STAT 431 — Applied Bayesian Analysis — Course Notes

Gibbs Sampling for a Hierarchical Model

Fall 2022

Review

Gibbs sampling: Cycle through the parameters, updating each by replacing it with a random draw from its posterior full conditional distribution (given the most recent draws of the other parameters). Record the parameter vector at the end of each cycle.

Semi-conjugacy for a parameter: The prior is chosen such that the parameter's *prior* full conditional and its *posterior* full conditional are of the same distributional family.

If that distributional family is easily sampled, the Gibbs step for that parameter is easy.

Discovering *natural* semi-conjugacy for a parameter θ_j is like discovering natural conjugacy for the one-parameter case:

1. Write out the likelihood. Keep only the factors that involve θ_j.
2. In θ_j, does the likelihood look like the kernel of an easily-sampled distribution type?
3. If so, arrange to make a distribution of that type the (conditional) prior for θ_j.

Hyperparameters

Recall: The prior for parameter vector

$$\boldsymbol{\theta} \;=\; (\theta_1, \ldots, \theta_p)$$

may be chosen from a distributional family having

hyperparameter $\boldsymbol{\phi}$

A frequently-encountered situation:

$$\theta_1, \ldots, \theta_p \mid \boldsymbol{\phi} \quad \sim \quad indep \;\; \text{from} \;\; \pi(\,\cdot \mid \boldsymbol{\phi})$$

Under conjugacy, a choice $\boldsymbol{\phi}_0$ to define the prior is updated by Bayes' rule to a value $\boldsymbol{\phi}_1$ defining the posterior.

Hierarchical Bayes

What if we are uncertain how to chose a value of ϕ for the prior?

One solution: Place a higher-level prior, a **hyperprior**, on ϕ.

This is sometimes called **hierarchical Bayes**.

The full prior density (for both mid-level parameters and hyperparameters) could then have the form

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi})\, \pi(\boldsymbol{\phi}) \quad = \quad \left(\prod_{j=1}^{p} \pi(\theta_j \mid \boldsymbol{\phi}) \right) \pi(\boldsymbol{\phi})$$

** See below

How do we choose the hyperprior (for $\boldsymbol{\phi}$)?

If possible, choose it to be conjugate (or perhaps semi-conjugate in each hyperparameter) if $\boldsymbol{\theta}$ were observed.

In that case, $\pi(\boldsymbol{\phi})$ will be of the same distributional family as

$$\pi(\boldsymbol{\phi} \mid \boldsymbol{\theta}) \;\; = \;\; p(\boldsymbol{\phi} \mid \boldsymbol{\theta}, \boldsymbol{y})$$

where $\boldsymbol{\phi}$ is conditionally independent of the data $\boldsymbol{y}$ because of the hierarchical structure of the model.

Thus, if this family is easy to sample, $\boldsymbol{\phi}$ can simply join the Gibbs sampler, along with the elements of $\boldsymbol{\theta}$.

1) OA
2) Monday class
3) RStudio update
4) STA 414 HW run code
5) 431 HW

6) Interview stack

To discover this kind of "conjugacy" for $\boldsymbol{\phi}$, try examining the form of

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi}) \quad = \quad \prod_{j=1}^{p} \pi(\theta_j \mid \boldsymbol{\phi})$$

Does it resemble the kernel of some type of density for $\boldsymbol{\phi}$ (or for each of its elements)?

$\pi(\underline{\phi} \mid \underline{\theta}) = p(\underline{\phi} \mid \underline{\theta}, \underline{y}) \propto_{m \underline{\phi}} p(\underline{\phi}, \underline{\theta} \mid \underline{y})$

$\propto_{in \underline{\phi}} f(\underline{y} \mid \underline{\theta}) \pi(\underline{\theta} \mid \underline{\phi}) \pi(\underline{\phi})$

$\propto_{in \underline{\phi}} \pi(\underline{\theta} \mid \underline{\phi}) \pi(\underline{\phi})$

