

Bayesian Statistics in Neuroimaging

Martin Lindquist
Department of Biostatistics
Johns Hopkins University

Bayesian Inference



- Most statistical methods covered in introductory statistics courses are **frequentist** (or classical) methods.
- **Bayesian inference** is an alternative approach that provides a somewhat different perspective.
- Bayesian methods are becoming increasingly common in the neuroimaging literature.

Classical vs Bayesian Approach



- The frequentist (or classical) point of view:
 - Probabilities describe long run relative frequencies.
 - Parameters are fixed unknown constants. Because they do not fluctuate, no useful probability statements can be made about them.
 - Statistical procedures should be designed to have well-defined long run frequency properties
 - Hypothesis tests and confidence intervals.

Classical vs Bayesian Approach



- The Bayesian point of view:
 - Probabilities describe a degree of belief.
 - Probability statements can be made about parameters, even though they are fixed constants.
 - Inferences are made about a parameter θ by producing a probability distribution for it.

The Bayesian Method

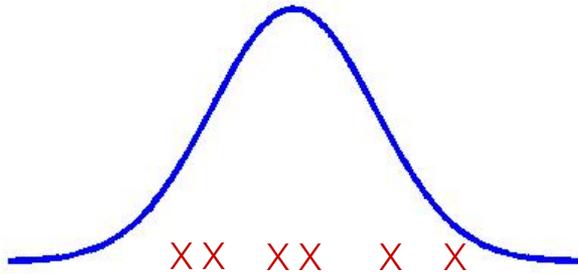


- Choose a probability density $p(\theta)$, the **prior distribution**, that expresses our beliefs about a parameter θ before we see any data.
- Choose a statistical model $p(y|\theta)$, the **likelihood**, that reflects our belief about the data y given θ .
- After observing y , update our beliefs and calculate the **posterior distribution** $p(\theta|y)$.

Example



- Let y_1, \dots, y_n be observations from a $N(\theta, \sigma^2)$ distribution, with θ unknown and σ^2 known.



Ex. θ is the task-induced change in brain activity and y_i the equivalent contrast for subject i .

Likelihood: $p(y_1, \dots, y_n | \theta) \sim N(\theta, \sigma^2)$

- Suppose we want to estimate the parameter θ .
 - A frequentist would use the sample mean.

Prior Distribution



- In the **Bayesian approach** θ can be described by a probability distribution (**Prior Distribution**).
- The prior distribution is a subjective distribution, based on the experimenter's belief and is formulated prior to viewing the data.
- In our example assume $p(\theta) \sim N(\mu_0, \tau_0^2)$

Posterior Distribution



- After a sample is taken from a population, the prior distribution can be updated using the information contained in the sample.
 - The updated prior is called the **posterior distribution**.
- Updating is done using **Bayes Rule**:

$$p(\theta | y) = \frac{p(y | \theta)p(\theta)}{p(y)}$$

Posterior Distribution



- Note that $p(y)$ does not depend on θ .
- Hence, the posterior density is often written:

$$p(\theta | y) \propto p(y | \theta) p(\theta)$$

Posterior Likelihood Prior

Posterior Inference



- The posterior distribution contains all current information about the parameter θ .
- Numerical summaries (e.g., mean, median, mode) of the distribution are used to obtain point estimates of the parameter.
- We can also make probability statements about the parameter of interest (e.g., $P(\theta > c | y)$) and create posterior intervals.

Example



- Let y_1, \dots, y_n be observations from a $N(\theta, \sigma^2)$ distribution, with θ unknown and σ^2 known.
- Suppose we take the prior distribution of θ to be $N(\mu_0, \tau_0^2)$ for some choice of μ_0 and τ_0^2 .

$$p(\theta | y_1, \dots, y_n) \propto p(y_1, \dots, y_n | \theta) p(\theta)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2\right\} \exp\left\{-\frac{1}{2\tau_0^2} (\theta - \mu_0)^2\right\}$$

Example



- It can be shown that

$$\theta \mid y_1, \dots, y_n \sim N(\mu_n, \tau_n^2)$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{n}{\sigma^2} \bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

