Bayesian Statistics - an Introduction

Dr Lawrence Pettit

School of Mathematical Sciences, Queen Mary, University of London

July 22, 2008

What is Bayesian Statistics?

- Based on an idea of subjective probability.

What is Bayesian Statistics?

- Based on an idea of subjective probability.
- Have knowledge, beliefs about matter in hand.

What is Bayesian Statistics?

- Based on an idea of subjective probability.
- Have knowledge, beliefs about matter in hand.
- Express these as a (prior) probability distribution.

What is Bayesian Statistics?

- Based on an idea of subjective probability.
- Have knowledge, beliefs about matter in hand.
- Express these as a (prior) probability distribution.
- Collect some data (likelihood).

What is Bayesian Statistics?

- Based on an idea of subjective probability.
- Have knowledge, beliefs about matter in hand.
- Express these as a (prior) probability distribution.
- Collect some data (likelihood).
- Use Bayes theorem to combine prior knowledge and new information to find new (posterior) probability distribution.

What is Bayesian Statistics?

- Based on an idea of subjective probability.
- Have knowledge, beliefs about matter in hand.
- Express these as a (prior) probability distribution.
- Collect some data (likelihood).
- Use Bayes theorem to combine prior knowledge and new information to find new (posterior) probability distribution.

- Today's posterior is tomorrow's prior.

- Today's posterior is tomorrow's prior.
- Although different people will start with different priors with enough data opinions will converge.

- Today's posterior is tomorrow's prior.
- Although different people will start with different priors with enough data opinions will converge.
- One coherent paradigm for all problems.

- Today's posterior is tomorrow's prior.
- Although different people will start with different priors with enough data opinions will converge.
- One coherent paradigm for all problems.
- Can handle more realistic complex models.

- Today's posterior is tomorrow's prior.
- Although different people will start with different priors with enough data opinions will converge.
- One coherent paradigm for all problems.
- Can handle more realistic complex models.
- Makes predictions.

- Today's posterior is tomorrow's prior.
- Although different people will start with different priors with enough data opinions will converge.
- One coherent paradigm for all problems.
- Can handle more realistic complex models.
- Makes predictions.
- The interpretation of interval estimates is more natural.

- Today's posterior is tomorrow's prior.
- Although different people will start with different priors with enough data opinions will converge.
- One coherent paradigm for all problems.
- Can handle more realistic complex models.
- Makes predictions.
- The interpretation of interval estimates is more natural.
- Nuisance parameter and constraints on parameters can be easily dealt with.

Bayes theorem

- Bayes Theorem is named after Rev. Thomas Bayes, a nonconformist minister who lived in England in the first half of the eighteenth century.

Bayes theorem

- Bayes Theorem is named after Rev. Thomas Bayes, a nonconformist minister who lived in England in the first half of the eighteenth century.
- The theorem was published posthumously in 1763 in 'An essay towards solving a problem in the doctrine of chances'.

Bayes theorem

- Let Ω be a sample space and $B_1, B_2, \ldots, B_k$ be mutually exclusive and exhaustive events in Ω (i.e. $B_i \cap B_j = \emptyset, i \neq j, \cup_{i=1}^{k} B_i = \Omega$; the B_i form a partition of Ω.)

Bayes theorem

- Let Ω be a sample space and $B_1, B_2, \ldots, B_k$ be mutually exclusive and exhaustive events in Ω (i.e. $B_i \cap B_j = \emptyset, i \neq j, \cup_{i=1}^{k} B_i = \Omega$; the B_i form a partition of Ω.)
- Let A be any event with $\Pr[A] > 0$.

Bayes theorem

- Let Ω be a sample space and $B_1, B_2, \ldots, B_k$ be mutually exclusive and exhaustive events in Ω (i.e. $B_i \cap B_j = \emptyset, i \neq j, \cup_{i=1}^{k} B_i = \Omega$; the B_i form a partition of Ω.)
- Let A be any event with $\Pr[A] > 0$.
- $$\Pr[B_i|A] = \frac{\Pr[B_i]\Pr[A|B_i]}{\Pr[A]} = \frac{\Pr[B_i]\Pr[A|B_i]}{\sum_{j=1}^{k}\Pr[B_j]\Pr[A|B_j]}$$

Example

- A diagnostic test for a disease gives a correct result 99% of the time. Suppose 2% of the population have the disease. If a person selected at random from the population is given the test and produces a positive result what is the probability that the person has the disease? Suppose they are given a second independent test and it is also positive, what is the probability that they have the disease now?

Example

- A diagnostic test for a disease gives a correct result 99% of the time. Suppose 2% of the population have the disease. If a person selected at random from the population is given the test and produces a positive result what is the probability that the person has the disease? Suppose they are given a second independent test and it is also positive, what is the probability that they have the disease now?
- Let '+' and '−' denote the events that a test is positive or negative, respectively. Let D denote the event that the person has the disease.
- We require $p(D|+)$ and $p(D|++)$.

Solution

- Let '+' and '−' denote the events that a test is positive or negative, respectively. Let D denote the event that the person has the disease.
- We require $p(D|+)$ and $p(D|++)$. We are told that

Solution

- Let '+' and '−' denote the events that a test is positive or negative, respectively. Let D denote the event that the person has the disease.
- We require $p(D|+)$ and $p(D|++)$. We are told that

$$p(+|D) = 0.99, \qquad p(-|\bar{D}) = 0.99, \qquad p(D) = 0.02.$$

Solution

- Let '+' and '−' denote the events that a test is positive or negative, respectively. Let D denote the event that the person has the disease.
- We require $p(D|+)$ and $p(D|++)$. We are told that

$$p(+|D) = 0.99, \qquad p(-|\bar{D}) = 0.99, \qquad p(D) = 0.02.$$

- By Bayes theorem

$$\begin{aligned} p(D|+) &= \frac{p(+|D)p(D)}{p(+|D)p(D) + p(+|\bar{D})p(\bar{D})} \\ &= \frac{0.99 \times 0.02}{0.99 \times 0.02 + 0.01 \times 0.98} \\ &= 0.6689 \end{aligned}$$

Solution for a second test

- Suppose that we have a second positive test result

$$\begin{aligned} p(D|++) &= \frac{p(++|D)p(D)}{p(++|D)p(D) + p(++|\bar{D})p(\bar{D})} \\ &= \frac{0.99^2 \times 0.02}{0.99^2 \times 0.02 + 0.01^2 \times 0.98} \\ &= 0.9950 \end{aligned}$$

Solution for a second test

- Suppose that we have a second positive test result

$$\begin{aligned} p(D|++) &= \frac{p(++|D)p(D)}{p(++|D)p(D) + p(++|\bar{D})p(\bar{D})} \\ &= \frac{0.99^2 \times 0.02}{0.99^2 \times 0.02 + 0.01^2 \times 0.98} \\ &= 0.9950 \end{aligned}$$

- Thus although the test is very accurate we need two positive results before we can say with confidence that the person has the disease.

