Lecture 12

Gibbs Sampling

and

Hierarchical Models

Last time: Bayes

- Globe toss bayesian updating
- to sample predictive, sample posterior and draw likelihood. y^* is predictive
- Decision Loss: posterior mean minimizes squared loss
- exchangeability $\implies$ IID
- poisson-gamma 2 class model

- normal-normal model
- prior mean regularizes sample mean
- flat prior as uninformative location prior
- Jeffreys prior as uninformative scale prior
- we will use weakly regularizing priors
- make posteriors be sensible, making sampling behave

Today

- the idea behind gibbs sampling
- examples of gibbs sampling
- gibbs is an always accepted MH
- hierarchical models as regularizers
- empirical bayes (for rat tumors)
- setting up full bayes for hierarchical models

What did Gibbs do?

He determined the energy states of gases at equilibrium by cycling through all the particles, drawing from each one of them conditionally given the energy levels of the others, taking the time average.

Geman and Geman used this idea to denoise images.

The idea of Gibbs

$$f(x) = \int f(x,y)dy = \int f(x|y)f(y)dy = \int dy f(x|y) \int dx' f(y|x')f(x')$$

Thus: $f(x) = \int h(x,x')f(x')dx'$ integral fixed point equation

where $h(x,x') = \int dy f(x|y)f(y|x')$.

Iterative scheme in which the "transition kernel" $h(x, x')$ is used to create a proposal for metropolis-hastings moves:

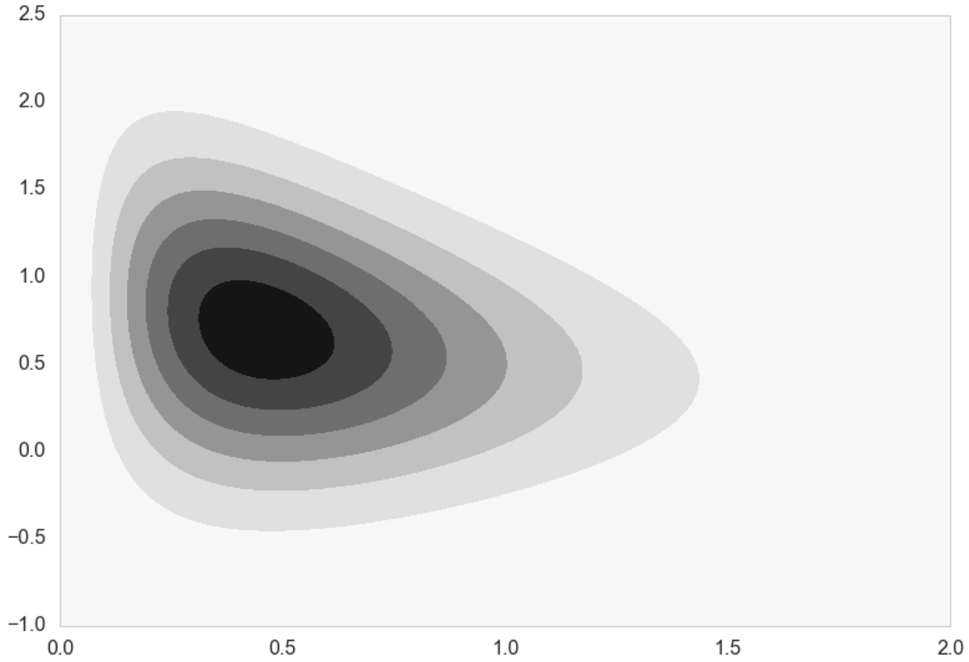
$$f(x_t) = \int h(x_t, x_{t-1}) f(x_{t-1}) dx_{t-1},$$ a Stationary distribution.

$$h(x, x') = \int dy f(x|y) f(y|x').$$: Sample alternately to get transitions.

Can sample x marginal and $x|y$ so can sample the joint x, y.

Example

Sample from $f(x, y) = x^2 \exp[-xy^2 - y^2 + 2y - 4x]$



Conditionals

$$f(x,y) = x^2 \exp[-xy^2 - y^2 + 2y - 4x]$$

$$f(x|y) = x^2 \exp[-x(y^2+4)] \exp[-y^2 + 2y]$$

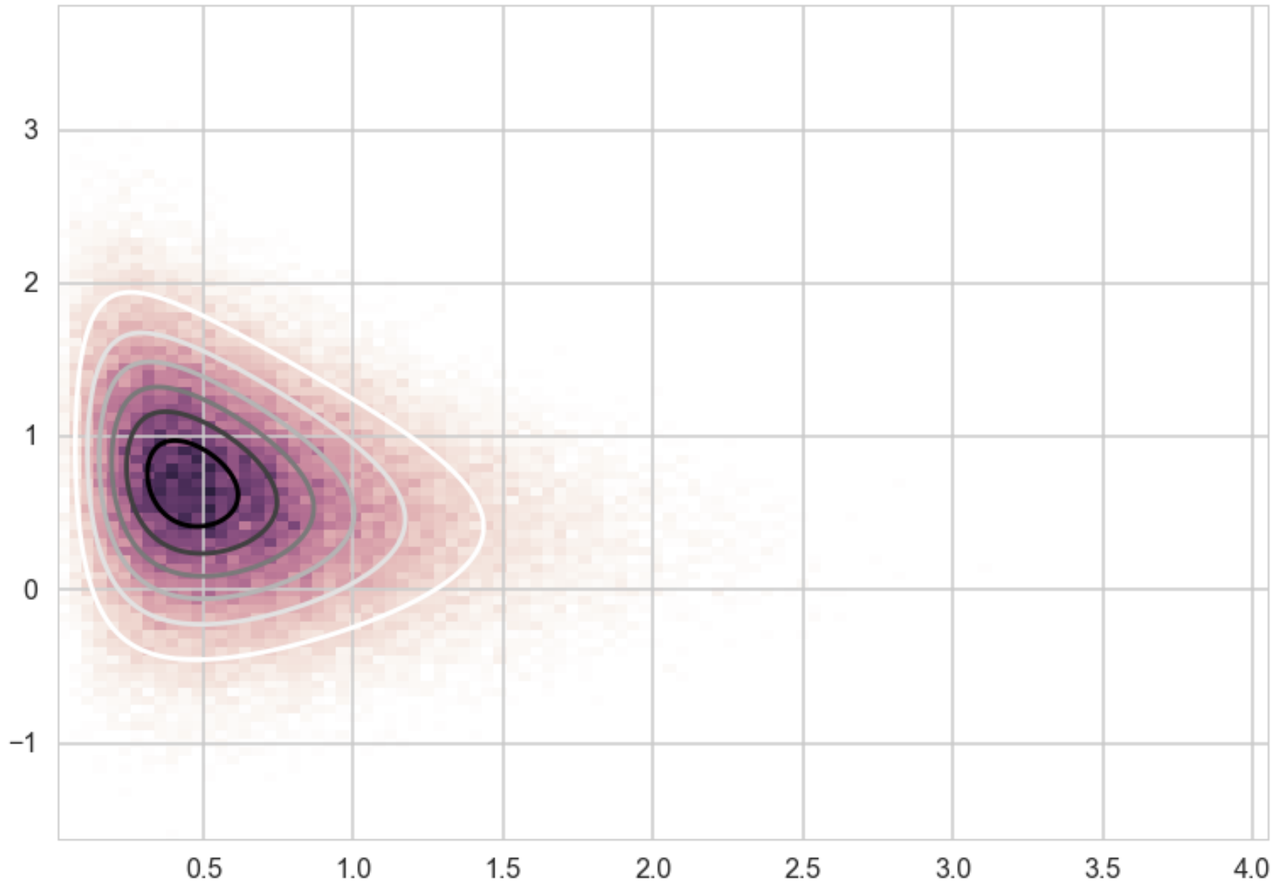
$$f(x|y) = g(y) \text{Gamma}(3, y^2+4)$$

