Tutorial on Bayesian Analysis (in Neuroimaging)

Christopher R. Genovese

Department of Statistics

Carnegie Mellon University

http://www.stat.cmu.edu/~genovese/

IPAM 20 July 2004 UCLA

Modeling Preliminaries

- Consider the simple, fixed-effects linear model

$$y = X\beta + \epsilon, \qquad (*)$$

  where X is $n \times p$, β is $p \times 1$ and unknown, and ϵ is an $n \times 1$ vector of independent Normal$(0, \sigma^2)$ random variables.

  This statement embodies an assumption that the observed data y have a Normal distribution with mean $X\beta$ and variance matrix $\sigma^2 I$ for *some* value of the quantities $\theta = (\beta, \sigma^2)$.

- In abstract terms, a *statistical model* is an indexed collection of probability distributions $f_\theta(y)$ for data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$.

  The index $\theta \in \Theta$ is called a *parameter* and may be finite or infinite-dimensional.

Modeling Preliminaries (cont'd)

The collection of possible parameters Θ allowed under the model is called the *parameter space*.

The model carries with it an assumption that $\boldsymbol{Y}$ has distribution f_θ for some θ.

- For whatever value y of $\boldsymbol{Y}$ we have observed, the function that maps $\theta \in \Theta$ to $f_\theta(y)$ is called the *likelihood function*. It is the most important object in statistical inference.

  In the linear model (*), the likelihood is given by

$$L(\beta, \sigma^2) = C \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2}\|y - X\beta\|^2\right).$$

  Note that the *data are fixed* here (and not even denoted on the left) and that we don't care about multiplicative constants. It is often convenient to work with the log of the likelihood.

Road Map

1. Philosophy (briefly)
2. Theory
3. Computation
4. Practice

Road Map

1. Philosophy (briefly)
   - 1306 Flavors, Subjectively Speaking
   - Why Bayes?
   - Why not Bayes?
2. Theory
3. Computation
4. Practice

1306 Flavors, Subjectively Speaking

To understand Bayesian inference, it helps to see how it differs from the *frequentist* (or classical) tradition and then in turn see how Bayesians differ amongst themselves.

The key distinctions lie in the definition and interpretation of probability.

(It's an especially slippery idea, and none of the approaches are entirely satisfactory.)

The Frequentist Tradition

For the frequentist, probability represents *limiting relative frequencies* over a series of (hypothetical) replications. Probabilities are objectively defined quantities. Classic example: coin flips.

This definition of probability has several implications:

- Probabilities can only stated for observations that are in principle replicable. (They need not actually be replicated, though.)
- Any parameters describing the probability distribution of a random quantity do not vary across replications. As such, no useful probability statements can be made about them.

  For example, we can't speak of the probability that a hypothesis is true.

- Statistical procedures should be chosen to have good long-run frequency performance.

  For example, a 95 percent confidence interval should cover the true value with limiting frequency 95 percent while being as short as possible *on average*.

The Bayesian Approach

For the Bayesian, probability represents a degree of belief. Probabilities are subjectively defined quantities.

This definition of probability has several implications:

- Probabilities can be stated for essentially any event, and relating to any quantity, random or non-random. We are using the classical calculus of probability without requiring "physical" randomness.
- In particular, we can make probability statements about parameters of interest in a statistical model.
- Statistical inference is a process by which the observed data update our beliefs.

  Our beliefs before seeing the data are described by a *prior distribution*; our beliefs after seeing the data are described by the *posterior distribution*.

  Bayesian inferences derive entirely from the posterior.

Illustration: Hypothesis Testing

Frequentist

- Based on the logic of surprise: if I see a result that is too surprising under my current world view, then I should change my view rather than think I'm that lucky. Rather convoluted.
- The threshold for surprise (significance level) is subjective but often chosen by convention (not always for the better).
- The quoted statistic, the p-value, is a very poor summary of evidence.
- Treats the candidate hypotheses differently (reject or retain H_0 but not accept!)
- Often difficult to test the hypotheses we want to test (ex: point null).

Bayesian

- Begin with prior beliefs expressed as $\mathsf{P}(H_0), \ldots, \mathsf{P}(H_m)$.
- Data update these prior beliefs to posterior beliefs $\mathsf{P}(H_0 \mid Y), \ldots, \mathsf{P}(H_m \mid Y)$.
- In two hypothesis case, can compute the Bayes factor B such that

$$\frac{\mathsf{P}(H_1 \mid Y)}{\mathsf{P}(H_0 \mid Y)} = B \frac{\mathsf{P}(H_1)}{\mathsf{P}(H_0)}.$$

  $\mathsf{P}(H_0 \mid Y)$ is what we would all like a p-value to be.

1306 Flavors (cont'd)

Bayesians are not monolithic. Among the most important points of disagreement is the question of how to choose a prior.

Two approaches in particular stand out.

- Subjective Bayes. The prior is "elicited" by careful assessment of belief and much hard work.
- Automatic Bayes. The prior is selected by a formal rule or algorithm.

Another issue is whether to consider frequentist performance in the selection of a prior. Some ignore it, others consider it.

Road Map

1. Philosophy (briefly)
   - 1306 Flavors, Subjectively Speaking
   - Why Bayes?
   - Why not Bayes?
2. Theory
3. Computation
4. Practice

Why Bayes?

1. Logic

- A single unified method applies to all problems.
- Uncertainty rather than randomness is the central focus.
- The data are used to obtain direct quantitative inferences about the parameters.
- Only concerned with the data in hand – not other possible data sets that could have been observed but weren't.
- All assumptions clearly stated up front and can be easily compared.

  Example: Inference about $\sum_{k=1}^{\infty} c_k z^k$.

Why Bayes? (cont'd)

2. Performance

Consider a simple version of the linear model (*):

$$y = \beta + \epsilon,$$

where $n = p \equiv \dim(\beta) \geq 3$.

The standard estimator (MLE, least squares) $\widehat{\beta} = y$ can be improved upon in mean squared error for every value of β, with, for instance, is the famous James-Stein estimator.