Density form of Bayes Theorem

- Let X, θ be two continuous r.v.'s (possibly multivariate).

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} = \frac{f(\theta)f(x|\theta)}{\int f(\theta')f(x|\theta')d\theta'}$$

Density form of Bayes Theorem

- Let X, θ be two continuous r.v.'s (possibly multivariate).

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} = \frac{f(\theta)f(x|\theta)}{\int f(\theta')f(x|\theta')d\theta'}$$

- Posterior $\propto$ Likelihood $\times$ Prior

Density form of Bayes Theorem

- Let X, θ be two continuous r.v.'s (possibly multivariate).

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} = \frac{f(\theta)f(x|\theta)}{\int f(\theta')f(x|\theta')d\theta'}$$

- Posterior $\propto$ Likelihood $\times$ Prior
-

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

The Likelihood Principle

- To illustrate the difference between the classical and Bayesian approaches we start with an example.

The Likelihood Principle

- To illustrate the difference between the classical and Bayesian approaches we start with an example.
- Suppose we toss a drawing pin and get 9 'ups' and 3 'downs'. We denote 'up' by U and 'down' by D. Is the pin unbiased?

The Likelihood Principle

- To illustrate the difference between the classical and Bayesian approaches we start with an example.
- Suppose we toss a drawing pin and get 9 'ups' and 3 'downs'. We denote 'up' by U and 'down' by D. Is the pin unbiased?
- Classically we might test $H_0 : p = \frac{1}{2}$ versus $H_1 : p > \frac{1}{2}$ where $p = p(U)$. The probability of the observed result or something more extreme (tail area) if H_0 is true is

$$\left\{ \binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0} \right\} \left(\frac{1}{2}\right)^{12}$$

which $= \frac{299}{4096} \approx 7.3\%$. Thus we would accept H_0 at the 5% level.

- However this *assumes* that we did the experiment by deciding to toss the drawing pin 12 times.

- However this *assumes* that we did the experiment by deciding to toss the drawing pin 12 times.
- What if we decided to toss the pin until we achieved 3 D's?

- However this *assumes* that we did the experiment by deciding to toss the drawing pin 12 times.
- What if we decided to toss the pin until we achieved 3 D's?
- Now the probability of the observed result or something more extreme if H_0 is true is

$$\binom{11}{2}\left(\frac{1}{2}\right)^{12} + \binom{12}{2}\left(\frac{1}{2}\right)^{13} + \binom{13}{2}\left(\frac{1}{2}\right)^{14} + \cdots$$

We may calculate the probability of the complement of this event by

$$\binom{10}{2}\left(\frac{1}{2}\right)^{11} + \binom{9}{2}\left(\frac{1}{2}\right)^{10} + \cdots + \binom{2}{2}\left(\frac{1}{2}\right)^{3}$$

It follows that the P-value is $\frac{134}{4096} \approx 3.25\%$. So we would reject H_0 at the 5% level.

- In order to perform a significance test we are required to specify a *sample space* i.e. the space of all possible outcomes.

- In order to perform a significance test we are required to specify a *sample space* i.e. the space of all possible outcomes.
- Possibilities for the drawing pin are:
  (i) $\{(u, d) : u + d = 12\}$,
  (ii) $\{(u, d) : d = 3\}$,
  or if I carry on tossing the pin until my coffee is ready, so that there is a random stopping point
  (iii) $\{\text{all } (u, d)\}$.

- The Bayesian analysis of this problem is somewhat different. Let θ be the chance that the pin lands up.

- The Bayesian analysis of this problem is somewhat different. Let θ be the chance that the pin lands up.
- θ is a "long run frequency" of U's. It is an objective property of the pin. It does not depend on You.

- The Bayesian analysis of this problem is somewhat different. Let θ be the chance that the pin lands up.
- θ is a "long run frequency" of U's. It is an objective property of the pin. It does not depend on You.
- You have beliefs about θ which you express in the form of a probability density function (pdf) $p(\theta)$. You use Bayes theorem to update your beliefs.

- The Bayesian analysis of this problem is somewhat different. Let θ be the chance that the pin lands up.
- θ is a "long run frequency" of U's. It is an objective property of the pin. It does not depend on You.
- You have beliefs about θ which you express in the form of a probability density function (pdf) $p(\theta)$. You use Bayes theorem to update your beliefs.
-

$$\begin{aligned} p(\theta|\text{data}) &\propto p(\text{data}|\theta)p(\theta) \\ &\quad \theta^9(1-\theta)^3 p(\theta) \end{aligned}$$

The sampling rule is irrelevant.

- In deriving the posterior distribution the only contribution from the data is through the likelihood $p(\mathrm{data}|\theta)$. Thus a Bayesian inference, which will depend only on the posterior distribution, obeys the *likelihood principle*.

- In deriving the posterior distribution the only contribution from the data is through the likelihood $p(\mathrm{data}|\theta)$. Thus a Bayesian inference, which will depend only on the posterior distribution, obeys the *likelihood principle*.
- This roughly says that if two experiments give the same likelihoods then the inferences we make should be the same in each case.

- In deriving the posterior distribution the only contribution from the data is through the likelihood $p(\mathrm{data}|\theta)$. Thus a Bayesian inference, which will depend only on the posterior distribution, obeys the *likelihood principle*.
- This roughly says that if two experiments give the same likelihoods then the inferences we make should be the same in each case.
- Classical hypothesis tests or confidence intervals violate the likelihood principle.

