Bayesian Learning

Lecture 2 - Poisson data. Prior elicitation. Invariant priors.

Mattias Villani 👨

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Lecture overview

- The **Poisson model**
- **Conjugate priors**
- **Prior elicitation**
- **Jeffreys' prior**

Poisson model

- **Model**

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

- **Poisson distribution**

$$p(y) = \frac{\theta^y e^{-\theta}}{y!}$$

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

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

- **Prior**

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

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

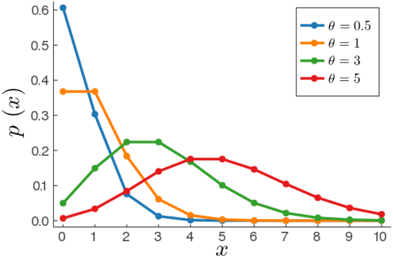
Poisson distribution

$X \sim \mathrm{Pois}(\theta)$
for $X \in 0, 1, 2, \ldots$

$$p(x) = \frac{\theta^x e^{-\theta}}{x!}$$

$$\mathbb{E}(X) = \theta$$

$$\mathbb{V}(X) = \theta$$



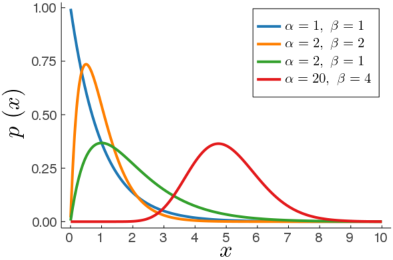
Gamma distribution

$$X \sim \mathrm{Gamma}(\alpha, \beta)$$

$$p(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$$\mathbb{E}(X) = \frac{\alpha}{\beta}$$

$$\mathbb{V}(X) = \frac{\alpha}{\beta^2}$$



Poisson posterior

- **Posterior**

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

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

- **Prior-to-Posterior mapping**

Model: $y_1, ..., y_n|\theta \overset{iid}{\sim} Pois(\theta)$

Prior: $\theta \sim \mathrm{Gamma}(\alpha, \beta)$

Posterior: $\theta|y_1, ..., y_n \sim \mathrm{Gamma}(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$.

Example - Number of bids in eBay auctions

- **Data**:
  - Number of placed bids in $n = 1000$ eBay coin auctions.
  - Sum of counts: $\sum_{i=1}^{n} y_i = 3635$.
  - Average number bids per auction: $\bar{y} = 3635/1000 = 3.635$.
- **Prior**: $\alpha = 2$, $\beta = 1/2$.

$$E(\theta) = \frac{\alpha}{\beta} = 4$$

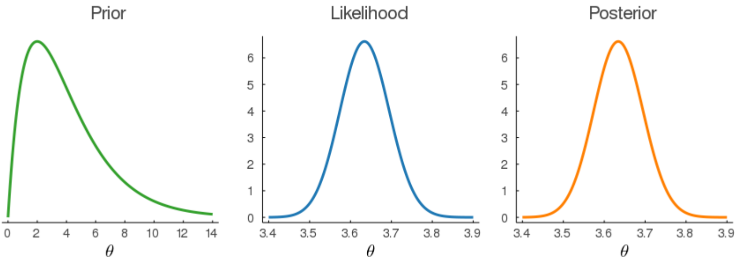
$$SD(\theta) = \frac{\alpha}{\beta^2} = 2.823$$

- **Posterior**

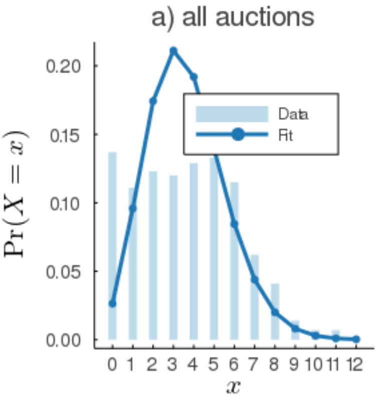
$$E(\theta|\mathbf{y}) = \frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n} = \frac{2 + 3635}{1/2 + 1000} \approx 3.635.$$

$$SD(\theta|\mathbf{y}) = \left(\frac{\alpha + \sum_{i=1}^{n} y_i}{(\beta + n)^2} \right)^{1/2} \approx 0.060.$$

eBay data - Posterior of θ



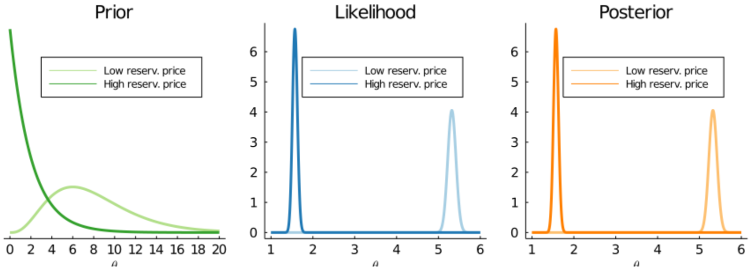
eBay data - model fit at $\theta = \mathbb{E}(\theta|\mathbf{x})$



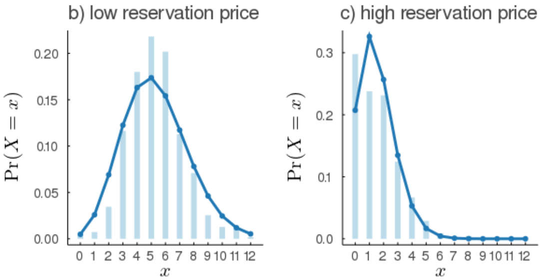
eBay - low/high seller's reservation price

- The data is very heterogenous. Some auctions start with very high reservations prices (lowest price accepted by the seller).
- Split the data into auctions with low/high reservation prices.
- **Low reservation price auctions**:
  - $n = 550$ eBay coin auctions.
  - Posterior mean: 5.321 bids.
- **High reservation price auctions**:
  - $n = 450$ eBay coin auctions.
  - Posterior mean: 1.576 bids.

eBay data split on reservation price



eBay data - model fit at $\mathbb{E}(\theta|\mathbf{x})$



- Better fits, but still not good enough.
- Lab 3: Fit **Poisson regression** with reservation price as continuous covariate.

