STATS 225: Bayesian Analysis
Hierarchical Bayesian Models

Babak Shahbaba

Department of Statistics, UCI

Reminder: Exchangeability

- We discussed exchangeability before.
- Informally, a set of observations $y = (y_1, ..., y_n)$ are exchangeable if in constructing their joint distribution, we believe that the indices are uninformative.
- We said that the exchangeability is important since according to deFinetti's representation theorem, if we can judge an infinite sequence of observations to be exchangeable, we can *model* any subset of them as independent and identically distributed (iid) samples from a distribution $P(y|\theta)$ with the parameter θ

$$P(y|\theta) = P(y_1, y_2, ..., y_n|\theta) \quad = \quad \prod_{i=1}^{n} P(y_i|\theta)$$

Reminder: Exchangeability

- Moreover, there exists a *prior* probability distribution $P(\theta)$ over the parameters of the model such that we can find the unconditional (or marginal) joint distribution of observations

$$P(y) = P(y_1, y_2, ..., y_n) \quad = \quad \int_\Omega \prod_{i=1}^{n} P(y_i|\theta)p(\theta)d\theta$$

- As we mentioned, the above theorem is an *existence* theorem. We still need to specify the form of these distributions.

Within-group exchangeability

- Now, consider the housing price, y_i, for a sample of 4 bedroom houses in the US.
- We might regard this sample as exchangeable if all we know is the price of each house.
- However, if we also know in which state the house is located, it might be more appropriate to assume exchangeability only within each group since the price distribution would probably be different from one state to another.
- In this case, the price is represented by y_{ij}, where j is an index for the states.
- Now the index is not completely uninformative anymore, since we expect different distributions for different j.
- We can still use the above theorem and consider each sub-sample, (i.e., for a fixed j) as iid given their own specific parametric model with parameter θ_j.

Within-group exchangeability

- Then, for each state j we have

$$P(y_{.j}|\theta_j) \quad = \quad P(y_{1j}, y_{2j}, ..., y_{n_j j}|\theta_j) = \prod_{i=1}^{n_j} P(y_{ij}|\theta_j)$$

- Therefore, the joint distribution of all samples is

$$P(y|\theta) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} P(y_{ij}|\theta_j)$$

- Assuming a normal $N(\mu_j, \sigma_j^2)$ for each state, we have

$$P(y|\mu, \sigma^2) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} N(y_{ij}|\mu_j, \sigma_j^2)$$

- We can assume all states have the same variance

$$P(y|\mu, \sigma^2) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} N(y_{ij}|\mu_j, \sigma^2)$$

Hyperprior

- Now, as we mentioned before, there exists a prior distribution over parameters, $\theta_1, \theta_2, ..., \theta_J$.
- Similar to y, if we could imagine the infinite sequence of such θ's being exchangeable, we can regard them as being iid samples given the prior distribution $P(\theta|\phi)$ with the parameter ϕ

$$P(\theta|\phi) = P(\theta_1, ..., \theta_J|\phi) = \prod_{j=1}^{J} P(\theta_j|\phi)$$

- ϕ is referred to as *hyperparameter*, for which we need to assume a *hyperprior* $P(\phi)$.