- We need to give a prior distribution for θ.

- We need to give a prior distribution for θ.
- Suppose we take

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

ie a beta distribution with mean $a/(a+b)$ and variance

$$\frac{a}{a+b}\frac{b}{a+b}\frac{1}{a+b+1}.$$

We shall write this distribution as $Be(a, b)$.

- We need to give a prior distribution for θ.
- Suppose we take

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

ie a beta distribution with mean $a/(a+b)$ and variance

$$\frac{a}{a+b}\frac{b}{a+b}\frac{1}{a+b+1}.$$

We shall write this distribution as $Be(a, b)$.

- It then follows that

$$p(\theta|\text{data}) \propto \theta^{9+a-1}(1-\theta)^{3+b-1}.$$

That is $Be(9+a, 3+b)$. Thus if we take a beta prior for θ we shall obtain a beta posterior.

- The choice $a = b = 1$ gives a uniform distribution as the prior, ie we think any value of θ is equally likely. More realistically for a pin we might take $a = b = 2$, reflecting a belief that we think θ is more likely to be near 0.5 than 0 or 1 but not being very sure.

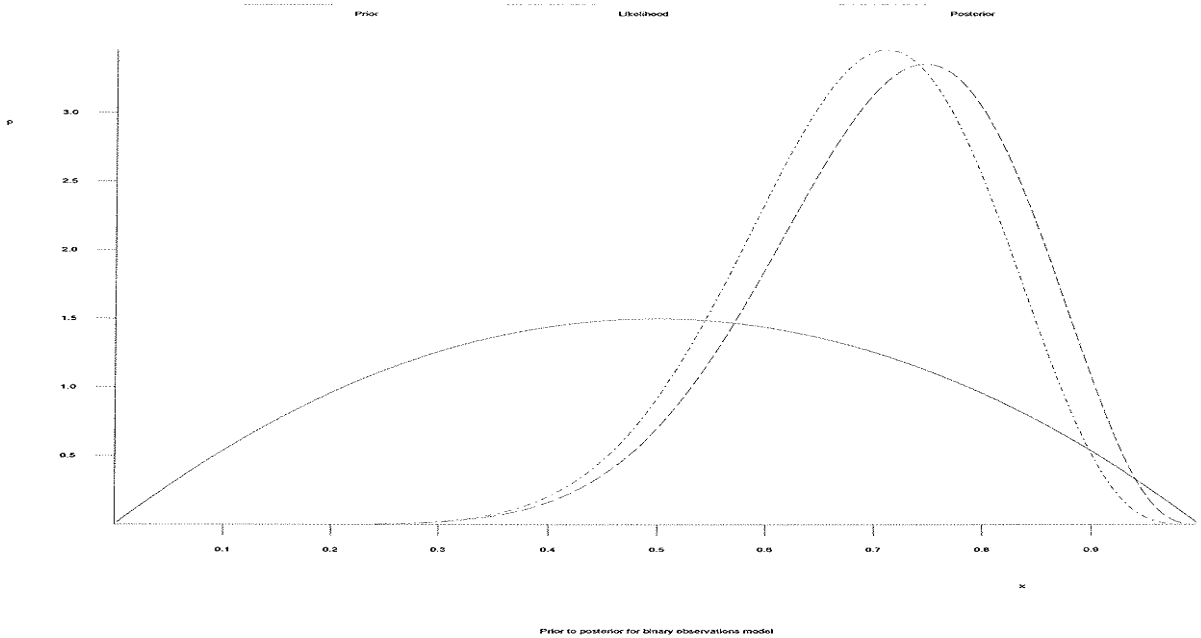
- The choice $a = b = 1$ gives a uniform distribution as the prior, ie we think any value of θ is equally likely. More realistically for a pin we might take $a = b = 2$, reflecting a belief that we think θ is more likely to be near 0.5 than 0 or 1 but not being very sure.
- Others might choose an asymmetric prior, perhaps arguing that a pin with a very long point would very likely land down so a pin with any point would land down more often than up. I am not convinced by this argument but it shows that different people do have different beliefs and one of the advantages of the Bayesian approach is that the analysis can reflect these.

- If we were throwing a coin rather than a pin then we would almost certainly choose a different prior. We have much more experience throwing coins than pins and are much more sure that θ, the chance the coin will land heads, is close to 0.5. Thus we might take $a = b = 50$ say.

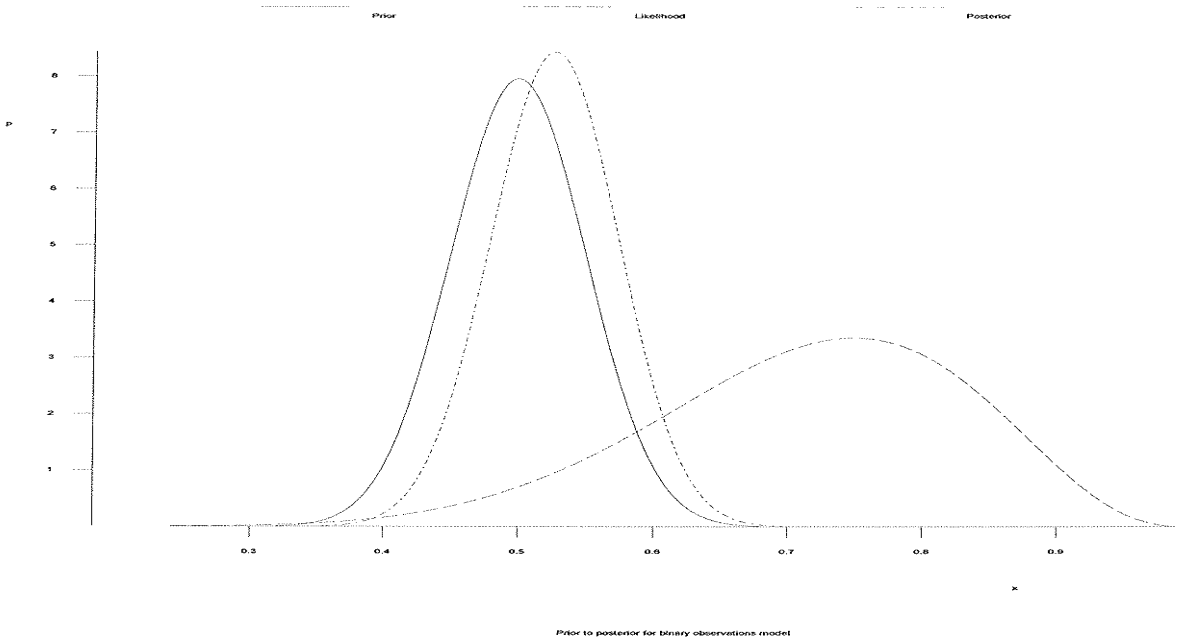
- If we were throwing a coin rather than a pin then we would almost certainly choose a different prior. We have much more experience throwing coins than pins and are much more sure that θ, the chance the coin will land heads, is close to 0.5. Thus we might take $a = b = 50$ say.
- The posteriors we get for the pin and the coin will be very different. For the pin the posterior mean will be $11/16 = 0.6875$ and the posterior variance $11/16 \times 5/16 \times 1/17 = 0.0126$. For the coin the posterior mean will be $59/112 = 0.527$ which is close to 0.5.

