BAYESIAN MODELS

Approach

The defining assumption of Bayesian statistics is that the distribution P which explains the data is *a random quantity* and itself has a distribution Q. The generative model for data $X_1, X_2, \ldots$ is

$$P \sim Q$$
$$X_1, X_2, \ldots \sim_{\text{iid}} P$$

Rationale

- In any statistical approach (Bayesian or classical), the distribution P is unknown.
- Bayesian statistics argues that any form of uncertainty should be expressed by probability distributions.
- We can think of the randomness in Q as a model of the statistician's lack of knowlegde regarding P.

An older name for Bayesian statistics is *inverse probability*.

Prior and Posterior

The distribution Q of P is called the **a priori distribution** (or the **prior** for short). Our objective is to determine the conditional probability of P given observed data,

$$\Pi[P|x_1, \ldots, x_n] \ .$$

This distribution is called the **a posteriori distribution** or **posterior**.

Parametric case

We can impose the modeling assumption that P is an element of a parametric model, e.g. that the density p of P is in a family $\mathcal{P} = \{p(x|\theta)|\theta \in \mathcal{T}\}$. If so, the prior and posterior can be expressed as distributions on $\mathcal{T}$. We write

$$q(\theta) \qquad \text{and} \qquad \Pi(\theta|x_1, \ldots, x_n)$$

for the prior and posterior density, respectively.

Remark

The posterior $\Pi[P|x_1, \ldots, x_n]$ is an abstract object, which can be rigorously defined using the tools of probability theory, but is in general (even theoretically) impossible to compute. However: In the parametric case, the posterior can be obtained using the Bayes equation.

Computing Parametric Posteriors

Parametric modeling assumption

Suppose $\mathcal{P} = \{p(x|\theta)|\theta \in \mathcal{T}\}$ is a model and q a prior distribution on $\mathcal{T}$. Our sampling model then has the form:

$$\theta \sim q$$
$$X_1, X_2, \ldots \sim_{\text{iid}} p(\,.\,|\theta)$$

Note that the data is *conditionally i.i.d.* given $\Theta = \theta$.

Bayes' Theorem

If $\mathcal{P}$ is a parametric Bayesian model and q a distribution on $\mathcal{T}$, the posterior under data $X_1, \ldots, X_n$ generated as above is

$$\Pi(\theta|x_1, \ldots, x_n) = \frac{\left(\prod_{i=1}^n p(x_i|\theta)\right) q(\theta)}{p(x_1, \ldots, x_n)} = \frac{\left(\prod_{i=1}^n p(x_i|\theta)\right) q(\theta)}{\int_{\mathcal{T}} \left(\prod_{i=1}^n p(x_i|\theta)\right) q(\theta) d\theta}$$

The individual terms have names:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

EXAMPLE: UNKNOWN GAUSSIAN MEAN

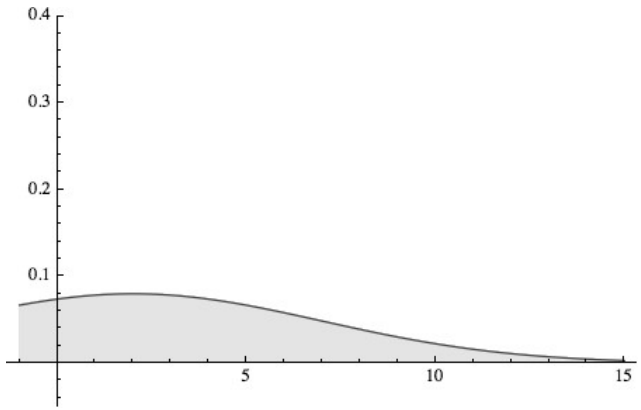
Model

We assume that the data is generated from a Gaussian with fixed variance σ^2. The mean μ is unknown. The model likelihood is $p(x|\mu,\sigma) = g(x|\mu,\sigma)$ (where g is the Gaussian density on the line).

Bayesian model

We choose a Gaussian prior on μ,

$$q(\mu) := g(\mu|\mu_0,\sigma_0) \ .$$



In the figure, $\mu_0 = 2$ and $\sigma_0 = 5$. Hence, we assume that $\mu_0 = 2$ is the most probable value of μ, and that $\mu \in [-3,7]$ with a probability ~ 0.68.