This extends to the full linear model (*) when $p \geq 3$. Put another way, the standard estimators in the linear model are known to be sub-optimal in an important sense, often nontrivially.

Estimators derived from the mean of a Bayesian posterior distribution do not have this problem.

In finite-dimensional problems, Bayesian estimators often have good frequentist performance.

Why Bayes? (cont'd)

3. "Coherence"

A small set of seemingly common sense principles for rational decision making, if accepted, lead inexorably to the Bayesian approach.

These also imply that two experiments with proportional likelihoods should give the same inference, which is not true of classical methods. (See Berger & Wolpert 1984 for details about these principles.)

This is a powerful argument for the subjective Bayesian paradigm, but it has important limits.

The claim that this requires a Bayesian approach – as one colleague puts it, "Either you're a Bayesian or you're a loser" – is easily (and often) oversold.

Why Bayes? (cont'd)

4. Flexibility

In order to use a statistic for inference in frequentist methods, it's necessary to compute the statistic's sampling distribution and relate that to the parameters of interest.

Even in conceptually simple situations, this can be difficult, sometimes prohibitively.

General Examples: truncated parameter ranges (bounded normal mean), discrete model components (variable selection), clustering (confidence statements concerning active regions in an image).

But for Bayesian methods, we need only be able to compute the likelihood and prior up to constants of proportionality. This is easy for a larger set of problems, and it's all handled using exactly the same approach.

Road Map

1. Philosophy (briefly)
   - 1306 Flavors, Subjectively Speaking
   - Why Bayes?
   - Why not Bayes?
2. Theory
3. Computation
4. Practice

Why Not Bayes?

What standards of performance should we demand of Bayesian procedures?

To the subjective Bayesian, the posterior is "right" whatever the prior, and the resulting inferences are not bound by frequentist notions of performance.

Freedman (1999) constructs an example in high dimensions – with a "reasonable" prior – in which a set with high posterior probability has arbitrarily small frequentist probability of containing the truth.

Should the likelihood alone be the basis of inference?

Robins and Ritov (1997) construct an example in which likelihood-based methods must produce useless inferences but for which another method performs well.

They argue that the "Coherence" principles need not make sense in high dimensions.

In high dimensional and nonparametric problems, Bayesian procedures often have poor frequentist performance. This forces the philosophy upon us.

Road Map

1. Philosophy (briefly)
   - 1306 Flavors, Subjectively Speaking
   - Why Bayes?
   - Why not Bayes?
2. Theory
3. Computation
4. Practice

Road Map

1. Philosophy (briefly)
2. Theory
   - Bayesian Inference
   - Posteriors and How To Use Them
   - The Impact of the Prior
   - Hierarchical Models
3. Computation
4. Practice

Bayesian Inference

The basic Bayesian method is the same in every problem:

1. Select a probability model $f(y \mid \theta)$ that reflects our beliefs about the data y for each value of the parameter. Note that this likelihood is now considered a conditional probability distribution, not just an indexed set of distributions.
2. Select a prior distribution $f(\theta)$ for the parameter.
3. Combine these to form a posterior distribution via Bayes Theorem:

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{\int f(y \mid \theta') f(\theta')\, d\theta'}. \qquad (*)$$

The first two steps specify a joint distribution for y and θ: $f(y, \theta) = f(y \mid \theta) f(\theta)$. The denominator in (*) is then $\int f(y, \theta')\, d\theta' = f(y)$.

Note that only the observed y is ever used and that $f(\theta \mid y)$ is proportional to likelihood times prior.

The integral in (*) can be very difficult to compute, but simulation-based methods exist for deriving posterior inferences without computing it explicitly.

Bayesian Inference: Examples

A simple example illustrates how Bayes theorem is used. In more complicated cases, the calculations are more challenging, but the idea is the same.

Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed (iid), each taking values 0 or 1 with probability $1-p$ or p respectively, as in the flips of a coin. Write $Y = (Y_1, \ldots, Y_n)$.

Likelihood: $f(Y \mid p) \propto p^{\sum_i Y_i}(1-p)^{n-\sum_i Y_i}$, for $0 < p < 1$.

Prior: Uniform distribution, $f(p) = 1$, $0 < p < 1$. (Note: this does *not* require that p be random.)

Then, $f(p \mid Y) \propto p^{\sum_i Y_i}(1-p)^{n-\sum_i Y_i}$, for $0 < p < 1$. By standard integrals, we can recover the constant of proportionality to get:

$$f(p \mid Y) = \frac{\Gamma(n+2)}{\Gamma(\sum_i Y_i + 1)\Gamma(n - \sum_i Y_i + 1)} p^{\sum_i Y_i + 1 - 1}(1-p)^{n-\sum_i Y_i + 1 - 1},$$

which is called a Beta$\langle \sum_i Y_i + 1, n - \sum_i Y_i + 1 \rangle$ distribution.

Bayesian Inference: Examples (cont'd)

If we were to then observe another datum Y_{n+1}, we would use that posterior as our new prior with likelihood $f(Y_{n+1} \mid p) = p^{Y_{n+1}}(1-p)^{1-Y_{n+1}}$. The posterior that results, Beta$\langle \sum_{i=1}^{n+1} Y_i + 1, n+1-\sum_{i=1}^{n+1} Y_i + 1\rangle$ is the same as if we had used $Y_1, \ldots, Y_{n+1}$ in the first place.

Suppose that we wanted to make inferences on the log odds ratio $\psi = \log(p/(1-p))$ instead of p. We can compute the posterior distribution of ψ directly from that for p:

$$\begin{aligned} H(u \mid Y) &\equiv \mathsf{P}\Big\{\psi \le u\Big\} \\ &= \mathsf{P}\Big\{\log(p/(1-p)) \le u \mid Y\Big\} \\ &= \mathsf{P}\left\{p \le \frac{e^u}{1+e^u} \mid Y\right\} \\ &= \int_0^{e^u/(1+e^u)} f(p \mid Y)\, dp. \end{aligned}$$

Bayesian Inference: Examples (cont'd)

The posterior density of ψ follows by taking derivatives with respect to u.

