Bayesian Inference

Ricardo Ehlers
ehlers@icmc.usp.br

Departamento de Matemática Aplicada e Estatística
Universidade de São Paulo

Introduction

A model is a simplification of reality
(and some are useful)

Observable quantities
(can be measured)

Unobservable quantities
(parameters and latent variables)

Approaches: classical and Bayesian

Data: the observed values of the observable quantities.

The Frequentist approach to statistics

- Parameters are fixed and unknown.
- Probability interpreted as long run relative frequency.
- Probabilities assigned to observable variables given the unknown parameters.
- Some procedures rely on the idea of an infinite number of hypothetical repetitions of an experiment.

The Bayesian approach to statistics

- Parameters are random variables.
- Probabilities assigned to parameters as well as observations.
- Probabilities on parameters are interpreted as "degree of belief" and can be subjective.
- Rules of probability are used to revise 'degree of beliefs' about parameters given the observed data.

Bayes Theorem

Consider an unknown quantity of interest θ (typically unobservable).

- The information we have about θ is probabilistically summarised in $p(\theta)$.
- This information can be updated by observing a random quantity X related to θ through $p(x|\theta)$.
- The idea that after observing $X = x$ the quantity of information about θ increases is quite intuitive.
- The Bayes theorem is the updating rule used to quantify this information increase.

Bayes Theorem

$$p(\theta|x) = \frac{p(x,\theta)}{p(x)} = \frac{p(x|\theta)\ p(\theta)}{p(x)} = \frac{p(x|\theta)\ p(\theta)}{\displaystyle\int p(\theta,x)d\theta}. \tag{1}$$

- Our goal is to infer about plausible value(s) of θ (or functions of θ).
- This is naturally based on the updated probabilistic information we have about θ, i.e. on $p(\theta|x)$.
- For a fixed value of x, $p(x|\theta)$ is the plausibility or likelihood of each possible value of θ while $p(\theta)$ is called the prior distribution of θ.
- $p(\theta|x)$ is called the posterior distribution of θ.

Note that $1/p(x)$ does not depend on θ and plays the role of a normalizing constant of $p(\theta|x)$. Then,

$$p(\theta|x) \propto p(x|\theta)p(\theta). \tag{2}$$

This is the unscaled posterior distribution which gives information on its shape.

The posterior mode can be obtained as,

$$\arg\max_{\theta} p(x|\theta)p(\theta),$$

or equivalently,

$$\arg\max_{\theta} \; [\log p(x|\theta) + \log p(\theta)],$$

Note also that,

$$p(x) = \int p(x, \theta) d\theta = \int p(x|\theta) p(\theta) d\theta = E_\theta[p(X|\theta)]$$

which is called the predictive distribution.

This is the expected distribution for x given θ. So,

- Before observing X we can check the adequacy of the prior making predictions using $p(x)$.
- If X is observed and it received low predictive probability then we should question the model.

In many applications (e.g. time series and geostatistics) we are interested in predicting a process in time or space.

Suppose that after observing $X = x$ we are interested in prediciting Y, which is also related to θ, and probabilistically described byr $p(y|x, \theta)$.

The predictive distribution of Y given x is obtained by integration as,

$$\begin{aligned} p(y|x) &= \int p(y, \theta|x) d\theta = \int p(y|\theta, x) p(\theta|x) d\theta \\ &= E_{\theta|x}[p(y|\theta, x)] \end{aligned}$$

In many applications we can assume conditional independence between X e Y given θ and the predictive distribution simplifies to,

$$p(y|x) = \int p(y|\theta)p(\theta|x)d\theta = E_{\theta|x}[p(y|\theta)].$$

Note that predictions are always verifiable since Y is observable.

Example. (Migon & Gamerman, 1999) John claims some discomfort and goes to the doctor. The doctor believes John may have a certain disease. Based on his expertise about this disease and information given by the patient, the doctor assigns a probability 0.7 that John has the disease.

The (unknown) quantity of interest here is the disease indicator,

$$\theta = \begin{cases} 1, & \text{se o paciente tem a doença} \\ 0, & \text{se o paciente não tem a doença.} \end{cases}$$

To increase the evidence about the disease, the doctor asks John to undertake an examination X related to θ through the following probability distribution,

$$\begin{aligned} P(X = 1 \mid \theta = 0) &= 0.4, \text{ positive result given no disease} \\ P(X = 1 \mid \theta = 1) &= 0.95, \text{ positive result given disease} \end{aligned}$$

Suppose that the exam resulted positive ($X = 1$).

