Harvard Department of Government 2017
Chapter 2, Specifying Bayesian Models

JEFF GILL
Visiting Professor, Spring 2025

Three General Steps (Review)

I. Specify a probability model for unknown parameter values that includes some prior knowledge about the parameters if available.

II. Update knowledge about the unknown parameters by conditioning this probability model on observed data.

III. Evaluate the fit of the model to the data and the sensitivity of the conclusions to the assumptions.

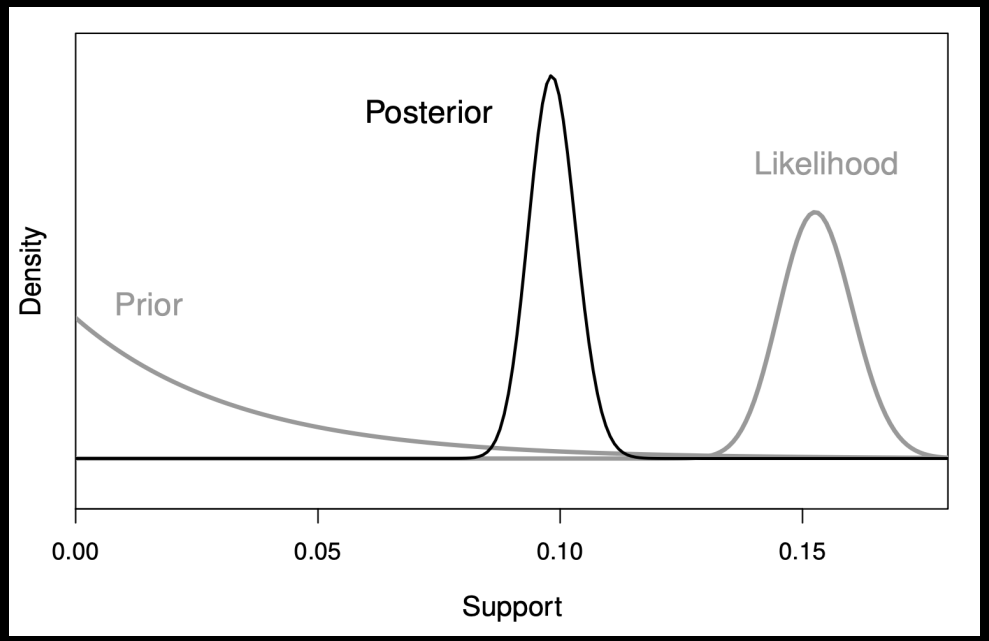
Simple Mechanics (Review)

$$\pi(\theta|\mathbf{x}) = \frac{p(\theta)L(\theta|\mathbf{x})}{\int_{\Theta} p(\theta)L(\theta|\mathbf{x})d\theta}$$

$$\propto p(\theta)L(\theta|\mathbf{x})$$

Posterior Probability $\propto$ Prior Probability $\times$ Likelihood Function

Bayesian Inference Illustration



Posterior $\propto$ Prior $\times$ Likelihood

The Denominator

- The "integrated likelihood" is the denominator of Bayes law calculated here by:

$$p(\mathbf{x}) = \int \underbrace{L(\theta|\mathbf{x})p(\theta)}_{\text{likelihood} \times \text{prior}} \, d\theta$$

- This is also called the "marginal likelihood," the "marginal probability of the data," or the predictive probability of the data".
- Why do we treat this as a constant?
- This quantity is often ignored since it can be recovered later, but it is important in Bayesian model comparison.

The Likelihood Function

- Assume that:

$$x_1, x_1, \ldots, x_n \sim \text{ iid } f(x|\theta),$$

  where θ is a parameter that is critical to the data generation process (DGP).

- Since these values are independent, the joint distribution of the observed data is just the product of their individual PDF/PMFs:

$$f(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

- But once we observe the the data $\mathbf{x}$ is fixed.

- It is θ that is unknown, so rewrite the joint distribution function according to:

$$f(\mathbf{x}|\theta) = L(\theta|\mathbf{x}).$$

- Note that this is a purely *notational* change, nothing is different mathematically.

The Likelihood Function

- Fisher (1922) justifies this because at this point we know $\mathbf{x}$.

$$f(\mathbf{x}|\theta) \longrightarrow L(\theta|\mathbf{x}).$$

- A semi-Bayesian justification works as follows, we want to perform:

$$p(\mathbf{x}|\theta) = \frac{p(\mathbf{x})}{p(\theta)} p(\theta|\mathbf{x}).$$

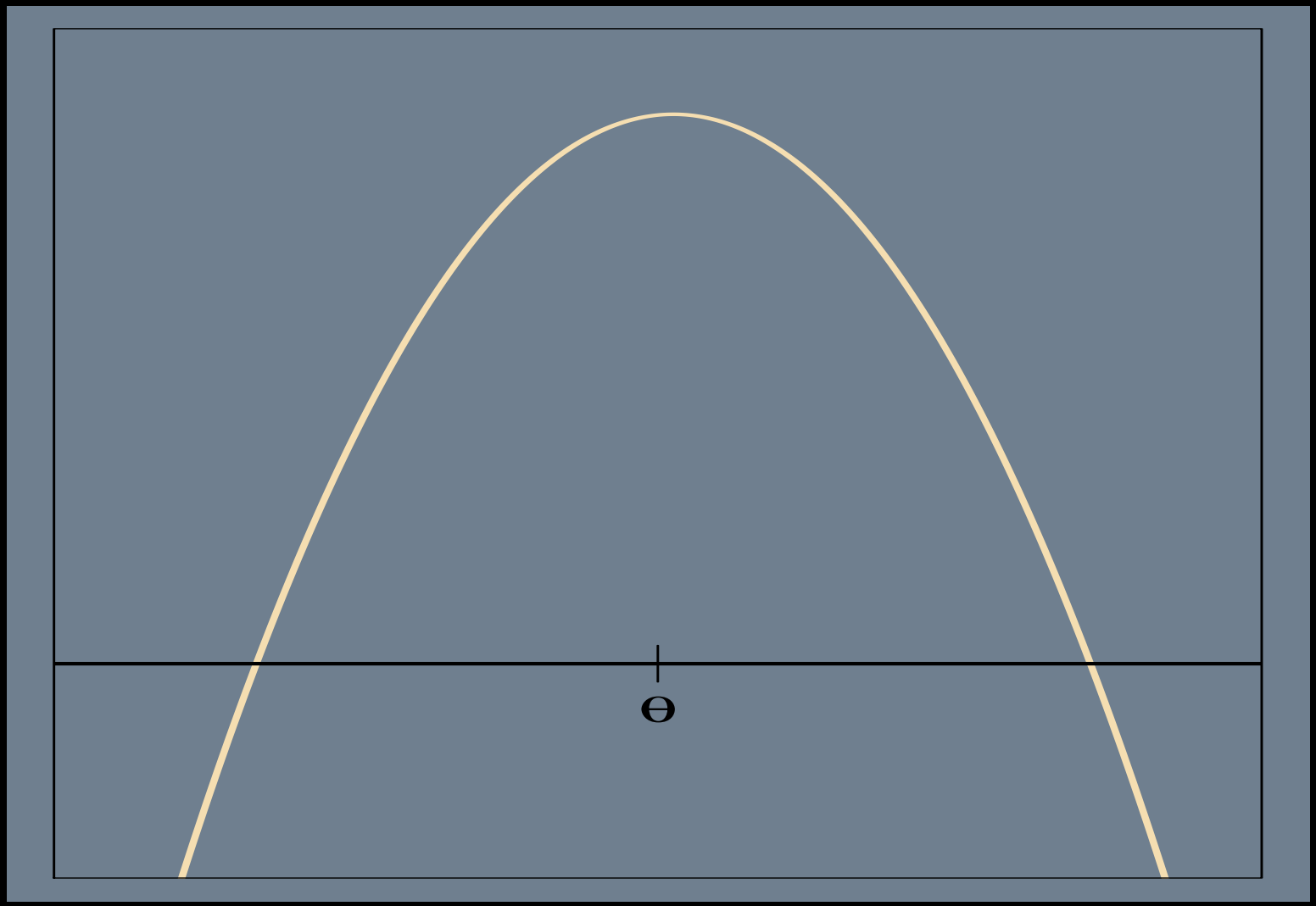
  but $p(\mathbf{x}) = 1$ since the data has already occurred, and if we put a finite uniform prior on θ over its finite allowable range (support), then $p(\theta) = 1$.

- Therefore:

$$p(\mathbf{x}|\theta) = \frac{1}{1} p(\theta|\mathbf{x}) = p(\theta|\mathbf{x}).$$

- The only caveat here is the finiteness of the support of θ.

Generic Likelihood Function Illustration



Poisson MLE

- Start with the Poisson PMF for x_i:

$$p(X = x_i) = f(x_i|\theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!},$$

which requires the assumptions: non-concurrence of arrivals, the number of arrivals is proportion to the time of study, this rate is constant over the time, and there is no serial correlation of arrivals.