Contrast this with the frequentist case, where making inference about a function of your parameter requires essentially a seperate analysis.

In the linear model $y = X\beta + \epsilon$, functions of the parameter β include most quantities of interest: parameters β_j, contrasts $\sum c_j \beta_j$ with $\sum c_j = 0$, indicators that β_j is positive or that $\beta_j > \beta_k$, and so forth.

Bayesian Inference: Examples (cont'd)

That trick above of taking a flat prior $f(\theta) = 1$ seems pretty handy.

Suppose that $Y_1, \ldots, Y_n$ are iid Normal$(\theta, 1)$. If we take prior $f(\theta) = 1$, we get

$$\begin{aligned} f(\theta \mid Y) &\propto e^{-\frac{1}{2}\sum_i (Y_i - \theta)^2} \\ &\propto e^{-\frac{1}{2}(\theta^2 - 2n\theta\overline{Y})} \\ &\propto e^{-\frac{n}{2}(\theta - \overline{Y})^2}, \end{aligned}$$

which we easily recognize as a Normal$(\overline{Y}, \theta)$ distribution. (Notice how we useed our ability to drop constants – anything not depending on a parameter – to make this easier.)

So the posterior is centered around the Maximum Likelihood Estimator (MLE) $\overline{Y}$ with variance equal to the variance of the MLE. This looks like a frequentist result but with all the advantages of Bayesian inference!

Bayesian Inference: Examples (cont'd)

But beware. While this "flat prior Bayes" can sometimes produce frequentist-like results, it is not the cure-all it might seem.

For instance, a flat prior is not the "noninformative" choice that it appears. In particular, it is not invariant to change of variables.

In the binary variable example above, a flat prior on p gives

$$f(\psi) = \frac{e^{\psi}}{(1+e^{\psi})^2},$$

which is not flat. Kass and Wasserman (1999) show similar and more extreme examples.

The "Jeffrey's prior" is invariant. This takes $f(\theta) \propto \sqrt{I(\theta)}$ where $I(\theta)$ is the Fisher information for θ under the model.

In the binary example, the Jeffrey's prior is

$$f(p) \propto p^{-1/2}(1-p)^{-1/2},$$

which is invariant but rather extreme at the endpoints.

Bayesian Inference: Examples (cont'd)

Example: Suppose $\Theta = \Theta_0 \cup \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$ and that each Θ_i corresponds to a hypothesis H_i.

Our prior probabilities are $\mathsf{P}(H_i) = \int_{\Theta_i} f(\theta)$.

Let $f_i(y) = \int_{\Theta_i} f(y) f(y \mid \theta) f(\theta) / P(H_i)$ be the marginal distribution of the data under hypothesis H_i.

Then our posterior probabilities $P(H_i \mid y)$ can be written as follows:

$$\begin{aligned} \mathsf{P}(H_i \mid y) &= \frac{\int_{\Theta_i} f(y \mid \theta) f(\theta)}{\int f(y \mid \theta) f(\theta)} \\ &= \frac{f_i(y)\mathsf{P}(H_i)}{f_0(y)\mathsf{P}(H_0) + f_1(y)\mathsf{P}(H_1)}. \end{aligned}$$

Hence,

$$\frac{\mathsf{P}(H_1 \mid y)}{\mathsf{P}(H_0 \mid y)} = \frac{f_1(y)}{f_0(y)} \frac{\mathsf{P}(H_1)}{\mathsf{P}(H_0)}.$$

This factor $B = f_1(y)/f_0(y)$ is called the Bayes factor.

Road Map

1. Philosophy (briefly)
2. Theory
   - Bayesian Inference
   - Posteriors and How To Use Them
   - The Impact of the Prior
   - Hierarchical Models
3. Computation
4. Practice

Posteriors and How to Use Them

The posterior is the end product of Bayesian inference, but the posterior can be a rather bulky summary of one's results.

On the plus side, because the posterior is a probability distribution, we can use all the calculus of probability to manipulate it.

In particular, it is straightforward to compute the posterior of derived quantities.

For example, with a parameter $\theta = (\theta_1, \ldots, \theta_d)$, we can find the posterior of θ_j by "marginalizing out" the other values

$$f(\theta_j \mid Y) = \int f(\theta \mid Y) d\theta_1 \cdots d\theta_{j-1} d\theta_{j+1} \cdots d\theta_d.$$

Here are some common ways to use the posterior.

Posteriors and How to Use Them (cont'd)

- Point estimators. Under a mean-squared error criterion, the posterior mean is commonly used:

$$\widehat{\theta} = \int \theta f(\theta \mid Y)\, d\theta = \frac{\int \theta f(y \mid \theta) f(\theta)\, d\theta}{\int f(y \mid \theta) f(\theta)\, d\theta}.$$

- Functions of the parameter. Use the posterior distribution of a few specified functions of θ.
- Posterior intervals. Find intervals (or more general sets) C such that $\mathrm{P}(\theta \in C \mid Y) = 1 - \alpha$ for a specified α.
- Summarize. Approximate the posterior by a simpler distribution such as a Gaussian or a few component mixture.
- Simulations. Numerically simulate draws from the posterior. This is often necessary and can be combined with the previous approaches.

Road Map

1. Philosophy (briefly)
2. Theory
   - Bayesian Inference
   - Posteriors and How To Use Them
   - The Impact of the Prior
   - Hierarchical Models
3. Computation
4. Practice

The Impact of the Prior

Although there is much hand-wringing about choice of prior, the prior is asymptotically dominated by the data as the sample size grows, at least in finite dimensional problems.

THEOREM. Let $\widehat{\theta}_n$ be the MLE and let $\widehat{s}_n = (nI(\widehat{\theta}_n))^{-1/2}$. Under mild regularity conditions, including that the prior does not put zero probability on the truth, the posterior is approximately Normal$(\widehat{\theta}_n, \widehat{s}_n^2)$.