- Note: $\mu_n = w\mu_0 + (1-w)\bar{y}$

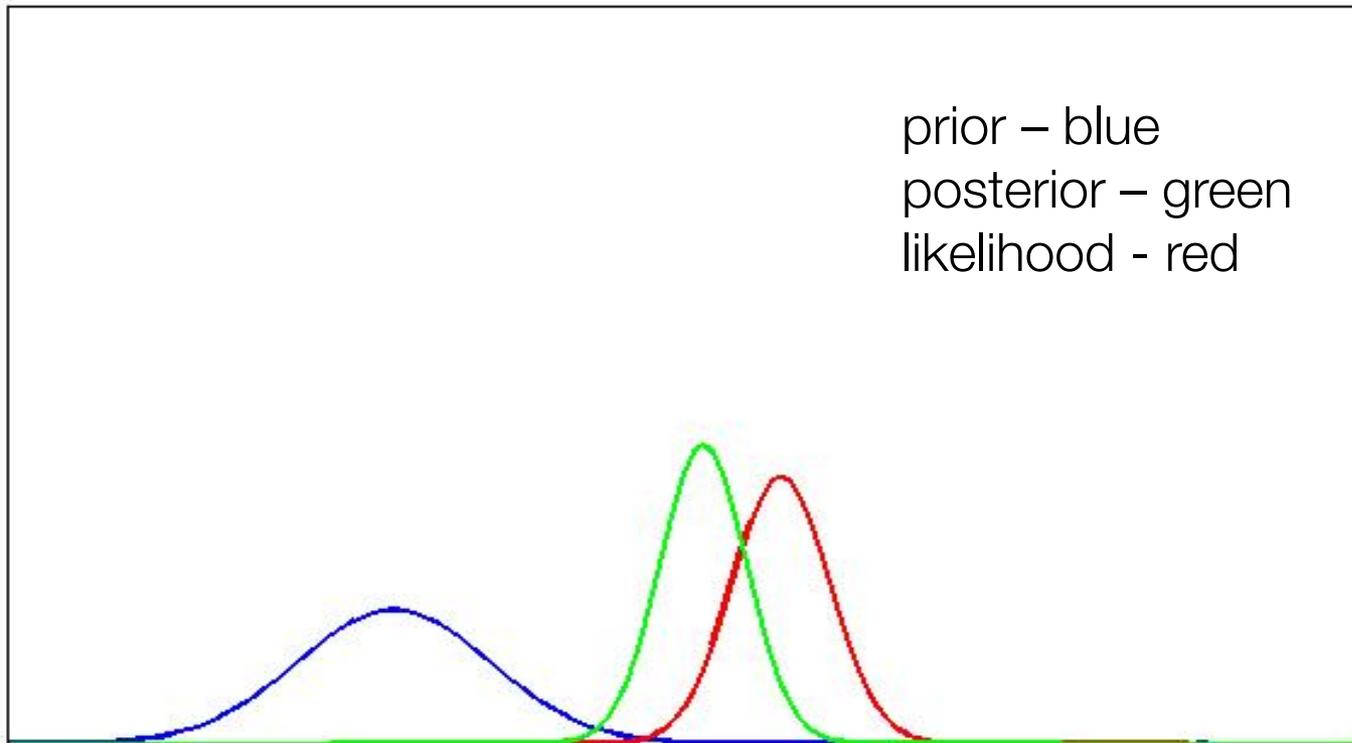
Can use μ_n to estimate θ .