- The likelihood function is created from the joint distribution:

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-\theta}\theta^{x_1}}{x_1!}\frac{e^{-\theta}\theta^{x_2}}{x_2!}\cdots\frac{e^{-\theta}\theta^{x_n}}{x_n!} = e^{-n\theta}\theta^{\sum x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1}.$$

- Suppose we have the data: $\mathbf{x} = \{5, 1, 1, 1, 0, 0, 3, 2, 3, 4\}$, then the likelihood function is:

$$L(\theta|\mathbf{x}) = \frac{e^{-10\theta}\theta^{20}}{207360},$$

which is the probability of observing *this* exact sample.

Poisson MLE

- It is often easier to deal the logarithm of the MLE:

$$\log L(\theta|\mathbf{x}) = \ell(\theta|\mathbf{x}) = \log\left(e^{-n\theta}\theta^{\sum x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1}\right) = -n\theta + \sum_{i=1}^{n} x_i \log(\theta) - \log\left(\prod_{i=1}^{n} x_i!\right).$$

- For our small example this is:

$$\ell(\theta|\mathbf{x}) = -10\theta + 20\log(\theta) - \underbrace{\log(207360)}_{12.242}.$$

- Importantly, for the family of functions that we will use the likelihood function and the log-likelihood function have the same mode (maximum of the function) for θ.
- They are both guaranteed to be concave to the x-axis.

Obtaining the Poisson MLE

- Freshman calculus: where is the maximum of the function? At the point when first derivative of the function equals zero.
- So take the first derivative, set it equal to zero, and solve.
- $\frac{d}{d\theta}\ell(\theta|\mathbf{x}) \equiv 0$ is called the likelihood equation.
- For the example:

$$\ell(\theta|\mathbf{x}) = -10\theta + 20\log(\theta) - \underbrace{\log(207360)}_{12.242}.$$

Taking the derivative, and setting equal to zero:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -10 + 20\theta^{-1} \equiv 0,$$

so that $20\theta^{-1} = 10$, and therefore $\hat{\theta} = 2$ (note the hat).

Obtaining the Poisson MLE

- More generally:

$$\ell(\theta|\mathbf{x}) = -n\theta + \sum_{i=1}^{n} x_i \log(\theta) - \log\left(\prod_{i=1}^{n} x_i!\right)$$

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i \equiv 0$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{\mathbf{x}}$$

- It is *not* true that the MLE is always the data mean.

General Steps

- This process is import to us:
  1. Identify the PMF or PDF.
  2. Create the likelihood function from the joint distribution of the observed data.
  3. Change to the log for convenience.
  4. Take the first derivative with respect to the parameter of interest.
  5. Set equal to zero.
  6. Solve for the MLE.

Poisson Example in R

```
# POISSON LIKELIHOOD AND LOG-LIKELIHOOD FUNCTION
llhfunc<-function(X,p,do.log=TRUE) {
        d <- rep(X,length(p))
        q.vec <- rep(length(y.vals),length(p)); p.vec <- rep(p,q.vec)
        print(q.vec)
        d.mat <- matrix(dpois(d,p.vec,log=do.log),ncol=length(p))
        print(d.mat)
        if (do.log==TRUE) apply(d.mat,2,sum)
        else apply(d.mat,2,prod)
}
```

Poisson Example in R

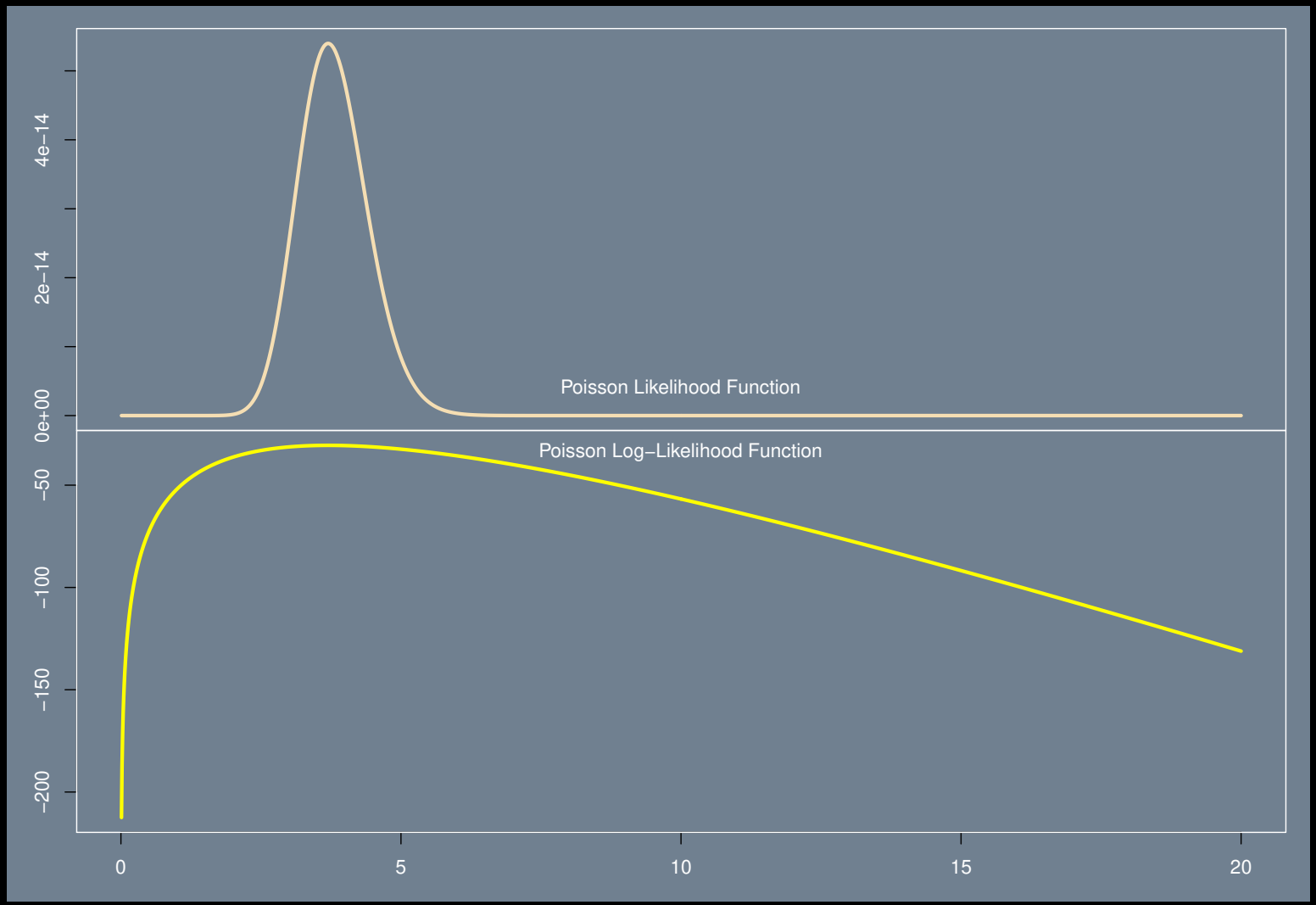
```
# HERE'S A TEST FUNCTION
y.vals<-c(1,3,1,5,2,6,8,11,0,0)
llhfunc(y.vals,c(4,30))
[1] 10 10
          [,1]    [,2]
 [1,] -2.6137 -26.599
 [2,] -1.6329 -21.588
 [3,] -2.6137 -26.599
 [4,] -1.8560 -17.782
 [5,] -1.9206 -23.891
 [6,] -2.2615 -16.172
 [7,] -3.5142 -13.395
 [8,] -6.2531 -10.089
 [9,] -4.0000 -30.000
[10,] -4.0000 -30.000
[1]  -30.666 -216.114
```

Poisson Example in R

```
# USE THE R CORE FUNCTION FOR OPTIMIZING, par=STARTING VALUES,
# control=list(fnscale=-1) INDICATES A MAXIMIZATION, bfgs=QUASI-NEWTON ALGORITHM
mle <- optim(par=1,fn=llhfunc,X=y.vals,control=list(fnscale=-1),method="BFGS")

# MAKE A PRETTY GRAPH OF THE LOG AND NON-LOG VERSIONS
ruler <- seq(from=.01, to=20, by= .01)
poison.ll <- llhfunc(y.vals,ruler)
poison.l <- llhfunc(y.vals,ruler,do.log=FALSE)

par(oma=c(3,3,1,1),mar=c(0,0,0,0),mfrow=c(2,1))
plot(ruler,poison.l,col="wheat",type="l",xaxt="n",lwd=3)
text(mean(ruler),mean(poison.l),"Poisson Likelihood Function")
plot(ruler,poison.ll,col="yellow",type="l",lwd=3)
text(mean(ruler),mean(poison.ll)/2,"Poisson Log-Likelihood Function")
```


Poisson Likelihood Function
Poisson Log−Likelihood Function
4e−14
2e−14
0e+00
−50
−100
−150
−200
0
5
10
15
20

Measuring the Uncertainty of the MLE

- The first derivative measures slope and the second derivative measures "curvature" of the function at a given point.
- The more peaked the function is at the MLE, the more "certain" the data are about this estimator.
- The square root of the negative inverse of the expected value of the second derivative is the SE of the MLE.
- In multivariate terms for vector $\boldsymbol{\theta}$, we take the negative inverse of the expected *Hessian*.
- Poisson example:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

$$\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{x}) = \frac{d}{d\theta}\left(\frac{d}{d\theta}\ell(\theta|\mathbf{x})\right) = -\theta^{-2}\sum_{i=1}^{n} x_i$$

- The expected value (estimate) of θ is the MLE, so:

$$SE(\hat{\theta}) = \frac{\hat{\theta}^2}{\sum_{i=1}^{n} x_i} = \frac{\bar{\mathbf{x}}^2}{n\bar{\mathbf{x}}} = \frac{\bar{\mathbf{x}}}{n}.$$

Multivariable MLE

- Now $\boldsymbol{\theta}$ is a vector of coefficients to be estimated (eg. regression).
- The Score Function is:

$$\dot{\ell}(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\mathbf{x})$$

  which we use to get the MLE $\hat{\boldsymbol{\theta}}$.