Hyperprior

- The joint prior distribution of all parameters is now

$$P(\phi, \theta) = P(\phi)P(\theta|\phi)$$

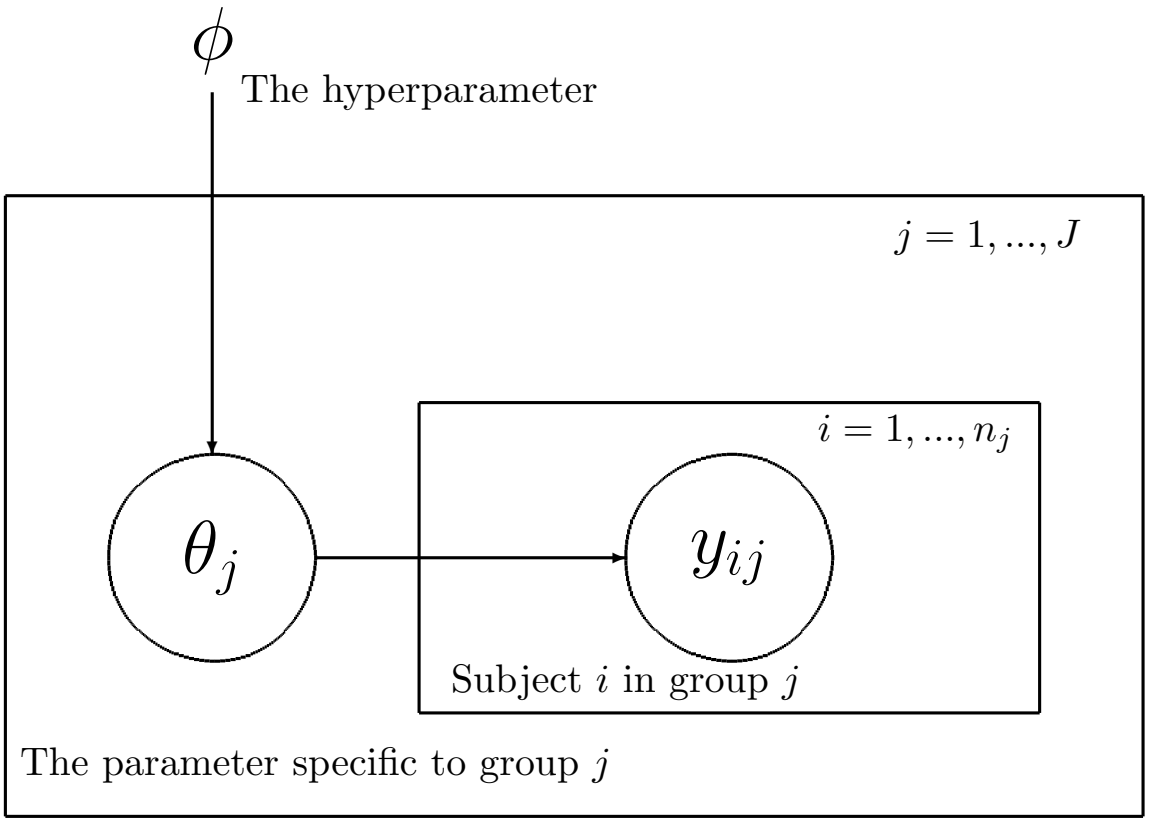
- The posterior distribution of parameters is

$$P(\phi, \theta|y) \propto P(\phi, \theta)P(y|\phi, \theta) = P(\phi)P(\theta|\phi)P(y|\theta)$$

- Note that given θ (i.e., if we fix θ), y becomes independent of ϕ.

Hierarchical Bayesian model

- The following figure shows a schematic representation of hierarchical models in general:



A schematic representation of hierarchical models.