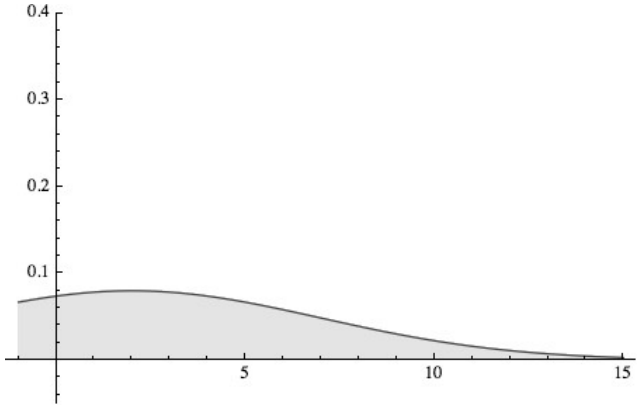
Posterior

Application of Bayes' formula to the Gaussian-Gaussian model shows

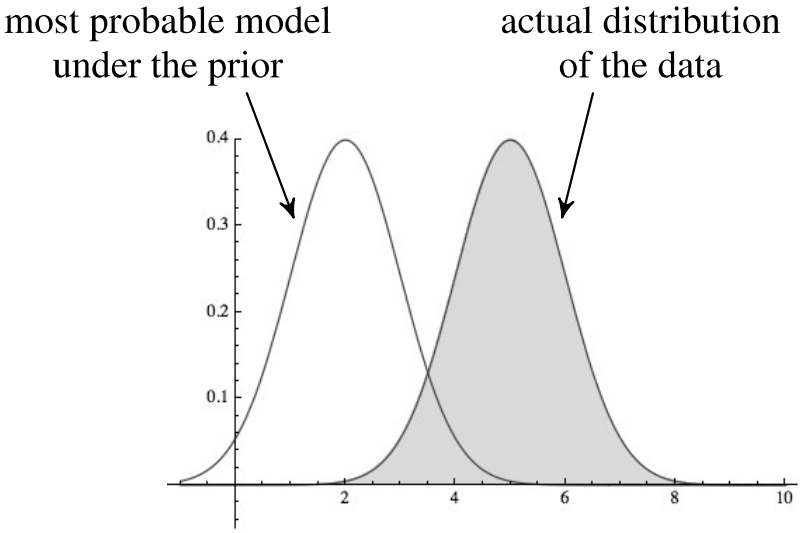
$$\Pi(\mu|x_{1:n}) = g(\mu|\mu_n,\sigma_n) \quad \text{where} \quad \mu_n := \frac{\sigma^2\mu_0 + \sigma_0^2\sum_{i=1}^n x_i}{\sigma^2 + n\sigma_0^2} \text{ and } \sigma_n := \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