$$f(y|x) = x^2 \exp[-y^2(1+x) + 2y] \exp[-4x]$$

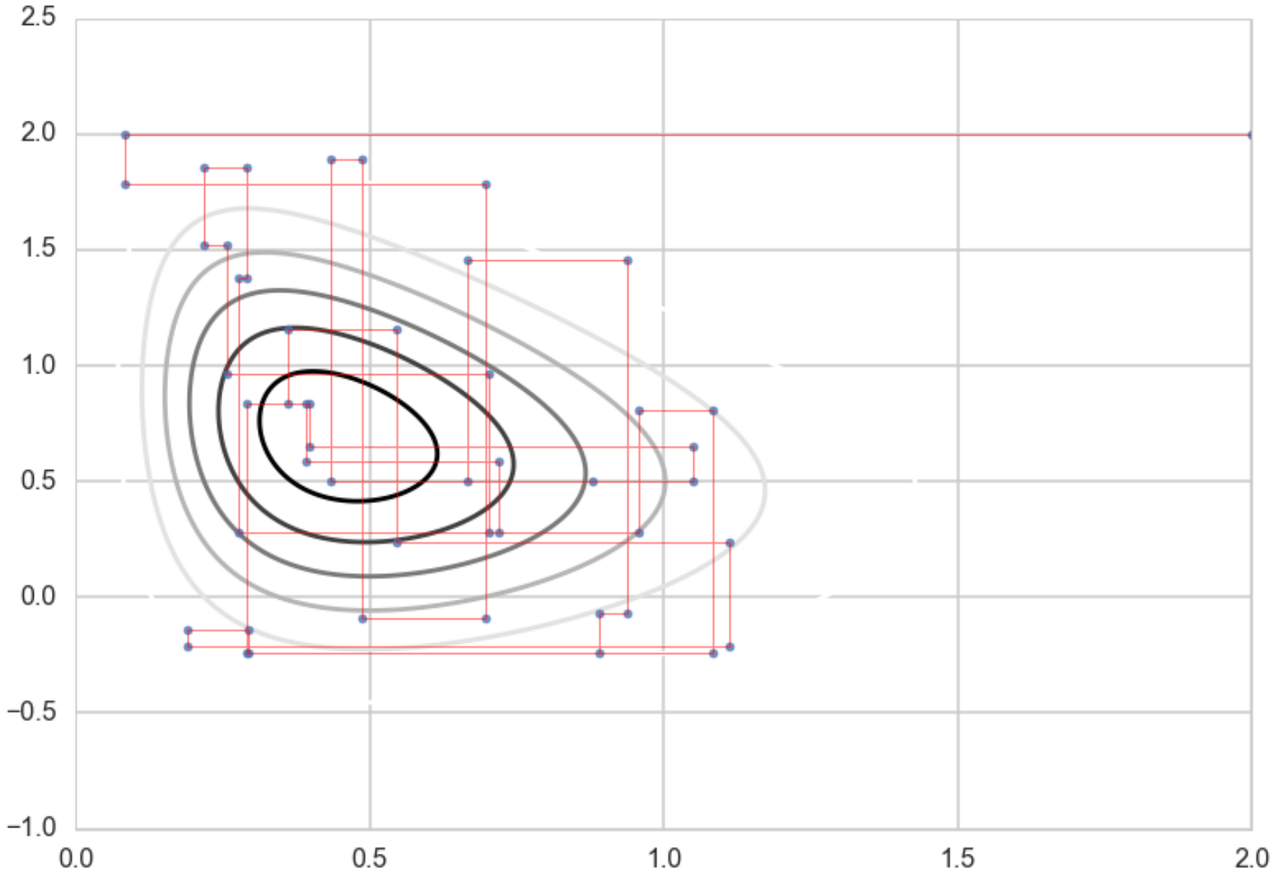
$$f(y|x) = N(\frac{1}{1+x}, \frac{1}{\sqrt{(2(1+x))}}) h(x)$$

Sampler

```
def xcond(y):
    return gamma.rvs(3, scale=1/(y*y + 4))
def ycond(x):
    return norm.rvs(1/(1+x), scale=1.0/np.sqrt(2*(x+1)))
def gibbs(xgiveny_sample, ygivenx_sample, N, start = [0,0]):
    x=start[0]
    y=start[1]
    samples=np.zeros((N+1, 2))
    samples[0,0]=x
    samples[0,1]=y
    for i in range(1,N,2):
        x=xgiveny_sample(y)
        samples[i,0]=x
        samples[i, 1]=y
        ######################
        y=ygivenx_sample(x)
        samples[i+1,0]=x
        samples[i+1,1]=y
    return samples
out=gibbs(xcond, ycond, 100000)
```

