Statistical Methods - Bayesian Inference

Lecture 2

Mattias Villani

Sveriges Riksbank and Stockholm University

March 18, 2010

Overview Lecture 2

- Conjugate priors
- Poisson model
- 'Non-Informative' priors
- Bayesian asymptotics
- Numerical optimization

Conjugate priors

- Normal likelihood: Normal prior$\rightarrow$Normal posterior. (posterior belongs to the same distribution family as prior)
- Binomial likelihood: Beta prior$\rightarrow$Beta posterior.
- *Conjugate priors*: Let $\mathcal{F} = \{p(y|\theta), \theta \in \Theta\}$ be a class of sampling distributions. A family of distributions $\mathcal{P}$ is conjugate for $\mathcal{F}$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|x) \in \mathcal{P}$$

  holds for all $p(x|\theta) \in \mathcal{F}$.
- *Natural conjugate prior*: $p(\theta) = c \cdot p(y_1, ..., y_n|\theta)$ for some constant c, i.e. the prior is of the same functional form as the likelihood.

Poisson model

- Likelihood from iid Poisson sample $y = (y_1, ..., y_n)$

$$p(y|\theta) = \left[\prod_{i=1}^{n} p(y_i|\theta)\right] \propto \theta^{(\sum_{i=1}^{n} y_i)} \exp(-\theta n),$$

  so that the sum of counts $\sum_{i=1}^{n} y_i$ is a sufficient statistic for θ.
- *Natural conjugate prior for Poisson parameter* θ

$$p(\theta) \propto \theta^{\alpha-1} \exp(-\theta\beta) \propto Gamma(\alpha, \beta)$$

  which contains the info: $\alpha - 1$ counts in β observations.

Poisson model, cont.

- *Posterior for Poisson parameter* θ. Multiplying the poisson likelihood and the Gamma prior gives the posterior

$$\begin{aligned} p(\theta|y_1,...,y_n) &\propto \left[\prod_{i=1}^n p(y_i|\theta)\right] p(\theta) \\ &\propto \theta^{\sum_{i=1}^n y_i} \exp(-\theta n)\theta^{\alpha-1}\exp(-\theta\beta) \\ &= \theta^{\alpha+\sum_{i=1}^n y_i-1}\exp[-\theta(\beta+n)], \end{aligned}$$

  which is proportional to the $Gamma(\alpha+\sum_{i=1}^n y_i, \beta+n)$ distribution.

- In summary

$$\begin{aligned} \text{Model:} \quad & y_1,...,y_n|\theta \overset{iid}{\sim} Po(\theta) \\ \text{Prior:} \quad & \theta \sim Gamma(\alpha,\beta) \\ \text{Posterior:} \quad & \theta|y_1,...,y_n \sim Gamma(\alpha+\sum_{i=1}^n y_i, \beta+n). \end{aligned}$$

Poisson example - Number of bomb hits in London

$n = 576$, $\sum_{i=1}^{n} y_i = 229 \cdot 0 + 211 \cdot 1 + 93 * 2 + 35 * 3 + 7 * 4 + 1 \cdot 5 = 537$.

Average number of hits per region=$\bar{y} = 537/576 \approx 0.9323$.