Example: Unknown Gaussian Mean

Model

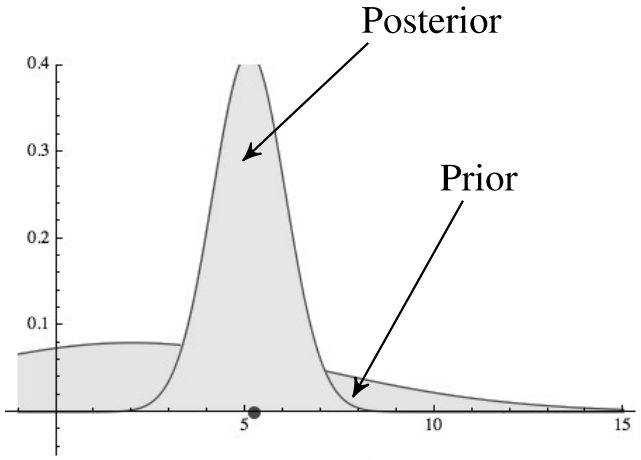


Prior

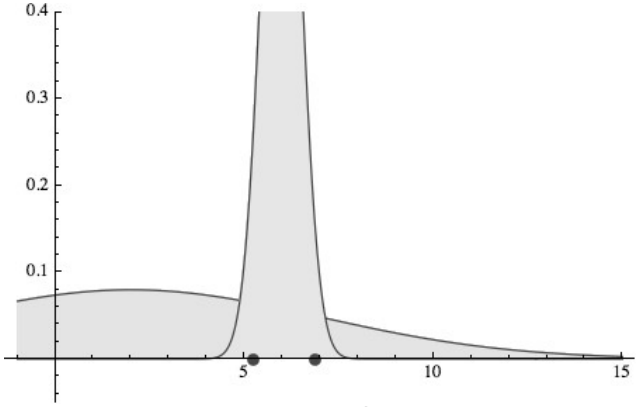


Sampling distribution

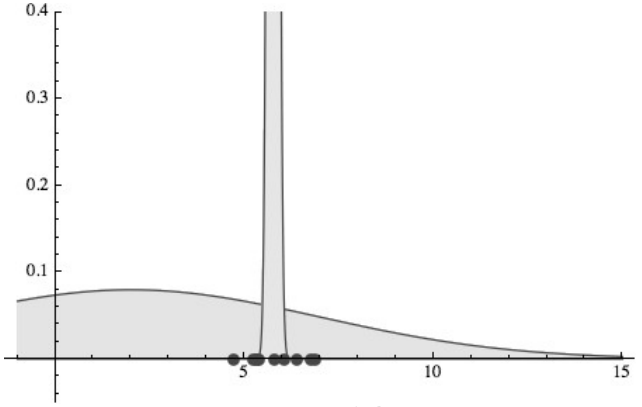
Posterior distributions



$n = 1$



$n = 2$



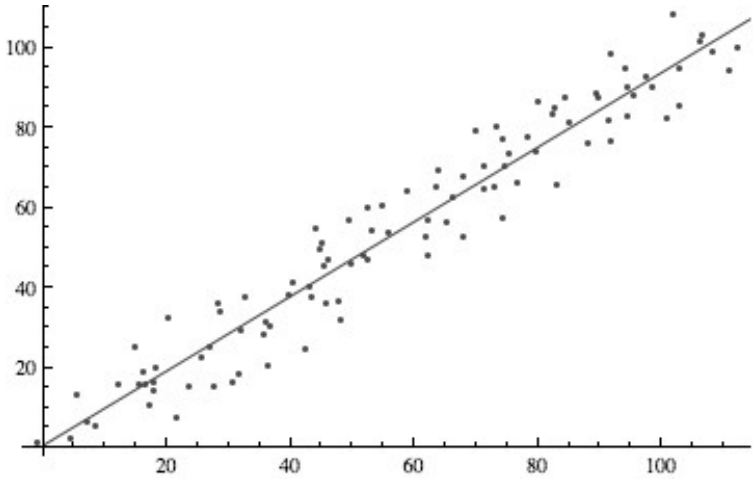
$n = 10$

A Slightly Different Perspective

Parameters

Intuitively, we can think of θ as the common pattern underlying the data:

$$P(X|\theta) = \text{Probability}[\text{data}|\text{pattern}]$$



Inference idea

$$\text{data} \quad = \quad \text{underlying pattern} \quad + \quad \text{independent randomness}$$

Broadly speaking, the goal of statistics is to extract the pattern from the data.
Bayesian statistics models the pattern as a random quantity.

MAP Estimation

Definition

Suppse $\Pi(\theta|x_{1:n})$ is the posterior of a Bayesian model. The estimator

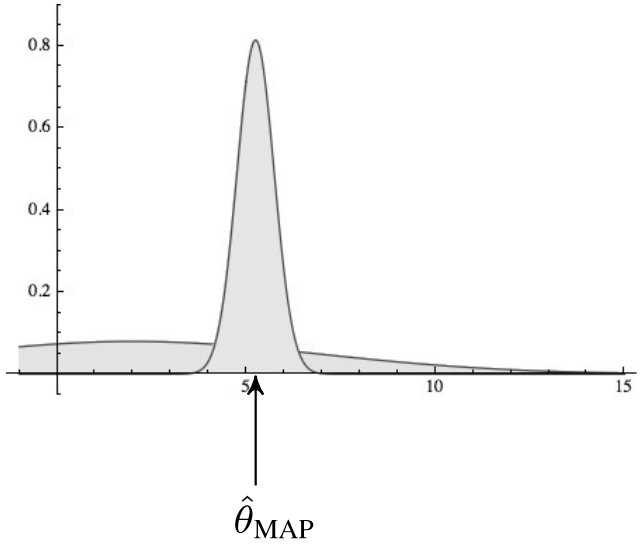
$$\hat{\theta}_{\text{MAP}} := \arg\max_{\theta \in \mathcal{T}} \Pi(\theta|x_{1:n})$$

is called the **maximum a posteriori** (or **MAP**) estimator for θ.

Point estimates

The goal of Bayesian inference is to compute the posterior distribution. Contrast this to classical statistics (e.g. maximum likelihood), where we typically estimate a single value for θ (a so-called **point estimate**).

MAP estimation combines aspects of Bayesian methodology (use of a prior) with aspects of classical methodology (since $\hat{\theta}_{\text{MAP}}$ is a point estimate).



