 1)$.

Poisson model.

- Data: $X_1, \ldots, X_n \mid \theta \sim \mathcal{P}(\theta)$ i.i.d.
- Prior family: $\theta \sim \mathrm{Exp}(\lambda)$.
- Empirical Bayes estimate: $\hat{\lambda} = 1/\bar{X}_n$, so the prior becomes $\theta \sim \mathrm{Exp}(1/\bar{X}_n)$.

Fusion of priors from multiple experts

Suppose we have M possible priors $\pi_1(\theta), \ldots, \pi_M(\theta)$ (e.g. from different experts), with weights $\omega_i \geqslant 0$, $\sum_{i=1}^{M} \omega_i = 1$.

Linear (arithmetic) pool.

$$\pi_{\text{lin}}(\theta) = \sum_{i=1}^{M} \omega_i \pi_i(\theta).$$

- Natural, but ⚠ posterior of π_{lin} is not the same as the weighted sum of posteriors $\pi_i(\theta \mid x)$.

Logarithmic (geometric) pool.

$$\pi_{\text{log}}(\theta) = \frac{\prod_{i=1}^{M} \pi_i(\theta)^{\omega_i}}{\int_{\Theta} \prod_{i=1}^{M} \pi_i(u)^{\omega_i} \, du}.$$

- Combining first, then updating, is coherent with updating each prior then combining.
- Note: it is the prior that minimizes a weighted sum of Kullback–Leibler divergences:

$$\pi^*(\theta) = \arg \min_{\pi} \sum_{i=1}^{M} \omega_i KL(\pi, \pi_i), \qquad KL(\pi, \pi_i) = \int \log(\pi(\theta)/\pi_i(\theta))\pi(\theta)d\theta$$

Different approaches

$$\alpha \sim \pi(\alpha)$$

$$\beta \mid \alpha \sim \pi(\beta \mid \alpha)$$

$$\theta \mid \alpha, \beta \sim \pi(\theta \mid \alpha, \beta)$$

$$X \mid \theta \sim p_\theta$$

$$\alpha \sim \pi(\alpha)$$

$$\beta \mid \alpha \sim \pi(\beta \mid \alpha)$$

$$\theta \mid \alpha, \beta \sim \pi(\theta \mid \alpha, \beta)$$

$$X \mid \theta \sim p_\theta \qquad \text{Likelihood / Frequentist model}$$

Different approaches

$$\alpha \sim \pi(\alpha)$$

$$\beta \mid \alpha \sim \pi(\beta \mid \alpha)$$

$$\theta \mid \alpha, \beta \sim \pi(\theta \mid \alpha, \beta) \qquad \text{Empirical Bayes / Bayesian model}$$

$$X \mid \theta \sim p_\theta$$

$$\alpha \sim \pi(\alpha)$$

$$\beta \mid \alpha \sim \pi(\beta \mid \alpha) \qquad \text{Hierarchical Bayes}$$

$$\theta \mid \alpha, \beta \sim \pi(\theta \mid \alpha, \beta)$$

$$X \mid \theta \sim p_\theta$$

$$\alpha \sim \pi(\alpha) \qquad \text{Hierarchical Bayes}$$

$$\beta \mid \alpha \sim \pi(\beta \mid \alpha)$$

$$\theta \mid \alpha, \beta \sim \pi(\theta \mid \alpha, \beta)$$

$$X \mid \theta \sim p_\theta$$