# Introduction to Bayesian Inference

Frank Wood

November 24, 2009

# Bayesian Analysis Recipe

Bayesian data analysis can be described as a three step process

1. Set up a full (generative) probability model
2. Condition on the observed data to produce a posterior distribution, the conditional distribution of the unobserved quantities of interest (parameters or functions of the parameters, etc.)
3. Evaluate the goodness of the model

# Philosophy

Gelman, "Bayesian Data Analysis"

> *A primary motivation for believing Bayesian thinking important is that it facilitates a common-sense interpretation of statistical conclusions. For instance, a Bayesian (probability) interval for an unknown quantity of interest can be directly regarded as having a high probability of containing the unknown quantity, in contrast to a frequentist (confidence) interval, which may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice.*

# Theoretical Setup

Consider a model with parameters $\Theta$ and observations that are independently and identically distributed from some distribution $X_i \sim F(\cdot, \Theta)$ parameterized by $\Theta$.
Consider a prior distribution on the model parameters $P(\Theta; \Psi)$

- What does

$$P(\Theta|X_1, \ldots, X_N; \Psi) \propto P(X_1, \ldots, X_N|\Theta; \Psi)P(\Theta; \Psi)$$

  mean?
- What does $P(\Theta; \Psi)$ mean? What does it represent?

# Example

Consider the following example: suppose that you are thinking about purchasing a factory that makes pencils. Your accountants have determined that you can make a profit (i.e. you should transact the purchase) if the percentage of defective pencils manufactured by the factory is less than 30%.

In your prior experience, you learned that, on average, pencil factories produce defective pencils at a rate of 50%.

To make your judgement about the efficiency of this factory you test pencils one at a time in sequence as they emerge from the factory to see if they are defective.

# Notation

Let $X_1, \ldots, X_N, X_i \in \{0, 1\}$ be a set of defective/not defective observations.

Let $\Theta$ be the probability of pencil defect.

Let $P(X_i|\Theta) = \Theta^{X_i}(1 - \Theta)^{1-X_i}$ (a Bernoulli random variable)

# Typical elements of Bayesian inference

Two typical Bayesian inference objectives are

1. The *posterior distribution* of the model parameters

$$P(\Theta|X_1, \ldots, X_n) \propto P(X_1, \ldots, X_n|\Theta)P(\Theta)$$

   This distribution is used to make statements about the distribution of the unknown or latent quantities in the model.

2. The *posterior predictive distribution*

$$P(X_n|X_1, \ldots, X_{n-1}) = \int P(X_n|\Theta)P(\Theta|X_1, \ldots, X_{n-1})d\Theta$$

   This distribution is used to make predictions about the population given the model and a set of observations.

# The Prior

Both the posterior and the posterior predictive distributions require the choice of a prior over model parameters $P(\Theta)$ which itself will usually have some parameters. If we call those parameters $\Psi$ then you might see the prior written as $P(\Theta; \Psi)$.

The prior encodes your prior belief about the values of the parameters in your model. The prior has several interpretations and many modeling uses

- Encoding previously observed, related observations (pseudocounts)
- Biasing the estimate of model parameters towards more realistic or probable values
- Regularizing or contributing towards the numerical stability of an estimator
- Imposing constraints on the values a parameter can take

# Choice of Prior - Continuing the Example

In our example the model parameter $\Theta$ can take a value in $\Theta \in [0, 1]$. Therefore the prior distribution's support should be $[0, 1]$

One possibility is $P(\Theta) = 1$. This means that we have no prior information about the value $\Theta$ takes in the real world. Our prior belief is uniform over all possible values.

Given our assumptions (that 50% of manufactured pencils are defective in a typical factory) this seems like a poor choice.

A better choice might be a non-uniform parameterization of the Beta distribution.

# Beta Distribution

The Beta distribution $\Theta \sim \text{Beta}(\alpha, \beta)$ ($\alpha > 0, \beta > 0, \Theta \in [0, 1]$) is a distribution over a single number between 0 and 1. This number can be interpreted as a probability. In this case, one can think of $\alpha$ as a pseudo-count related to the number of successes (here a success will be the failure of a pencil) and $\beta$ as a pseudo-count related to the number of failures in a population. In that sense, the distribution of $\Theta$ encoded by the Beta distribution can produce many different biases.