Logarithmic view

Since the logarithm leaves the maximum invariant,

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in \mathcal{T}} \Pi(\theta|x_{1:n}) = \arg\max_{\theta \in \mathcal{T}} \log \Pi(\theta|x_{1:n})$$

Substituting in the Bayes equation gives

$$\log \Pi(\theta|x_{1:n}) = \sum_{i=1}^{n} \log p(x_i|\theta) + \log q(\theta) - \log p(x_1, \ldots, x_n) \ .$$

MAP as regularized ML

Since log-evidence does not depend on θ,

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta \in \mathcal{T}} \Big\{ \sum_{i=1}^{n} \log p(x_i|\theta) + \log q(\theta) \Big\}$$

Thus, the MAP estimate can be regarded as a regularized version of a maximum likelihood estimator. The regularization term $\log q(\theta)$ favors values where q (and hence $\log q$) is large.

Parametric Prior Families

Families of priors

The prior has to be expressed by a specific distribution. In parametric Bayesian models, we typically choose q as an element of a standard parametric family (e.g. the Gaussian in the previous example).

Hyperparameters

If we choose q as an element of a parametric family

$$\mathcal{Q} = \{q(\theta|\phi)|\phi \in \mathcal{H}\}$$

on $\mathcal{T}$, selecting the prior comes down to choosing ϕ. Hence, ϕ becomes a tuning parameter of the model.

Parameter of the prior familiy are called **hyperparameters** of the Bayesian model.

Natural Conjugate Priors

Exponential family likelihood

We now assume the parametric model $\mathcal{P} = \{p(x|\theta)|\theta \in \mathcal{T}\}$ is an exponential family model, i.e.

$$p(x|\theta) = \frac{h(x)}{Z(\theta)} e^{\langle S(x)|\theta \rangle} \ .$$

Natural conjugate prior

We define a prior distribution using the density

$$q(\theta|\lambda, y) = \frac{1}{K(\lambda, y)} \exp\Big(\langle \theta | y \rangle - \lambda \cdot \log Z(\theta) \Big)$$

- Hyperparameters: $\lambda \in \mathbb{R}_+$ and $y \in \mathcal{T}$.
- Note that the choice of $\mathcal{P}$ enters through Z.
- K is a normalization function.

Clearly, this is itself an exponential family (on $\mathcal{T}$), with $h \equiv Z^{-\lambda}$ and $Z \equiv K$.

Ugly Computation

Substitution into Bayes' equation gives

$$\begin{aligned}\Pi(\theta|x_1,\dots,x_n) &= \frac{\prod_{i=1}^n p(x_i|\theta)}{p(x_1,\dots,x_n)} \cdot q(\theta) \\ &= \frac{\frac{\prod_{i=1}^n h(x_i)}{Z(\theta)^n} \exp\left\langle \sum_i S(x_i)|\theta \right\rangle}{p(x_1,\dots,x_n)} \cdot \frac{\exp\big(\langle\theta|y\rangle - \lambda \log Z(\theta)\big)}{K(\lambda,y)}\end{aligned}$$

If we neglect all terms which do not depend on θ, we have

$$\Pi(\theta|x_1,\dots,x_n) \propto \; = \frac{\exp\left\langle \sum_i S(x_i)|\theta \right\rangle}{Z(\theta)^n} \exp\big(\langle\theta|y\rangle - \lambda \log Z(\theta)\big) = \frac{\exp\Big(\big\langle y + \sum_i S(x_i)|\theta \big\rangle\Big)}{Z(\theta)^{\lambda+n}}$$

Up to normalization, this is precisely the form of an element of $\mathcal{Q}$:

$$\dots = \exp\left(\left\langle y + \sum_i S(x_i)|\theta \right\rangle - (\lambda+n)\log Z(\theta)\right) \propto q(\theta|\lambda+n, y+\sum_{i=1}^n S(x_i))$$

Conclusion

If $\mathcal{P}$ is an exponential family model with sufficient statistic S, and if $q(\theta|\lambda, y)$ is a natural conjugate prior for $\mathcal{P}$, the posterior under observations $x_1, \ldots, x_n$ is

$$\Pi(\theta|x_1, \ldots, x_n) = q(\theta|\lambda + n, y + \sum_{i=1}^{n} S(x_i))$$

Remark

The form of the posterior above means that we can *compute the posterior by updating the hyperparameters*. This property motivates the next definition.