Precision

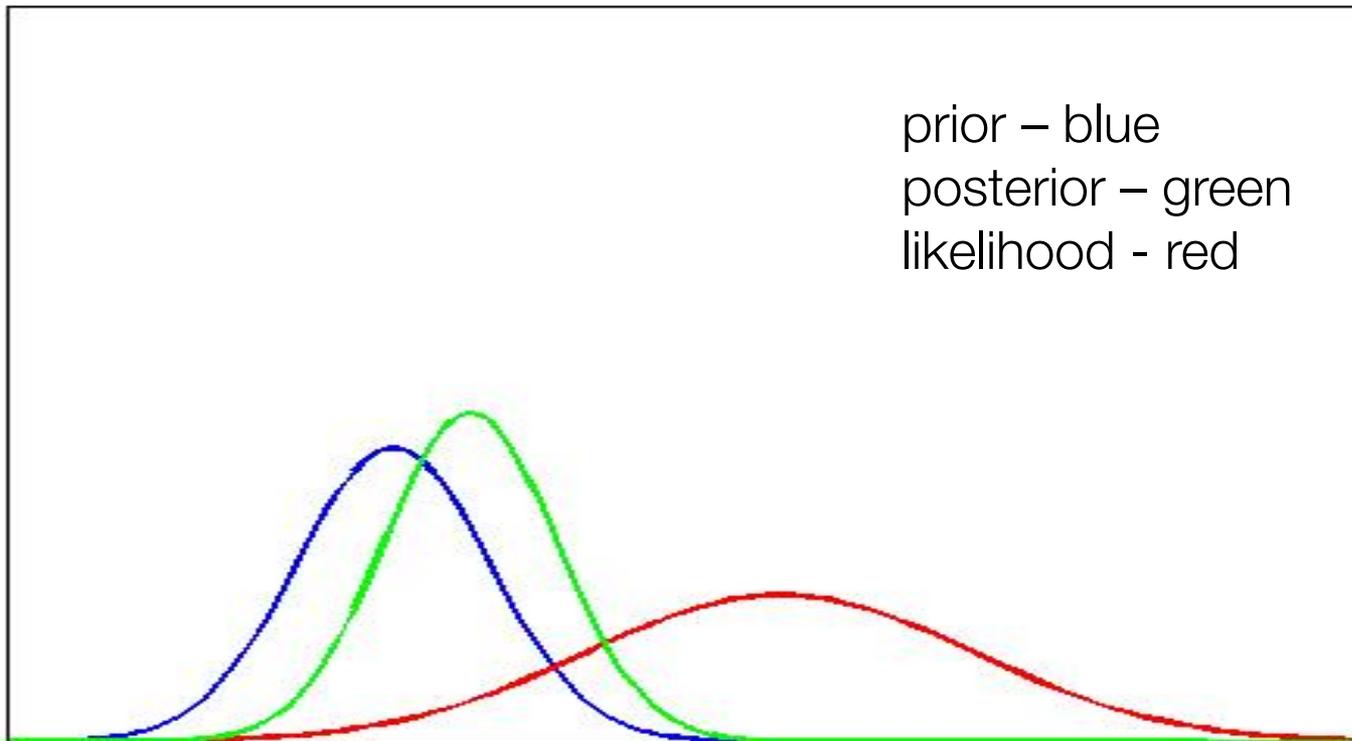


- The inverse of the variance is called the **precision**.
- The posterior mean is expressed as a weighted average of the prior mean and sample mean.
- The weights are proportional to the precisions.
- The posterior mean lies between the prior mean and the sample mean.