Example 1: House prices in the US

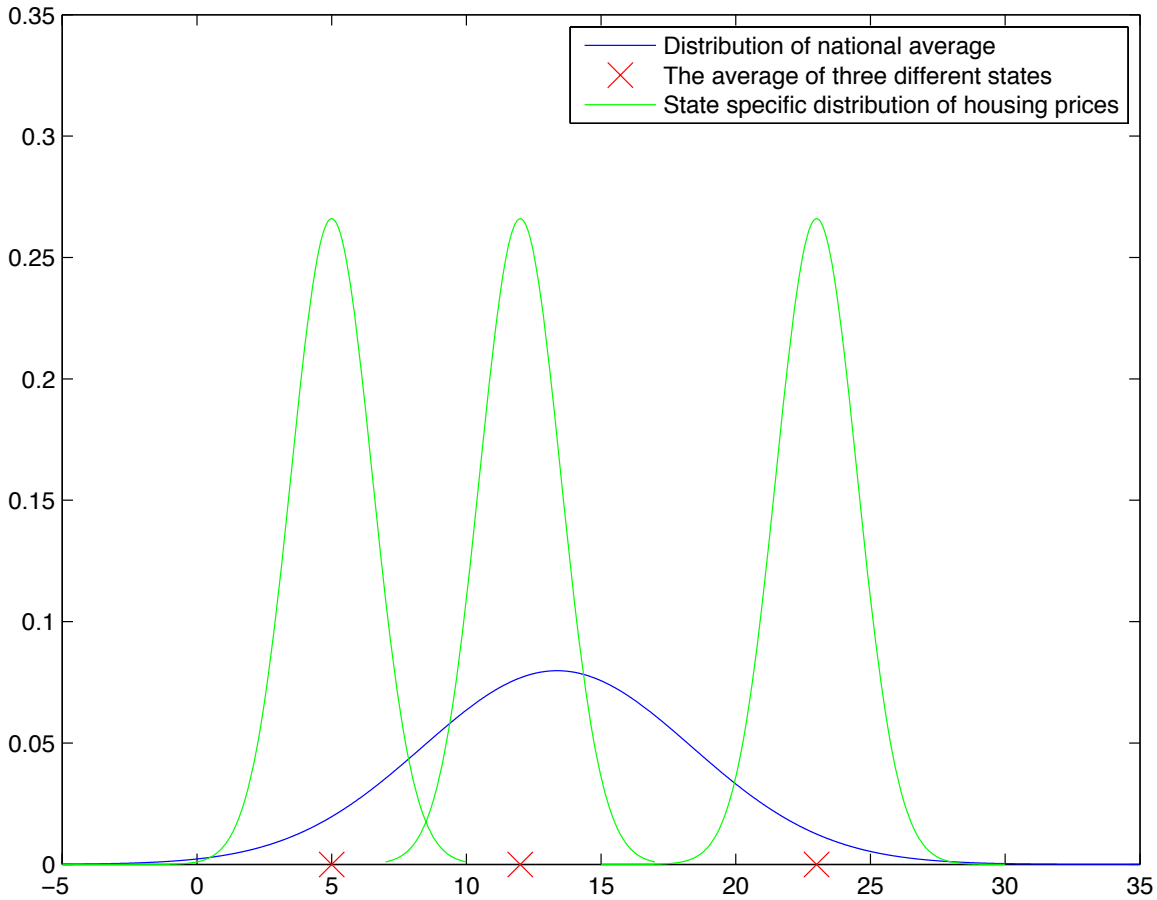
- For the house prices example, we can assume the following priors

$$\begin{aligned} \mu_0 &\sim N(M, V^2) \\ \mu_j &\sim N(\mu_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

- Here, we are assuming that τ_0^2 is fixed and only μ_0 is the hyperparameter.
- Moreover, we are assuming that the variance σ^2 is the same for all states for simplicity.