- Intuitively, the disease probability must have increased after this result.
- We want to quantify this increase.

This can be accomplished using the Bayes theorem,

$$P(\theta = 1 \mid X = 1) \propto P(X = 1 | \theta = 1)\, P(\theta = 1) = (0.95)(0.7) = 0.665$$

$$P(\theta = 0 \mid X = 1) \propto P(X = 1 | \theta = 0)\, P(\theta = 0) = (0.4)(0.3) = 0.12.$$

The normalizing constant k is easily obtained since $0.665k + 0.12k = 1$ and then $k = 1/0.785$.

The posterior distribution of θ is given by,

$$
\begin{aligned}
P(\theta = 1 \mid X = 1) &= 0.665/0.785 = 0.847 \\
P(\theta = 0 \mid X = 1) &= 0.12/0.785 = 0.153
\end{aligned}
$$

The information $X = 1$ increases the disease probability from 0.70 to 0.847.

Now John undertakes a second test Y which relates to θ as follows,

$$P(Y = 1 \mid \theta = 0) = 0.04 \qquad \text{and} \qquad P(Y = 1 \mid \theta = 1) = 0.99.$$

Before observing Y it is interesting to obtain its predictive distribution.

Since θ is discrete, it follows that,

$$p(y|x) = \sum_{\theta=0}^{1} p(y|x,\theta)p(\theta|x)$$

and note that $p(\theta|x)$ is a prior probability with respect to Y.

Now, assuming that X and Y are conditionally independent given θ,

$$p(y|x) = \sum_{\theta=0}^{1} p(y|\theta)p(\theta|x)$$

The discrete predicitve distribution of Y is then given by,

$$\begin{aligned} P(Y=1 \mid X=1) &= P(Y=1 \mid \theta=0)\, P(\theta=0 \mid X=1) \\ &+ P(Y=1 \mid \theta=1)\, P(\theta=1 \mid X=1) \\ &= (0.04)(0.153)+(0.99)(0.847)=0.845 \end{aligned}$$

$$P(Y=0 \mid X=1) = 1-P(Y=1 \mid X=1)=0.155.$$

Suppose the second test resulted negative $Y = 0$.

This value had little predicitve probability (0.155) which might lead the doctor to rethink the model in the first place.

- Was $P(\theta = 1) = 0.7$ a reasonable prior?
- Is test X really so unreliable? Is test Y that powerful?
- Have the tests been carried out properly?

Anyway, it is intuitive that the disease probability must have decreased and this can be quantified with a second application of Bayes theorem,

$$\begin{aligned} P(\theta = 1 \mid X = 1, Y = 0) &\propto l(\theta = 1; Y = 0) P(\theta = 1 \mid X = 1) \\ &\propto (0.01)(0.847) = 0.0085 \end{aligned}$$

$$\begin{aligned} P(\theta = 0 \mid X = 1, Y = 0) &\propto l(\theta = 0; Y = 0) P(\theta = 0 \mid X = 1) \\ &\propto (0.96)(0.153) = 0.1469. \end{aligned}$$

The normalizing constant is 1/(0.0085+0.1469)=1/0.1554 so that the posterior distribution of θ is given by,

$$P(\theta = 1 \mid X = 1, Y = 0) = 0.0085/0.1554 = 0.055$$

$$P(\theta = 0 \mid X = 1, Y = 0) = 0.1469/0.1554 = 0.945.$$

So, disease probability evolves along time like,

$$P(\theta = 1) = \begin{cases} 0.7, & \text{before the tests,} \\ 0.847, & \text{after test } X, \\ 0.055, & \text{after } X \text{ and } Y. \end{cases}$$

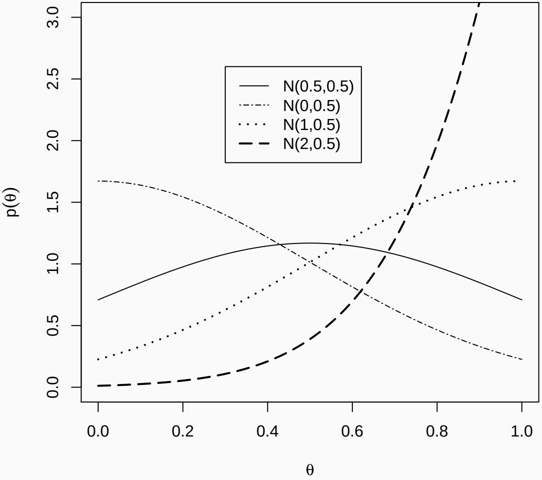
Example. Suppose we want to estimate the proportion θ of defective itens in a large shipment. Which probability distribution can be assigned to probabilistically encode our knowledge about $\theta \in (0, 1)$?