Also if $C_n = (\widehat{\theta}_n - z_{\alpha/2}\widehat{s}_n, \widehat{\theta}_n + z_{\alpha/2}\widehat{s}_n)$ is the asymptotic frequentist $1-\alpha$ confidence interval, then

$$\mathsf{P}\{\theta \in C_n \mid Y\} \to 1-\alpha,$$

as $n \to \infty$.

This need not be true in nonparametric (infinite-dimensional) problems. In finite samples, the choice of prior can have a large impact, and these approximations need not be especially good.

Road Map

1. Philosophy (briefly)
2. Theory
   - Bayesian Inference
   - Posteriors and How To Use Them
   - The Impact of the Prior
   - Hierarchical Models
3. Computation
4. Practice

Hierarchical Models

Constructing a prior for a high-dimensional parameter is difficult – there is a lot room in high-dimensional space.

But fortunately, because we are dealing with probability distributions, we have the freedom to specify the prior in more intuitive pieces.

There are two basic tricks:

1. Conditioning

   Suppose $\theta = (\theta_1, \ldots, \theta_m)$. The prior $f(\theta) \equiv f(\theta_1, \ldots, \theta_m)$ gives the joint distribution of these components. But by standard probability theory, we can write

   $$f(\theta_1, \ldots, \theta_m) = f(\theta_1)\, f(\theta_2 \mid \theta_1)\, f(\theta_3 \mid \theta_2, \theta_1) \cdots f(\theta_m \mid \theta_1, \ldots, \theta_{m-1}).$$

   That is, we can arrange the components in any order, and specify the distribution of each given some of the others. (The θ_is above can be vector blocks.) Other variants are possible.

Hierarchical Models (cont'd)

2. Hyperparameters

   Write the prior $f(\theta)$ as a mixture by *introducing* hyperparameters λ:

$$f(\theta) = \int f(\theta \mid \lambda) f(\lambda)\, d\lambda.$$

   The hyperparameters can be anything we specify.

By combining these two techniques, we get a *hierarchical model*.

Example: Normal-Normal

$$\begin{aligned} Y_i \mid \theta, \lambda &\leftarrow \text{Normal}(\theta, \sigma^2) \\ \theta \mid \lambda &\leftarrow \text{Normal}(0, \lambda^2) \\ \frac{1}{\lambda^2} &\leftarrow \text{Gamma}(a_0, b_0), \end{aligned}$$

where $a_0, b_0 > 0$ are fixed and pre-specified.

Hierarchical Models (cont'd)

This can be quite elaborate and flexible. Consider an $m \times m$ grid of sites $\mathcal{G}$, as on an image. Let $\mathcal{N}_{i,j}$ be the set of direct neighbors of the point (i, j). Formally.

$$\mathcal{N}_{i,j} = \{(k, \ell) \in \mathcal{G} \colon\ i = k \text{ and } |j - \ell| = 1 \text{ or } j = \ell \text{ and } |i - i| = 1\}\,.$$

We want a model for a smooth field θ on the grid $\mathcal{G}$. To specify a prior $f(\theta \mid \lambda)$ up to a constant, it is sufficient to define the conditional distributions $\theta_{i,j} \mid \theta_{-(i,j)}$.

An example, called a Markov Random Field, determines $f(\theta \mid \lambda)$ by

$$\theta_{i,j} \mid \theta_{-(i,j)}, \lambda \leftarrow \text{Normal}\left(\frac{\sum_{(k,\ell)\in\mathcal{N}_{i,j}} \theta_{k,\ell}}{\#\mathcal{N}_{(i,j)}}, \lambda^2\right).$$

This is a somewhat crude model, but it can be elaborated in interesting ways.

And we can use the methods coming up to simulate directly from the posterior.

Road Map

1. Philosophy (briefly)
2. Theory
   - Bayesian Inference
   - Posteriors and How To Use Them
   - The Impact of the Prior
   - Hierarchical Models
3. Computation
4. Practice

Road Map

1. Philosophy (briefly)
2. Theory
3. Computation
   - Simulation-Based Methods
   - Markov Chain Monte Carlo
   - Model Jumping and Averaging
4. Practice

Simulation-Based Methods

While applying Bayes theorem to get a posterior is conceptually simple, it can be computationally demanding.

The biggest difficulty lies in computing the "normalizing constant"

$$f(y) = \int f(y \mid \theta) f(\theta)\, d\theta,$$

which is often high-dimensional and complicated, making standard numerical integration problematic.

This is needed to compute the posterior probabilities or posterior means: e.g.,

$$\mathsf{P}(\mathcal{A} \mid Y) = \frac{\int_{\mathcal{A}} f(y \mid \theta) f(\theta)\, d\theta}{f(y)}.$$

Three approaches:

1. Basic Monte Carlo
2. Importance Sampling and its extensions
3. Markov Chain Monte Carlo

Importance Sampling

Suppose we want to compute the integral $I = \int h(x)f(x)\,dx$. If we could simulate draws from f, we could compute

$$\widehat{I} = \frac{1}{N}\sum_{j=1}^{N} h(X_j) \approx \mathsf{E}_f h(X) = I,$$

to estimate I. This is basic Monte Carlo.

But typically we will not know how to obtain draws from f. In *importance sampling*, we find a distribution g that we can draw from and draw N iid samples. We then compute

$$\widehat{I} = \frac{1}{N}\sum_{j=1}^{N} \frac{h(X_j)f(X_j)}{g(X_j)} \approx \mathsf{E}_g \frac{h(X)f(X)}{g(X)} = \mathsf{E}_f h(X) = I.$$

Importance Sampling (cont'd)

The basic rule of importance sampling is to sample from a density g with *thicker* tails than f. Otherwise, the estimate will have large variance, and may even blow up.

The choice of g that minimizes variance is

$$g_*(x) = \frac{|h(x)|f(x)}{\int |h(t)|f(t)\,dt}$$

Unfortunately, this is only of theoretical interest.

Making a good choice of g becomes challenging in high dimensions. Gelman and Meng (1998) give a nice review of importance sampling and its extensions.