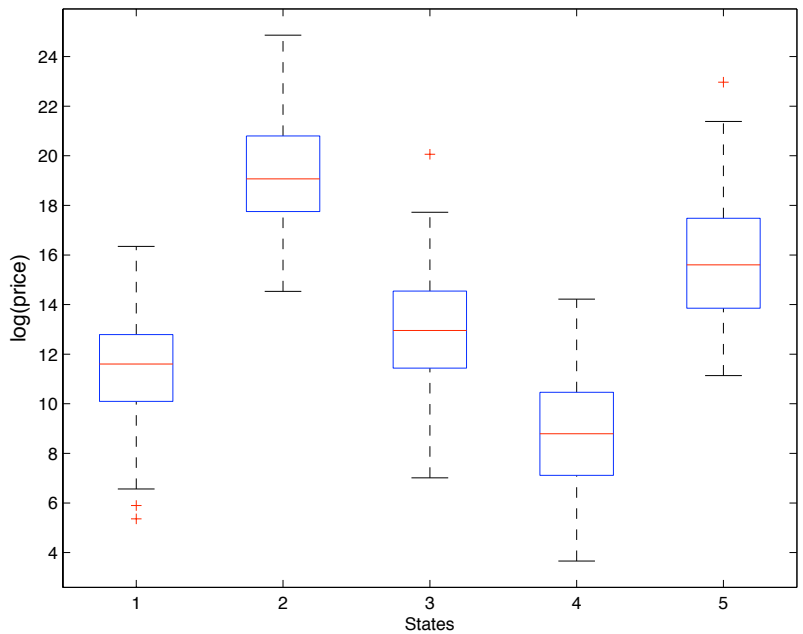
Example 1: House prices in the US

- This graph shows the relation between the distribution of the national average and state specific distributions for three states.



Example 1: House prices in the US

- For this problem, we can use MCMC to obtain samples from the posterior distribution of σ, μ_j, and μ_0.
- For simplicity, we consider only 5 states. We have sampled 100 houses in each states.
- This graph shows the box plot of the observed values.



The boxplots of the observed price of 100 houses for 5 states.

Example 1: House prices in the US

- We use the log transformation of prices and assume the following broad priors for model parameters

$$\begin{aligned} \sigma^2 &\sim \text{Inv-}\chi^2(1, 0.5^2) \\ \mu_j &\sim N(\mu_0, 25^2) \\ \mu_0 &\sim N(0, 50^2) \end{aligned}$$

- Note that these priors are conditionally conjugate so we can use the Gibbs sampler.

Example 1: House prices in the US

- Given μ_0 and σ^2 the problem reduces to 5 independent normal models with known variance. Given μ_0 and σ^2 at each iteration, we can sample from the posterior distribution of μ_j.
- Given μ_j's, we also have a conditional conjugate situation for σ^2 with Inv-χ^2 posterior distribution. So we can sample a new σ^2.
- Note that since σ^2 is common between all states, we use all the y's from the 5 states to update σ^2.
- Next, given the current samples of μ_j, we again have a normal model with conditional conjugacy for μ_0 (taking μ_j's as observations) so we can sample a new μ_0.
- We repeat the above steps to obtain MCMC samples. (The code is available from the course website).

Example 1: House prices in the US

- At each iteration, given the value of μ_0 and σ^2 we sample μ_j from the following normal distribution:

$$\mu_j|y,\mu_0,\sigma^2 \quad \sim \quad N(\frac{\frac{\mu_0}{\tau_0^2}+\frac{n_j\overline{y}_{.j}}{\sigma^2}}{\frac{1}{\tau_0^2}+\frac{n_j}{\sigma^2}},\frac{1}{\frac{1}{\tau_0^2}+\frac{n_j}{\sigma^2}})$$

- Given $\mu=(\mu_1,...,\mu_J)$, we sample a new σ^2 from

$$\sigma^2|y,\mu \sim \text{Inv-}\chi^2(\nu_0+n,\frac{\nu_0\sigma_0^2+\nu n}{\nu_0+n})$$

