$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{\int p(D|\Theta)p(\Theta)d\Theta}$$

Bayesian Learning 732A46: Lecture 2

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

March 2016

Lecture overview

- The Poisson model
- Conjugate priors
- Prior elicitation
- Non-informative priors

The Poisson model with a Gamma prior

- **Model**:

$$y_1, ..., y_n|\theta \overset{iid}{\sim} \mathrm{Poisson}(y_i|\theta) = \frac{1}{y_i!}\theta^{y_i} \exp(-\theta), \quad \theta > 0.$$

- **Likelihood**

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

- **Prior**

$$p(\theta) \propto \theta^{\alpha_0 - 1} \exp(-\theta\beta_0) \propto \mathrm{Gamma}(\theta|\alpha_0, \beta_0)$$

Interpretation: contains the info: $\alpha_0 - 1$ counts in β_0 observations.

- **Posterior**

$$\begin{aligned} p(\theta|y) &\propto \left[\prod_{i=1}^{n} p(y_i|\theta)\right] p(\theta) \\ &\propto \theta^{\sum_{i=1}^{n} y_i} \exp(-\theta n)\theta^{\alpha_0 - 1} \exp(-\theta\beta_0) \\ &= \theta^{(\alpha_0 + \sum_{i=1}^{n} y_i) - 1} \exp[-\theta(\beta_0 + n)] \propto \mathrm{Gamma}(\theta|\underbrace{\alpha_0 + \sum_{i=1}^{n} y_i}_{\alpha_n}, \underbrace{\beta_0 + n}_{\beta_n}). \end{aligned}$$

Poisson example - Bomb hits in London

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

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

$$p(\theta|y) \propto \theta^{\alpha_0 + 537 - 1} \exp[-\theta(\beta_0 + 576)]$$

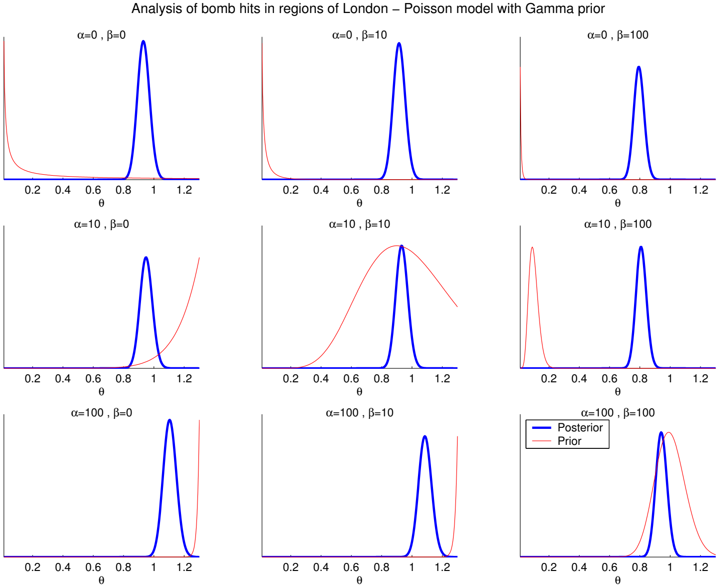
$$E(\theta|y) = \frac{\alpha_0 + \sum_{i=1}^{n} y_i}{\beta_0 + n} \approx \bar{y} \approx 0.9323,$$

and

$$SD(\theta|y) = \left(\frac{\alpha_0 + \sum_{i=1}^{n} y_i}{(\beta_0 + n)^2} \right)^{1/2} = \frac{(\alpha_0 + \sum_{i=1}^{n} y_i)^{1/2}}{(\beta_0 + n)} \approx \frac{(537)^{1/2}}{576} \approx 0.0402.$$

if α_0 and β_0 **are small compared** to $\sum_{i=1}^{n} y_i$ and n.