But importance sampling has been mostly subsumed by Markov Chain Monte Carlo.

Road Map

1. Philosophy (briefly)
2. Theory
3. Computation
   - Simulation-Based Methods
   - Markov Chain Monte Carlo
   - Model Jumping and Averaging
4. Practice

Markov Chain Monte Carlo (MCMC)

A (discrete time) Markov Chain is a random process $X = (X_0, X_1, X_2, \ldots)$ with the so-called Markov property: at every time, the future is conditionally independent of the past given the present. Loosely: the distribution of X_{n+1} depends only on X_n.

Under certain conditions, a Markov Chain will settled down into an equilibrium, where the distribution of X_n approaches a fixed distribution π. Loosely: the Markov Chain "forgets" its initial conditions.

The idea of MCMC is to design a Markov Chain whose limiting distribution is the desired posterior. (See Gelman et al. 1995, Gilks et al. 1998, Robert and Casella 1999.)

Then, we run the chain until it is approximately in equilibirium (how long?) and read off the sequence of states as a sample from our posterior (iid?).

The Metropolis-Hastings Algorighm

The Metropolis-Hastings algorithm is a general method for constructing Markov Chains for posterior sampling. (See Tierney 1994.)

We start with a "proposal distribution" $q(\cdot \mid x)$ which generates a proposed move given that we are state x. We must know how to draw from each $q(\cdot \mid x)$.

The algorithm then creates a sequence $X_0, X_1, \ldots,$ whose limiting (equlibrium) distribution is the desired posterior.

The construction is designed to ensure that with the target distribution the chain satisfies the "detailed balance" condition: $f(s)p(s,t) = f(t)p(t,s)$. That is, for any pair of states s and t, the rate at which the chain moves from state s to t equals that rate at which it moves from t to s.

If this were not true, the chain could not be in equilibrium with that distribution.

The Metropolis-Hastings Algorighm (cont'd)

The algorithm is as follows:

0. Choose X_0 arbitrarily.
1. Having generated $X_0, \ldots, X_i$, draw Z from $q(\cdot \mid X_i)$.
2. Evaluate $r \equiv r(X_i, Z)$ where

$$r(x, z) = \min \left\{ \frac{f(z)}{f(x)} \frac{q(x \mid z)}{q(z \mid x)}, 1 \right\}.$$

3. Let X_{i+1} equal Z with probability r and X_i with probability $1 - r$.

The possibility that $X_{i+1} = X_i$ is essential and such repeated states cannot be simply ignored.

The Metropolis-Hastings Algorighm (cont'd)

Different choices proposal distribution lead to quite varied methods.

1. Independence Metropolis. (Simple but inflexible.)
   Let $q(z \mid x) = q(z)$, a fixed density. Then,
   $$r(x, z) = \min \left\{ \frac{f(z)}{f(x)} \frac{q(x)}{q(z)}, 1 \right\}.$$

2. Random Walk Metropolis. (Commonly used.)
   Let $q(z \mid x) = g(z - x)$ for fixed, symmetric density g, such as a Normal$(0, \tau^2)$.
   $$r(x, z) = \min \left\{ \frac{f(z)}{f(x)}, 1 \right\}.$$
   The trick is to choose τ so that the chain moves around nontrivially. A very rough empirical guideline is to target around 50% move acceptance. Note that this is designed for walks on full Euclidean spaces. On subsets like $(0, \infty)$, one should do a random walk in transformed (e.g., log) coordinates.

The Metropolis-Hastings Algorighm (cont'd)

3. Gibbs Sampling. (Good if you can get it.)

   Write $X_n = (X_n^1, X_n^2)$. We cycle through draws of each component given the others:

$$X_{n+1}^1 \leftarrow f_{X^1|X^2}(x^1 \mid X_n^2)$$
$$X_{n+1}^2 \leftarrow f_{X^2|X^1}(x^2 \mid X_{n+1}^1)$$

   Repeat

   This generalizes to any number of components, which may be scalars or vector blocks.

   Gibbs sampling is a form of Metropolis-Hastings where the move is always accepted. (What is the proposal distribution?)

   If we don't know how to sample from the conditional distributions, we can use a separate Metropolis-Hastings step for that. This is called *Metropolis within Gibbs*.

Road Map

1. Philosophy (briefly)
2. Theory
3. Computation
   - Simulation-Based Methods
   - Markov Chain Monte Carlo
   - Model Jumping and Averaging
4. Practice

Model Jumping and Averaging

Consider a linear model $y = X\beta + \epsilon$ with many predictors (and thus parameters), but many of the parameters should be zero in any one fit.

Fitting too large a model leads to estimates with high variance. Fitting too small a model leads to biased estimates. This is called the bias-variance trade-off, a fundamental reality of statistical inference.

Without some form of "shrinkage" toward a good bias-variance balance, the estimators will be statistically inefficient.

Model Jumping and Averaging (cont'd)

One approach is model selection: use the data to pick which parameters should be allowed to be non-zero. Unfortunately, it is difficult to account for model uncertainty in this case, rendering the resulting inferences optimistic.

What we want to do is allow variation in the model (i.e., which β_js non-zero) while accounting for model uncertainty. This is very hard to do with classical methods, but the Bayesian approach makes this straighforward.

This leads to the idea of model averaging: make the model a parameter.

(This idea also arises in neuroimaging in the problem of making inferences about regions that borrow strength across related voxels.)

Model Jumping and Averaging (cont'd)

Suppose we have models $\mathcal{M}_1, \ldots, \mathcal{M}_K$. For any event $\mathcal{A}$, we can write it's posterior probability as the average over the posterior probability in each model:

$$\mathrm{P}(\mathcal{A} \mid Y) = \sum_{k=1}^{K} \mathrm{P}(\mathcal{A} \mid Y, \mathcal{M}_k)\, \mathrm{P}(\mathcal{M}_k \mid Y).$$

The model probabilities are given, as you might expect, by

$$\mathrm{P}\Big\{\mathcal{M}_k \mid Y\Big\} = \frac{f(y \mid \mathcal{M}_k) f_k}{\sum_{k'=1}^{K} f(y \mid \mathcal{M}'_k) f'_k},$$

where the model k likelihood is defined by

$$f(y \mid \mathcal{M}_k) = \int f(y \mid \theta_k, \mathcal{M}_k) f(\theta_k \mid \mathcal{M}_k)\, d\theta_k.$$

Using the model averaged probabilities, gives better performance than any single model and maintains a full accounting of the uncertainty. Hoeting, Madigan, Raftery, and Volinksy (1999) gives an excellent review of model averaging techniques.