The formula for the Beta distribution is

$$P(\Theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \Theta^{\alpha-1}(1 - \Theta)^{\beta-1}$$

Run introduction_to_bayes/main.m

# $\Gamma$ function

In the formula for the Beta distribution

$$P(\Theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Theta^{\alpha-1}(1-\Theta)^{\beta-1}$$

The gamma function (written $\Gamma(x)$) appears.

It can be defined recursively as $\Gamma(x) = (x-1)\Gamma(x-1) = (x-1)!$ with $\Gamma(1) = 1$.

This is just a generalized factorial (to real and complex numbers in addition to integers). It's value can be computed. It's derivative can be taken, etc.

Note that, by inspection (and definition of distribution)

$$\int \Theta^{\alpha-1}(1-\Theta)^{\beta-1}d\Theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

# Beta Distribution

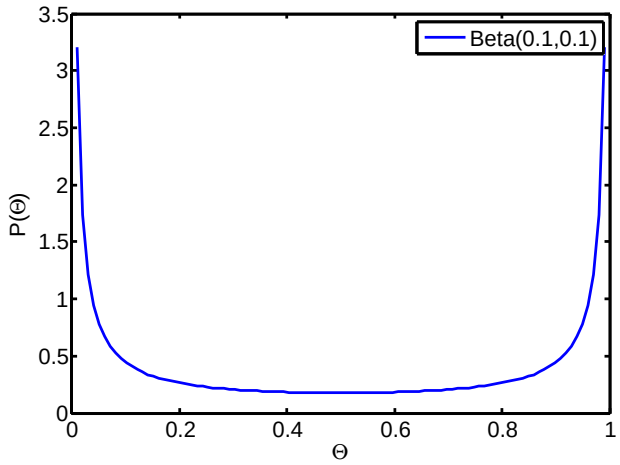


Figure: Beta(.1,.1)

# Beta Distribution

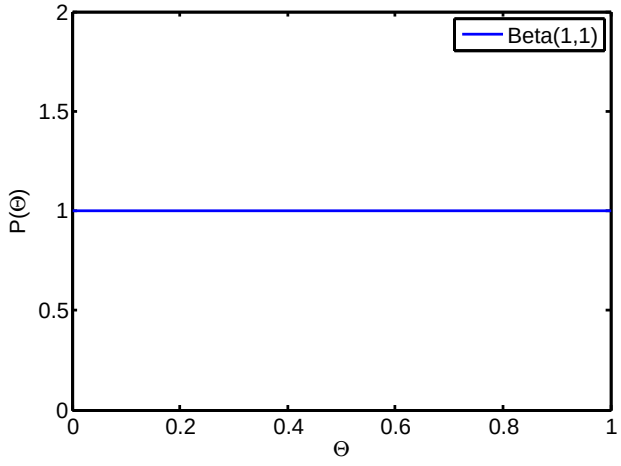


Figure: Beta(1,1)

# Beta Distribution

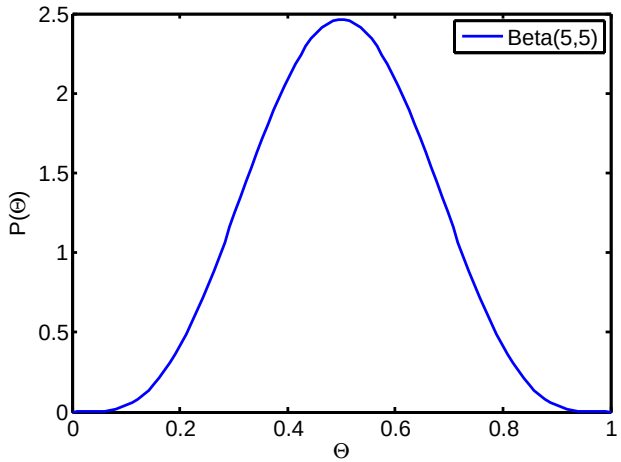


Figure: Beta(5,5)