Poisson bomb hits in London



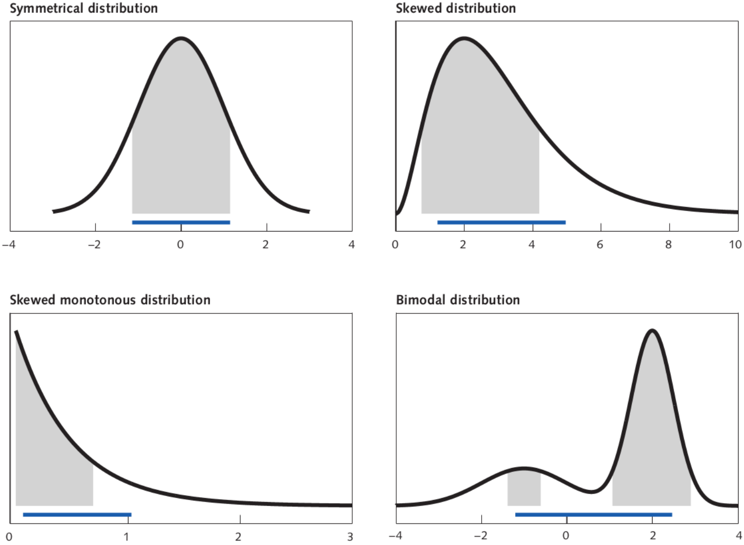
Poisson example - posterior intervals

- **Bayesian 95% interval**: the probability that the **unknown parameter** θ lies in the interval is 0.95. **What an easy and logical interpretation**!
- *Approximate* 95% **credible interval** for θ (for small α_0 and β_0):

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

  Assumes that $p(\theta|y)$ is (approximately) normal.
- An exact 95% **equal-tail interval** is [0.8550; 1.0125] (assuming $\alpha_0 = \beta_0 = 0$)
- **Highest Posterior Density** (**HPD**) interval contains the θ values with highest pdf. Here [0.8525; 1.0144], assuming $\alpha = \beta = 0$.

Illustration of different interval types



Conjugate priors

- **Models** we have see

Model	**Prior**	$\rightarrow$	**Posterior**
Bernoulli	$\theta \sim \text{Beta}(\alpha_0, \beta_0)$	$\rightarrow$	$\theta \vert y \sim \text{Beta}(\alpha_n, \beta_n)$
Normal (σ^2 known)	$\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$	$\rightarrow$	$\theta \vert y \sim \mathcal{N}(\mu_n, \tau_n^2)$
Poisson	$\theta \sim \text{Gamma}(\alpha_0, \beta_0)$	$\rightarrow$	$\theta \vert y \sim \text{Gamma}(\alpha_n, \beta_n)$

- **Conjugate priors**: A prior is conjugate to a model (likelihood) if the prior and posterior belong to the same distributional family.
- **Formally**: 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

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

  holds for all $p(y|\theta) \in \mathcal{F}$.
- A Conjugate prior is **computationally convenient**.

Prior elicitation

- The prior should (ideally) be elicited by an **expert** ($\neq$ statistician, often)
- Elicit the prior on a **quantity that she knows well** (maybe log odds $\log \frac{\theta}{1-\theta}$ when the model is $\mathrm{Bern}(\theta)$).
- The statistician can compute the **implied prior** on θ by transformation of variables.
  Recall: Let $p_u(u)$ be continuous and let $v = h(u)$ be a one-to-one transform.

$$p_v(v) = p_u(h^{-1}(v))|J|, \quad |J| = \text{determinant of } h^{-1}(v) \left[1\mathrm{-dim}: \frac{d}{dv}h^{-1}(v)\right].$$

- **Example**: expert believes $\phi = \log \frac{\theta}{1-\theta} \sim \mathcal{N}(0, 20)$. The implied prior on θ is $[u = \phi,\ v = \theta,\ h^{-1}(v) = \log \frac{v}{1-v}]$

$$p_\theta(\theta) = \mathcal{N}\left(\log \frac{\theta}{1-\theta}\middle| 0, 20\right)\frac{1}{\theta(1-\theta)}, \quad 0 < \theta < 1.$$

- The example works out a **full distribution**.

Prior elicitation, cont.