$$\nu=\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\mu_j)^2$$

Example 1: House prices in the US

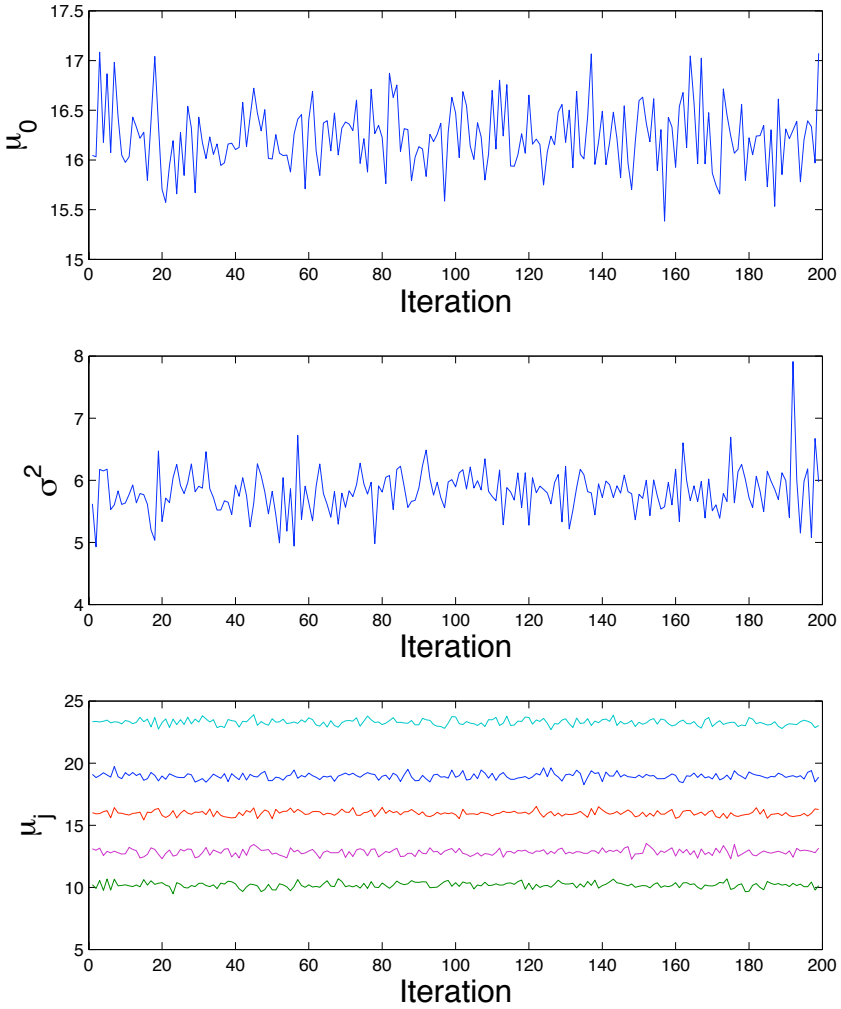
- Given $\mu = (\mu_1, ..., \mu_J)$, we sample μ_0 from

$$\mu_0|\mu \quad \sim \quad N\left(\frac{\frac{M}{V^2} + \frac{J\overline{\mu}}{\tau_0^2}}{\frac{1}{V^2} + \frac{J}{\tau_0^2}}, \frac{1}{\frac{1}{V^2} + \frac{J}{\tau_0^2}}\right)$$

- Notice how using the conditional independence reduces the complexity of the model.

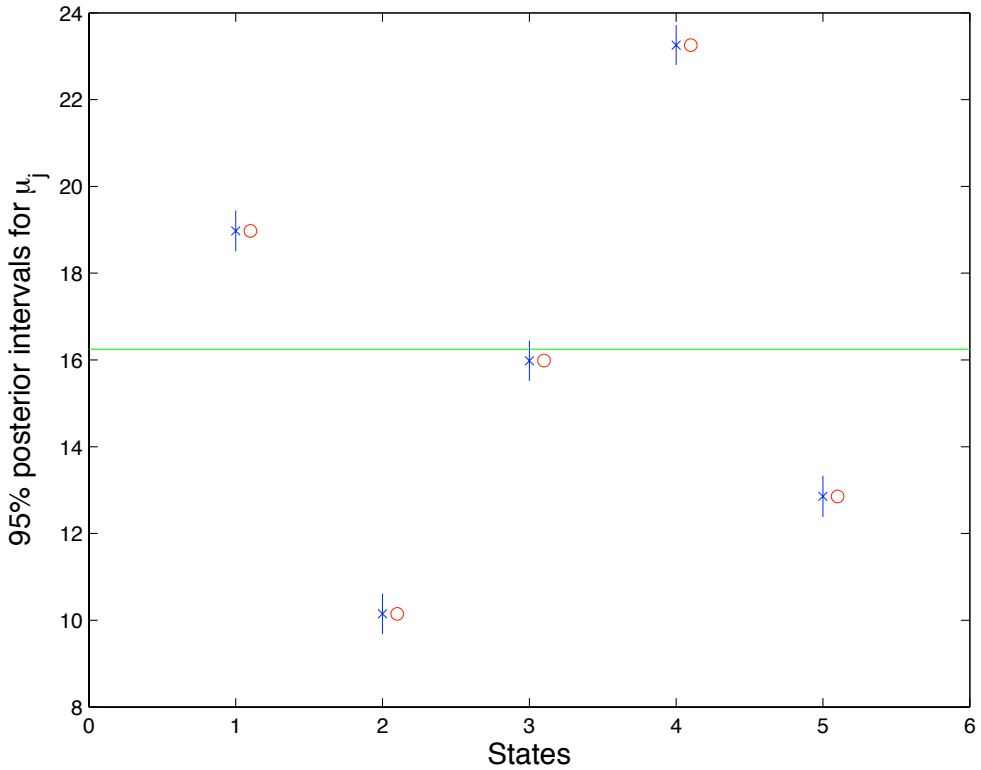
- For this reason, hierarchical Bayesian models are quite powerful.