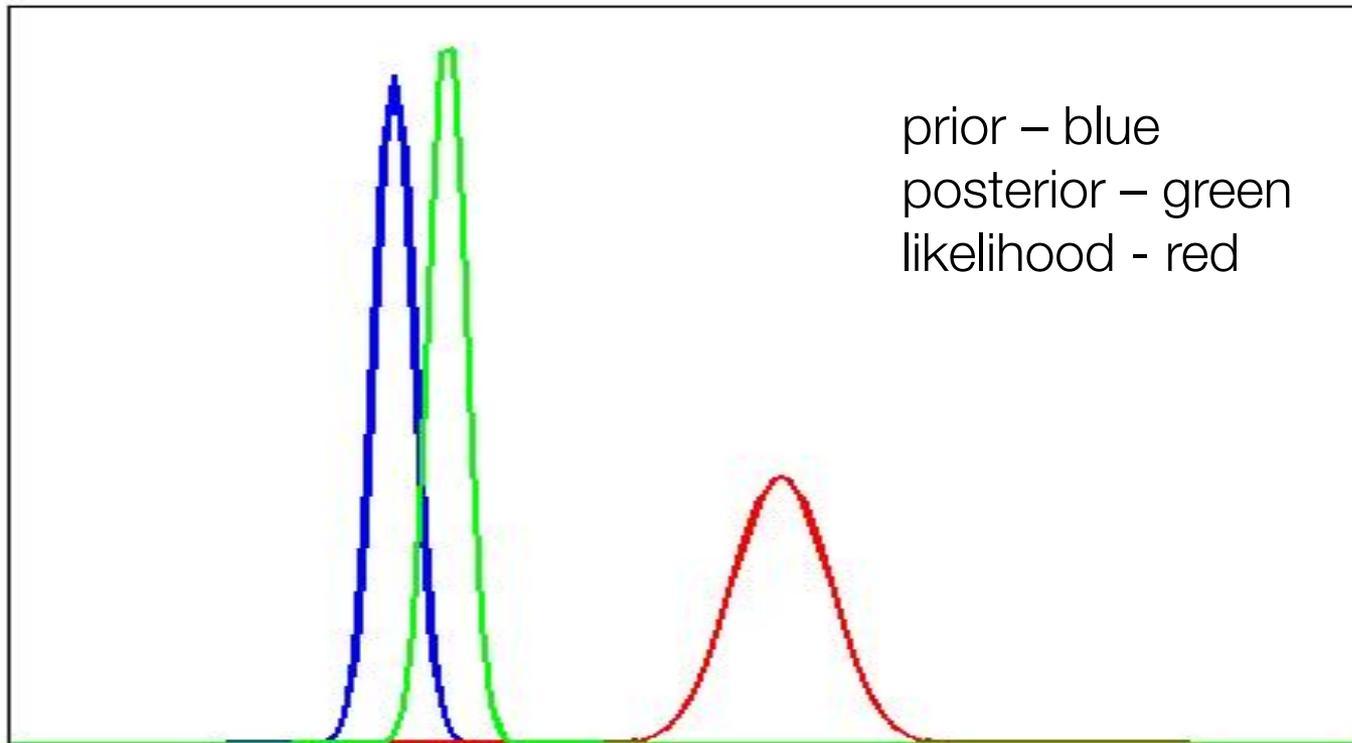
Illustration



Illustration



Illustration



Priors



- In the Bayesian framework the choice of prior is crucial.
- If no prior information about the parameters are available, **non-informative priors** can be used.
 - These types of priors let the data ‘speak for itself’ .
- One can also choose the priors in such a way that the posterior lies in the same family of distributions as the prior (**conjugate priors**).

Posterior Computation



- The posterior distribution is the basis for all Bayesian inference.
- Even if the posterior is known, it is sometimes difficult to obtain exact values of certain posterior quantities.
- By generating random samples from the posterior, all quantities of interest can be approximated using [Monte Carlo methods](#).

Monte Carlo Method



- Generate an i.i.d sequence $\theta_1, \dots, \theta_N$ from the posterior distribution of θ .
- Compute numerical summarizes of the distribution:

$$E(\theta | y) \approx \bar{\theta} = \frac{1}{N} \sum_{i=1}^N \theta_i$$

$$\text{Var}(\theta | y) \approx \frac{1}{N-1} \sum_{i=1}^N (\theta_i - \bar{\theta})^2$$

$$P(\theta \leq c | y) \approx \frac{\#(\theta_i \leq c)}{N}$$

Bayesian Computations



- Sampling from the posterior is effective when it can be implemented.
- However, it is often difficult in practice.
- For most probability distributions there is no simple way to simulate random variables of that particular distribution.



- Markov-chain Monte-Carlo (MCMC) is a method for sampling from a posterior distribution.
 - A Markov chain is generated that has the desired distribution as its **stationary distribution**.
 - The state of the chain after a large number of steps is used as a sample from the desired distribution.
 - Can be extremely computationally expensive.

Variational Bayes



- Variational Bayes (VB) is an approach towards approximating the posterior density which is less computationally intensive than MCMC.
 - Received a lot of attention in fMRI research.
 - It allows one to approximate the posterior density with another density that has a more analytically tractable form.