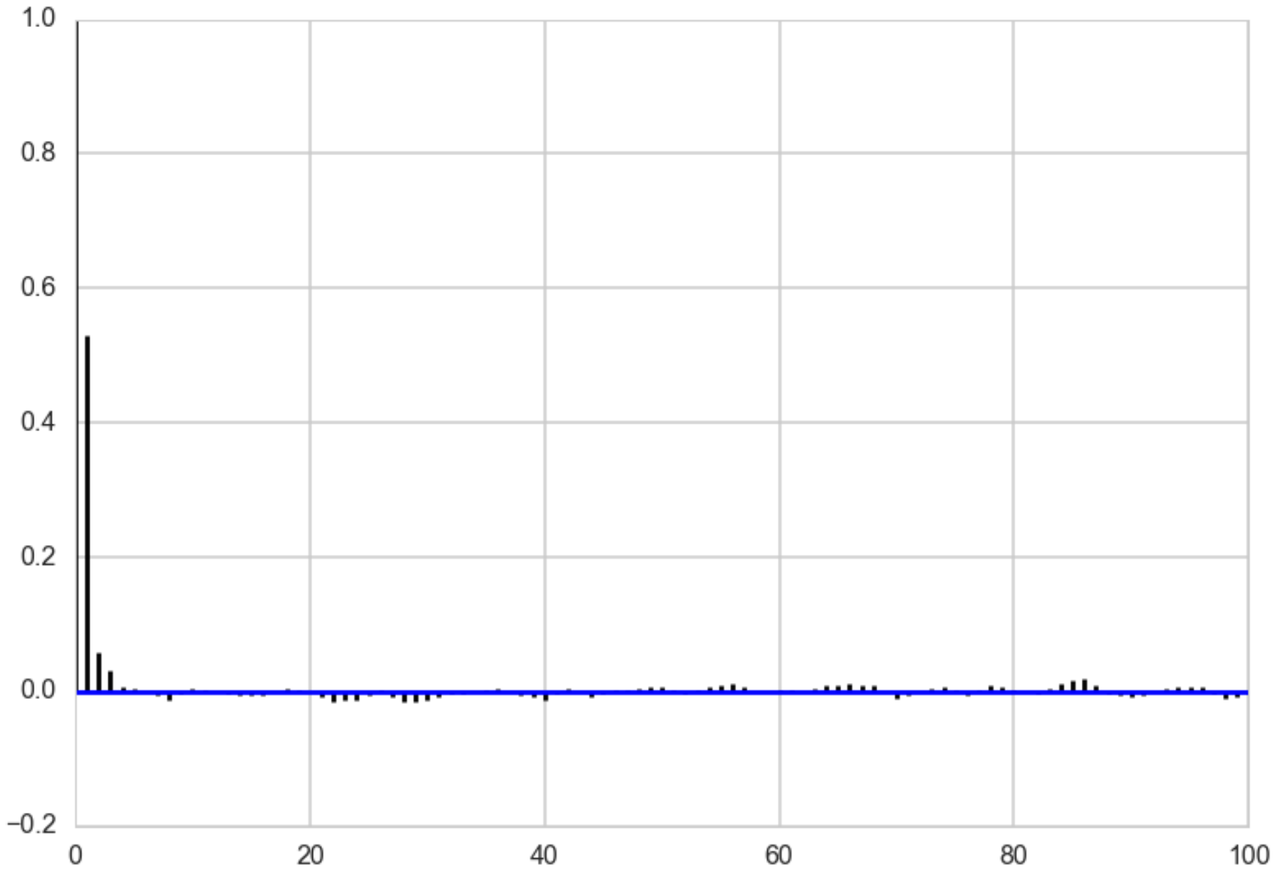


More about gibbs



- easiest is to know how to sample directly from conditionals
- moves one compoment (or one block) at a time
- all is not lost if thats not the case: can use a MH-step once stationarity has been reached
- this makes gibbs a very general idea

Autocorrelation



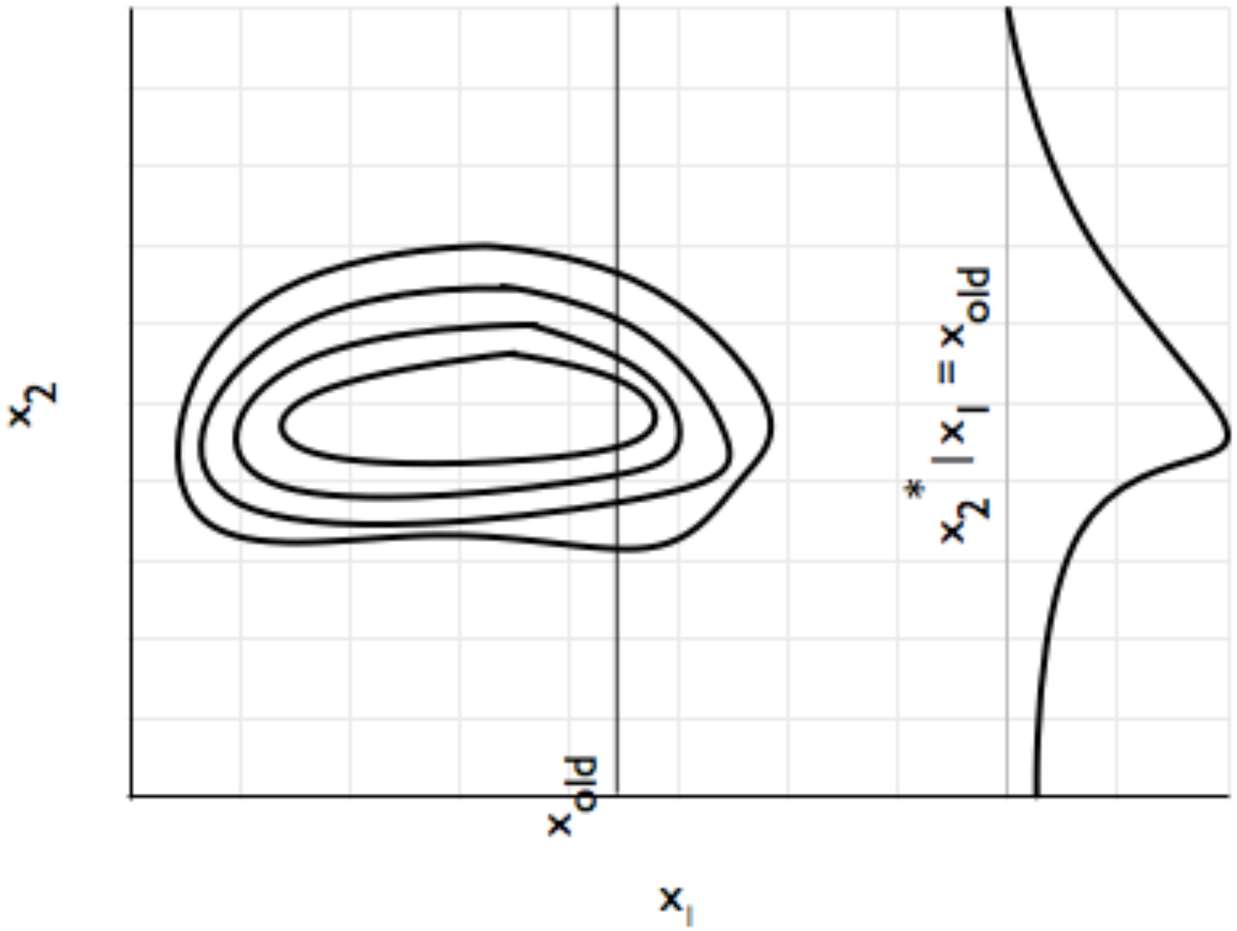
- this joint has very little autocorrelation
- highly correlated joints will have lots of autocorrelation
- thinning may be required, but as usual it depends on what you are trying to calculate.
- expectations require far fewer samples
- complete posterior characterization require many more

