Bayesian Statistics in Neuroimaging

Martin Lindquist
Department of Statistics
Columbia University
Bayesian Inference

• Most statistical methods covered in introductory statistics courses are so called frequentist (or classical) methods.

• **Bayesian inference** is another approach that provides a somewhat different perspective.

• Bayesian methods have recently received a great deal of attention in fMRI research.
Classical vs Bayesian Approach

• The frequentist (or classical) point of view:

  – Probability refers to limiting 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.
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 $\psi$ 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 $\theta$ before we see any data.

• Choose a statistical model $p(y|\theta)$, the likelihood, that reflects our belief about $y$ given $\theta$.

• After observing $y$, update our beliefs and calculate the posterior distribution $p(\theta|y)$. 

Example

- Suppose we observe a sequence of observations $y_1, \ldots, y_n$ from a $N(\mu, \sigma^2)$ distribution, where $\mu$ is unknown and $\sigma^2$ known.

Assume $\mu$ is the task-induced change in brain activity and $y_i$ is the equivalent contrast for subject $i$.

**Likelihood:** $p(y_1, \ldots, y_n | \mu) = N(\mu, \sigma^2)$

- We are interested in estimating the parameter $\mu$.
Prior Distribution

• In the Bayesian approach \( \mathcal{W} \) 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 \( p(\mathcal{W}) = \mathcal{N}(\mu_0, \sigma_0^2) \)
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:

\[ p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \]
• Note that $p(y)$ does not depend on $\theta$.

• Hence, the posterior density is often written:

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$$
Posterior Inference

• The posterior distribution contains all current information about the parameter $\theta$.

• 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 and create posterior intervals.
Example

- Let $y_1,\ldots,y_n$ be observations from a $\mathcal{N}(\mu, \sigma^2)$ distribution, with $\mu$ unknown and $\sigma^2$ known.

- Suppose we take the prior distribution of $\mu$ to be $\mathcal{N}(\mu_0, \sigma_0^2)$ for some choice of $\mu_0$ and $\sigma_0^2$.

$$p(\theta \mid y_1, \ldots, y_n) \propto p(y_1, \ldots, y_n \mid \theta)p(\theta)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2\right\} \exp\left\{-\frac{1}{2\tau_0^2} (\theta - \mu_0)^2\right\}$$
• It can be shown that $$\mathcal{W} \mid y_1, \ldots, y_n \sim N(\mathcal{W}_n, \mathcal{W}_n^2)$$

where

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

and

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

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

• The inverse of the variance is called the precision.

• The posterior mean is expressed as a weighted average of the prior mean and the sample mean.

• The weights are proportional to the precisions.

• The posterior mean lies between the prior mean and the sample mean.
Illustration

- prior – blue
- posterior – green
- likelihood - red
Illustration

prior – blue
posterior – green
likelihood - red
Illustration

prior – blue
posterior – green
likelihood - red
In the Bayesian framework the choice of prior is crucial.

If we have no prior information about the parameters, we may use non-informative priors. 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).
• The posterior distribution is the basis for all Bayesian inference.

• Even if the posterior is known, it can be difficult to obtain exact values of certain posterior quantities (e.g., $E(\mathbb{X}_1/\mathbb{X}_2)$).

• By generating random samples from the posterior, all quantities of interest can be approximated using Monte Carlo methods.
Monte Carlo Method

• Let $g(\theta)$ be some function of $\theta$ (e.g., $\log(\theta)$).

• Suppose we want to estimate $E(g(\theta)|y)$.

• Generate an i.i.d sequence $\theta_1, \ldots, \theta_N$ from the posterior distribution of $\theta$.

• Estimate $E(g(\theta)|y)$ using $\bar{g} = \frac{1}{N} \sum_{i=1}^{N} g(\theta_i)$.
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.
GLM with Priors

• Consider the standard GLM:

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

• Suppose we place a prior on \( \beta \), e.g.

\[ \beta \sim N(\beta_0, \Sigma_0) \]

• It can be shown that the posterior distribution of \( \beta \) follows a normal distribution. We can use this distribution to perform inference.
• The posterior mean provides a point estimate of $\mathbf{\beta}$:

$$\hat{\mathbf{\beta}} = (X^T V^{-1}X + \Sigma_0^{-1})^{-1}(X^T V^{-1}y + \Sigma_0^{-1}\mathbf{\beta}_0)$$

• If $\Sigma_0$ is large then $\hat{\mathbf{\beta}} \approx (X^T V^{-1}X)^{-1}X^T V^{-1}y$

  GLS estimate

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

  Shrinkage
Posterior Probability Maps

• The Posterior distribution is the probability of getting an effect, given the data $p(\mathbf{X} | y)$.

• Posterior probability maps are images of the probability or confidence that an activation exceeds some specified threshold, given the data $p(\beta > \gamma | y) > \alpha$.
**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.*]
Frequentist and Bayesian methods are answering different questions.

To combine prior beliefs with data in a principled way use Bayesian inference.

To construct procedures with guaranteed long run performance use frequentist methods.
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 allows us to derive probabilities about hypothesis of interest.
  - Not restricted to disproving the null hypothesis.
  - One can argue that it is more interesting to prove a hypothesis, than disprove a hypothesis of no effect.
Multilevel Models

• Data sets where there is a hierarchy of nested populations are often called multilevel.
  – Voxels nested within subjects nested within groups.

• Multilevel models are extensions of regression in which data are structured in groups and coefficients can vary by group.
  – Allows information to be shared across groups.
Y = Xβ + ε  \quad ε \sim N(0, V)

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

β = β_g + η  \quad η \sim N(0, Σ)

p(β|β_g) = N(β_g, Σ) represents variability across subjects

β_g \sim p(β_g)

p(β_g) represents information about a fixed but unknown quantity.
Illustration

\[ N(\beta_g, \sigma_g^2) \]

\[ \beta_g \sim p(\beta_g) \]
Empirical Bayes

• It is common in neuroimaging to use so-called empirical Bayes methods.

• Here the parameters of the prior are estimated directly from the data, rather than being subject to prior specification of their own as is the case in a fully Bayesian model.
• Multilevel models allow for heterogeneity across subjects, but still consider values observed in other subjects.

• Each subject-specific estimate gets shrunk towards the overall 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.
Example

- Radon levels of houses in 85 counties in Minnesota.

- Most differences no longer statistically significant

Gelman and Hill, 2007
Model Comparison

• Model comparison can be performed to determine whether the data favors one model over another.

• The model evidence is defined as

\[ p(y \mid m) = \int p(y \mid \theta, m) p(\theta \mid m) d\theta \]

• The Bayes factor for comparing model i to j:

\[ B_{ij} = \frac{p(y \mid m = i)}{p(y \mid m = j)} \]

If \( B_{ij} \) is large than i more likely than j.
Example

Use Bayes factors to compare three different candidate models.
Bayesian methods are gaining in popularity in fMRI research.

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.