$$p(\theta|y) \propto \theta^{\alpha+537-1} \exp[-\theta(\beta + 576)]$$

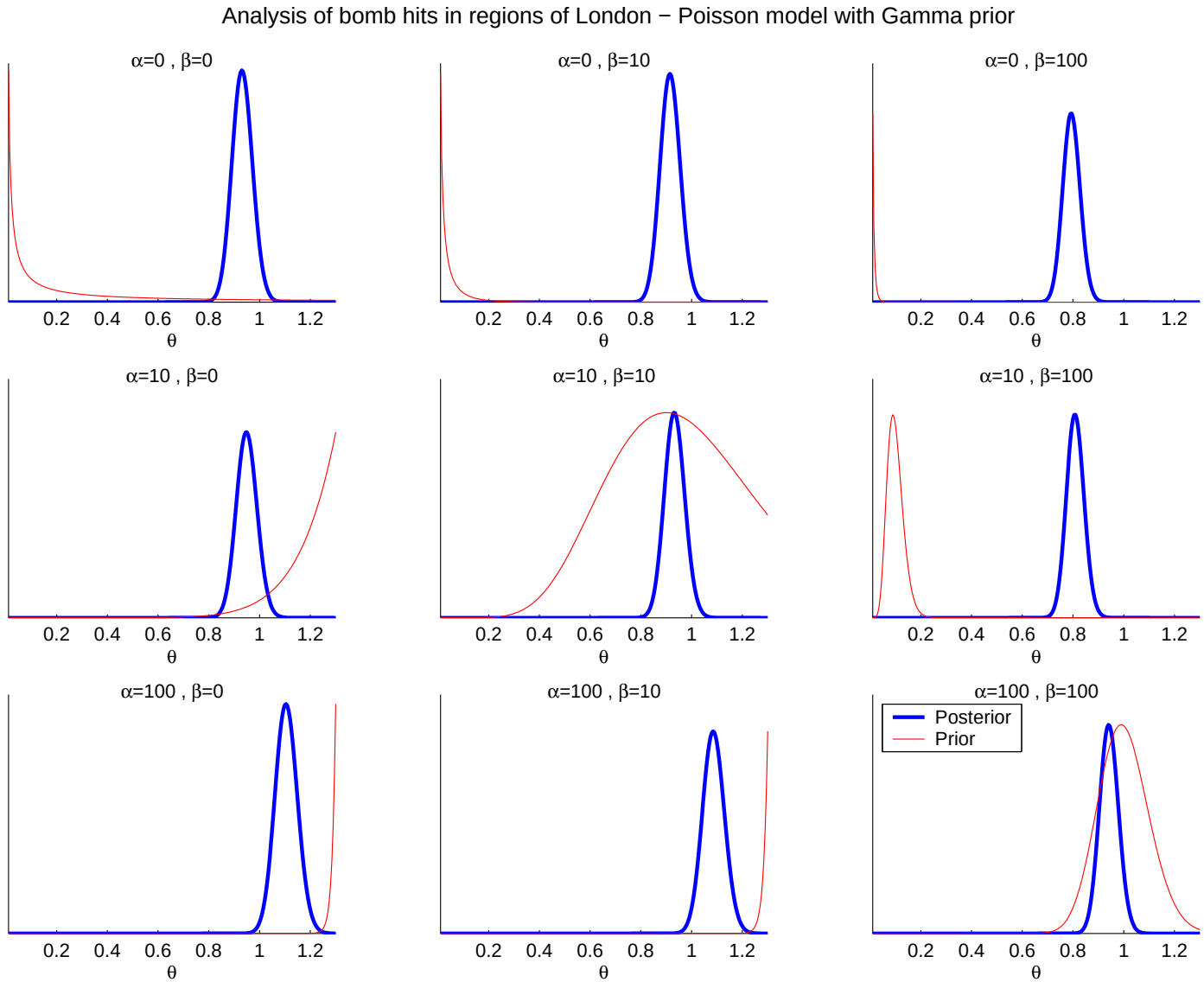
$$E(\theta|y) = \frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n} \approx \bar{y} \approx 0.9323,$$

and

$$SD(\theta|y) = \left(\frac{\alpha + \sum_{i=1}^{n} y_i}{(\beta + n)^2} \right)^{1/2} = \frac{(\alpha + \sum_{i=1}^{n} y_i)^{1/2}}{(\beta + n)} \approx \frac{(537)^{1/2}}{576} \approx 0.0402.$$

if α and β are small compared to $\sum_{i=1}^{n} y_i$ and n.

Poisson example - Bomb hits



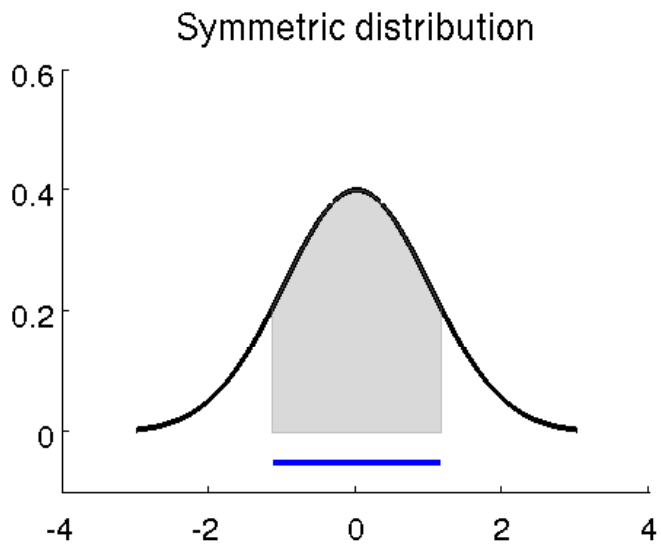
Poisson example - posterior probability intervals