$\propto_{in \underline{\phi}} \pi(\underline{\phi} \mid \underline{\theta})$

Example: Airliner Fatalities

$$Y_i = \text{number of passenger-fatal events for airliners of type } i$$

$$N_i = \text{millions of flights (total) by airliners of type } i$$

By the Poisson approximation to the binomial,

$$Y_i \mid \lambda_i \sim indep\ \text{Poisson}(N_i \lambda_i)$$

where

$$\lambda_i = \text{passenger-fatal event rate per million flights for airliners of type } i$$

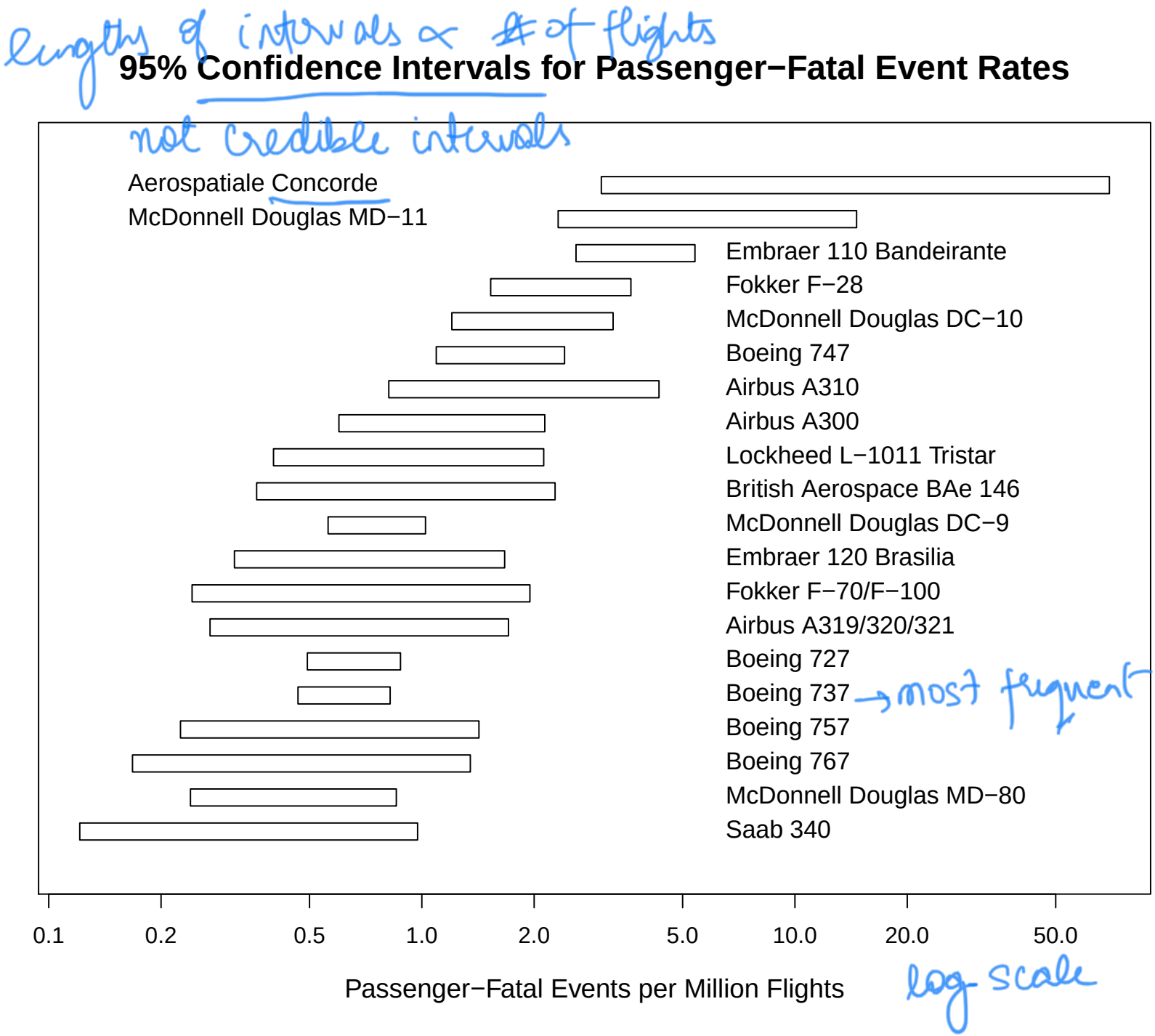
Since

$$\mathrm{E}(Y_i \mid \lambda_i) \quad = \quad N_i \lambda_i$$

a natural unbiased frequentist estimator is the sample rate

$$\hat{\lambda}_i \quad = \quad \frac{Y_i}{N_i}$$

There are also individual 95% confidence intervals for the λ_is ...