Example 1: House prices in the US



Example 1: House prices in the US

- The following graph shows the 95% posterior intervals (i.e., credible region), the posterior expectations ($\times$), the maximum likelihood estimations (circles), and the posterior expectation of overall mean μ_0 (the green horizontal line).

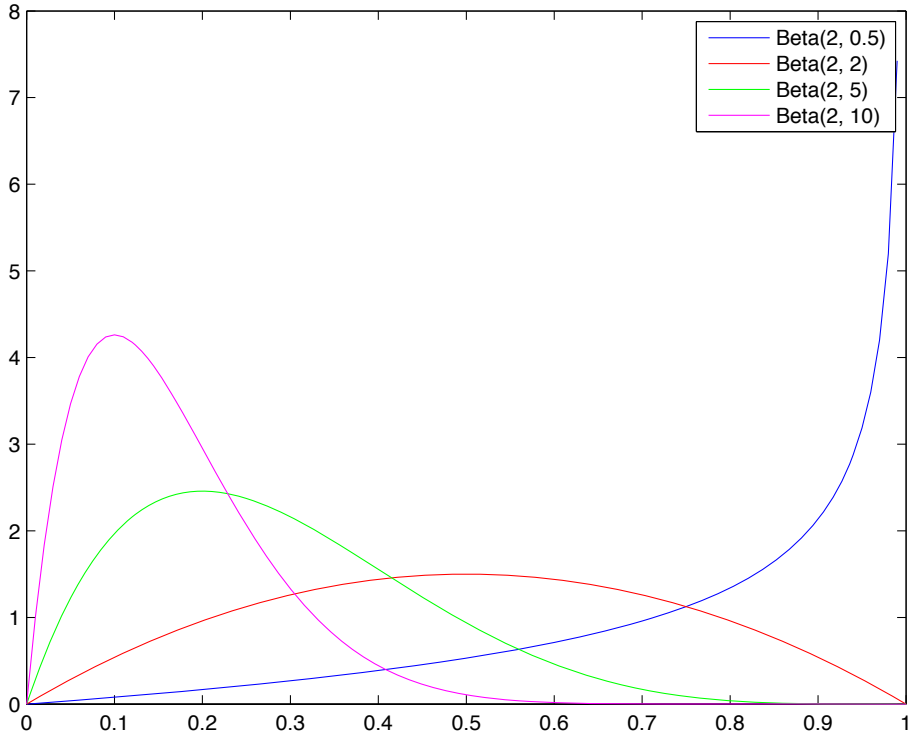


Example 2: Rat tumors

- Another example, which is discussed in Gelman et. al., is related to evaluating the risk of tumor in 71 groups of rats.
- The number of rats that develop tumor, y_j, in each group j with sample size n_j is assumed to have a binomial$(y_j|n_j, \theta_j)$ distribution.
- In a higher level, θ_j are assumed to have a Beta(α, β) distribution.
- α and β are now hyperparameters and we need to specify their own priors.
- For simplicity, we fix $\alpha = 2$ and let only β be a hyperparameter with a Gamma$(2, 1)$ prior.