More Gibbs Theory

The transition kernel corresponds to this proposal:

$$q_k(x^*|x^i) = \begin{cases} p(x_k^*|x_{-k}^i) & for\ x_{-k}^* = x_{-k}^i, \\ 0 & otherwise \end{cases}$$

where x_k^i is the kth component (or block) of x at ith step, while x_{-k}^i is all other components of x at the same step



Gibbs=MH with no rejection

$$A = min(1, \frac{p(x^*)}{p(x^i)} \frac{q_k(x^i|x^*)}{q_k(x^*|x^i)})$$

$$p(x^*) = p(x^*_{-k}, x^*_k) = p(x^*_k|x^*_{-k})p(x^*_{-k})$$

$$A = min(1, \frac{p(x^*_k|x^*_{-k})p(x^*_{-k})}{p(x^i_k|x^i_{-k})p(x^i_{-k})} \frac{q_k(x^i|x^*)}{q_k(x^*|x^i)}) = min(1, \frac{p(x^*_k|x^*_{-k})p(x^*_{-k})}{p(x^i_k|x^i_{-k})p(x^i_{-k})} \frac{p(x^i_k|x^*_{-k})}{p(x^*_k|x^i_{-k})})$$

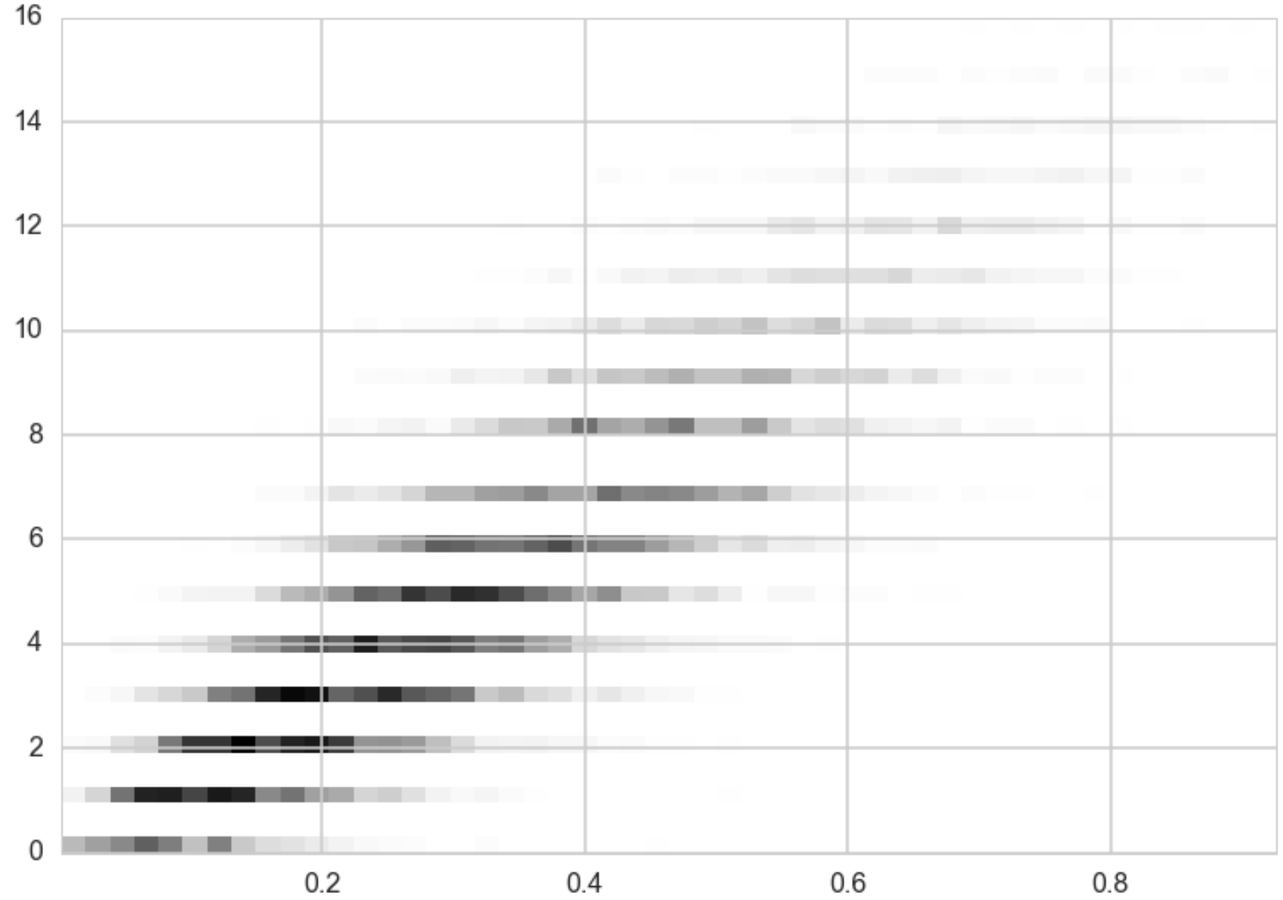
Componentwise update, $\implies x^*_{-k} = x^i_{-k}$ and A is 1!