- If we were throwing a coin rather than a pin then we would almost certainly choose a different prior. We have much more experience throwing coins than pins and are much more sure that θ, the chance the coin will land heads, is close to 0.5. Thus we might take $a = b = 50$ say.
- The posteriors we get for the pin and the coin will be very different. For the pin the posterior mean will be $11/16 = 0.6875$ and the posterior variance $11/16 \times 5/16 \times 1/17 = 0.0126$. For the coin the posterior mean will be $59/112 = 0.527$ which is close to 0.5.
- The classical unbiased estimate is $9/12 = 0.75$ if the number of throws is fixed, or $9/11 = 0.818$ if we continue until we have three 'failures'. The classical answers are the same whether for pins or coins, ignoring the extra information that we have.

Plot for pins



Plot for coins



- By taking mixtures of conjugate priors we can represent more realistic beliefs.

- By taking mixtures of conjugate priors we can represent more realistic beliefs.
- As an example consider the result when a coin is spun on its edge. Experience has shown that when spinning a coin the proportion of heads is more likely to be 1/3 or 2/3 than 1/2. Therefore a bimodal prior seems appropriate. Since spinning coins will be Bernoulli trials the beta distribution will be conjugate.

- By taking mixtures of conjugate priors we can represent more realistic beliefs.
- As an example consider the result when a coin is spun on its edge. Experience has shown that when spinning a coin the proportion of heads is more likely to be $1/3$ or $2/3$ than $1/2$. Therefore a bimodal prior seems appropriate. Since spinning coins will be Bernoulli trials the beta distribution will be conjugate.
- Therefore we take the following prior

$$p(\theta) = \frac{1}{2} Be(10, 20) + \frac{1}{2} Be(20, 10)$$

i.e

$$p(\theta) = \frac{1}{2} \frac{\Gamma(30)}{\Gamma(10)\Gamma(20)} \theta^{10-1}(1-\theta)^{20-1} + \frac{1}{2} \frac{\Gamma(30)}{\Gamma(20)\Gamma(10)} \theta^{20-1}(1-\theta)^{10-1}$$

where θ is the chance that the coin comes down heads

- If we observe 3 heads and 7 tails in 10 spins the posterior based on the first prior is $Be(13, 27)$ and on the second prior is $Be(23, 17)$.

- If we observe 3 heads and 7 tails in 10 spins the posterior based on the first prior is $Be(13, 27)$ and on the second prior is $Be(23, 17)$.
- The weight on the first posterior, α', is

$$(0.5p_1(\underline{x}))(0.5p_1(\underline{x}) + 0.5p_2(\underline{x}))^{-1}$$

Now

$$p_1(\underline{x}) = \frac{\Gamma(30)}{\Gamma(10)\Gamma(20)} \frac{\Gamma(13)\Gamma(27)}{\Gamma(40)}$$

and similarly

$$p_2(\underline{x}) = \frac{\Gamma(30)}{\Gamma(20)\Gamma(10)} \frac{\Gamma(23)\Gamma(17)}{\Gamma(40)}$$

A little calculation shows that $\alpha' = 0.89$ and $\beta' = 0.11$.

- If we observe 3 heads and 7 tails in 10 spins the posterior based on the first prior is $Be(13, 27)$ and on the second prior is $Be(23, 17)$.
- The weight on the first posterior, α', is

$$(0.5p_1(\underline{x}))(0.5p_1(\underline{x}) + 0.5p_2(\underline{x}))^{-1}$$

Now

$$p_1(\underline{x}) = \frac{\Gamma(30)}{\Gamma(10)\Gamma(20)} \frac{\Gamma(13)\Gamma(27)}{\Gamma(40)}$$