- The Hessian Matrix is:

$$\mathbf{H} = \frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

  which we use to get the SE of the MLE.

- The information matrix is:

$$\mathbf{I} = -\mathbb{E}(f)\left[\left.\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}}\right] \equiv \mathbb{E}(bbbh)\left[\left.\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}'}\right|_{\hat{\boldsymbol{\theta}}}\right]$$

  where the equivalence of these forms is called the *information equality*.

- The variance-covariance of $\hat{\boldsymbol{\theta}}$ is produced by:

$$\boldsymbol{\Sigma} = \mathbf{I}^{-1}$$

Properties of the MLE (Birnbaum 1962)

- Consistency:

$$\text{plim}\hat{\theta} = \theta.$$

- Asymptotic Normality:

$$\hat{\theta} \underset{a}{\sim} N\left(\theta, I(\theta)^{-1}\right) \quad \text{where } I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\right].$$

- Asymptotic Efficiency: no other estimator has lower variance, the variance of the MLE meets the Crámer-Rao Lower Bound.
- Invariance To Reparameterization:

$$\gamma = c(\theta) \implies \hat{\gamma} = c(\hat{\theta}).$$

Exponential PDF MLE

- Assume: $x_i, i = 1, \ldots, n$ iid with $f(x|\theta) = \frac{1}{\theta}\exp[-x/\theta]$.
- This gives:

$$L(\theta|\mathbf{x}) = \theta^{-n}\exp\left[-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right] \qquad \ell(\theta|\mathbf{x}) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

- Taking the derivative, setting equal to zero, and solving, gives:

$$\frac{d}{d\theta}\ell(\theta|\mathbf{x}) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i \equiv 0 \quad \Longrightarrow \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

- The curvature for this estimate is:

$$\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{x}) = \frac{d}{d\theta}\left(-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i\right) = n\theta^{-2} - 2\theta^{-3}\sum_{i=1}^{n} x_i = n\theta^{-2}\left[1 - 2\theta^{-1}\bar{x}\right] = \mathbf{H}.$$

- So the variance of the MLE is given by:

$$-\mathbf{H}^{-1} = -\frac{\theta^2}{n(1-2\theta^{-1}\bar{x})} = \frac{\theta^2/n}{2\theta^{-1}\bar{x} - 1} \quad \Longrightarrow \quad \mathbb{E}\left[-\mathbf{H}^{-1}\right] = \frac{\bar{x}^2/n}{2\bar{x}^{-1}\bar{x} - 1} = \frac{\bar{x}^2}{n}.$$

Bayes' Law for Multiple Events

- It would be extremely limiting if Bayes' Law only applied to two alternative events.
- Suppose we observe some data $\mathbf{D}$ and are interested in the relative probabilities of three events A, B, and C conditional on these data.
- Thinking just about event A, although any of the three could be selected, we know from Bayes' Law that:

$$p(A|\mathbf{D}) = \frac{p(\mathbf{D}|A)p(A)}{p(\mathbf{D})}$$

  where D is a generic name for some data.
- From the Law of Total Probability:

$$\begin{aligned} p(\mathbf{D}) &= p(A \cap \mathbf{D}) + p(B \cap \mathbf{D}) + p(C \cap \mathbf{D}) \\ &= p(\mathbf{D}|A)p(A) + p(\mathbf{D}|B)p(B) + p(\mathbf{D}|C)p(C). \end{aligned}$$

Bayes' Law for Multiple Events

- Substituting this into Bayes' Law:

$$p(A|\mathbf{D}) = \frac{p(\mathbf{D}|A)p(A)}{p(\mathbf{D}|A)p(A) + p(\mathbf{D}|B)p(B) + p(\mathbf{D}|C)p(C)},$$

  which demonstrates that the conditional distribution for any of the rival hypotheses can be produced as long as there exist unconditional distributions for the three rival hypotheses, $p(A)$, $p(B)$, and $p(C)$, and three statements about the probability of the data given these three hypotheses, $p(\mathbf{D}|A)$, $p(\mathbf{D}|B)$, $p(\mathbf{D}|C)$.

- So $p(A|\mathbf{D})$, can be determined through Bayes' Law to look at the weight of evidence for any one of several rival hypotheses or claims.

More General Notation

- Denote the three hypotheses as $\theta_i,\ i = 1, 2, 3$, so more generally:

$$p(\theta_i|\mathbf{D}) = \frac{p(\mathbf{D}|\theta_i)p(\theta_i)}{\sum_{j=1}^{3} p(\mathbf{D}|\theta_j)p(\theta_j)}$$

  for the posterior distribution of θ_i.

- This is much more in line with standard Bayesian models in the social and behavioral sciences because it allows us to compactly state Bayes' Law for any number of discrete outcomes/hypotheses, say k, for instance:

$$p(\theta_i|\mathbf{D}) = \frac{p(\theta_i)p(\mathbf{D}|\theta_i)}{\sum_{j=1}^{k} p(\theta_j)p(\mathbf{D}|\theta_j)}.$$

- Consider also that the denominator of this expression averages over the θ variables and therefore just produces the *marginal distribution of the sample data*, which we could overtly label as $p(\mathbf{D})$.

- Doing this provides a form that very clearly looks like the most basic form of Bayes' Law: $p(\theta_i|\mathbf{D}) = p(\theta_i)p(\mathbf{D}|\theta_i)/p(\mathbf{D})$.

Standard Bayesian Conventions

- Uncertainty always described with probability.
- The use of *precisions* rather than *variances*.
- Posterior description with quantiles.
- Required statement of all statistical assumptions.
- Much less emphasis on asymptotics.

Reporting Posterior Results

- Consider a single parameter θ and some generic (unspecified structure) data $\mathbf{D}$.
- Bayesians generally report $p(\theta|\mathbf{D})$ to readers via distributional summaries such as means, modes, quantiles, probabilities over regions, traditional-level probability intervals, and graphical displays.
- Once the posterior distribution has been calculated, everything about is known and it is entirely up to the researcher to highlight features of interest.
- Often it is convenient to report the posterior mean and variance in papers and reports since this is what non-Bayesians do by default.
- The posterior mean:

$$E[\theta|\mathbf{D}] = \int_{-\infty}^{\infty} \theta \pi(\theta|\mathbf{D}) d\theta$$

Reporting Posterior Results

- The posterior variance:

$$
\begin{aligned}
\text{Var}[\theta|\mathbf{D}] &= E\left[(\theta - E[\theta|\mathbf{D}])^2|\mathbf{D}\right] \\
&= \int_{-\infty}^{\infty} (\theta - E[\theta|\mathbf{D}])^2 \pi(\theta|\mathbf{D}) d\theta \\
&= \int_{-\infty}^{\infty} \left(\theta^2 - 2\theta E[\theta|\mathbf{D}]) + E[\theta|\mathbf{D}]^2\right) \pi(\theta|\mathbf{D}) d\theta \\
&= \int_{-\infty}^{\infty} \theta^2 \pi(\theta|\mathbf{D}) d\theta - 2E[\theta|\mathbf{D}] \int_{-\infty}^{\infty} \theta \pi(\theta|\mathbf{D}) d\theta + \left(\int_{-\infty}^{\infty} \theta \pi(\theta|\mathbf{D}) d\theta\right)^2 \\
&= E[\theta^2|\mathbf{D}] - E[\theta|\mathbf{D}]^2
\end{aligned}
$$

Summarizing a Posterior

- Suppose we had data, $\mathbf{D}$, distributed $p(\mathbf{D}|\theta) = \theta e^{-\theta \mathbf{D}}$, which can be either a single scalar or a vector for our purposes.
- Thus $\mathbf{D}$ is exponentially distributed with the support $(0{:}\infty)$.
- The prior distribution for θ is $p(\theta) = 1$, where $\theta \in (0{:}\infty)$.
- The resulting posterior distribution is:

$$\pi(\theta|\mathbf{D}) \propto p(\theta)p(\mathbf{D}|\theta) = (1)\theta e^{-\theta \mathbf{D}} = \theta e^{-\theta \mathbf{D}}.$$

- This posterior distribution has mean:

$$E[\theta|\mathbf{D}] = \int_0^\infty (\theta) \left(\theta e^{-\theta \mathbf{D}}\right) d\theta = \frac{2}{\mathbf{D}^3},$$

which is found easily with two iterations of integration-by-parts.