- Consider the standard GLM:

$$Y = X\beta + \varepsilon \quad \varepsilon \sim N(0, V)$$

- Suppose we place a prior on β , e.g.

$$\beta \sim N(\beta_0, \Sigma_0)$$

- It can be shown that the posterior distribution of β follows a normal distribution.
 - This distribution can be used to perform inference.



- The posterior mean provides an estimate of β :

$$\hat{\beta} = (X^T V^{-1} X + \Sigma_0^{-1})^{-1} (X^T V^{-1} y + \Sigma_0^{-1} \beta_0)$$

- If Σ_0 large then, $\hat{\beta} \approx (X^T V^{-1} X)^{-1} X^T V^{-1} y$

GLS estimate

- If $\beta_0=0$ then, $\hat{\beta} = (X^T V^{-1} X + \Sigma_0^{-1})^{-1} X^T V^{-1} y$

Shrinkage

Multilevel GLM



$$Y = X\beta + \varepsilon \quad \varepsilon \sim N(0, V)$$

$p(y|\beta) = N(X\beta, V)$ represents variability within a subject.

$$\beta = \beta_g + \eta \quad \eta \sim N(0, \Sigma)$$

$p(\beta|\beta_g) = N(\beta_g, \Sigma)$ represents variability across subjects

$$\beta_g \sim p(\beta_g)$$

$p(\beta_g)$ represents information about a fixed but unknown quantity.

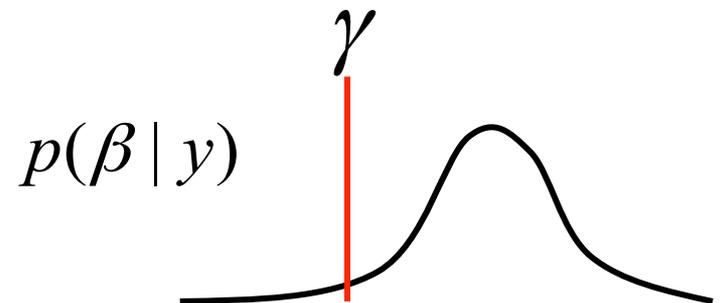


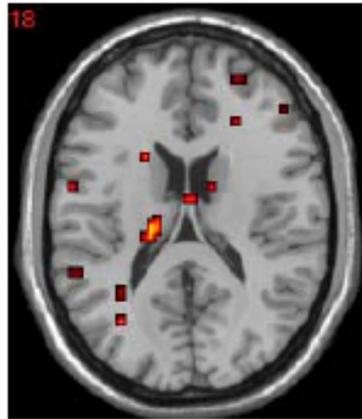
- It is common in neuroimaging to use so-called **empirical Bayes** methods.
 - Parameters of the prior estimated directly from the data, rather than being subject to prior specification of their own as in a fully Bayesian model.
- In the multilevel setting we often estimate β_g using the group mean.
 - Shrinks subject-specific estimates towards the group estimate.
 - The greater the uncertainty, the more shrinkage.
 - The less the uncertainty, the more we trust that individual estimate and the less it gets shrunk.

Posterior Probability Maps



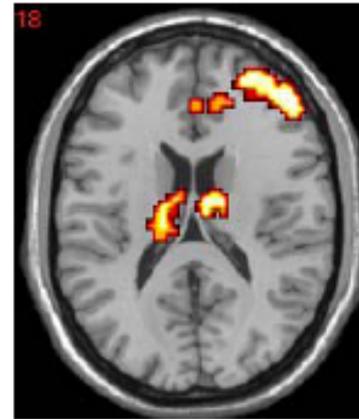
- The Posterior distribution $p(\beta|y)$ describes the probability of getting an effect, given the data.
- Posterior probability maps
 - Images of the probability that an activation exceeds some specified value γ , given the data, thresholded at some value α , i.e. $p(\beta > \gamma | y) > \alpha$
 - Question remains how to choose γ and α .





Frequentist

Thresholded t-statistic map
($p=0.005$, uncorrected)



Bayesian PPM

Voxels with probabilities of
task-related increases in
activity exceeding $\alpha=0.85$.