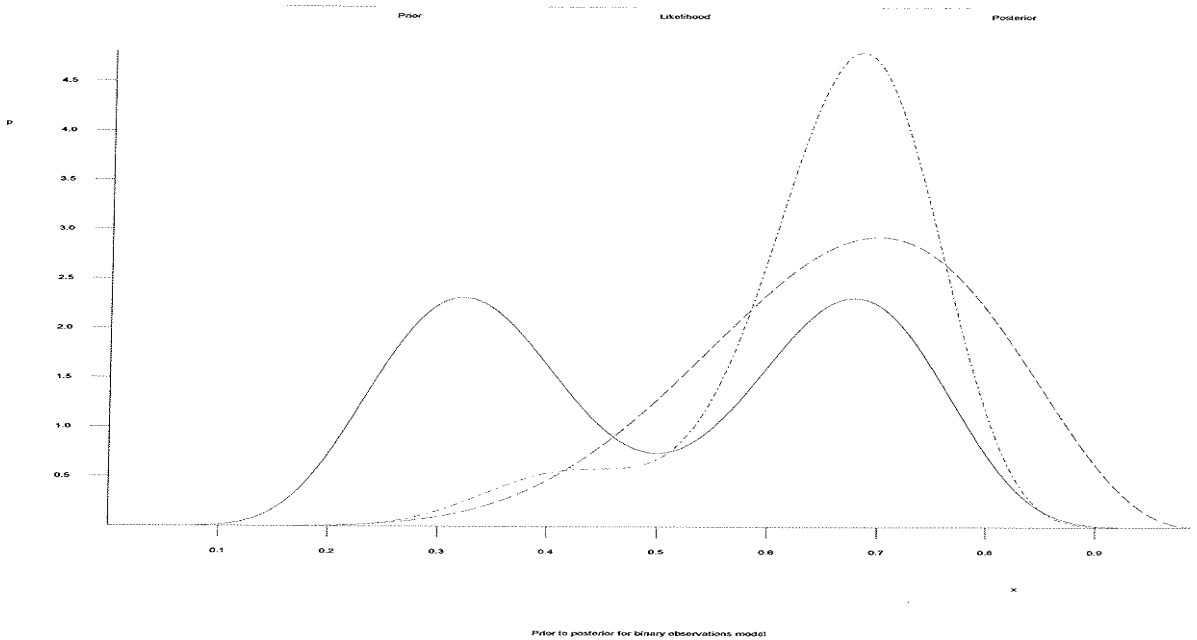
and similarly

$$p_2(\underline{x}) = \frac{\Gamma(30)}{\Gamma(20)\Gamma(10)} \frac{\Gamma(23)\Gamma(17)}{\Gamma(40)}$$

A little calculation shows that $\alpha' = 0.89$ and $\beta' = 0.11$. Hence the posterior is

$$p(\theta|x) = 0.89Be(13, 27) + 0.11Be(23, 17)$$

Plot for mixture prior



Example

- Suppose that we are about to introduce a new insurance policy. Interested in the mean number of claims per month.

Example

- Suppose that we are about to introduce a new insurance policy. Interested in the mean number of claims per month.
- Prior?

Example

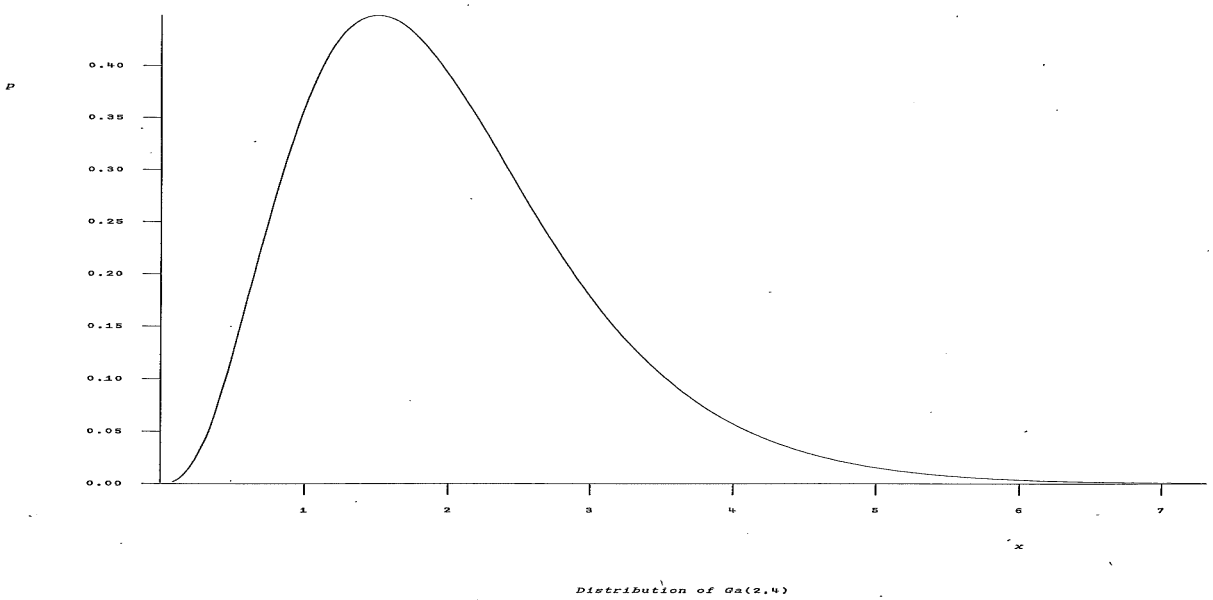
- Suppose that we are about to introduce a new insurance policy. Interested in the mean number of claims per month.
- Prior?
- May expect that on average there would be 2 claims per month, with a variance of 1.

Example

- Suppose that we are about to introduce a new insurance policy. Interested in the mean number of claims per month.
- Prior?
- May expect that on average there would be 2 claims per month, with a variance of 1.
- Conjugate prior for θ is Gamma(4,2):

$$p(\theta) \propto \theta^{4-1} \exp(-2\theta)$$

Prior density - Gamma(4,2)



Calculation of the posterior

- After 6 months there have been 18 claims.

Calculation of the posterior

- After 6 months there have been 18 claims.
- Assume that data follow a Poisson distribution with mean θ.

$$p(x|\theta) = \frac{\exp(-\theta)\theta^x}{x!}$$

Calculation of the posterior

- After 6 months there have been 18 claims.
- Assume that data follow a Poisson distribution with mean θ.

$$p(x|\theta) = \frac{\exp(-\theta)\theta^x}{x!}$$

- The likelihood is given by

$$p(\underline{x}|\theta) \propto \theta^{\sum x_i} \exp(-n\theta)$$

where $n = 6$ and $\sum x_i = 18$.

Calculation of the posterior

- After 6 months there have been 18 claims.
- Assume that data follow a Poisson distribution with mean θ.

$$p(x|\theta) = \frac{\exp(-\theta)\theta^x}{x!}$$

- The likelihood is given by

$$p(\underline{x}|\theta) \propto \theta^{\sum x_i} \exp(-n\theta)$$

where $n = 6$ and $\sum x_i = 18$.

- Posterior

$$p(\theta|\underline{x}) \propto \theta^{22-1} \exp(-8\theta)$$

which we can recognise is a $Ga(22, 8)$.