Another example

$$p(x,y) = \binom{16}{y} x^{y+1}(1-x)^{19-y}$$

$$p(y|x) = g(x)\binom{16}{y} x^{y+1}(1-x)^{16-y}$$

$$p(y|x) \sim Binom(16, y)$$

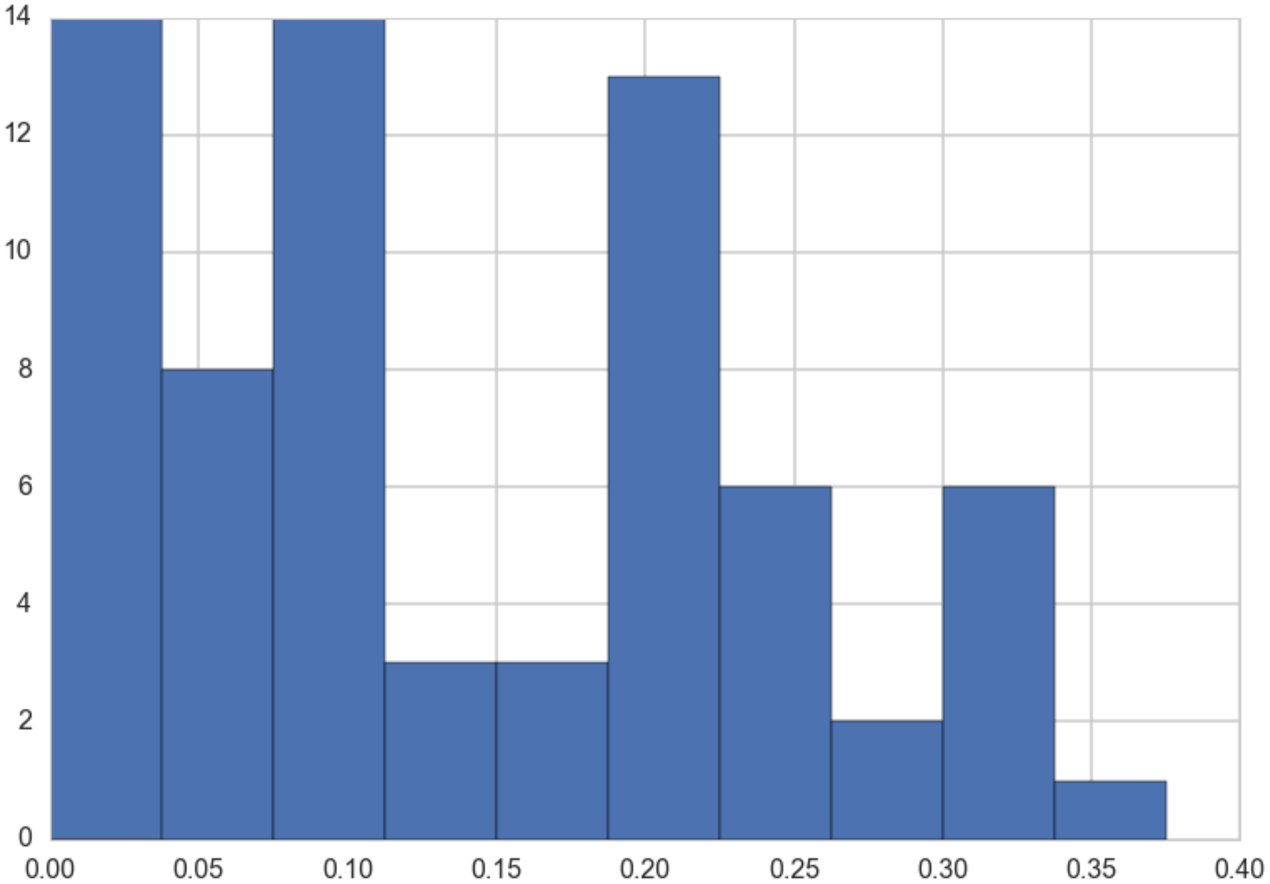


$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

$$p(x|y) = g(y)x^{\alpha+y-1}(1-x)^{\beta+n-y-1}$$

$$p(x|y) \sim Beta(y+2, n-y+4)$$

Rat Tumors



- tumors in female rats of type "F344" that recieve a particular drug, in 70 different experiments.
- mean and variance of tumor incidence: `0.13600653889043893`, `0.010557640623609196`
- 71st experiment done: 4 out of 14 rats develop tumors. Estimate the risk of tumor in the rats in the 71st experiment

Modeling

$$p(y_i|\theta_i; n_i) = Binom(n_i, y_i, \theta_i)$$

$$p(Y|\Theta; \{n_i\}) = \prod_{i=1}^{70} Binom(n_i, y_i, \theta_i)$$

Need to choose a prior $p(\Theta)$.

No Pooling

Separate priors on each θ_i:

$$\theta_i \sim Beta(\alpha_i, \beta_i).$$

$$p(\Theta|\{\alpha_i\}, \{\beta_i\}) = \prod_{i=1}^{70} Beta(\theta_i, \alpha_i, \beta_i),$$

Very overfit model with 210 parameters. VARIANCE!