Bayesian Spatial Hierarchical Model
[Bowman et al., 2008, *NeuroImage*.]

Comments



- Frequentist and Bayesian methods are answering different questions.
- To combine prior beliefs with data in a principled manner use Bayesian inference.
- To construct procedures with guaranteed long run performance use frequentist methods.

Hypothesis Testing



- In classical hypothesis testing we seek to determine whether we can reject a null hypothesis of no effect.
 - The p -value is the probability of obtaining a result as or more extreme under the assumption that the null hypothesis is true.
 - We can never accept the null hypothesis.
 - Given enough data every voxel would be significant.
- Bayesian methods allow us to derive probabilities about hypothesis of interest.
 - Not restricted to disproving the null hypothesis.
 - It may be more interesting to compute the probability of some hypothesis, than to disprove a hypothesis of no effect.
 - Does not necessarily avoid problems with multiple comparisons.

Model Comparison



- Model comparison can be performed to determine whether the data favors one model over another.
- The **model evidence** is defined as

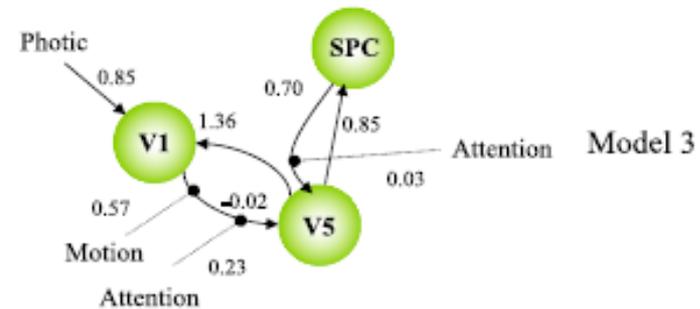
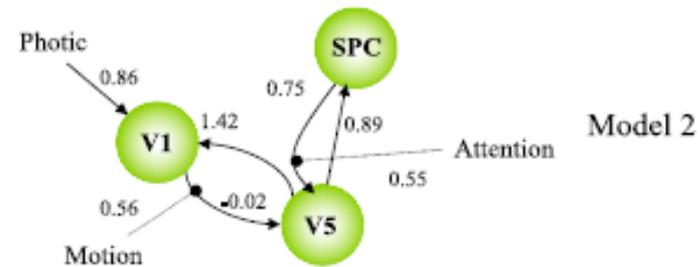
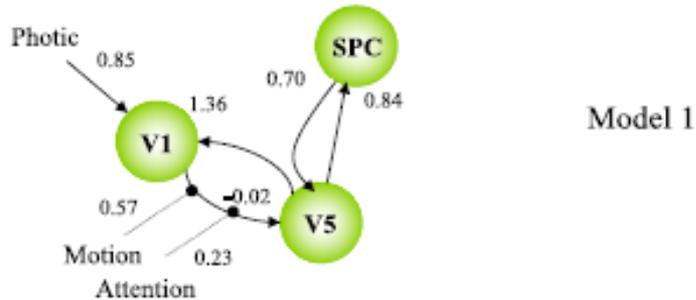
$$p(y | m) = \int p(y | \theta, m) p(\theta | m) d\theta$$

- The **Bayes factor** for comparing model i to j :

$$B_{ij} = \frac{p(y | m = i)}{p(y | m = j)}$$

If B_{ij} is large than i more likely than j .

Example



Use Bayes factors to compare three different candidate DCMs.

Table 6
Attention data—comparing modulatory connectivities

	B_{12}	B_{13}	B_{32}
AIC	3.56	2.81	1.27
BIC	3.56	19.62	0.18

Bayes factors provide consistent evidence in favor of the hypothesis embodied in model 1, that attention modulates (solely) the bottom-up connection from V1 to V5. Model 1 is preferred to models 2 and 3.

Summary



- Bayesian methods have made a large impact on neuroimaging in the past decade.
- They allow us to calculate the probability that an activation exceeds some specific threshold, given the data.
 - Not restricted to disproving a null hypothesis.
- Requires specifying prior distributions and can be computationally expensive.