We can assume that $\theta \sim N(\mu, \sigma^2)$ truncated to $\theta \in (0, 1)$

Denoting by $f_N(\cdot|\mu, \sigma^2)$ the density function of a $N(\mu, \sigma^2)$ distribution it follows that the prior density of θ is given by,

$$p(\theta) = \frac{f_N(\theta|\mu, \sigma^2)}{\int_0^1 f_N(\theta|\mu, \sigma^2) d\theta} = \frac{(2\pi\sigma^2)^{-1/2} \exp(-0.5(\theta - \mu)^2/\sigma^2)}{\Phi\left(\frac{1-\mu}{\sigma}\right) - \Phi\left(\frac{-\mu}{\sigma}\right)}.$$

Truncated normal prior densities for θ.



Another possibility is to find a map from (0,1) to the real line and assign a prior on $\mathbb{R}$.

Assume that $\delta \sim N(\mu, \sigma^2)$ and consider the transformation,

$$\theta = \frac{\exp(\delta)}{1 + \exp(\delta)}.$$

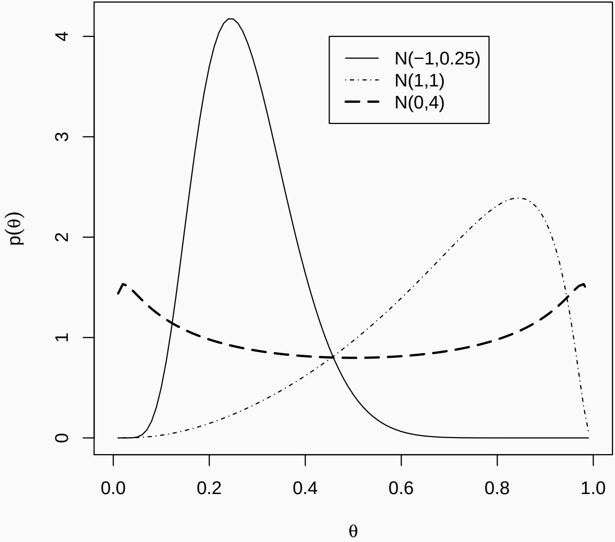
The inverse transformation is simply,

$$\delta = \log\left(\frac{\theta}{1-\theta}\right)$$

and the prior density of θ becomes,

$$\begin{aligned} p(\theta) &= f_N(\delta(\theta)|\mu, \sigma^2)\left|\frac{d\delta}{d\theta}\right| \\ &= (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\log\left(\frac{\theta}{1-\theta}\right)-\mu\right)^2\right\}\frac{1}{\theta(1-\theta)} \end{aligned}$$

Logistic-type prior densities for θ.



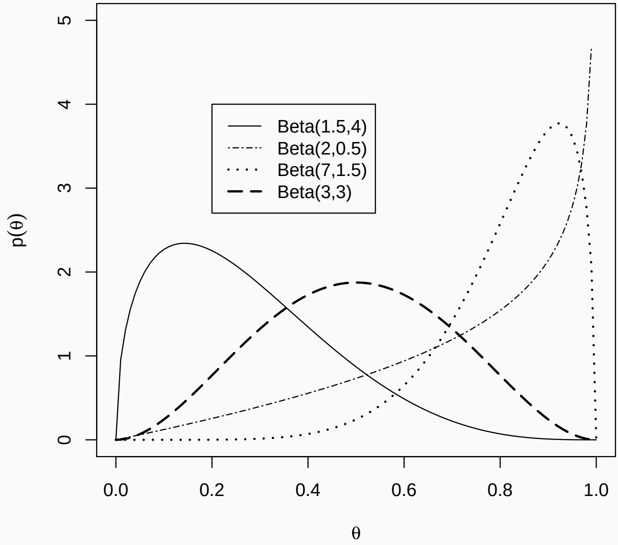
Finally, we can assign the prior $\theta \sim Beta(a, b)$

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},\ a > 0,\ b > 0,\ \theta \in (0,1).$$

The Beta distribution is symmetric about $1/2$ when $a = b$ and skewed when $a \neq b$.