Calculation of the posterior

- Posterior mean ($\frac{22}{8}$) is a weighted average of the prior mean (2) and data mean (3).

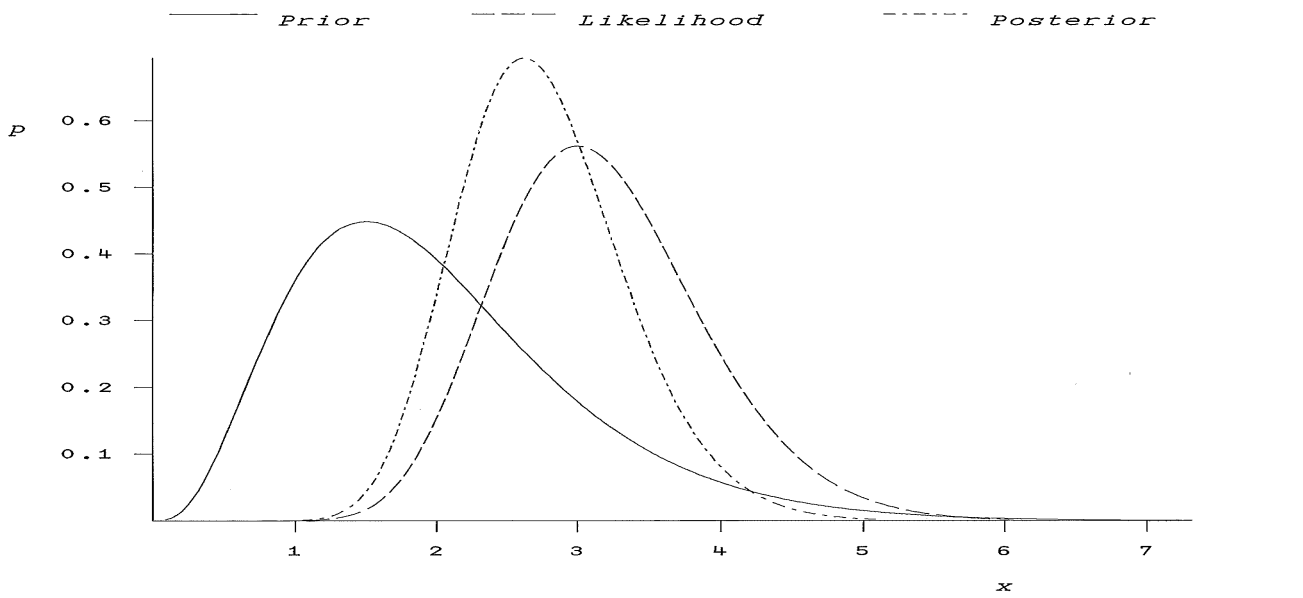
Calculation of the posterior

- Posterior mean ($\frac{22}{8}$) is a weighted average of the prior mean (2) and data mean (3).
- Weights are related to the prior and data variances.

Calculation of the posterior

- Posterior mean ($\frac{22}{8}$) is a weighted average of the prior mean (2) and data mean (3).
- Weights are related to the prior and data variances.
- This holds for other common distributions.

Prior/Posterior density



Prior to posterior for Poisson observations (time parameter 1) model

Predictive distributions

- Distribution of a future observation y?

Predictive distributions

- Distribution of a future observation y?
- $\underline{x} = (x_1, \ldots, x_n)$ is a random sample from $p(x_i|\theta)$

Predictive distributions

- Distribution of a future observation y?
- $$\underline{x} = (x_1, \ldots, x_n) \text{ is a random sample from } p(x_i|\theta)$$
- Prior $p(\theta)$. Posterior $p(\theta|\underline{x})$.

Predictive distributions

- Distribution of a future observation y?
- $$\underline{x} = (x_1, \ldots, x_n) \text{ is a random sample from } p(x_i|\theta)$$
- Prior $p(\theta)$. Posterior $p(\theta|\underline{x})$.
- y - independent observation from the same distribution.

Predictive distributions

- Distribution of a future observation y?
- $$\underline{x} = (x_1, \ldots, x_n) \text{ is a random sample from } p(x_i|\theta)$$
- Prior $p(\theta)$. Posterior $p(\theta|\underline{x})$.
- y - independent observation from the same distribution. Want $p(y|\underline{x})$ - predictive distribution of y.

Predictive density

Now by extension of the argument

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

since y is independent of $\underline{x}$ given θ.

Poisson predictive example

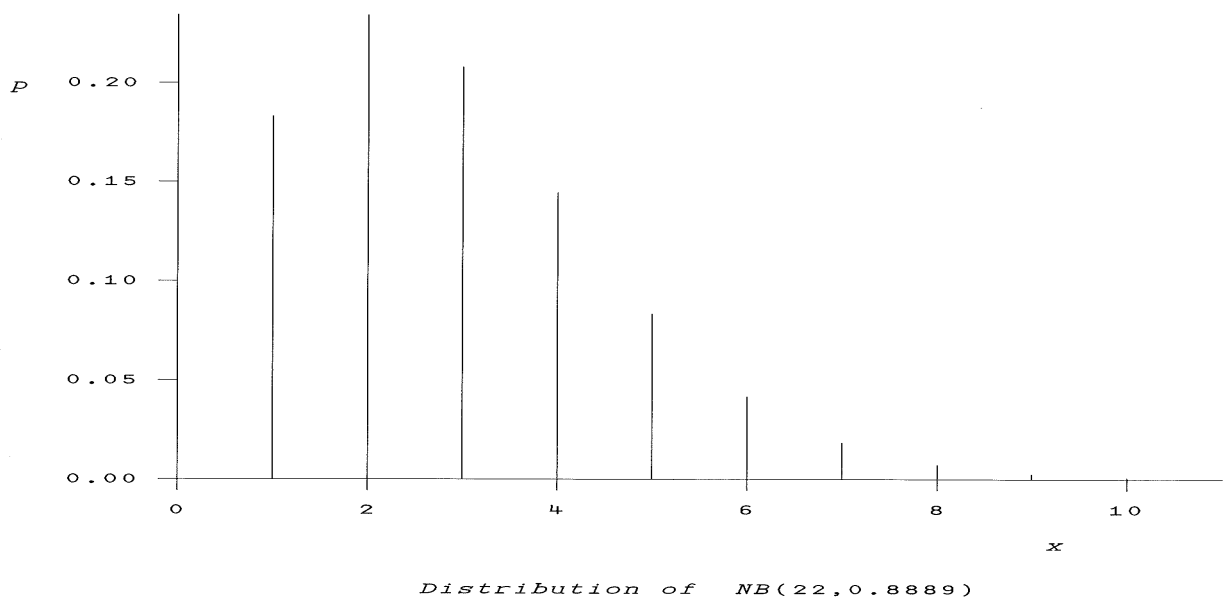
- $x_1, \ldots, x_n \sim$ Poisson(θ), prior for θ is Gamma(4,2)
  Posterior $\Rightarrow$ Gamma(22,8)