# Beta Distribution

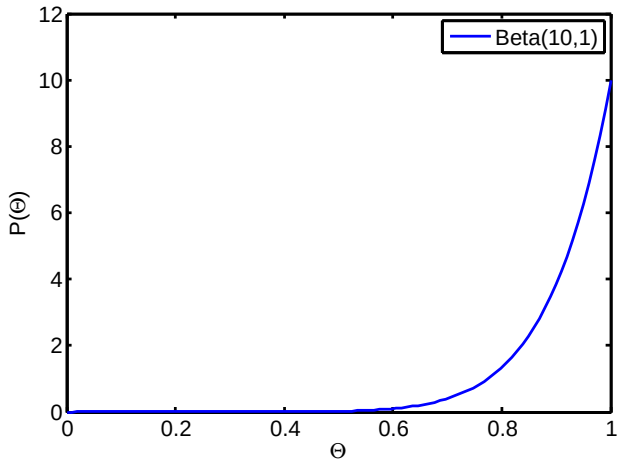


Figure: Beta(10,1)

# Generative Model

With the introduction of this prior we now have a full generative model of our data (given $\alpha$ and $\beta$, the model's hyperparameters). Consider the following procedure for generating pencil failure data:

- Sample a failure rate parameter $\Theta$ for the "factory" from a Beta$(\alpha, \beta)$ distribution. This yields the failure rate for the factory.
- Given the failure rate $\Theta$, sample $N$ defect/no-defect observations from a Bernoulli distribution with parameter $\Theta$.

Bayesian inference involves "turning around" this generative model, i.e. uncovering a distribution over the parameter $\Theta$ given both the observations and the prior.

# Inferring the Posterior Distribution

Remember that the *posterior distribution* of the model parameters is given by

$$P(\Theta|X_1, \ldots, X_n) \propto P(X_1, \ldots, X_n|\Theta)P(\Theta)$$

Let's consider what the posterior looks like after observing a single observation (in our example).

Our likelihood is given by

$$P(X_1|\Theta) = \Theta^{X_1}(1-\Theta)^{1-X_1}$$

Our prior, the Beta distribution, is given by

$$P(\Theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Theta^{\alpha-1}(1-\Theta)^{\beta-1}$$

# Posterior Update Computation

Since we know that

$$P(\Theta|X_1) \propto P(X_1|\Theta)P(\Theta)$$

we can write

$$P(\Theta|X_1) \propto \Theta^{X_1}(1-\Theta)^{1-X_1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Theta^{\alpha-1}(1-\Theta)^{\beta-1}$$

but since we are interested in a function (distribution) of $\Theta$ and we are working with a proportionality, we can throw away terms that do not involve $\Theta$ yielding

$$P(\Theta|X_1) \propto \Theta^{\alpha+X_1-1}(1-\Theta)^{1-X_1+\beta-1}$$

# Bayesian Computation, Implicit Integration

From the previous slide we have

$$P(\Theta|X_1) \propto \Theta^{\alpha+X_1-1}(1-\Theta)^{1-X_1+\beta-1}$$

To make this proportionality an equality (i.e. to construct a properly normalized distribution) we have to integrate this expression w.r.t. $\Theta$, i.e.

$$P(\Theta|X_1) = \frac{\Theta^{\alpha+X_1-1}(1-\Theta)^{1-X_1+\beta-1}}{\int \Theta^{\alpha+X_1-1}(1-\Theta)^{1-X_1+\beta-1} d\Theta}$$

But in this and other special cases like it (when the likelihood and the prior form a conjugate pair) this integral can be solved by recognizing the form of the distribution, i.e. note that this expression looks exactly like a Beta distribution but with updated parameters, $\alpha_1 = \alpha + X_1, \beta_1 = \beta + 1 - X_1$

# Posterior and Repeated Observations

This yields the following pleasant result

$$\Theta|X_1, \alpha, \beta \sim \text{Beta}(\alpha + X_1, \beta + 1 - X_1)$$

This means that the posterior distribution of $\Theta$ given an observation is in the same parametric family as the prior. This is characteristic of conjugate likelihood/prior pairs.

Note the following decomposition

$$P(\Theta|X_1, X_2, \alpha, \beta) \propto P(X_2|\Theta, X_1)P(\Theta|X_1, \alpha, \beta)$$

This means that the preceding posterior update procedure can be repeated. This is because $P(\Theta|X_1, \alpha, \beta)$ is in the same family (Beta) as the original prior. The posterior distribution of $\Theta$ given two observations will still be Beta distributed, now just with further updated parameters.