Varying a and b we can define a rich family of distributions for θ, including the Uniform (0,1) when $a = b = 1$.

Beta prior densities for θ.



Suppose now that,

$$\begin{aligned} X|\theta &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\mu_0, \tau_0^2) \end{aligned}$$

where σ^2, μ_0 and τ_0^2 are known.

What is the posterior distribution of θ?

We have that,

$$\begin{aligned} p(x|\theta) &\propto \exp\{-\sigma^{-2}(x-\theta)^2/2\}, \text{ and} \\ p(\theta) &\propto \exp\{-\tau_0^{-2}(\theta-\mu_0)/2\} \end{aligned}$$

Then,

$$\begin{aligned} p(\theta|x) &\propto \exp\left\{-\frac{1}{2}[\sigma^{-2}(\theta^2-2x\theta)+\tau_0^{-2}(\theta^2-2\mu_0\theta)]\right\} \\ &\propto \exp\left\{-\frac{1}{2}[\theta^2(\sigma^{-2}+\tau_0^{-2})-2\theta(\sigma^{-2}x+\tau_0^{-2}\mu_0)]\right\}. \end{aligned}$$

The terms that do not depend on θ were incorporated into the proportionality constant.

Defining the following quantities,

$$\begin{aligned}\tau_1^{-2} &= \sigma^{-2} + \tau_0^{-2}\\ \tau_1^{-2}\mu_1 &= \sigma^{-2}x - \tau_0^{-2}\mu_0\end{aligned}$$

it follows that,

$$\begin{aligned}p(\theta|x) &\propto \exp\left\{-\frac{\tau_1^{-2}}{2}(\theta^2 - 2\theta\mu_1)\right\}\\ &\propto \exp\left\{-\frac{\tau_1^{-2}}{2}(\theta - \mu_1)^2\right\}\end{aligned}$$

since μ_1 does not depend on θ.

Then, the posterior density function has the same form (up to a constant) of a normal density with mean μ_1 and variance τ_1^2, i.e.

$$\theta|x \sim N(\mu_1, \tau_1^2).$$

- Note that defining precision as the inverse of variance, the posterior precision is the sum of prior and likelihood precisions and does not depend on x.
- We can interpret precision as a measure of information.
- Defining
$$w = \tau_0^{-2}/(\tau_0^{-2} + \sigma^{-2}) \in (0,1)$$
then w measures the relative information contained in the prior with respect to the total information.
- We can write,
$$\mu_1 = w\mu_0 + (1-w)x,$$
i.e. μ_1 is a convex linear combination of μ_0 and x so that,
$$\min\{\mu_0, x\} \leq \mu_1 \leq \max\{\mu_0, x\}.$$

The predictive distribution of X is easily obtained by noting that,

$$\begin{aligned} X &= \theta + \epsilon,\ \epsilon \sim N(0, \sigma^2) \\ \theta &= \mu_0 + w,\ w \sim N(0, \tau_0^2) \end{aligned}$$

such that $Cov(\theta, \epsilon) = Cov(\mu_0, w) = 0$.

The unconditional distribution of X is then normal as it results of a sum of two normal random variables.

Also,

$$\begin{aligned} E(X) &= E(\theta) + E(\epsilon) = \mu_0 \\ Var(X) &= Var(\theta) + Var(\epsilon) = \tau_0^2 + \sigma^2 \end{aligned}$$

so that, $X \sim N(\mu_0, \tau_0^2 + \sigma^2)$.

Example. (Box & Tiao, 1992) Two physicists A and B wish to determine a physical constant θ. They specify the following prior distributions,

$$\text{Physicist A (more experienced): } \theta \sim N(900, 20^2),$$
$$\text{Physicist B (not so experienced): } \theta \sim N(800, 80^2).$$

It is not difficult to obtain for example that,

$$\text{for Physicist } A\text{:} \quad P(860 < \theta < 940) \approx 0.95$$
$$\text{for Physicist } B\text{:} \quad P(640 < \theta < 960) \approx 0.95.$$

Using a calibrated device in a laboratory a measurement X of θ is made. The device has a sampling distribution $X|\theta \sim N(\theta, 40^2)$ and $X = 850$ was observed.