Poisson predictive example

- $x_1, \ldots, x_n \sim$ Poisson(θ), prior for θ is Gamma(4,2)
  Posterior $\Rightarrow$ Gamma(22,8)
- Suppose $y \sim$ Poisson(θ)

$$
\begin{aligned}
p(y|\underline{x}) &= \int p(y|\theta)p(\theta|\underline{x})d\theta \\
&= \int \frac{\theta^y \exp(-\theta)}{y!} \frac{8^{22}\theta^{21}\exp\{-8\theta\}}{\Gamma(22)} d\theta \\
&= \frac{8^{22}}{\Gamma(22)y!} \int \theta^{21+y} \exp\{-\theta(9)\} d\theta \\
&= \frac{8^{22}}{\Gamma(22)y!} \frac{\Gamma(y+22)}{9^{22+y}} \qquad y = 0, 1, 2, 3, \ldots
\end{aligned}
$$

Predictive density of y



- The posterior distribution represents all our knowledge after seeing the data.

- The posterior distribution represents all our knowledge after seeing the data.
- Why quote posterior mean and variance so much?

- The posterior distribution represents all our knowledge after seeing the data.
- Why quote posterior mean and variance so much?
- Point estimates are justified by decision theory.

- The posterior distribution represents all our knowledge after seeing the data.
- Why quote posterior mean and variance so much?
- Point estimates are justified by decision theory.
- As an extra ingredient we specify a loss function and choose an estimate to minimise expected loss.

- For example, quadratic loss

$$l(d, \theta) = (d - \theta)^2$$

leads to the posterior mean as the estimator and posterior variance as the expected loss.

- For example, quadratic loss

$$l(d, \theta) = (d - \theta)^2$$

leads to the posterior mean as the estimator and posterior variance as the expected loss.

- Absolute error loss

$$l(d, \theta) = |d - \theta|$$

leads to the posterior median as the estimator.

Stochastic Simulation - Gibbs sampling

Stochastic Simulation - basic idea:

- simulate from the distribution we are interested in

Stochastic Simulation - Gibbs sampling

Stochastic Simulation - basic idea:

- simulate from the distribution we are interested in
- use the simulated sample to make inferences.

Stochastic Simulation - Gibbs sampling

Stochastic Simulation - basic idea:

- simulate from the distribution we are interested in
- use the simulated sample to make inferences.
- Gibbs sampling is one of these methods.

Stochastic Simulation - Gibbs sampling

Stochastic Simulation - basic idea:

- simulate from the distribution we are interested in
- use the simulated sample to make inferences.
- Gibbs sampling is one of these methods.

Illustration of Gibbs

Stochastic Simulation - Gibbs sampling

Stochastic Simulation - basic idea:

- simulate from the distribution we are interested in
- use the simulated sample to make inferences.
- Gibbs sampling is one of these methods.

Illustration of Gibbs

- Suppose we have three parameters θ, δ and ϕ.

Stochastic Simulation - Gibbs sampling

Stochastic Simulation - basic idea:

- simulate from the distribution we are interested in
- use the simulated sample to make inferences.
- Gibbs sampling is one of these methods.

Illustration of Gibbs

- Suppose we have three parameters θ, δ and ϕ.
- Likelihood - $p(\underline{x}|\theta, \delta, \phi)$.

Stochastic Simulation - Gibbs sampling

Stochastic Simulation - basic idea:

- simulate from the distribution we are interested in
- use the simulated sample to make inferences.
- Gibbs sampling is one of these methods.

Illustration of Gibbs

- Suppose we have three parameters θ, δ and ϕ.
- Likelihood - $p(\underline{x}|\theta, \delta, \phi)$.
- Prior - $p(\theta, \delta, \phi)$.

Stochastic Simulation - Gibbs sampling

Stochastic Simulation - basic idea:

- simulate from the distribution we are interested in
- use the simulated sample to make inferences.
- Gibbs sampling is one of these methods.

Illustration of Gibbs

- Suppose we have three parameters θ, δ and ϕ.
- Likelihood - $p(\underline{x}|\theta, \delta, \phi)$.
- Prior - $p(\theta, \delta, \phi)$.
- Suppose we can find the conditional distributions:

$$p(\theta|\delta, \phi, \underline{x}), \quad p(\delta|\phi, \theta, \underline{x}), \quad p(\phi|\theta, \delta, \underline{x})$$

that we can simulate from.

Gibbs Sampling

- The Gibbs sampler is an iterative scheme:
  - Step 1 Choose initial estimates $\theta^0, \delta^0, \phi^0$
  - Step 2 Given current estimates $\theta^i, \delta^i, \phi^i$ simulate new values
    θ^{i+1} from $p(\theta|\delta^i, \phi^i, \underline{x})$
    δ^{i+1} from $p(\delta|\phi^i, \theta^{i+1}, \underline{x})$
    ϕ^{i+1} from $p(\phi|\theta^{i+1}, \delta^{i+1}, \underline{x})$
  - Step 3 Return to step 2

Gibbs Sampling

- The Gibbs sampler is an iterative scheme:
  - Step 1 Choose initial estimates $\theta^0, \delta^0, \phi^0$
  - Step 2 Given current estimates $\theta^i, \delta^i, \phi^i$ simulate new values
    θ^{i+1} from $p(\theta|\delta^i, \phi^i, \underline{x})$
    δ^{i+1} from $p(\delta|\phi^i, \theta^{i+1}, \underline{x})$
    ϕ^{i+1} from $p(\phi|\theta^{i+1}, \delta^{i+1}, \underline{x})$
  - Step 3 Return to step 2
- The sequence $(\theta, \delta, \phi)^i \quad i = 1, 2, \ldots$ is a realisation of a Markov chain which, under mild regularity conditions, has equilibrium distribution $p(\theta, \delta, \phi|\underline{x})$, the joint posterior distribution of θ, δ, ϕ.