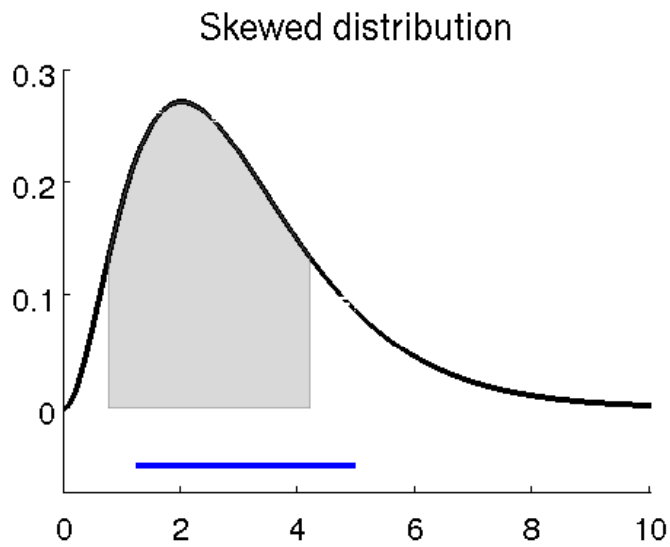
- Bayesian 95% interval: the probability that the unknown parameter θ lies in the interval is 0.95. What a relief!
- Approximate 95% credible interval for θ (for small α and β):

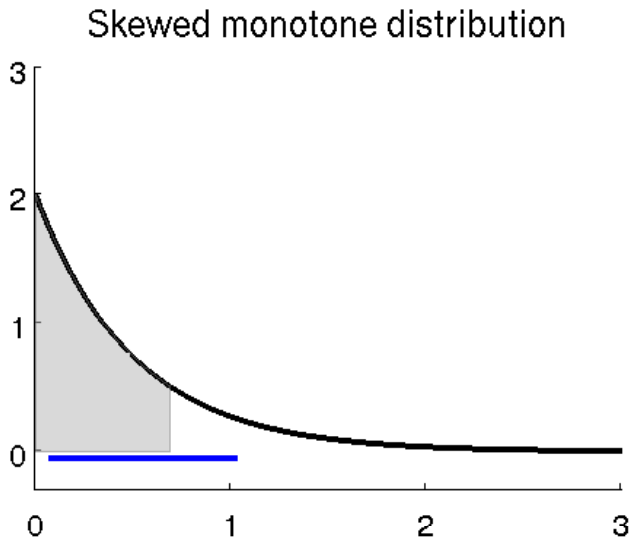
$$E(\theta|y) \pm 1.96 \cdot SD(\theta|y) = [0.8535; 1.0111]$$

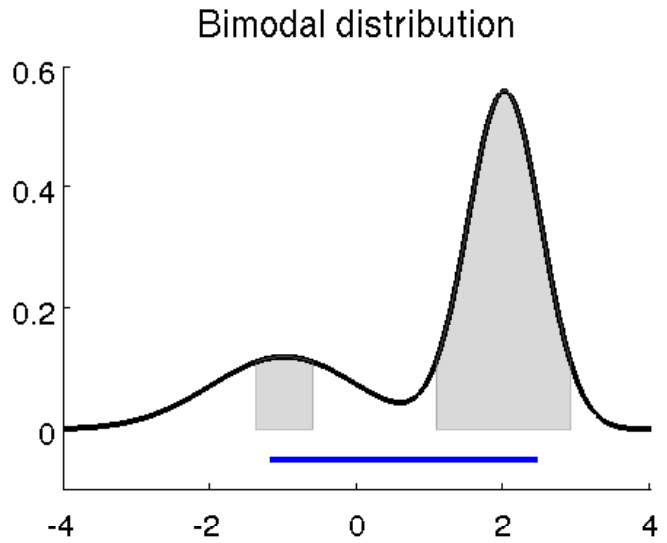
- An exact 95% equal-tail interval is $[0.8550; 1.0125]$ (assuming $\alpha = \beta = 0$)
- An exact Highest Posterior Density (HPD) interval is $[0.8525; 1.0144]$. Obtained numerically, assuming $\alpha = \beta = 0$.