- Working out **hyper-parameters from expert information**.
- Elicit the prior by asking the expert simple questions: What is $E(\theta)$? or $V(\theta)$?
- The hyper-parameters are "backed out". **Example:** The prior is

$$p(\theta) = \mathrm{Gamma}(\theta|\alpha_0, \beta_0), \quad \text{expert believes} \quad E(\theta) = 2 \text{ and } V(\theta) = 0.25.$$

$$E(\theta) = \frac{\alpha_0}{\beta_0}, \quad V(\theta) = \frac{\alpha_0}{\beta_0^2} \implies p(\theta) = \mathrm{Gamma}(\theta|16, 8).$$

- **Show the expert some consequences** of her elicitated prior.

Prior elicitation - AR(p) example

- **Autoregressive process** of order p

$$y_t \quad = \quad \mu + \phi_1 \cdot (y_{t-1} - \mu) + ... + \phi_p \cdot (y_{t-p} - \mu) + \varepsilon_t,\ \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

- **Informative prior** on the unconditional mean: $\mu \sim N(\mu_0, \tau_0^2)$.
- **"Non-informative"** prior on σ^2:

$$p(\sigma^2) \propto 1/\sigma^2 \quad \text{[uniform in the parameterization } p(\log(\sigma^2)) \propto c\text{]}$$

- **Assume** for simplicity that all $\phi_i, i = 1, ..., p$ are independent a priori, and $\phi_i \sim N(\mu_i, \psi_i^2)$.
- Prior on $\phi = (\phi_1, ..., \phi_p)$ centered on a persistent AR(1) process:

$$\mu_1 = 0.8, \mu_2 = ... = \mu_p = 0.$$

- **Prior variance** ψ_i^2 of the ϕ_i decay towards zeros: $Var(\phi_i) = \frac{c}{i^\lambda}$, so that "longer" lags are **more concentrated around zero** (less likely a priori).
- λ is a parameter that can be used to determine the rate of decay. **Shrinkage/regularization/smoothness** prior.

Different types of prior information

- Real **expert information**. Combo of previous studies and experience.
- Vague prior information, or even **non-informative priors**. **Beware of improper priors - make sure the posterior is proper!**
- **Smoothness priors**. Regularization. Shrinkage. Big thing in modern statistics/machine learning.
- **Hierarchical priors**. Model the uncertainty in the hyper-parameters. **Bayesian estimation of hyper-parameters**.

Non-informative priors

- **Do not exist**! The "flatness" depends on the parametrization of the model.
- Can be improper but still lead to a **proper posterior**.
- **Reference prior**: A prior that plays a "minimal role". "Let the data speak for themselves".
- Jeffreys' **invariance principle**: The prior should contain the same information **regardless of the parametrization** of the model.
- **Jeffreys'** prior (1-dim)

$$p(\theta) \propto |I(\theta)|^{1/2}, \quad I(\theta) = -E_y\left(\frac{d^2}{d\theta^2}\log p(y|\theta)\right),$$

  where $I(\theta)$ is the **Fisher information** for θ.
- The expectation **is w.r.t data**... an **unconditional** (frequentist) feature!
- ... consequently, Jeffreys' prior **does not respect** the likelihood principle.
- Can give **dubious results** in multivariate (parameter) models.

Jeffreys' prior for Bernoulli trial data

Let $y = (y_1, ..., y_n)$

$$y_1, ..., y_n|\theta \overset{iid}{\sim} \mathrm{Bern}(\theta) \quad \text{and} \quad \log p(y|\theta) = s\log\theta + f\log(1-\theta).$$

$$\begin{aligned}
\frac{d\log p(y|\theta)}{d\theta} &= \frac{s}{\theta} - \frac{f}{(1-\theta)} \\
\frac{d^2\log p(y|\theta)}{d\theta^2} &= -\frac{s}{\theta^2} - \frac{f}{(1-\theta)^2} \\
I(\theta) &= \frac{E_y(s)}{\theta^2} + \frac{E_{y|\theta}(f)}{(1-\theta)^2} \\
&= \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}
\end{aligned}$$

Thus, **the Jeffreys' prior** is

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

Non-informative priors - my two cents

- **Overrated**. Likelihood **dominates the prior** as more data becomes available.
- **State-of-the-art** models are **very complex** these days. **Regularization/shrinkage/smoothness priors** to avoid over-fitting.
- Non-informative priors **do not shrink**.

Non-informative prior $\implies$ **no shrinkage** $\implies$ **no fun**.