# Incremental Posterior Inference

Starting with

$$\Theta|X_1, \alpha, \beta \sim \text{Beta}(\alpha + X_1, \beta + 1 - X_1)$$

and adding $X_2$ we can almost immediately identify

$$\Theta|X_1, X_2, \alpha, \beta \sim \text{Beta}(\alpha + X_1 + X_2, \beta + 1 - X_1 + 1 - X_2)$$

which simplifies to

$$\Theta|X_1, X_2, \alpha, \beta \sim \text{Beta}(\alpha + X_1 + X_2, \beta + 2 - X_1 - X_2)$$

and generalizes to

$$\Theta|X_1, \ldots, X_N, \alpha, \beta \sim \text{Beta}(\alpha + \sum X_i, \beta + N - \sum X_i)$$

# Interpretation, Notes, and Caveats

- The posterior update computation performed here is unusually simple in that it is analytically tractable. The integration necessary to normalize the posterior distribution is more often not analytically tractable than it is analytically tractable. When it is not analytically tractable other methods must be utilized to get an estimate of the posterior distribution – numerical integration and Markov chain Monte Carlo (MCMC) amongst them.
- The posterior distribution can be interpreted as the distribution of the model parameters given both the structural assumptions made in the model selection step and the selected prior parameterization. Asking questions like, "What is the probability that the factory has a defect rate of less than 10%?" can be answered through operations on the posterior distribution.

# More Interpretation, Notes, and Caveats

The posterior can be seen in multiple ways

$$\begin{aligned} P(\Theta|X_{1:N}) &\propto P(X_1,\ldots,X_N|\Theta)P(\Theta) \\ &\propto P(X_N|X_{1:N-1},\Theta)P(X_{N-1}|X_{1:N-2},\Theta)\cdots P(X_1|\Theta)P(\Theta) \\ &\propto P(X_N|\Theta)P(X_{N-1}|\Theta)\cdots P(X_1|\Theta)P(\Theta) \end{aligned}$$

(when $X$'s are iid given $\Theta$ or exchangeable) and

$$\begin{aligned} P(\Theta|X_1,\ldots,X_N) &\propto P(X_N,\Theta|X_1\ldots,X_{N-1}) \\ &\propto P(X_N|\Theta)P(\Theta|X_1\ldots,X_{N-1}) \end{aligned}$$

The first decomposition highlights the fact that the posterior distribution is influenced by each observation.

The second recursive decomposition highlights the fact that the posterior distribution can be interpreted as the full characterization of the uncertainty about the hidden parameters after having accounted for all observations to some point.

# Posterior Predictive Inference

Now that we know how to update our prior beliefs about the state of latent variables in our model we can consider posterior predictive inference.

Posterior predictive inference performs a weighted average prediction of future values over all possible settings of the model parameters. The prediction is weighted by the posterior probability of the model parameter setting, i.e.

$$P(X_{N+1}|X_{1:N}) = \int P(X_{N+1}|\Theta)P(\Theta|X_{1:N})d\Theta$$

Note that this is just the likelihood convolved against the posterior distribution having accounted for $N$ observations.

# More Implicit Integration

If we return to our example we have the updated posterior distribution

$$\Theta|X_1,\ldots,X_N,\alpha,\beta \sim \text{Beta}(\alpha+\sum_{i=1}^{N} X_i, \beta + N - \sum_{i=1}^{N} X_i)$$

and the likelihood of the $(N+1)^{th}$ observation

$$P(X_{N+1}|\Theta) = \Theta^{X_{N+1}}(1-\Theta)^{1-X_{N+1}}$$

Note that the following integral is similar in many ways to the posterior update

$$P(X_{N+1}|X_{1:N}) = \int P(X_{N+1}|\Theta)P(\Theta|X_{1:N})d\Theta$$

which means that in this case (and in all conjugate pairs) this is easy to do.