Model Jumping and Averaging (cont'd)

But one additional method has proved very important, this is the Reversible Jump MCMC of Green (1995), called loosely "Model Jumping".

The idea is to construct a meta-Metropolis-Hastings chain that jumps across model spaces – usually of different dimensions – linking chains that are running on the separate spaces.

Instead of one proposal distribution, there are now several $q_1, \ldots, q_r$, with one chosen randomly at each stage.

Some of the proposal distributions are standard Metropolis-Hastings moves within the current model.

Some carry the chain from one model to the other. The key requirement is that these moves satisfy detailed balance. Green (1995) gives a recipe for such moves.

Model Jumping and Averaging (cont'd)

To get the gist, a simple illustration is sufficient. Suppose we want to move between a two parameter space (θ_1, θ_2) and a one parameter space θ_0.

From the two-parameter space, we could move by dropping the second coordinate $(\theta_1, \theta_2) \rightarrow \theta_1$ and move back by adding a fixed component $\theta_0 \rightarrow (\theta_0, 0)$.

But this doesn't satisfy detailed balance because the chain moves to any given θ_0 from $(\theta_0, 3.4)$ but never from a one-dimensional state to $(\theta_0, 3.4)$.

Green's solution is to introduce auxilliary random variables, say a one-dimensional variable U such that the moves are (θ_1, θ_2) to $\theta_1 + \theta_2$ and θ_0 to $\theta_0 + U, \theta_0 - U$. By careful choice of distribution for U, detailed balance can be satisfied.

Road Map

1. Philosophy (briefly)
2. Theory
3. Computation
   - Simulation-Based Methods
   - Markov Chain Monte Carlo
   - Model Jumping and Averaging
4. Practice

Road Map

1. Philosophy (briefly)
2. Theory
3. Computation
4. Practice
   - Example: A Basic Bayesian Neuroimaging Model

Example: A Basic Bayesian Neuroimaging Model

BRAIN (Bayesian Response Analysis and Inference for Neuroimaging) is a software package that implements a variety of Bayesian models for fMRI data. (See Genovese 2000.)

Handles:

– block, event-related, and mixed designs;

– a variety of noise models, response shapes, and priors;

– spatial inferences.

Here, I'll describe a basic version that illustrates some of today's topics.

BRAIN software in public domain
(`http://www.stat.cmu.edu/~genovese/brain/`)

Example Model (cont'd)

Parameters grouped in blocks, each related to one source of variation

μ	Baseline level of signal
Drift	Coefficients of drift profile in current basis
Response	Amplitude of response in an epoch/trial ($\theta^{\text{Response}}_{c,k}$) Average amplitude of response in a condition ($\theta^{\text{Response}}_{c}$)
Shape	Shape of response curve (θ^{Shape})
Noise	Noise Level

Blocks at each level of the hierarchy are taken as independent

Voxelwise Hierarchy

$$Y(t) \;=\; \mu + d(t;\theta^{\text{Drift}}) + a(t;\theta^{\text{Response}}_{c(t),k(t)},\theta^{\text{Shape}},\mu) + \epsilon(t;\theta^{\text{Noise}})$$

Baseline Drift Profile Activation Profile Noise

Across-Epoch Variations: (Optional)

$$\theta^{\text{Response}}_{c,k} \mid \theta^{\text{Response}}_{-(c,k)},\ldots \;\leftarrow\; \pi_{\text{Epoch}}(\theta^{\text{Response}}_{c,k} \mid \theta^{\text{Response}}_{c})$$

Voxelwise Variations:

$$\begin{aligned}
\theta^{\text{Drift}} \mid \lambda,\theta^{\text{Noise}},\ldots &\leftarrow A\exp(-Q(d;\lambda)) \\
\lambda \mid \theta^{\text{Noise}},\ldots &\leftarrow \text{Exponential}(\theta^{\text{Noise}}/\lambda_0) \\
\theta^{\text{Response}}_{c} \mid \theta^{\text{Response}}_{-c},\ldots &\leftarrow \text{Gamma/Point-Mass Mixture} \\
\theta^{\text{Shape}},\ldots &\leftarrow \text{Gamma} \quad \text{[Independent Components]} \\
\theta^{\text{Noise}} &\leftarrow \text{Inverse Gamma} \quad \text{[Proper and diffuse]}
\end{aligned}$$

Priors

- $\pi_{\text{Epoch}}\left(\theta_{c,k}^{\text{Response}} \mid \theta_c^{\text{Response}}\right) = \begin{cases} N_+(\theta_c^{\text{Response}}, \tau_0^2) & \text{if } \theta_c^{\text{Response}} > 0 \\ \delta_0 & \text{o.w.} \end{cases}$
- Model drift profile $d(t)$ as a spline, but constrained to be smooth (λ, a hyperparameter).

  Knots and coefficients of splines are model parameters.