Summarizing a Posterior

- The expectation of $\theta^2|\mathbf{D}$ is:

$$E[\theta^2|\mathbf{D}] = \int_0^\infty (\theta^2)\left(\theta e^{-\theta\mathbf{D}}\right) d\theta = \frac{6}{\mathbf{D}^4},$$

  which is found with three iterations of integration-by-parts now.

- So the posterior variance is:

$$\text{Var}[\theta|\mathbf{D}] = E[\theta^2|\mathbf{D}] - E[\theta|\mathbf{D}]^2 = 6\mathbf{D}^{-4} - 4\mathbf{D}^{-6}.$$

Credible Intervals and Sets

- The Bayesian analogue to the confidence interval is the credible interval and more generally the credible set, which does not have to be contiguous.
- Most of the time in practice, it is calculated in *exactly the same way* as the confidence interval.
- For instance calculating a 95% credible interval under the Gaussian normal assumption means marching out 1.96 standard errors from the mean in either direction, just like the analagous confidence interval is created. (The difference lies in the interpretation.)
- A $100(1-\alpha)\%$ credible interval gives the region of the parameter space where the probability of covering $\boldsymbol{\theta}$ is at least $1-\alpha$.
- In contrast, applying this new definition to the confidence interval means that the probability of coverage is either zero or one, since it either covers the true $\boldsymbol{\theta}$ or it doesn't.

Credible Intervals and Sets

- Define C as a *contiguous* subset of the parameter space, $\boldsymbol{\Theta}$, such that a $100(1-\alpha)$ credible interval meets the condition:

$$1-\alpha = \int_C \pi(\boldsymbol{\theta}|\mathbf{X})d\theta$$

  for some chosen α level.

- Conventions: centered at mean or mode, equal tails.

- So credible intervals are *not* unique!

Credible Intervals and Sets, Example

- Suppose we have duration data, $\mathbf{X}$, exponentially distributed $p(X|\theta) = \theta e^{-\theta X}$ defined over $(0, \infty)$, where interest is in the posterior distribution of the unknown parameter θ.
- Specify the prior distribution of $p(\theta) = 1/\theta$, for $\theta \in (0{:}\infty)$.
- The posterior is:

$$\pi(\theta|\mathbf{X}) \propto p(\theta)L(\theta|\mathbf{X}) = \frac{1}{\theta}\theta^n \exp\left[-\theta \sum_{i=1}^{n} x_i\right] = \theta^{n-1} \exp\left[-\theta \sum_{i=1}^{n} x_i\right].$$

- This means that $\theta|\mathbf{X} \sim \mathcal{G}(\theta|n, \sum x_i)$, where putting the constants back in front to recover the full form of this gamma posterior distribution produces:

$$\pi(\theta|\mathbf{X}) = \frac{(\sum x_i)^n}{\Gamma(n)}\theta^{n-1} \exp\left[-\theta \sum x_i\right].$$

- Since we know everything about this posterior distribution, we are free to choose any desired credible interval.

Credible Intervals and Sets, Example

- Browne, Freidreis, and Gleiber (1986) tabulate complete cabinet duration for eleven Western European countries from 1945 to 1980:

Table 1: European Cabinet Duration Annualized, 1945-1980

Country	N	Average Duration
Austria	15	2.114
Belgium	27	1.234
Denmark	20	1.671
Finland	28	1.070
Iceland	15	2.080
Ireleand	14	2.629
Italy	38	0.833
Netherland	12	2.637
Norway	17	2.065
Sweden	15	2.274

Credible Intervals and Sets, Example

- Country averages from the third column of the table are weighted by N in the second column to reflect the number of such events: $\mathbf{X}_i N_i$.
- For a chosen α the end-points of an equal-tail credible interval can be calculated with:

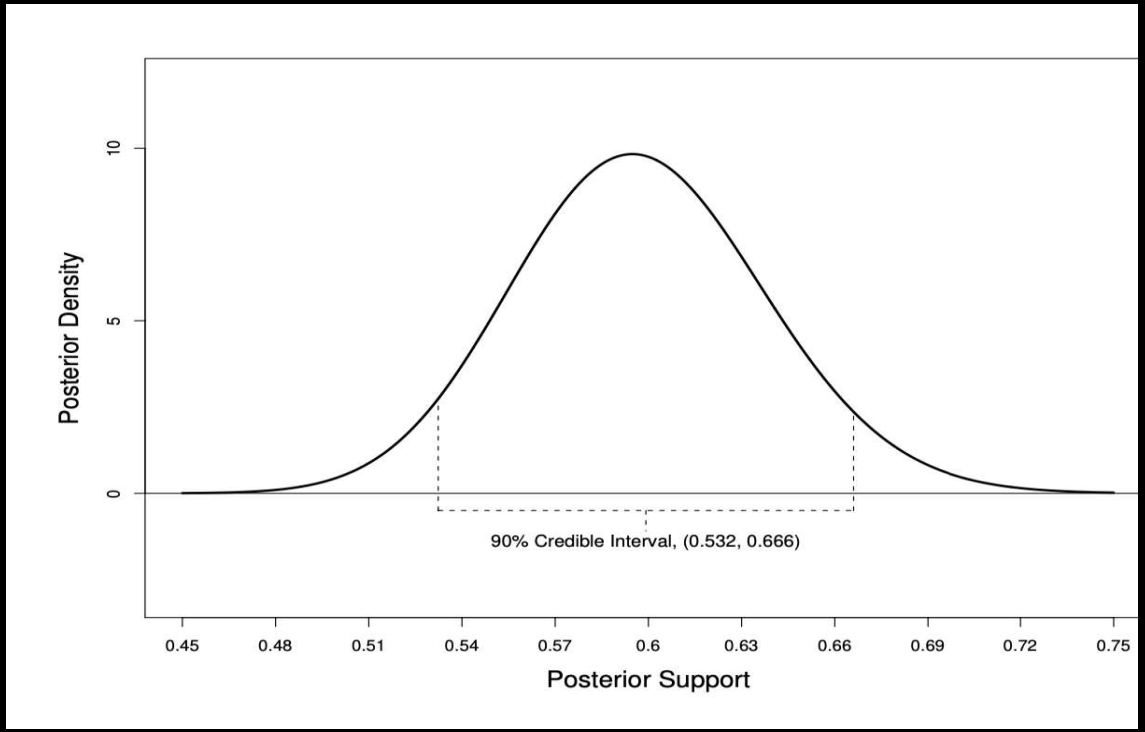
$$\frac{\alpha}{2} = \int_0^L \pi(\boldsymbol{\theta}|\mathbf{X})d\theta \qquad \frac{\alpha}{2} = \int_H^\infty \pi(\boldsymbol{\theta}|\mathbf{X})d\theta$$

or we could simply use the following R commands for a 95% credible interval:

```
dur <- c(2.114,1.234,1.671,1.070,2.168,2.080, 2.629,0.833,2.637,2.065,2.274)
N <- c(15,27,20,28,15,15,14,38,12,17,15)
qgamma(0.025,shape=sum(N),rate=sum(N*dur))
[1] 0.52056
qgamma(0.975,shape=sum(N),rate=sum(N*dur))
[1] 0.67988
```

Credible Intervals and Sets, Example

Equal Tail Credible Interval for Cabinet Duration



Credible Interval, Fifty U.S. States Time to Adoption for Health Bills

- Boehmke (2009) counts bills passed in the fifty states between 1998 and 2005 that contain policy implications for the increasing obesity rates in the U.S.
- These include limits on sugary drinks at schools, requiring insurers to cover particular medical procedures, as well as limitations on lawsuits from consumer groups on the fast food industry.
- Define duration data, $\mathbf{X}$, to be the time in years through this period for a bill to be passed.
- Assume that $\mathbf{X}$ is exponentially distributed $p(X|\theta) = \theta e^{-\theta X}$ over $[0, \infty)$.
- Specify the prior distribution as $p(\theta) = 1/\theta$, for $\theta \in [0{:}\infty)$.
- The resulting posterior is given by:

$$\pi(\theta|\mathbf{X}) \propto p(\theta)L(\theta|\mathbf{X}) = \left(\frac{1}{\theta}\right)\theta^n \exp\left[-\theta\sum_{i=1}^{n} x_i\right] = \theta^{n-1}\exp\left[-\theta\sum_{i=1}^{n} x_i\right]$$