95% Confidence Intervals for Passenger−Fatal Event Rates
Aerospatiale Concorde
McDonnell Douglas MD−11
Embraer 110 Bandeirante
Fokker F−28
McDonnell Douglas DC−10
Boeing 747
Airbus A310
Airbus A300
Lockheed L−1011 Tristar
British Aerospace BAe 146
McDonnell Douglas DC−9
Embraer 120 Brasilia
Fokker F−70/F−100
Airbus A319/320/321
Boeing 727
Boeing 737
Boeing 757
Boeing 767
McDonnell Douglas MD−80
Saab 340
0.1
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0
Passenger−Fatal Events per Million Flights

Some possible research questions:

- What is the "average" passenger-fatal event rate over the different airliner types, after taking estimation uncertainties into account?

through hyperparams

- How would "pooling" information across different types of aircraft affect estimates of their individual passenger-fatal event rates λ_i?

- If these are representative of a "population" of airliner types, what range of passenger-fatal event rates might we expect for a "new" type of airliner?

Should we just analyze different airliner types separately, with a separate "noninformative" prior on each λ_i?

No, for at least these reasons:

shrinkage - towards each other

- A Bayesian framework that links the λ_is might allow better estimation ("shrinkage") of poorly-estimated cases.
- Some research questions refer to a "population" of airliner types, so the λ_is should be modeled flexibly with a distribution over that population.

Without prior information about differences between the types of airliner, we choose to model the λ_is as (conditionally) iid.

Objective

given some hyperparam

Suppose we give the λ_is a common exponential prior:

$$\lambda_i \mid \gamma \sim iid \text{ Exponential}(\gamma)$$

Why? We know that the gamma distribution is conjugate for a single Poisson mean, and the exponential is a special case of the gamma:

$$\text{Exponential}(\gamma) = \text{Gamma}(1, \gamma)$$

(Why not use the full gamma? Later.)

Shari Tomar

We now have a model with mid-level parameters and a hyperparameter:

$$\boldsymbol{\theta} \;=\; (\lambda_1, \ldots, \lambda_{20}) \qquad\qquad \boldsymbol{\phi} \;=\; \gamma$$

For hierarchical Bayes, we need a (marginal) prior for γ.

We will choose one later, based on a kind of conjugacy.

For now, just assume it has density $\pi(\gamma)$.

The full prior density has the form

$$\pi(\lambda_1, \ldots, \lambda_{20}, \gamma) = \left(\prod_{i=1}^{20} \pi(\lambda_i \mid \gamma) \right) \pi(\gamma)$$

so the full posterior density is, by Bayes' rule,

$$\begin{aligned} & p(\lambda_1, \ldots, \lambda_{20}, \gamma \mid y_1, \ldots, y_{20}) \\ & \qquad \propto \quad f(y_1, \ldots, y_{20} \mid \lambda_1, \ldots, \lambda_{20}) \, \pi(\lambda_1, \ldots, \lambda_{20}, \gamma) \\ & \qquad = \quad \left(\prod_{i=1}^{20} f(y_i \mid \lambda_i) \right) \left(\prod_{i=1}^{20} \pi(\lambda_i \mid \gamma) \right) \pi(\gamma) \\ & \qquad = \quad \left(\prod_{i=1}^{20} f(y_i \mid \lambda_i) \, \pi(\lambda_i \mid \gamma) \right) \pi(\gamma) \end{aligned}$$

To obtain the (posterior) full conditional of λ_i, we pick out the factors that involve λ_i

$$\begin{aligned} f(y_i \mid \lambda_i)\, \pi(\lambda_i \mid \gamma) \quad &\propto \quad \lambda_i^{y_i}\, e^{-N_i \lambda_i} \cdot e^{-\gamma \lambda_i} \\ &= \quad \lambda_i^{(y_i+1)-1}\, e^{-(N_i+\gamma)\lambda_i} \end{aligned}$$

which we recognize as the kernel of $\mathrm{Gamma}(y_i + 1,\, N_i + \gamma)$.