  Drift profile penalized with weighted Sobolev prior, e.g.,

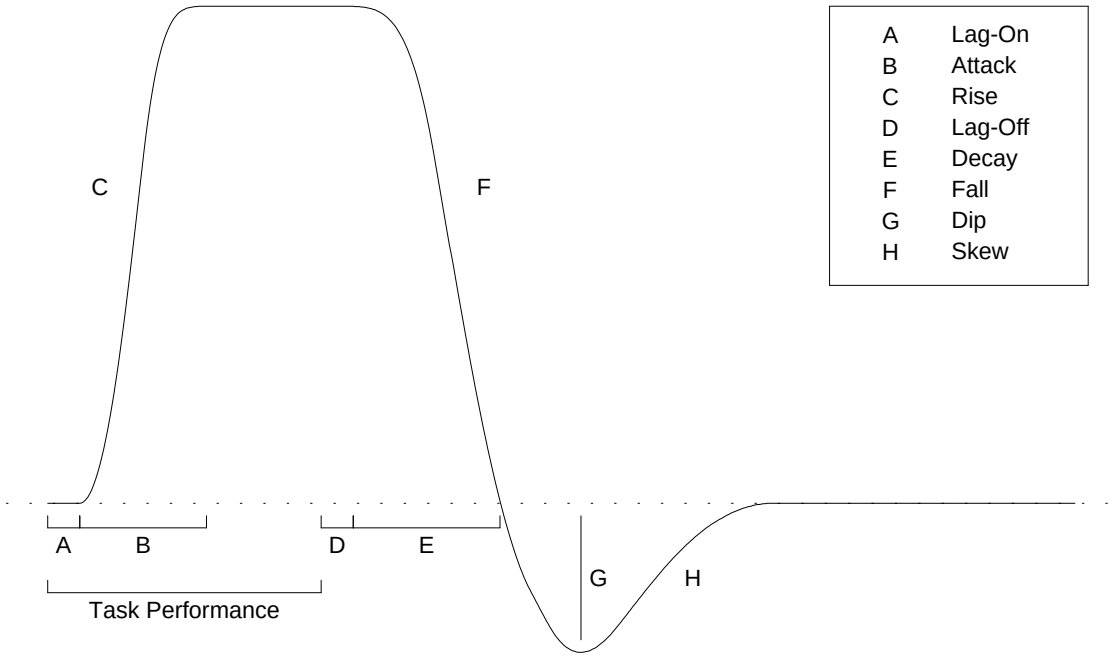
$$Q(d;\lambda) \propto -\frac{1}{2\lambda}\left[\rho_{\text{nc}} \int |d(t)|^2 + \int |d''(t)|^2\right]$$

- Response amplitude by default $\approx$ 1–5% of baseline for active voxel, constrained to be non-negative.
- Prior mean for noise level usually well constrained from prior data.

Structure of the Response

- Parameterize shape of response function by a smooth, nonlinear family of piecewise polynomials.

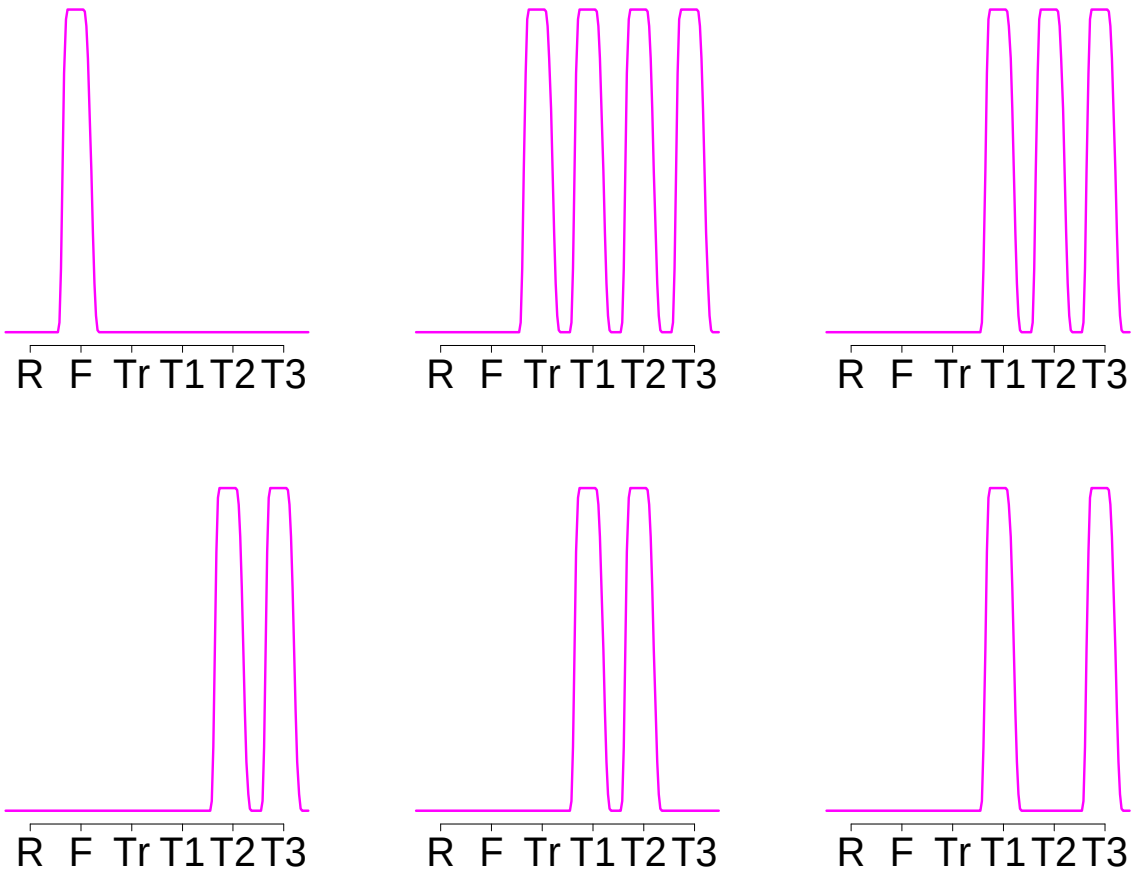
  Offers flexibility, but still constrains the shape.



(This function can be replaced by any desired alternative.)

Model Averaging

- Posterior inferences average over submodels based on subsets of conditions.