Definition

Assume that $\mathcal{P}$ is a parametric family and $\mathcal{Q}$ a family of priors. Suppose, for each sample size $n \in \mathbb{N}$, there is a function $T_n : \mathbf{X}^n \times \mathcal{H} \to \mathcal{H}$ such that

$$\Pi(\theta|x_1, \ldots, x_n) = q(\theta|\hat{\phi}) \qquad \text{with} \qquad \hat{\phi} := T_n(x_1, \ldots, x_n, \phi) \ .$$

Then $\mathcal{P}$ and $\mathcal{Q}$ are called **conjugate**.

CONJUGATE PRIORS

Closure under sampling

If the posterior is an element of the prior family, i.e. if

$$\Pi(\theta|x_1,\ldots,x_n) = q(\theta|\tilde{\phi})$$

for *some* $\tilde{\phi}$, the model is called **closed under sampling**. Clearly, every conjugate model is closed under sampling.

Remark

Closure under sampling is a weaker property than conjugacy; for example, any Bayesian model with

$$\mathcal{Q} = \{ \text{ all probability distributions on } \mathcal{T}\}$$

is trivially closed under sampling, but not conjugate.
Warning: Many Bayesian texts use conjugacy and closure under sampling equivalently.

Which models are conjugate?

It can be shown that, up a few "borderline" cases, the only paramteric models which admit conjugate priors are exponential family models.

Natural Conjugate Posteriors

Generic posterior updates

For an exponential family $\mathcal{P}$ with natural conjugate family $\mathcal{Q}$, the posterior is computed as the hyperparameter update

$$T_n(x_1, \ldots, x_n, \lambda, y) = (\lambda + n, y + \sum_{i=1}^{n} S(x_i)) \ .$$

Effect of hyperparameters

The natural conjugate prior $q(\theta|\lambda, y)$ has expected value $\mathbb{E}[\Theta] = y$. The parameter λ is a concentration, i.e.

$$\text{large } \lambda \quad \leftrightarrow \text{prior peaks sharply around } y \ .$$

Interpretation of posterior updates

The posterior mean is

$$\mathbb{E}[\Theta] = y + \sum_{i=1}^{n} S(x_i) \ ,$$

i.e. we linearly interpolate the prior guess and the sufficient statistics of the data in parameter space. The more data we observe, the larger the posterior concentration $\lambda + n$, which reflects increasing certainty regarding Θ given more data.

HIERARCHICAL MODEL

Motivation

Choosing a prior means we have to choose a distribution on Q (or a density q) on the parameter space $\mathcal{T}$. How?

Recall

A Bayesian model with prior q can be regarded as a decomposition of the data distribution $p(x)$ into a mixture $p(x_{1:n}) = \int_{\mathcal{T}} \prod_i p(x_i|\theta) q(\theta) d\theta$.

Hierarchical modeling idea

Split the prior up further into a mixture

$$q(\theta) = \int q(\theta|\phi)\tilde{q}(\phi)d\phi \ .$$

The generative model for the data is then:

$$\begin{aligned} \Phi &\sim \tilde{q} \\ \Theta &\sim q(\,.\,|\Phi) \\ X_1, \ldots, X_n &\sim p(\,.\,|\Theta) \end{aligned}$$

Additional levels in hierarchy

If we are so inclined, we can recursively split further:

$$q(\theta) = \int q(\theta|\phi_1)\tilde{q}_1(\phi_1)d\phi_1 \qquad \tilde{q}_1(\phi_1) = \int \tilde{q}_1(\phi_1|\phi_2)\tilde{q}_2(\phi_2)d\phi_2$$

etc.

Why?

- If the "intermediate" parameters ϕ (or ϕ_1,ϕ_2, etc) have a well-defined meaning, this can be a very useful way to derive a meaningful prior q.
- In problems with several related parameters, this permits "sharing" of information.

Practical aspects

- Permits use of well-studied, available models as "building blocks".
- Note: The parameters θ, ϕ can be regarded as layers of latent variables.
- Inference is possible using Markov chain sampling (later).
- Warning: Inference becomes more difficult with each additional layer.

Hierarchical models: Interpretation

- In practice, we start with a prior on θ, which has hyperparameters; we then add a "hyperprior" on the hyperparameters, etc.
- It is easy to get confused and regard the various distributions involved as "multiple priors".
- Keep in mind: This is a way to construct a single, *overall* prior q, which is given by

$$q(\theta) = \int q(\theta|\phi)\tilde{q}(\phi)d\phi$$

(or multiple integrals for multiple layers).