Therefore, applying our previous results it follows that,

$$\begin{aligned}
(\theta | X = 850) &\sim N(\mu_{1A}, \tau_{1A}^2) \quad \text{for Physicist } A \\
(\theta | X = 850) &\sim N(\mu_{1B}, \tau_{1B}^2) \quad \text{for Physicist } B.
\end{aligned}$$

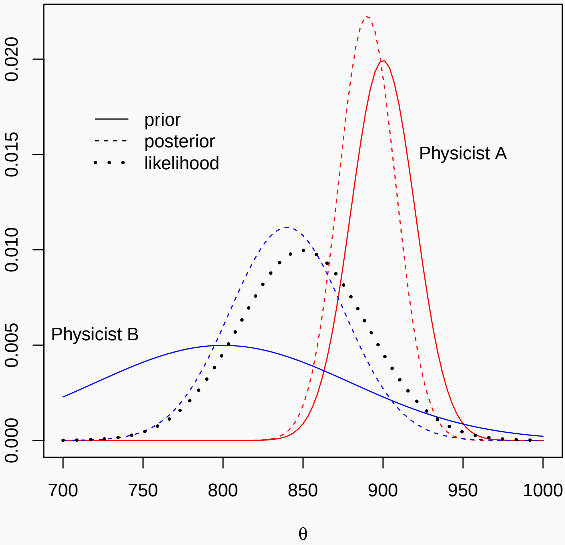
where

$$\begin{aligned}
\tau_{1A}^{-2} &= \tau_{0A}^{-2} + \sigma^{-2} = 0.003125 \\
w_A &= \tau_{0A}^{-2} / \tau_{1A}^{-2} = 0.8 \\
\mu_{1A} &= w\mu_{0A} + (1 - w)x = 890
\end{aligned}$$

$$\begin{aligned}
\tau_{1B}^{-2} &= \tau_{0B}^{-2} + \sigma^{-2} = 0.00078125 \\
w_B &= \tau_{0B}^{-2} / \tau_{1B}^{-2} = 0.2 \\
\mu_{1B} &= w\mu_{0B} + (1 - w)x = 840
\end{aligned}$$

Note how the posterior precisions increased with respect to prior precisions.

- For Physicist A: precision(θ) went up from $\tau_0^{-2} = 0.0025$ to $\tau_1^{-2} = 0.00312$ (25% increase).
- For Physicist B: precision(θ) went up from $\tau_0^{-2} = 0.000156$ to $\tau_1^{-2} = 0.000781$ (400% increase).



prior
posterior
likelihood
Physicist A
Physicist B
0.000
0.005
0.010
0.015
0.020
700
750
800
850
900
950
1000
θ

Example. Suppose again that $X|\theta \sim N(\theta, \sigma^2)$, with σ^2 known, but now $p(\theta) \propto 1$.

This is not even a density function since,

$$\int_{-\infty}^{\infty} p(\theta) d\theta = \infty.$$

and $p(\theta)$ is called an improper prior.

Even so we have that,

$$p(\theta|x) \propto \exp\{-(\theta - x)^2/2\sigma^2\}$$

and it can be verified that $\theta|x \sim N(x, \sigma^2)$ which is a proper posterior density function.

This is the limiting case of the previous result when $\tau_0^{-2} \to 0$ which implies that $\mu_1 \to x$ and $\tau_1^2 \to \sigma^2$.

Example. Suppose that $P(\text{obtain head after tossing a coin}) = \theta$ and the possible values of θ are 0.5 and 0.95 with probabilities,

$$\begin{aligned} P(\theta = 0.5) &= w \\ P(\theta = 0.95) &= 1 - w \end{aligned}$$

Suppose that we assign probabilities $w = 1 - w = 1/2$. Defining,

$$X = \begin{cases} 1, & \text{if the result is head} \\ 0, & \text{otherwise,} \end{cases}$$

Then,

$$P(X = x|\theta) = \theta^x(1-\theta)^{1-x},\ x \in \{0, 1\}.$$

The predictive distribution of X is given by,

$$
\begin{aligned}
P(X = x) &= w\ 0.5^x(1 - 0.5)^{1-x} + (1 - w)\ 0.95^x(1 - 0.95)^{1-x} \\
&= 0.5\left[0.5^x(1 - 0.5)^{1-x} + 0.95^x(1 - 0.95)^{1-x}\right].
\end{aligned}
$$

so that,

$$P(X = 0) = 0.275 \quad \text{and} \quad P(X = 1) = 0.725.$$

Example. In the previous example suppose we now have $\theta \in \{0.2, 0.4, 0.6, 0.8, 1\}$ with equal probabilities $1/5$.