Illustration of different interval types









Non-informative priors

- ... do not exist!
- ... may be improper and still lead to proper posterior
- Regularization priors
- Ideal communication. Present the posterior distributions for all possible priors.
- Practical communication - Reference priors.
- Cannot demand that users specify priors on high-dimensional in detail. Model the prior in terms of a few hyperparameters.

Bayesian asymptotics

- General result:

$$p(\theta|y) \longrightarrow N\left[\hat{\theta}, J^{-1}(\hat{\theta})\right] \text{ for all } p(\theta) \text{ as } n \longrightarrow \infty,$$

where

$$J(\theta) = -E_{y|\theta}\left[\frac{\partial^2 \ln p(y|\theta)}{\partial \theta^2}\right]$$

is the **expected** Fisher information.

- Subjective consensus: eventually all priors give the same posterior.
- Similarly, we have the large-sample approximation

$$\theta|y \overset{approx}{\sim} N\left[\hat{\theta}, I^{-1}(\hat{\theta})\right]$$

where

$$I(\theta) = -\frac{\partial^2 \ln\left[p(y|\theta)p(\theta)\right]}{\partial \theta^2}$$

is the **observed** Fisher information.

Numerical optimization

- The approximation

$$\theta|y \overset{approx}{\sim} N\left[\hat{\theta}, I^{-1}(\hat{\theta})\right]$$

  can often be used with both $\hat{\theta}$ and $I(\hat{\theta})$ obtained from off-the-shelf numerical optimization routines.
- Re-parametrization $\phi = g(\theta)$ may improve normal approximation. If $\theta \geq 0$ use logs. If $0 \leq \theta \leq 1$, use $\text{Logit}(\theta) = \ln[\theta/(1-\theta)]$.
- Standard (e.g. gradient-based) optimization routines may be used (e.g. optim in R or fminunc in Matlab).
  - Input: $p(y|\theta)$ and $p(\theta)$ and initial values.
  - Output: Posterior mode and Hessian matrix (minus observed information).

Jeffreys' prior

- A common non-informative prior is Jeffreys' prior

$$p(\theta) = |J(\theta)|^{1/2},$$

  where $J(\theta)$ is the **expected** Fisher information.
- Invariant to 1:1 transformations of θ. Doesn't matter which parametrization we derive the prior, it always contains the same info.
- Two models with identical likelihood functions (up to constant) can yield different Jeffreys' prior. Jeffreys' prior does not respect the likelihood principle. The crux of the matter is the expectation with respect to the sampling distribution.
- Jeffreys' prior may be a very complicated (non-conjugate) distribution.
- Problematic in multivariate problems. Dubious results in many standard models.

Jeffreys' prior for Bernoulli trial data

$$y_1, ..., y_n|\theta \overset{iid}{\sim} Bern(\theta).$$

$$\ln p(y|\theta) = s \ln \theta + f \ln(1-\theta)$$

$$\frac{d \ln p(y|\theta)}{d\theta} = \frac{s}{\theta} - \frac{f}{(1-\theta)}$$

$$\frac{d^2 \ln p(y|\theta)}{d\theta^2} = -\frac{s}{\theta^2} - \frac{f}{(1-\theta)^2}$$

$$J(\theta) = \frac{E_{y|\theta}(s)}{\theta^2} + \frac{E_{y|\theta}(f)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Thus, the Jeffreys' prior is

$$p(\theta) = |J(\theta)|^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto Beta(\theta|1/2, 1/2).$$

Jeffreys' prior Binomial vs Negative binomial sampling

- Bernoulli experiment: Perform n independent trials with success probabilty θ and count the number of successes. Here

$$y|\theta \sim Bin(\theta)$$

- Inverse Bernoulli experiment: Perform independent trials with success probabilty θ until you have observed y successes. Here

$$y|\theta \sim NegBin(\theta)$$

- Exercise: Suppose you performed both of the two experiments and that in both cases you ended up doing n trials and observed y successes. Show that the likelihood function conveys the same information on θ in both cases, but that Jeffreys prior is not the same in both models. Is this reasonable?