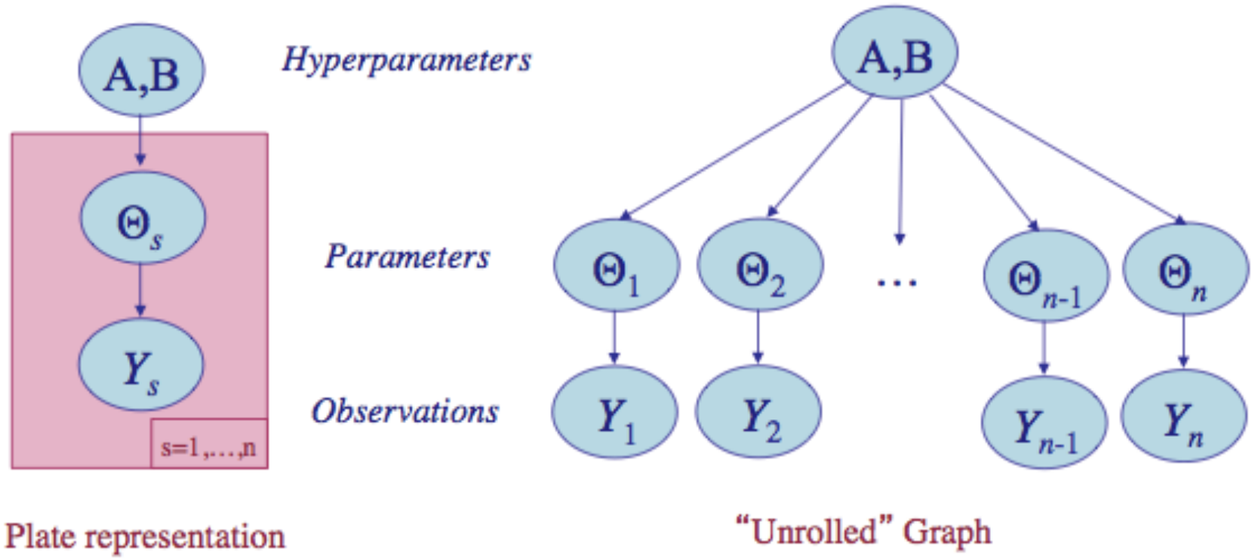
Full Pooling

Assume that there is only one θ in the problem, and set an prior on it.

Ignores any variation amongst the sampling units other than sampling variance.

Underfit model with 3 params. BIAS

Partial pooling: Hierarchical Model



θ_is drawn from "population distribution" given by a conjugate Beta prior $Beta(\alpha, \beta)$ with **hyperparameters** α and β.

$$\theta_i \sim Beta(\alpha, \beta).$$

$$p(\Theta|\alpha, \beta) = \prod_{i=1}^{70} Beta(\theta_i, \alpha, \beta).$$

Priors from data

Where do α and β come from?

Why are we calling them hyperparameters?

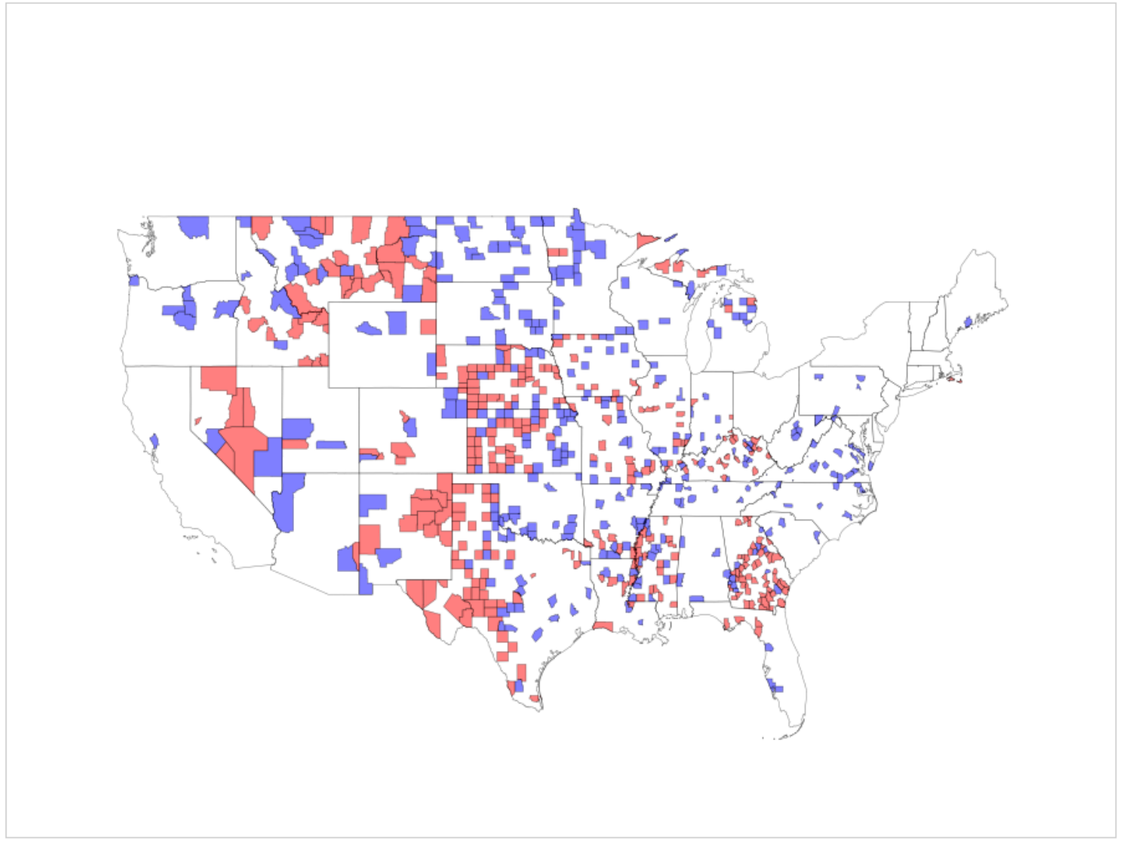
So far have assumed α and β known in priors to be weakly informative.