Estimation from Gibbs Output

- Burn in period of length L.

Estimation from Gibbs Output

- Burn in period of length L.
- Posterior mean of θ can be estimated for large t by

$$\frac{1}{t}\sum_{i=L+1}^{L+t}\theta^i$$

Estimation from Gibbs Output

- Burn in period of length L.
- Posterior mean of θ can be estimated for large t by

$$\frac{1}{t}\sum_{i=L+1}^{L+t}\theta^i$$

- To see a picture of the posterior density of θ - use *Kernel density estimation*.

Estimation from Gibbs Output

- Burn in period of length L.
- Posterior mean of θ can be estimated for large t by

$$\frac{1}{t}\sum_{i=L+1}^{L+t}\theta^i$$

- To see a picture of the posterior density of θ - use *Kernel density estimation*.
- The predictive density of a new observation y can be estimated as

$$p(y|\underline{x}) = \frac{1}{t}\sum_{i=L+1}^{L+t} P(y|\theta^i, \delta^i, \phi^i)$$

Poisson hierarchical model

- 1st stage - observations $s_j \sim$ Poisson with mean $t_j \lambda_j$ for $j = 1, 2, \ldots, p$, t_j known

Poisson hierarchical model

- 1st stage - observations $s_j \sim$ Poisson with mean $t_j\lambda_j$ for $j = 1, 2, \ldots, p$, t_j known
- 2nd stage - prior $\lambda_j \sim Ga(\alpha, \beta)$ iid, α known

Poisson hierarchical model

- 1st stage - observations $s_j \sim$ Poisson with mean $t_j \lambda_j$ for $j = 1, 2, \ldots, p$, t_j known
- 2nd stage - prior $\lambda_j \sim Ga(\alpha, \beta)$ iid, α known
- 3rd stage - hyperprior $\beta \sim Ga(\gamma, \delta)$, where γ and δ are known.

Poisson hierarchical model

- 1st stage - observations $s_j \sim$ Poisson with mean $t_j\lambda_j$ for $j = 1, 2, \ldots, p$, t_j known
- 2nd stage - prior $\lambda_j \sim Ga(\alpha, \beta)$ iid, α known
- 3rd stage - hyperprior $\beta \sim Ga(\gamma, \delta)$, where γ and δ are known.
- $p + 1$ unknown parameters - λ's and β.

Poisson hierarchical model

- 1st stage - observations $s_j \sim$ Poisson with mean $t_j\lambda_j$ for $j = 1, 2, \ldots, p$, t_j known
- 2nd stage - prior $\lambda_j \sim Ga(\alpha, \beta)$ iid, α known
- 3rd stage - hyperprior $\beta \sim Ga(\gamma, \delta)$, where γ and δ are known.
- $p + 1$ unknown parameters - λ's and β.
- Joint posterior of all the parameters:

$$\prod_{j=1}^{p} \frac{(\lambda_j t_j)^{s_j} \exp(-\lambda_j t_j)}{s_j!} \times \prod_{j=1}^{p} \Gamma(\alpha)^{-1} \beta^{\alpha} \lambda_j^{\alpha-1} \exp(-\beta\lambda_j)$$

$$\times \Gamma(\gamma)^{-1} \delta^{\gamma} \beta^{\gamma-1} \exp(-\delta\beta)$$

Conditional distributions

- Conditional distributions of λ's and β - pick out the relevant terms from this posterior.

Conditional distributions

- Conditional distributions of λ's and β - pick out the relevant terms from this posterior.
-

$$p(\lambda_k|\beta, \lambda_j(j \neq k), \text{data}) \propto \lambda_k^{s_k} \exp(-\lambda_k t_k) \lambda_k^{\alpha-1} \exp(-\lambda_k \beta)$$

$$\Rightarrow Ga(\alpha + s_k, \beta + t_k)$$

Conditional distributions

- Conditional distributions of λ's and β - pick out the relevant terms from this posterior.
-

$$p(\lambda_k|\beta, \lambda_j(j \neq k), \text{data}) \propto \lambda_k^{s_k} \exp(-\lambda_k t_k)\lambda_k^{\alpha-1} \exp(-\lambda_k \beta)$$

$$\Rightarrow Ga(\alpha + s_k, \beta + t_k)$$

-

$$p(\beta|\underline{\lambda}, \text{data}) \propto \beta^{p\alpha} \exp(-\sum \lambda_j \beta)\beta^{\gamma-1} \exp(-\delta\beta)$$

$$\Rightarrow Ga(p\alpha + \gamma, \sum \lambda_j + \delta)$$

Gibbs runs

- For suitable starting values for the unknown parameters the Gibbs sampler proceeds by drawing

$$\begin{aligned} \lambda_j^{i+1} &\sim Ga(\alpha + s_j, \beta^i + t_j) \quad j = 1, \ldots, p \\ \beta^{i+1} &\sim Ga(p\alpha + \gamma, \sum \lambda_j^{i+1} + \delta) \end{aligned}$$

Gibbs runs

- For suitable starting values for the unknown parameters the Gibbs sampler proceeds by drawing

$$\begin{aligned} \lambda_j^{i+1} &\sim Ga(\alpha + s_j, \beta^i + t_j) \quad j = 1, \ldots, p \\ \beta^{i+1} &\sim Ga(p\alpha + \gamma, \sum \lambda_j^{i+1} + \delta) \end{aligned}$$

- Use large sample of λ's and β's to estimate posterior and predictive quantities of interest.