Credible Interval, Fifty U.S. States Time to Adoption for Health Bills

State	N	Mean Duration	State	N	Mean Duration	State	N	Mean Duration
AL	2	7.500	LA	14	5.571	OK	12	6.583
AK	12	6.667	ME	2	5.500	OH	0	NaN
AZ	12	6.250	MD	11	6.455	OR	1	8.000
AR	6	6.167	MA	7	7.143	PA	12	7.083
CA	46	6.000	MI	4	7.000	RI	7	7.000
CO	11	6.636	MN	2	7.000	SC	6	6.333
CT	2	7.000	MS	7	7.143	SD	1	7.000
DE	4	7.000	MO	18	5.556	TN	17	7.235
FL	11	6.364	MT	2	7.000	TX	16	6.250
GA	7	5.857	NE	5	7.400	UT	3	7.667
HI	8	6.375	NV	4	8.000	VT	8	6.625
ID	6	6.000	NH	1	5.000	VA	15	6.533
IL	4	6.750	NJ	6	7.333	WA	12	6.083
IN	31	7.065	NM	6	6.500	WV	2	7.500
IA	3	5.000	NY	9	6.556	WI	4	7.750
KS	4	8.000	NC	8	7.250	WY	1	8.000
KY	4	7.500	ND	9	6.111			

Calculating the Credible Interval

- Note that $\theta|\mathbf{X} \sim \mathcal{G}(\theta|n, \sum x_i)$ has the "rate" specification for the second parameter.
- Putting the constants back in front to recover the full form of this gamma posterior distribution produces:

$$\pi(\theta|\mathbf{X}) = \frac{(\sum x_i)^n}{\Gamma(n)}\theta^{n-1}\exp\left[-\theta\sum x_i\right]$$

- The complete data are given above for annualized periods, as well as in the `R` package `BaM`.
- Note the "`NaN`" value for the Ohio mean duration given by `R` since there is nothing to average (censored from us)
- The state averages from the third column of the table are weighted by N in the second column to reflect the number of such events: $\mathbf{X}_i N_i$.
- Since the sufficient statistic in the posterior distribution is a sum, there is no loss of information from not having the full original data from the authors (sums of means times n equal the total sum).

Calculating the 90% Credible Interval

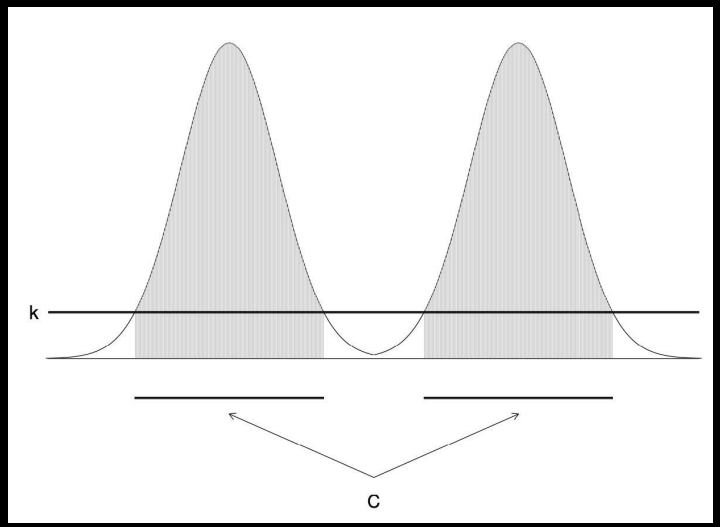
- The end-points of the equal tail credible interval are created by solving for the limits (L and H) in the two integrals:

$$\frac{\alpha}{2} = \int_0^L \pi(\boldsymbol{\theta}|\mathbf{X})d\theta \qquad \frac{\alpha}{2} = \int_H^\infty \pi(\boldsymbol{\theta}|\mathbf{X})d\theta$$

```
state.df <- state.df[-35,]          # REMOVES OHIO
qgamma(0.05,shape=sum(state.df$N),rate=sum(state.df$N*state.df$dur))
[1] 0.14034
qgamma(0.95,shape=sum(state.df$N),rate=sum(state.df$N*state.df$dur))
[1] 0.16528
```

Highest Posterior Density Intervals and Sets

- When looking at posterior distributions, we really care where the highest density exists on the support of the posterior density, regardless of whether it is contiguous or not.
- HPD created such that that no region outside of the interval will have higher posterior density than any region inside the the HPD.
- Therefore HPDs are not necessarily contiguous.



Highest Posterior Density Intervals and Sets

- A $100(1-\alpha)\%$ highest posterior density (HPD) is the subset of the support of the posterior distribution for some parameter, θ, that meets the criteria:

$$C = \{\theta : \pi(\theta|\mathbf{x}) \geqslant k\},$$

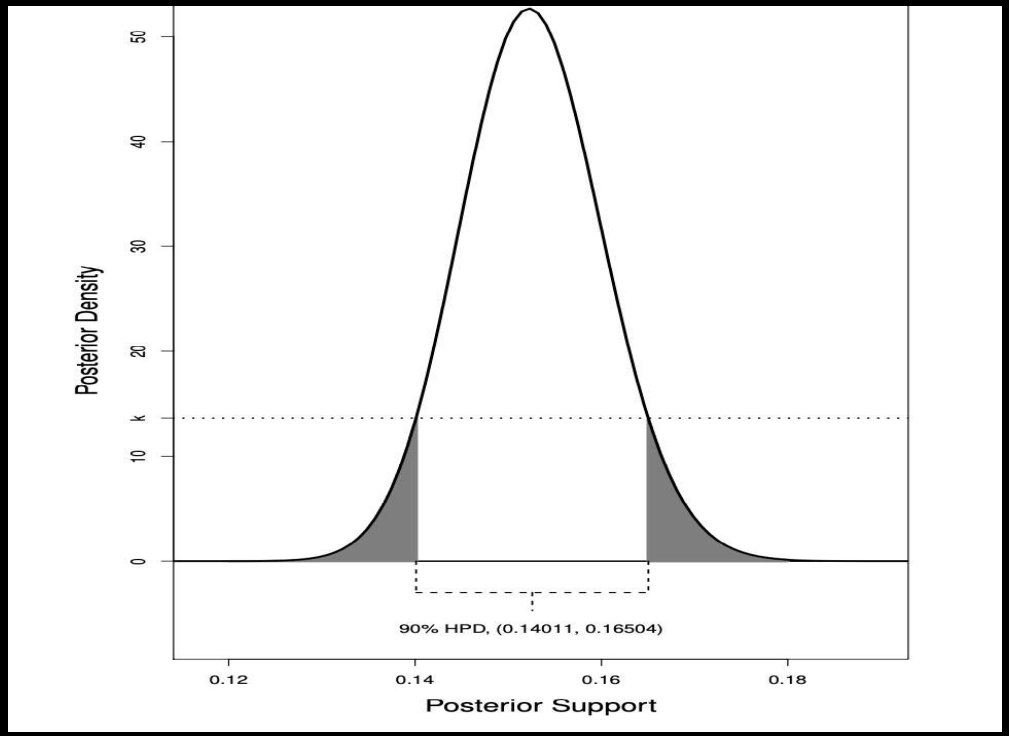
  where k is the largest number such that:

$$1 - \alpha = \int_{\theta:\pi(\theta|\mathbf{x})>k} \pi(\theta|\mathbf{x}) d\theta$$

- The important difference is $\theta : \pi(\theta|\mathbf{x}) > k$ instead of a single contiguous interval as with the credible interval.
- Sometimes this can be done analytically.
- The `R` code for this example is in Chapter 2.

Highest Posterior Density Intervals and Sets, Example

HPD Interval for a Different Dataset



Bayesian Updating: Overview

- Start with the prior distribution: $p(\theta)$, on an unknown variable θ.
- Observe a first set of iid data, $\mathbf{x}_1$, and calculate the posterior: $\boxed{\pi_1(\theta|\mathbf{x}_1) \propto p(\theta)L(\mathbf{x}_1|\theta).}$
- Now observe a second set of iid data, $\mathbf{x}_2$ from the same data-generating process and *update* the posterior and therefore improve our state of knowledge by treating the previous posterior as a prior:

$$\pi_2(\theta|\mathbf{x}_1, \mathbf{x}_2) \propto \pi_1(\theta|\mathbf{x}_1)L(\mathbf{x}_2|\theta) = p(\theta)L(\mathbf{x}_1|\theta)L(\mathbf{x}_2|\theta) = p(\theta)L(\mathbf{x}_1, \mathbf{x}_2|\theta).$$

- This is exactly the same result we would have obtained if all the data had arrived at once: $\pi(\theta|\mathbf{x})$.
- This process can continue *ad infinitum* and the model will continue to update the posterior conclusions as new information continues to roll in.