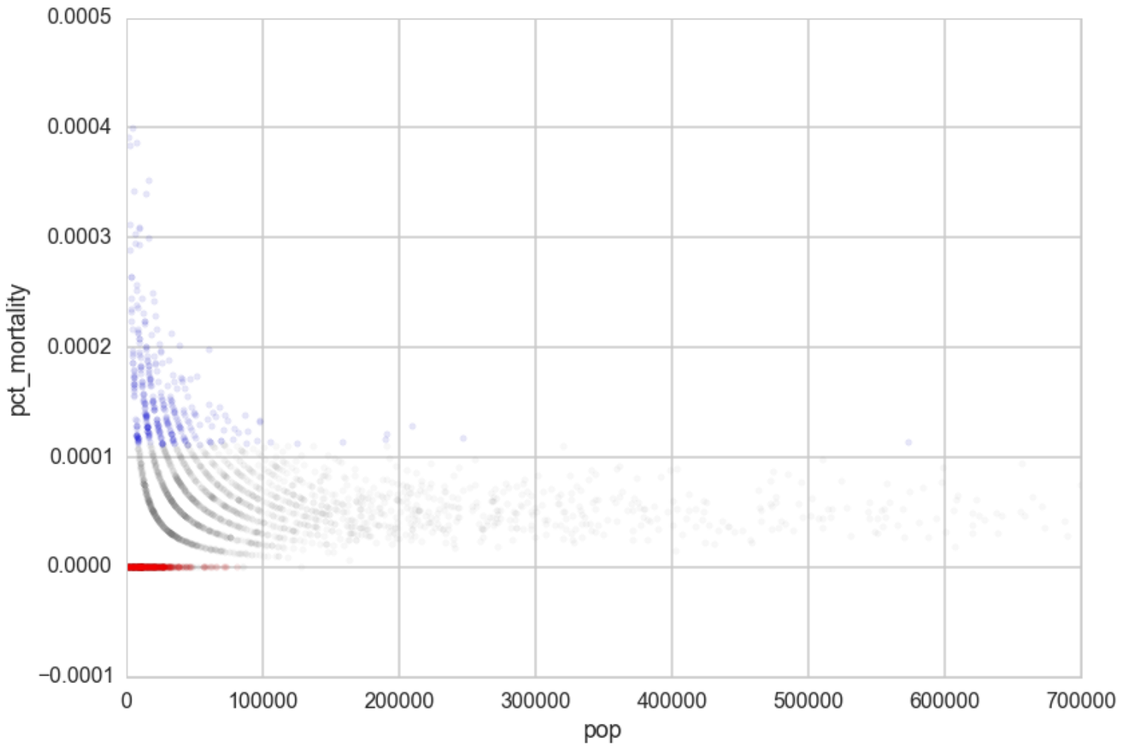
New idea: estimate priors from data. Looks like a cross-validation like setup.

Key Idea: Share statistical strength

- Some **units** (experiments) statistically more robust
- Non-robust experiments have smaller samples or outlier like behavior
- Borrow strength from all the data as a whole through the estimation of the hyperparameters
- **regularized partial pooling model** in which the "lower" parameters (θs) tied together by "upper level" hyperparameters.

Another Example: Kidney cancers





First idea: estimate directly from data

Posterior-predictive distribution, as a function of upper level parameters $\eta = (\alpha, \beta)$.

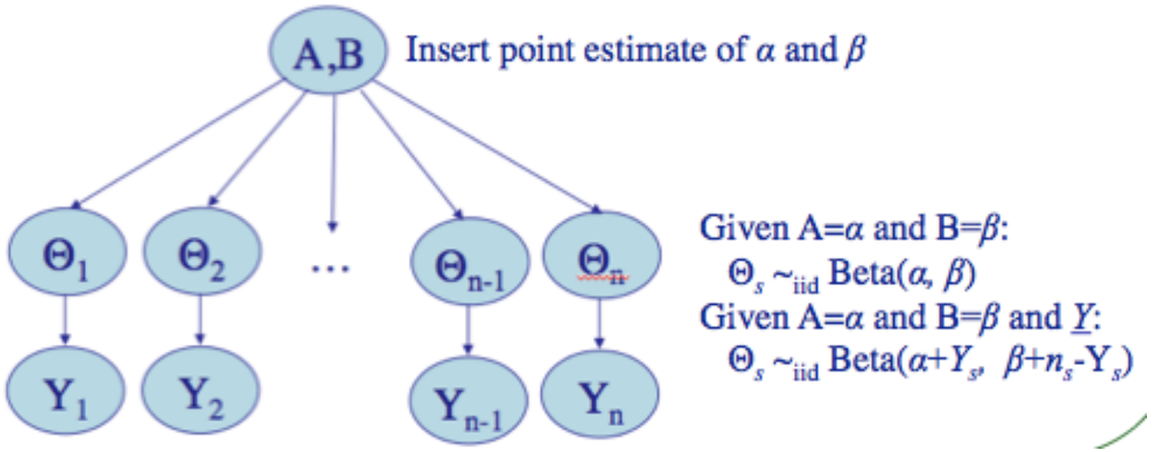
$$p(y^*|D, \eta) = \int d\theta \, p(y^*|\theta) \, p(\theta|D, \eta)$$

A likelihood with parameters η and simply use maximum-likelihood with respect to η to estimate these η using our "data" y^*

Called Empirical Bayes or Type-2 MLE

- MLE with respect to η
- involves an optimization
- unlike cross-validation, θs not-yet estimated on training set.
- indeed we marginalize over θs so can use training set.
- in practice often match moments of predictive or posterior