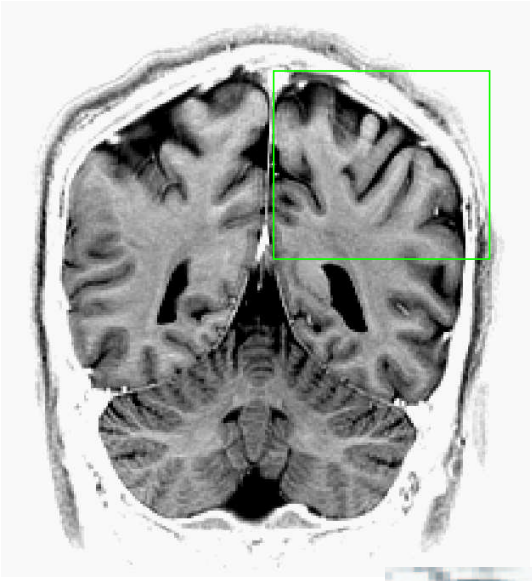


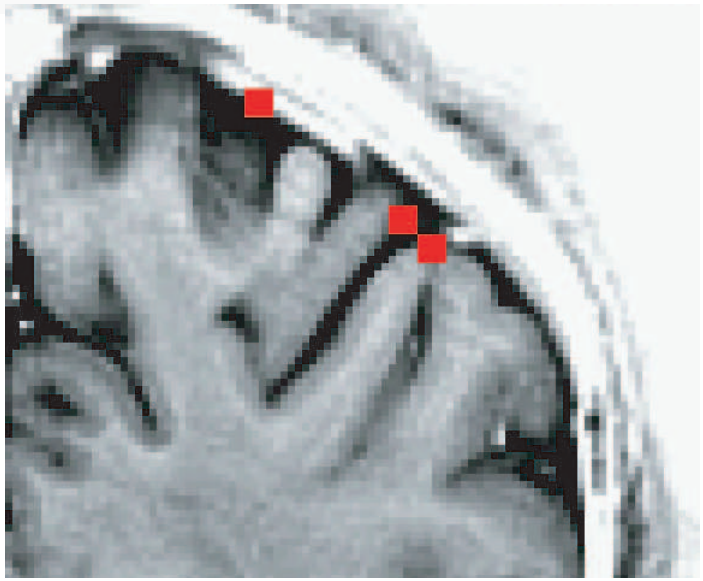
- Estimate posterior probabilities of the sub-models.

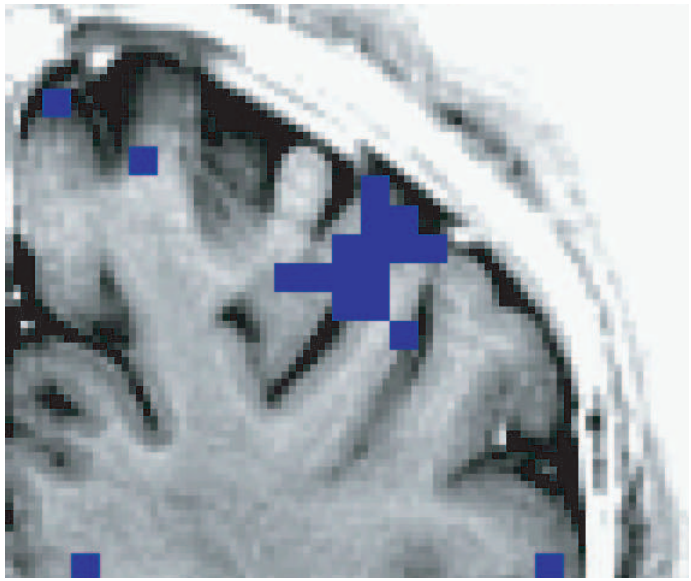
Computational Techniques

- Posterior Maximization
  - Direct numerical optimization
  - Standard Errors derived from normal approximation at the mode
  - Posterior probabilities of submodels derived from approximate Bayes Factors
- Posterior Sampling
  - Mix of Metropolis and Hastings steps.
  - Reversible jump MCMC to move across submodels.
  - Automatically tune jumping parameters during prescan phase.
- With many thousands of voxels, require automated tuning of fitting algorithms.

A Quick Look







Road Map

1. Philosophy (briefly)
2. Theory
3. Computation
4. Practice
   - Example: A Basic Bayesian Neuroimaging Model

Take-Home Points

- Bayesian inference differs in fundamental ways from classical inference even when the procedures are similar.
- Bayesian inferences derive entirely from the posterior and are determined in turn by the likelihood.
- This approach has substantial appeal, and it has become an important part of mainstream statistics.
- Among the features that recommend it for neuroimaging models are the flexibility with which one can handle discrete components in models, including variable selection and spatial clustering.
- Computational methods have improved markedly in recent years, but they can still be costly relative to a simple regression.

  In the end, the gains in efficiency and inferential freedom may be worth the cost.

References

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. Springer Verlag, NY.

Berger, J. and Wolpert, R. (1984) *The Likelihood Principle*, 2nd edition. Institute of Mathematical Statistics Lecture Notes – Monographs Volume 6.

Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite dimensional parameters. *The Annals of Statistics*, **27**, 1119–1141.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall.

Gelman, A. and Meng, X. (1998). Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling, *Statistical Science*, **13**, 163–185.

Genovese, C. R. (2000). A Bayesian Time-Course Model for Functional Magnetic Resonance Imaging Data (with discussion), *Journal of the American Statistical Association*, **95**, 691–703.

References (cont'd)

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1998) *Markov Chain Monte Carlo in Practice*, Chapman& Hall.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, *Biometrika*, **82**, 711–732.

Hoeting, Madigan, Raftery, and Volinksy (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, **14**, No. 4, 382–412.

Kass, R. and Wasserman L. (1999). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Meng, X. and Wong, W. (1996). Simulating Ratios of Normalizing Constants Via a Simple Identity: A Theoretical Exploration, *Statistica Sinica*, **6**, 831–860.

Robert and Casella (1999). *Monte Carlo Statistical Methods*. Springer Verlag.

Robins and Ritov (1997). Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, **16**, 285–319.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions (Disc: P1728-1762), *The Annals of Statistics*, **22**, 1701–1728.

Wasserman, L. (2004). *All of Statistics*. Springer Verlag, NY.