# More Implicit Integration

$$
\begin{aligned}
P(X_{N+1}|X_{1:N}) &= \int \Theta^{X_{N+1}}(1-\Theta)^{1-X_{N+1}} \\
&\times \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum_{i=1}^{N} X_i)\Gamma(\beta+N-\sum_{i=1}^{N} X_i)} \\
&\times \Theta^{\alpha+\sum_{i=1}^{N} X_i-1}(1-\Theta)^{\beta+N-\sum_{i=1}^{N} X_i)-1} d\Theta
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum_{i=1}^{N} X_i)\Gamma(\beta+N-\sum_{i=1}^{N} X_i)} \\
&\times \frac{\Gamma(\alpha+\sum_{i=1}^{N} X_i+X_{N+1})\Gamma(\beta+N+1-\sum_{i=1}^{N} X_i-X_{N+1})}{\Gamma(\alpha+\beta+N+1)}
\end{aligned}
$$

# Interpretation

$$
\begin{aligned}
& P(X_{N+1}|X_{1:N}) \\
&= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum_{i=1}^{N} X_i)\Gamma(\beta+N-\sum_{i=1}^{N} X_i)} \\
&\times \frac{\Gamma(\alpha+\sum_{i=1}^{N} X_i + X_{N+1})\Gamma(\beta+N+1-\sum_{i=1}^{N} X_i - X_{N+1})}{\Gamma(\alpha+\beta+N+1)}
\end{aligned}
$$

Is a ratio of Beta normalizing constants.

This a distribution over $[0, 1]$ which averages over all possible models in the family under consideration (again, weighted by their posterior probability).

# Caveats again

In posterior predictive inference many of the same caveats apply.

- Inference can be computationally demanding if conjugacy isn't exploited.
- Inference results are only as good as the model and the chosen prior.

But Bayesian inference has some pretty big advantages

- Assumptions are explicit and easy to characterize.
- It is easy to plug and play Bayesian models.

# Beta Distribution

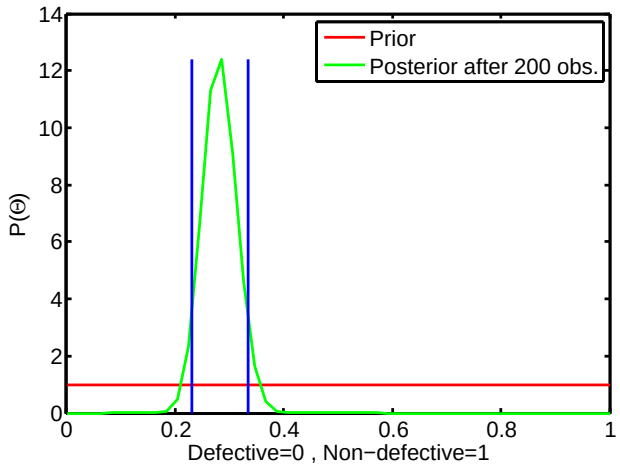


Figure: Posterior after 1000 observations.

# Beta Distribution

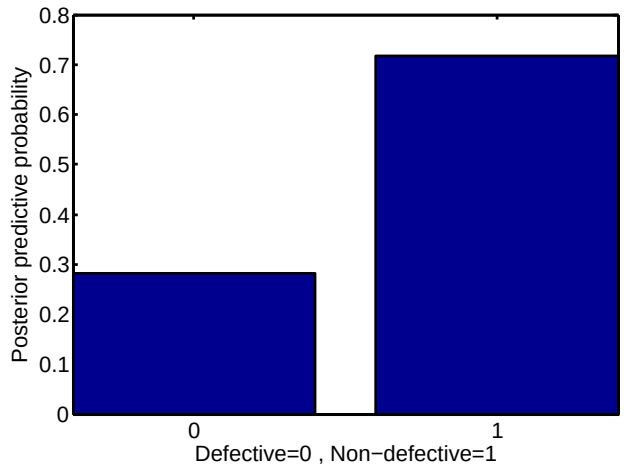


Figure: Posterior predictive after 1000 observations.

# Bayesian Regression

But this prior is not very smart.
newline
We've seen how one choice of noninformative prior gives rise to analytic posterior and posterior predictive distributions.
newline
What about different choices of prior? How do we do inference if the posterior distribution isn't tractable?
newline

*Answer: Sampling*