The predictive distribution of X is given by,

$$P(X = x) = \frac{1}{5} \sum_{\theta} \theta^x (1-\theta)^{1-x},$$

so that,

$$P(X = 0) = 0.4 \quad \text{and} \quad P(X = 1) = 0.6.$$

In general, if $\theta \in \{\theta_1, \ldots, \theta_k\}$ with probabilities $w_1, \ldots, w_k$ then,

$$\begin{aligned} P(X = x) &= \sum_{i=1}^{k} \theta_i^x (1-\theta_i)^{1-x} w_i \\ P(X = 1) &= \sum_{i=1}^{k} \theta_i w_i \\ P(X = 0) &= \sum_{i=1}^{k} (1-\theta_i) w_i. \end{aligned}$$

Sequential Bayes

Let $x_1, \ldots, x_n$ be the observed values of $X_1, \ldots, X_n$ which are independent given θ and are related to θ through $p_i(x_i|\theta)$. Then,

$$
\begin{aligned}
p(\theta|x_n, x_{n-1}, \cdots, x_1) &\propto p(\theta)p_1(x_1|\theta)\cdots p_n(x_n|\theta) \\
&\propto p(\theta|x_1)p_2(x_2|\theta)\cdots p_n(x_n|\theta) \\
&\propto p(\theta|x_1, x_2)p_3(x_3|\theta)\cdots p_n(x_n|\theta) \\
&\vdots \\
&\propto p(\theta|x_1, \ldots, x_{n-1})p_n(x_n|\theta)
\end{aligned}
$$

- The concepts of prior and posterior are relative to the observation that is being considered.
- $p(\theta|x_1)$ is the posterior distribution of θ with respect to x_1 but,
- It is the prior distribution of θ with respect to $x_2, \ldots, x_n$ (before they are observed).

The Likelihood Principle

The following example (DeGroot, 1970, pages 165–166) illustrates this property.

Imagine that each item from a population of manufactured items is classified into either defective or nondefective. The proportion θ of defective items in the population is unknown and a sample of items will be selected according to one of the following methods.

- n items will be selected at random.
- Items will be selected at random until y defective are obtained.
- Items will selected at random until the inspector is called to solve another problem.
- Items will be selected at random until the inspector decides that enough information about θ has been gathered.

Whatever sampling scheme is chosen, if n items $x_1, \cdots, x_n$ are inspected y of which are defective, then

$$p(x|\theta) \propto \theta^y (1-\theta)^{n-y}.$$

The Likelihood Principle postulates that in order to make inferences about a parameter θ it only matters what was really observed and not what could have occured but has not.

To sum up

- Bayesian statistics follows the rules of probability.
- Bayesian statistics is based on a single tool, the Bayes theorem.
- Finding the posterior distribution using Bayes theorem is easy in theory, but generally hard in practice.

Model Uncertainty

Suppose there are different competing models which can be enumerated and represented by a set $M = \{M_1, M_2, \ldots\}$. We assume that the true model is in M.

- *a priori* we assign probabilities $p(M_i)$ to each model.
- For each model there is a vector of parameters $\boldsymbol{\theta}_i \in \mathbb{R}^{n_i}$ with,

  a prior distribution $p(\boldsymbol{\theta}_i|M_i)$, and
  a likelihood function given the observations $\mathbf{x}$, $p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)$.

Applications of Bayes theorem

- Within-model posterior,

$$p(\boldsymbol{\theta}_i|x, M_i) = \frac{p(x|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)}{p(\mathbf{x}|M_i)}$$

- Within-model marginal likelihood,

$$p(\mathbf{x}|M_i) = \int p(\mathbf{x}|\boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i$$

- Joint posterior distribution,

$$\pi(M_i, \boldsymbol{\theta}_i) = \frac{p(x|\boldsymbol{\theta}_i, M_i)\ p(\boldsymbol{\theta}_i|M_i)\ p(M_i)}{\sum_{M_i \in M} \int p(x|\boldsymbol{\theta}_i, M_i)\ p(\boldsymbol{\theta}_i|M_i)\ p(M_i) d\boldsymbol{\theta}_i}$$