Application of Bayesian Updating: A Meta-Estimate of Deaths in Stalin's Gulags

- It is difficult to estimate of the number of people that perished in the Soviet Gulags during Stalin's era as dictator (1924-1953).
- Apparently there will never be a *definitive* answer (Solzhenitsyn 1997).
- But we can look at the sequential estimates, $\mathbf{X}$, of historical scholars over time as an updating.
- Uses the qualitative statements made by these authors as probabilistic assessments.

Application of Bayesian Updating: A Meta-Estimate of Deaths in Stalin's Gulags

- Assertions:

$$\begin{aligned}
\text{Wiles, 1965:} \quad & p(\theta_1|\mathbf{X}) \sim \mathcal{U}(0{:}4.6) \\
\text{Kurganov, 1973:} \quad & p(\theta_2|\theta_1, \mathbf{X}) \sim \mathcal{U}(20{:}30) \\
\text{Conquest, 1978:} \quad & p(\theta_3|\theta_2, \theta_1, \mathbf{X}) \sim \mathcal{N}(18.2, 8.5) \\
\text{Medvedev, 1989:} \quad & p(\theta_4|\theta_3, \theta_2, \theta_1, \mathbf{X}) \sim \mathcal{N}(12, 9).
\end{aligned}$$

- Therefore the likelihood from these "data" is:

$$L(\theta|\theta_4, \theta_3, \theta_2, \theta_1, \mathbf{X}) = p(\theta_4|\theta_3, \theta_2, \theta_1, \mathbf{X}) p(\theta_3|\theta_2, \theta_1, \mathbf{X}) p(\theta_2|\theta_1, \mathbf{X}) p(\theta_1|\mathbf{X}).$$

Application of Bayesian Updating: A Meta-Estimate of Deaths in Stalin's Gulags

- The posterior is modeled as a normal weighted by the precisions with the assumption that the intermediate conditionals are normal, $\mathcal{N}(\mu_i, \sigma_i^2)$, and this posterior form is therefore given by $\mathcal{N}(\mu_\pi, \sigma_\pi^2)$ where $\sigma_\pi^2 = \left(\sum_{i=1}^4 \sigma_i^{-2})\right)^{-1}$ and $\mu_\pi = \sigma_\pi^2 \sum_{i=1}^4 (\mu_i/\sigma_i^2)$.
- Start with the diffuse normal prior $\mathcal{N}(8, 12)$, so:

$$\pi(\theta|\theta_4, \theta_3, \theta_2, \theta_1, \mathbf{X}) \propto p(\theta)L(\theta|\theta_4, \theta_3, \theta_2, \theta_1, \mathbf{X}) \sim N(13.2, 3.2).$$

- This translates to a 95% credible interval of $[9.7 : 16.7]$ million deaths, which is a compromise between the four experts and the author of the final meta-analysis (Blyth).

Bayes Factor

- 2 competing models, M_1: $f_1(\mathbf{x}|\theta_\mathbf{1})$ M_2: $f_2(\mathbf{x}|\theta_\mathbf{2})$
- $\theta_\mathbf{1}$ and $\theta_\mathbf{2} \in \mathbf{\Theta}$ or $\mathbf{\Theta_1}$ and $\mathbf{\Theta_2}$
- specify parameter priors: $\pi_1(\theta_\mathbf{1})$ and $\pi_2(\theta_\mathbf{2})$ and model priors: $p(M_1)$ and $p(M_2)$.
- Note that $p(\mathbf{x}|M_i) = \int_{\theta_i} f_i(\mathbf{x}|\theta_\mathbf{i})\pi_i(\theta_\mathbf{i})d\theta_i$
- Thus:

$$\underbrace{\frac{p(M_1|\mathbf{x})}{p(M_2|\mathbf{x})}}_{\text{posterior odds}} = \underbrace{\frac{p(M_1)/p(\mathbf{x})}{p(M_2)/p(\mathbf{x})}}_{\text{prior odds/data}} \times \underbrace{\frac{\int_{\theta_1} f_1(\mathbf{x}|\theta_\mathbf{1})\pi_1(\theta_\mathbf{1})d\theta_1}{\int_{\theta_2} f_2(\mathbf{x}|\theta_\mathbf{2})\pi_2(\theta_\mathbf{2})d\theta_2}}_{\text{Bayes factor}}.$$

$$\text{posterior odds ratio} = \text{prior odds ratio} \times \text{integrated likelihood ratio}$$

- Rearranging this and canceling out $p(\mathbf{x})$ gives:

$$B(\mathbf{x}) = \frac{p(M_1|\mathbf{x})/p(M_1)}{p(M_2|\mathbf{x})/p(M_2)}$$

$$= \text{“posterior to prior odds ratio”}$$

Jeffreys' Typology

$(\mathbf{x}) \geqslant 1$	model 1 supported
$1 > B(\mathbf{x}) \geqslant 10^{-\frac{1}{2}}$	minimal evidence against model 1
$10^{-\frac{1}{2}} > B(\mathbf{x}) \geqslant 10^{-1}$	substantial evidence against model 1
$10^{-1} > B(\mathbf{x}) \geqslant 10^{-2}$	strong evidence against model 1
$10^{-2} > B(\mathbf{x})$	decisive evidence against model 1

Kass and Raftery, 1995, JASA

$\log_{10}(B_{10})$	B_{10}	Evidence against H_0
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

Bayes Factor Example

- $H_0 : \theta = 0.5$, vs. $H_1 : \theta = 0.7$
- priors: $p(H_0) = p(H_1) = \frac{1}{2}$ (so prior odds ratio just one)
- data: $n = 10, x = 7$
- Bayes Factor between "model 0 supported" and "minimal evidence against model 0":

$$B(\mathbf{x}) = \frac{p(M_1|\mathbf{x})/p(M_1)}{p(M_2|\mathbf{x})/p(M_2)} = \frac{p(x = 7|\theta = 0.5)}{p(x = 7|\theta = 0.7)} = \frac{\binom{10}{7}(0.5)^7(0.5)^3}{\binom{10}{7}(0.7)^7(0.3)^3} = 0.44$$

- Now calculate the corresponding p-value:

$$\begin{aligned} p &= p(\text{observed data or more extreme}|H_0) = p(x = \{7, 8, 9, 10\}|\theta = 0.5) \\ &= \sum_{x=7}^{10} \binom{10}{x} (0.5)^x (0.5)^{10-x} = 0.172, \text{fail to reject } H_0 \end{aligned}$$

Problems (challenges) with Bayes Factors

- Sensitivity to priors
- No directional hypotheses
- Improper priors...
  $\pi(\theta) \propto h(\theta)$, set $\pi(\theta) = ch(\theta) \ldots$ Okay for posteriors:

$$\begin{aligned} p(\theta|\mathbf{x}) &= \frac{\pi(\theta)p(\mathbf{x}|\theta)}{\int_\theta \pi(\theta)p(\mathbf{x}|\theta)d\theta} \\ &= \frac{cg(\theta)p(\mathbf{x}|\theta)}{c\int_\theta g(\theta)p(\mathbf{x}|\theta)d\theta} \\ &= \frac{g(\theta)p(\mathbf{x}|\theta)}{\int_\theta g(\theta)p(\mathbf{x}|\theta)d\theta} \end{aligned}$$

  Bad for Bayes factors:

$$B(\mathbf{x}) = \frac{c_1}{c_2}\frac{\int_{\theta_1} g_1(\theta_1)p(\mathbf{x}|\theta_1)d\theta_1}{\int_{\theta_2} g_2(\theta_2)p(\mathbf{x}|\theta_2)d\theta_2}$$

Bayes Factors for the Linear Model

- We want to compare two, not necessarily nested, different right-hand-side specifications in $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is an $n \times k$, rank k matrix of explanatory variables with a leading vector of ones, $\boldsymbol{\beta}$ is a $k \times 1$ unknown vector of coefficients, $\mathbf{y}$ is an $n \times 1$ vector of outcomes, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of residuals with $\mathcal{N}(0, \sigma^2 I)$ for a constant σ^2 (homoscedasticity).
- The likelihood function for model j is:

$$L_j(\boldsymbol{\beta}_j, \sigma_j^2 | \mathbf{X}_j, \mathbf{y}) = (2\pi\sigma_j^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma_j^2}(\mathbf{y} - \mathbf{X}_j\boldsymbol{\beta}_j)'(\mathbf{y} - \mathbf{X}_j\boldsymbol{\beta}_j)\right]$$

  where $j = 0, 1$ providing models M_0 and M_1.
- $\mathbf{y}$ is not indexed here since both models intend to explain the structure of the same outcome.
- Make the definitions $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})/(n - k)$.