From prior to posterior

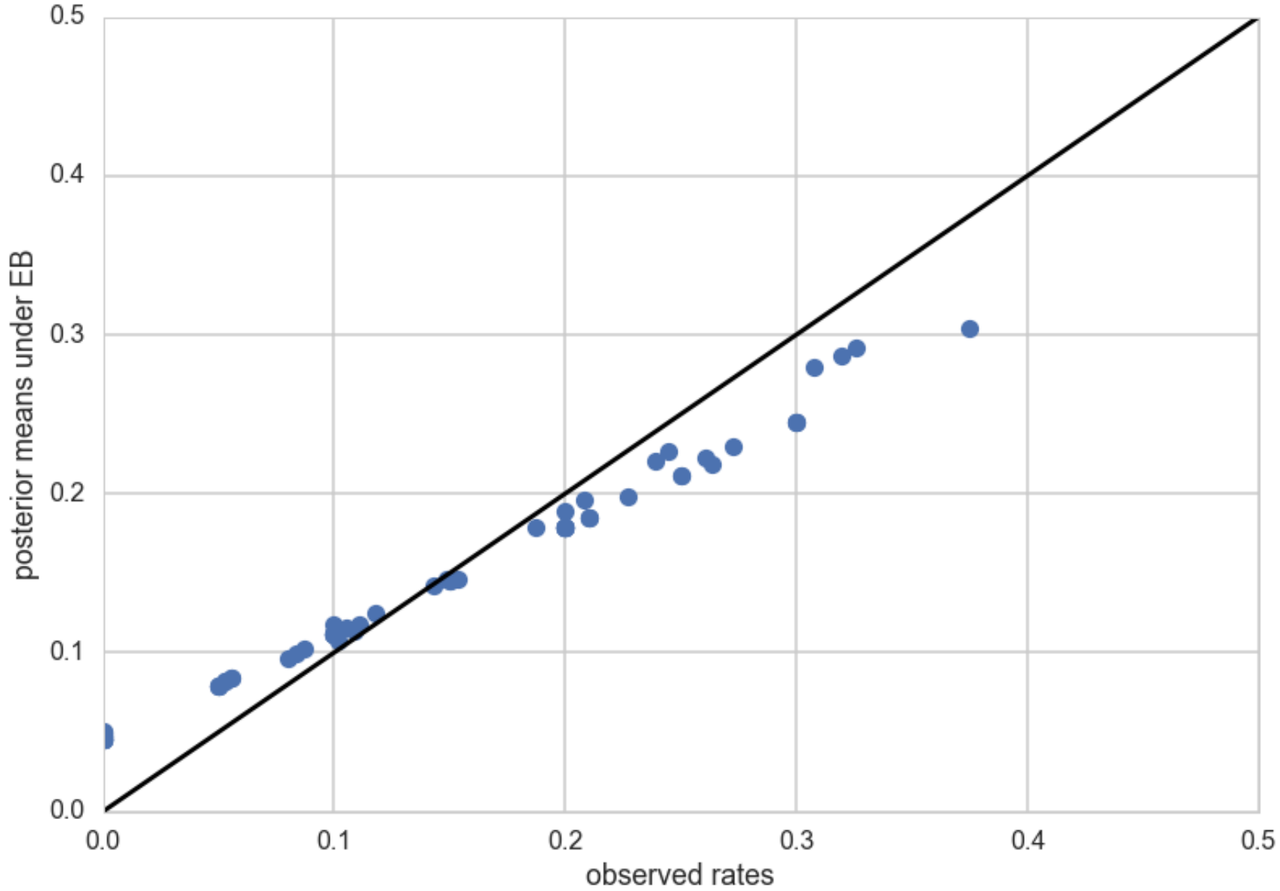


- (α, β) = (1.3777748392916778, 8.752435447153129)
- Conditional posterior distribution for each of the θ_i, given everything else is Beta:.

$$p(\theta_i|y_i, n_i, \alpha, \beta) = Beta(\alpha + y_i, \beta + n_i - y_i)$$

$$\bar{\theta}_{post,i} = \frac{\alpha + y_i}{\alpha + \beta + n_i}$$

Shrinkage in rat (tumors)



Posterior estimates shrink towards full pooling.

Now, for the 71st experiment, we have 4 out of 14 rats having tumors. The posterior estimate for this would be

$$\frac{\alpha + y_{71}}{\alpha + \beta + n_{71}}$$

```
4/14, (4+a_est)/(14+a_est+b_est)
= (0.2857142857142857, 0.2228648144982493)
```

Full Bayesian

- every optimization is a chance to overfit, would like to use integration all the way
- specify a **hyper-prior** $p(\eta)$ ($p(\alpha, \beta)$) on these hyperparameters η (α, β)
- helps us develop a computational strategy of gibbs sampling
- allows estimates of the probabilities of any one unit to borrow strength from all the data as a whole

Fully Bayesian Rat tumors

Joint Posterior:

$$p(\Theta, \alpha, \beta | Y, \{n_i\}) \propto p(\alpha, \beta) \prod_{i=1}^{70} Beta(\theta_i, \alpha, \beta) \prod_{i=1}^{70} Binom(n_i, y_i, \theta_i)$$

Conditionals:

$$p(\theta_i | y_i, n_i, \alpha, \beta) = Beta(\alpha + y_i, \beta + n_i - y_i)$$

More Conditionals

$$p(\alpha|Y,\Theta,\beta) \propto p(\alpha,\beta)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\right)^N \prod_{i=1}^{N}\theta_i^{\alpha}$$

$$P(\beta|Y,\Theta,\alpha) \propto p(\alpha,\beta)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\right)^N \prod_{i=1}^{N}(1-\theta_i)^{\beta}$$

These depend on Y and $\{n\}$ via the θ's

Sampling (sampler done in lab)

- Fix α and β, we have a Gibbs step for all of the θ_is
- For α and β, everything else fixed, use stationary metropolis step, as conditionals are not isolatable to simply sampled distributions
- when we sample for α, we will propose a new value using a normal proposal, while holding all the θs and β constant at the old value. ditto for β.

Usefulness of Hierarchical models

- a directed acyclic graph, with the observations layer at the bottom of a tree, the next layer being the intermediate parameters, and the upper layers being the hyper-parameters
- sample conditionals from parents up the tree.
- sampling downward usually requires MH steps
- simplifies further for conjugacy as many conditionals are super easy to sample

Hierarchy organizes exchangeability

- we use the notion of exchangeability at the level of 'units'.
- for our rats, the y_j were exchangeable since we had no additional information about experimental conditions.
- if specific groups of experiments came from specific laboratories, assume experiments interchangeable if from the same lab.
- lab specific α_{lab} and β_{lab} parameters
- add another level of hierarchy to draw these from hyperprior.

The levels of Bayesian analysis

Method	Definition
Maximum Likelihood	$\hat{\theta} = argmax_\theta p(D\|\theta)$
MAP estimation	$\hat{\theta} = argmax_\theta p(D\|\theta)p(\theta\|\eta)$
ML-2 (Empirical Bayes)	$\hat{\eta} = argmax_\eta \int d\theta\, p(D\|\theta)p(\theta\|\eta) = argmax_\eta p(D\|\eta)$
MAP-2	$\hat{\eta} = argmax_\eta \int d\theta\, p(D\|\theta)p(\theta\|\eta)p(\eta) = argmax_\eta p(D\|\eta)p(\eta)$
Full Bayes	$p(\theta, \eta\|D) \propto p(D\|\theta)p(\theta\|\eta)p(\eta)$