- Posterior model probabilities,

$$p(M_i|\mathbf{x}) = \frac{p(\mathbf{x}|M_i)p(M_i)}{\sum_{M_j \in M} p(\mathbf{x}|M_j)\ p(M_j)}$$

- Overall prior predictive distribution,

$$p(\mathbf{x}) = \sum_{M_j \in M} p(\mathbf{x}|M_j)\ p(M_j)$$

Pairwise comparison of models

The posterior odds of model M_i relative to M_j is given by,

$$\underbrace{\frac{p(M_i|\mathbf{x})}{p(M_j|\mathbf{x})}}_{\text{Posterior Odds}} = \underbrace{\frac{p(\mathbf{x}|M_i)}{p(\mathbf{x}|M_j)}}_{\text{Bayes Factor}} \times \underbrace{\frac{p(M_i)}{p(M_j)}}_{\text{Prior Odds}}$$

Posterior model probabilities can be obtained as,

$$p(M_i|\mathbf{x}) = \left[\sum_{j=1}^{K} B_{ji}\frac{p(M_j)}{p(M_i)}\right]^{-1}$$

where $B_{ji} = \dfrac{p(\mathbf{x}|M_j)}{p(\mathbf{x}|M_i)}$.

Searching for the "Best" Model(s)

- How to compare competing models?
- What if the number of alternative models is quite large? E.g. linear model with 19 possible covariates: $2^{19} = 524288$ alternative models (with no interations).
- Enumerate, estimate and associate a measure of fit and parsimony to each possible model may not be the best strategy.
- How to make average inference using the competing models (or a subset of this)?

Bayes factor to compare models

Some rules of thumb to decide between models j and k based on Bayes factors.

Jeffreys (1961) recommendations.

$\log_{10} B_{jk}$	B_{jk}	Evidence against k
0.0 to 0.5	1.0 to 3.2	Not worth more than a bare mention
0.5 to 1.0	3.2 to 10	Substantial
1.0 to 2.0	10 to 100	Strong
> 2	> 100	Decisive

Kass and Raftery (1995) recommendation.

$2 \ln B_{jk}$	B_{jk}	Evidence against k
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Substantial
6 to 10	20 to 150	Strong
> 10	> 150	Decisive

Rationale: $2 \ln B_{jk}$ is on the same scale as the deviance and likelihood ratio test statistics.

The Marginal Likelihood

For a model M, recall that the predictive distribution of $\mathbf{x}$ is given by,

$$\begin{aligned} p(x|M) &= \int p(x|\theta, M)p(\theta|M)d\theta \\ &= E_\theta[p(x|\theta, M)] \end{aligned}$$

which is the normalizing constant in the posterior distribution.

This predictive density can now be viewed as the likelihood of model M (or marginal likelihood) and is a basic ingredient for model assessment.

Bayesian Computation

After observing the data, $p(\theta|x)$ summarizes all we know about θ.

Most features of the posterior distribution have the form of an expectation,

$$E[g(\theta)|\mathbf{x}] = \int g(\theta)p(\theta|\mathbf{x})d\theta.$$

Also, if $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ then,

$$p(\boldsymbol{\theta}_1|\mathbf{x}) = \int p(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}_2.$$

Some examples,

- Normalizing constant. $g(\theta) = 1$ and $p(\theta|\mathbf{x}) = kq(\theta)$, it follows that,
$$k = \left[\int q(\theta)d\theta\right]^{-1}.$$
- If $g(\theta) = \theta$, then $\mu = E(\theta|\mathbf{x})$ is the posterior mean.
- When $g(\theta) = (\theta - \mu)^2$, then $\sigma^2 = E((\theta - \mu)^2|\mathbf{x})$ is the posterior variance.
- If $g(\theta) = I_A(\theta)$, where $I_A(x) = 1$ if $x \in A$ and zero otherwise, then
$$P(A \mid \mathbf{x}) = \int_A p(\theta|\mathbf{x})d\theta.$$
- If $g(\theta) = p(y|\theta)$, where $y \perp \mathbf{x}|\theta$ we obtain $E[p(y|\mathbf{x})]$, the predictive distribution of a future observation y.

- In most interesting applications $E[g(\theta)|\mathbf{x}]$ cannot be worked out analytically.
- Unless otherwise noted, we assume that $E[g(\theta)|\mathbf{x}]$ exists.
- Exceptions which do fall in this framework are: the marginal likelihood and quantiles of the posterior distribution.