Bayes Factors for the Linear Model

- Now specify conjugate priors for each of these models with k_j columns of $\mathbf{X}$ according to:

$$p(\boldsymbol{\beta}_j|\sigma^2) = (2\pi)^{-\frac{k_j}{2}}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{\beta}_j - \mathbb{B}_j)'\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\beta}_j - \mathbb{B}_j)\right],$$

and:

$$p(\sigma_j^2) \propto {\sigma_j}^{-(a_j-k_j)}\exp\left[-\frac{b_j}{\sigma_j^2}\right]$$

- We add a multiplier h_j on the variance term in the normal prior for $\boldsymbol{\beta}_j$: $\boldsymbol{\Sigma}_j = h_j\sigma_j^2\mathbf{I}$.
- Make the common choice of prior mean for $\boldsymbol{\beta}$ to be $\mathbb{B} = 0$ in both models.
- The marginal likelihood for model j from this setup is:

$$p_j(\mathbf{y}|\mathbf{X}_j, M_j) = \frac{|\mathbf{X}_j'\mathbf{X}_j + h|^{-\frac{1}{2}}|h_j|^{\frac{1}{2}}b_j^{a_j}\Gamma(a_j + \frac{a_j}{2})}{\pi^{\frac{n}{2}}\Gamma(a_j)}\left(2b_j + (n-k_j){\hat{\sigma}_j}^2\right).$$

Bayes Factors for the Linear Model

- This means that the Bayes Factor for Model 1 over Model 0 is given by:

$$BF_{(1,0)} = \frac{p_1(\mathbf{y}|\mathbf{X}_1, M_1)}{p_0(\mathbf{y}|\mathbf{X}_0, M_0)} = \frac{\frac{|\mathbf{X}_1'\mathbf{X}_1+h|^{-\frac{1}{2}}|h_1|^{\frac{1}{2}}b_1^{a_1}\Gamma(a_1+\frac{a_1}{2})}{\pi^{\frac{n}{2}}\Gamma(a_1)}\left(2b_1+(n-k_1)\hat{\sigma_1}^2\right)}{\frac{|\mathbf{X}_0'\mathbf{X}_0+h|^{-\frac{1}{2}}|h_0|^{\frac{1}{2}}b_0^{a_0}\Gamma(a_0+\frac{a_0}{2})}{\pi^{\frac{n}{2}}\Gamma(a_0)}\left(2b_0+(n-k_0)\hat{\sigma_0}^2\right)}.$$

- This is a long expression but a relatively simple form due to the elegance of the linear model.

New and Old Ways to Look at Model Fit

- Akaike Information Criterion.
  minimizes the negative likelihood penalized by the number of parameters:

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + 2p$$

  where $\ell(\hat{\boldsymbol{\beta}}|\mathbf{y})$ is the maximized model log likelihood value and p is the number of explanatory variables in the model (including the constant). (AIC has a bias towards models that overfit with extra parameters since the penalty component is obviously linear with increases in the number of explanatory variables, and the log likelihood often increases more rapidly.)

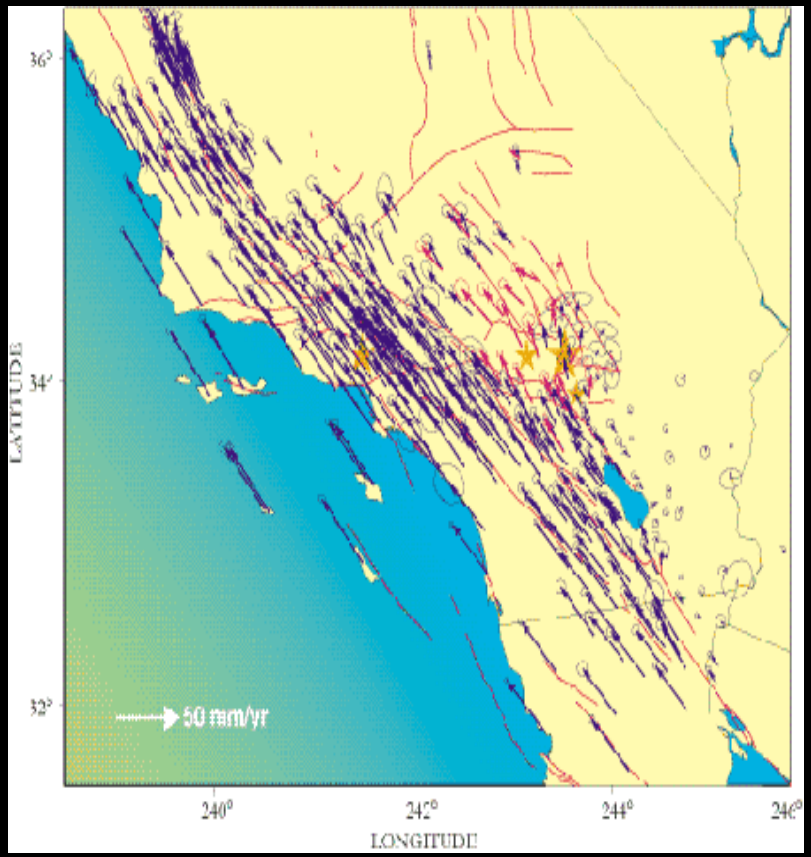
- Schwartz Criterion/Bayesian Information Criterion (BIC).

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\beta}}|\mathbf{y}) + p\log(n)$$

  where n is the sample size.

- There is also a Deviance Information Criterion (DIC) used in Bayesian MCMC estimation.

Multivariate Application: the Predicting Earthquake Aftershocks



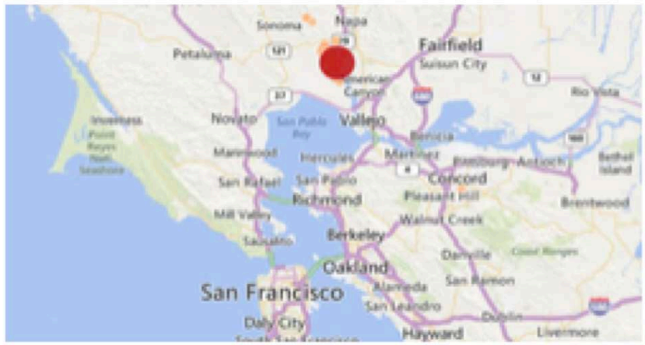
- Topical.
- Immediately after a powerful earthquake in a high population density area decisions must be made about operating powerplants, schools, and transportation facilities.
- A series of aftershocks can be equally deadly and destructive as a mainshock.
- Predicting aftershocks based on empirical evidence is far reliable than predicting mainshocks.

Multivariate Application: the Predicting Earthquake Aftershocks (cont.)

- Why is this relevant?
- Some geopolitical events are very hard to predict, but their after-effects may be much more reliably anticipated.
- Examples: terrorist attacks, unannounced nuclear tests, civil wars, coups.
- Bayesian learning may (over time) increase our knowledge.
- The need for real-time analysis parallels necesary government reactions after such events.

How Aftershocks Are Described

Infographic



Bay Area's 6.0 quake and aftershocks

READ THE STORY >

A little more than two hours after the quake, a shallow magnitude 3.6 tremor was reported by the USGS. The aftershock occurred at 5:47 a.m. at a depth of five miles. The National California Seismic System put the chance of a strong aftershock in the next week at 54%. Scientists at UC Berkely released a video showing an early-warning system that sent an alert 10 seconds before the earthquake.

Multivariate Application: the Predicting Earthquake Aftershocks (cont.)

- Model aftershocks as a *non-homogeneous* Poisson process with the intensity parameter:

$$N(t) \propto \frac{1}{(t+c)^p}.$$

  This is actually called "Omori's Law" where t is time, and the rest are constants: c is a time offset, p is a rate of decay.

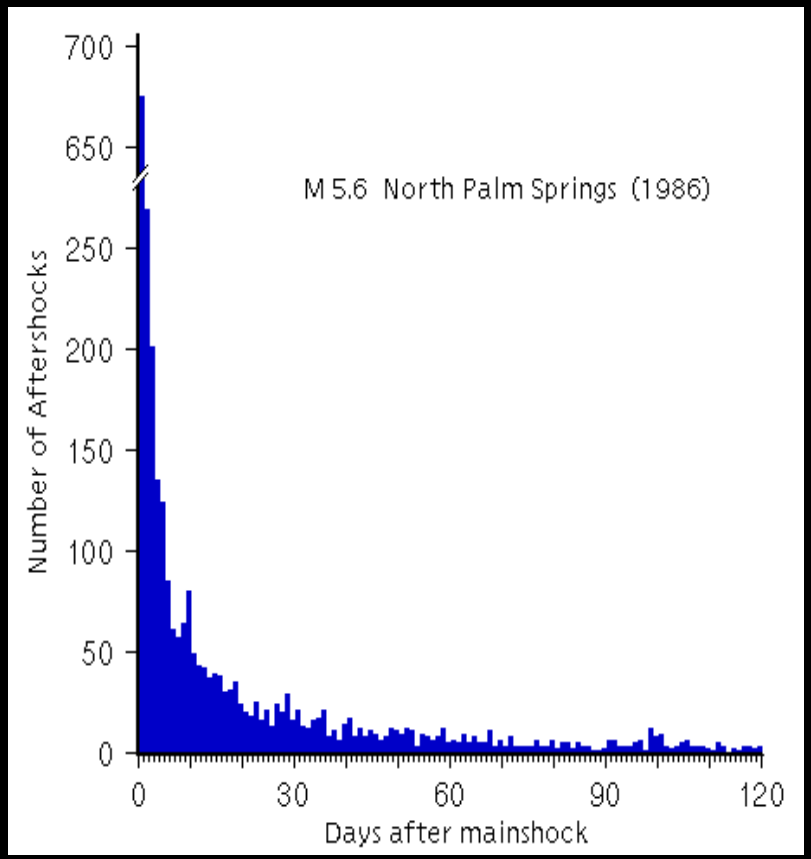
- So the probability of n aftershocks at time period t is: $\boxed{P(n|t) = \frac{N(t)^n e^{-N(t)}}{n!}.}$