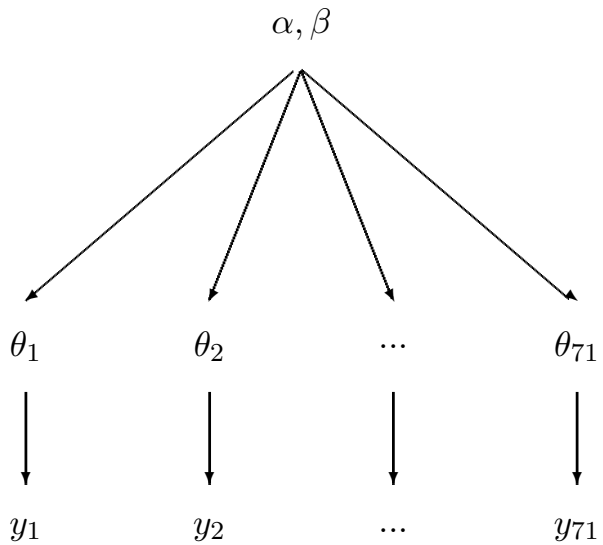
Example 2: Rat tumors

The following plot shows different Beta distributions with fixed $\alpha = 2$.



Example 2: Rat tumors

- The solution in Gelman's book is rather complicated since this problem is discussed before MCMC sampling.
- Since we know MCMC by now, we can obtain posterior distributions in a much simpler way.
- The following figure shows the schematic representation of this model:



Example 2: Rat tumors

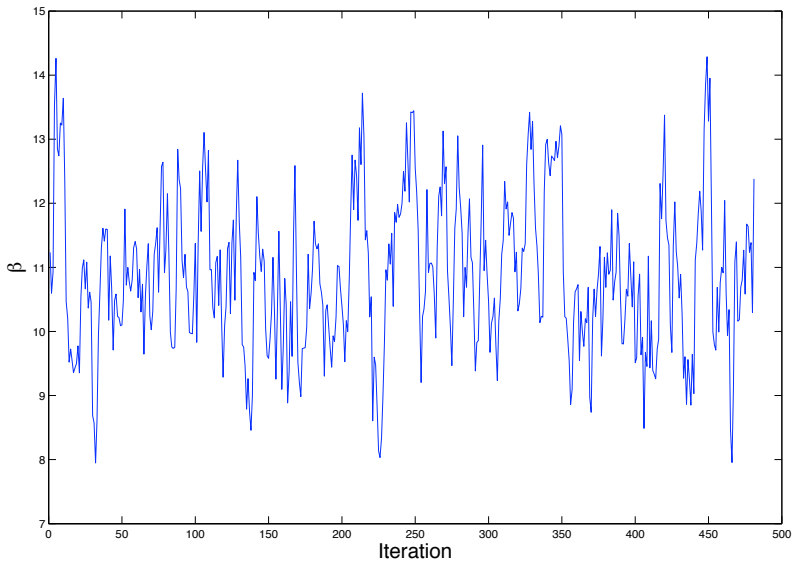
- Given α and β, we have 71 simple binomial models with conditionally conjugate priors for which we can use the Gibbs sampler.

$$\theta_j|\alpha,\beta,y_j \sim \text{Beta}(\alpha+n_j,\beta+n_j-y_j)$$

- After sampling new values for θ's, we take them as observations and update β using Metropolis or MH algorithm. (The code is available from the course website.)

Example 2: Rat tumors

- The following figure shows the trace plot of posterior samples for β.
- The posterior expectation of β is 10.9.



Example 2: Rat tumors

- This graph shows the 95% posterior intervals for θ's, with their posterior expectations ($\times$). As we can see, these values are shrunk towards the overall mean (the horizontal green line shows the posterior expectation of the Beta distribution) compared to the maximum likelihood estimates (circles).

- This is called *shrinkage*. (Note that we didn't see this in the previous example.)