Moreover, since this does not depend on any of the other λs, we conclude that $\lambda_1, \ldots, \lambda_{20}$ are independent in their (joint) posterior full conditional.

The (posterior) full conditional of γ would be obtained by picking out the factors that involve γ

$$\left(\prod_{i=1}^{20} \pi(\lambda_i \mid \gamma)\right) \pi(\gamma) = \left(\prod_{i=1}^{20} \gamma e^{-\gamma \lambda_i}\right) \pi(\gamma)$$

$$= \gamma^{20} \exp\left(-\left(\sum_{i=1}^{20} \lambda_i\right)\gamma\right) \pi(\gamma)$$

Q: As a function of γ, what kind of density kernel do the factors other than $\pi(\gamma)$ resemble?

Trying

$$\gamma \sim \text{Gamma}(\alpha, \beta)$$

we find the (posterior) full conditional density of γ to be proportional to

$$\gamma^{20} \exp\left(-\left(\sum_{i=1}^{20} \lambda_i\right)\gamma\right) \cdot \gamma^{\alpha-1} e^{-\beta\gamma}$$

$$= \gamma^{20+\alpha-1} \exp\left(-\left(\sum_{i=1}^{20} \lambda_i + \beta\right)\gamma\right)$$

which is the kernel of $\text{Gamma}\left(20+\alpha, \sum_{i=1}^{20} \lambda_i + \beta\right)$.

(Note: α and β may be chosen to reflect prior information).

Non-informative

Dangers of $\alpha, \beta \to 0$: Here none, (maybe improper).

So the Bayesian hierarchical model ($i = 1, \ldots, 20$)

$$\begin{aligned} Y_i \mid \lambda_i, \gamma &\sim \textit{indep}\ \text{Poisson}(N_i \lambda_i) \\ \lambda_i \mid \gamma &\sim \textit{iid}\ \text{Exponential}(\gamma) \\ \gamma &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

leads to these easily-sampled posterior full conditionals:

$$\lambda_i \mid \gamma, \boldsymbol{y} \sim \textit{indep}\ \text{Gamma}(y_i + 1,\ N_i + \gamma)$$

$$\gamma \mid \lambda_1, \ldots, \lambda_{20}, \boldsymbol{y} \sim \text{Gamma}\left(20 + \alpha,\ \sum_{i=1}^{20} \lambda_i + \beta\right)$$

Upper limit of highly informative prior here here $\alpha = 1$,

R Example 3.5:

Gibbs Sampler for Hierarchical Model (Poisson Rates)

$\alpha = \beta = 0.001$ less informative hyper prior

Remark:

The joint sampling of the vector of λs as a single Gibbs step is an example of a **block** Gibbs update.

Block updating can sometimes improve the convergence rate of Gibbs MCMC.

Remark:

What about trying a gamma prior (rather than an exponential prior) for the λs?

$$\lambda_i \mid \delta, \gamma \sim iid \text{ Gamma}(\delta, \gamma)$$

Then we would need a hyperprior for δ.

You may verify that, though there is still a "conjugate" gamma prior for γ, there does not seem to be any easily-recognized PDF kernel that would give conjugacy for δ.

Beyond Gibbs

What if no available type of conjugacy is satisfactory?

Then Gibbs sampling is not as easy, but there are other MCMC options.

The **Metropolis-Hastings (MH)** sampling algorithm provides another type of MCMC that is more generally applicable and does not require conjugacy.

MH can also be used to implement a Gibbs step that does not permit easy sampling from the full conditional.

(See BSM, Sec. 3.2.2, if interested.)