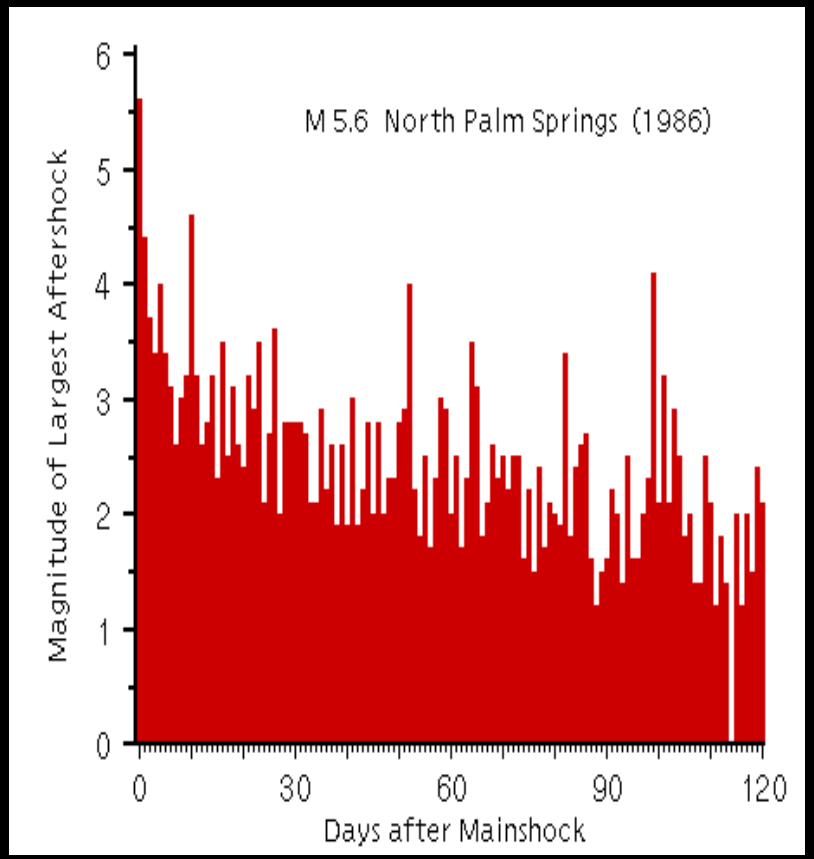
- Use the Gutenberg-Richter relation (an empirical law), aftershock version:

$$\log_{10} N(M) = a + b(M_{\text{mainshock}} - M_{\text{aftershock}})$$

  where $N(M)$ is the number per year of aftershocks of magnitude greater than $M_{\text{aftershock}}$ following a mainshock of magnitude $M_{\text{mainshock}}$, a and b are constants.

Multivariate Application: the Predicting Earthquake Aftershocks (cont.)





Multivariate Application: the Predicting Earthquake Aftershocks (cont.)

- Putting these two principles together gives the rate of aftershocks of magnitude $M_{\text{aftershock}}$ or larger at time t following a mainshock:

$$\lambda(t, M) = 10^{a+b(M_{\text{mainshock}}-M_{\text{aftershock}})}(t+c)^{-p}$$

- More usefully, the probability of an aftershock between M_1 and M_2, both less than $M_{\text{mainshock}}$, and between time t_1 and t_2 after the mainshock:

$$p(t, M) = 1 - \exp\left[-\int_{M_1}^{M_2}\int_{t_1}^{t_2}\lambda(t, M)dtdM\right]$$

  under the assumption that the joint instantaneous rate is distributed exponential (see the figure!)

- What we need now is a posterior distribution for $\boldsymbol{\mu} = (a, b, p, c)$ conditional on the mainshock.

Multivariate Application: the Predicting Earthquake Aftershocks (cont.)

- Start with some (regionalized) data, calculate posteriors with a Bayesian gaussian model and update as new data (earthquakes) occur.
- Multivariate priors:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}_k\left(\mathbf{m}, \frac{\boldsymbol{\Sigma}}{n_0}\right), \qquad \boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\alpha, \boldsymbol{\beta}),$$

  where n_0/n measures our belief in the representativeness prior data.
- The Weibull distribution is described by: $w(x|\alpha,\beta) = \frac{\alpha}{\beta}x^{\alpha-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$ if $x \geqslant 0$ and 0 otherwise, where: $\alpha, \beta > 0$.
- This produces posteriors:

$$\hat{\boldsymbol{\mu}}|\boldsymbol{\Sigma} \sim \mathcal{N}_k\left(\frac{n_0\mathbf{m} + n\bar{\mathbf{x}}}{n_0 + n}, \frac{\boldsymbol{\Sigma}}{n_0 + n}\right)$$

$$\widehat{\boldsymbol{\Sigma}^{-1}} \sim \mathcal{W}_k\left(\alpha + n, \boldsymbol{\beta}^{-1} + S^2 + \frac{n_0 n}{n_0 + n}(\bar{x} - \mathbf{m})(\bar{x} - \mathbf{m})'\right).$$

Multivariate Application: the Predicting Earthquake Aftershocks (cont.)

- Some information to build "Generic California" priors:
  - 62 aftershock sequences with $M_{\text{mainshock}} \geqslant 5$, occurring from 1933 to 1987 in California (exclusive of two unusual events),
  - Omori's Law parameters (a, p) from $M_{\text{mainshock}} - M_{\text{aftershock}} \geqslant 3$,
  - b from $M_{\text{mainshock}} - M_{\text{aftershock}} \geqslant 2$,
  - c picked to get maximum distinction between mainshock "coda" and aftershocks using post-1970 data.
- Reasenberg and Jones (1989) assume $\boldsymbol{\Sigma}^{-1}$ is diagonal and produce normal priors with means:

$$\bar{a} = -1.67, \quad \bar{b} = 0.91, \quad \bar{p} = 1.08, \quad c = 0.05$$

($\sigma_a = 0.0.7$, $\sigma_b = 0.02$, $\sigma_p = 0.03$, c deterministic).

Multivariate Application: the Predicting Earthquake Aftershocks (cont.)

- Data taken from real-time sequence of aftershocks for two excluded events:
  - ▻ Coalinga (1983), $M_{\text{mainshock}} = 6.5$
  - ▻ Whittier-Narrows (1987), $M_{\text{mainshock}} = 5.9$

  and updated *during* aftershock times.
- Thus probabilities are Bayesianly improved during risk period for an event greater than the mainshock.
- Updating the "Generic California" priors with the conjugate-normal Bayesian model gives any desired set of probabilities over a period of time after the mainshock by integrating some region of the posterior.

Probability of $M_{\text{aftershock}} > M_{\text{mainshock}} - 1$

Within	**Time After Mainshock, Coalinga**								
$t_2 - t_1$	15 min.	6 hrs.	12 hrs.	1 day	3 days	7 days	15 days	30 days	60 days
1 Day	0.330	0.176	0.125	0.081	0.033	0.015	0.007	0.003	0.002
3 Days	0.413	0.265	0.209	0.153	0.077	0.039	0.020	0.010	0.005
7 Days	0.467	0.330	0.276	0.218	0.129	0.074	0.040	0.022	0.011
30 Days	0.545	0.427	0.378	0.324	0.234	0.165	0.109	0.069	0.039
60 Days	0.577	0.466	0.420	0.370	0.283	0.214	0.154	0.105	0.066

Within	**Time After Mainshock, Whittier-Narrows**								
$t_2 - t_1$	15 min.	6 hrs.	12 hrs.	1 day	3 days	7 days	15 days	30 days	60 days
1 Day	0.393	0.141	0.084	0.044	0.012	0.004	0.001	0.000	0.000
3 Days	0.431	0.185	0.123	0.074	0.026	0.010	0.004	0.001	0.000
7 Days	0.488	0.208	0.146	0.095	0.040	0.017	0.007	0.003	0.001
30 Days	0.465	0.232	0.171	0.120	0.062	0.034	0.017	0.009	0.004
60 Days	0.470	0.238	0.178	0.127	0.069	0.040	0.023	0.012	0.006

Example for Whitter-Narrows, if the main shock happened within the last 15 minutes then the probablity of a serious aftershoock in the next 24 hours is 0.393.