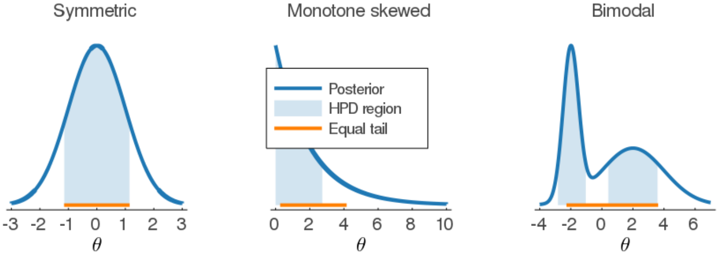
Posterior intervals

- **Bayesian 95% credible interval**: the probability that the unknown parameter θ lies in the interval is 0.95.
- 95% **equal-tail interval**: from 2.5% to 97.5% percentile.
- Approximate 95% **credible interval**

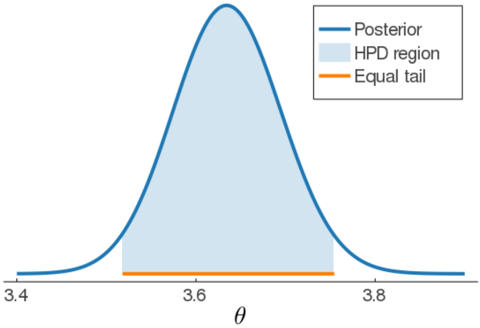
$$E(\theta|y) \pm 1.96 \cdot SD(\theta|y)$$

- **Highest Posterior Density** (**HPD**) interval contains the θ values with highest pdf.

Illustration of different interval types



Credible intervals - eBay auction data



Conjugate priors

- Normal likelihood: Normal prior $\rightarrow$ Normal posterior.
- Bernoulli likelihood: Beta prior $\rightarrow$ Beta posterior.
- Poisson likelihood: Gamma prior $\rightarrow$ Gamma posterior.

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

> a family of prior distributions $\mathcal{P}$ is **conjugate** for
> a family of likelihoods $\mathcal{L} = \{p(\mathbf{x}|\theta), \theta \in \Theta\}$ if
>
> $$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|\mathbf{x}) \in \mathcal{P} \qquad \text{for all } p(\mathbf{x}|\theta) \in \mathcal{L}$$

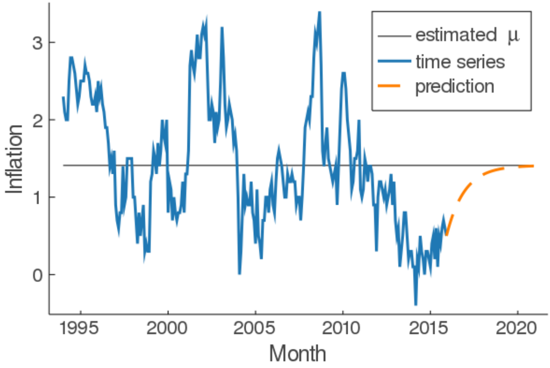
Autoregressive time series model

- **Autoregressive process** or order p - AR(p)

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

- Unconditional mean: $\mathbb{E}(y_t) = \mu$. Long run forecast attraction.

$$\mathbb{E}(y_{T+h} | y_{1:T}) \to \mu \text{ as } h \to \infty.$$



Prior elicitation - AR(p)

- **Autoregressive process**

$$y_t = \mu + \phi_1(y_{t-1} - \mu) + ... + \phi_p(y_{t-p} - \mu) + \varepsilon_t$$

- **Expert prior** on the unconditional mean: $\mu \sim N(\mu_0, \tau_0^2)$.
- **Regularization prior** on $\phi_1, \ldots \phi_p$

$$\phi_k \sim N\left(\mu_k, \frac{\tau^2}{k^2}\right) \text{ independently apriori}$$

  - Prior mean on persistent AR(1): $\mu_1 = 0.8, \mu_2 = ... = \mu_p = 0$
  - $\mathbb{V}(\phi_k) = \frac{\tau^2}{k^2}$. Coeff on "longer" lags more likely to be small.

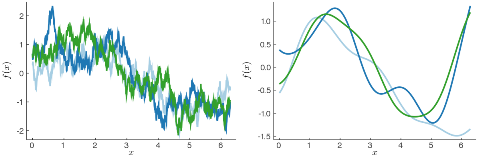
- **Hierarchical prior**
  - Hard to specify τ^2? Put a prior on it!
  - $\phi_k | \tau^2 \sim N\left(\mu_k, \frac{\tau^2}{k^2}\right)$ and $\tau^2 \sim \chi^2_\nu$.
  - Gives a posterior on global shrinkage τ^2.

Prior elicitation

Smoothness priors

- a version of regularization priors
- nonlinear regression function $f(\cdot)$ is believed to be smooth

$$y = f(x) + \varepsilon$$



Noninformative priors

- **Uniform**: $\theta \sim \mathrm{Beta}(1,1)$.
  Issue 1: same as prior sample with one success and one failure.
  Issue 2: not uniform for $\phi = \log \frac{\theta}{1-\theta}$.
- **Zero prior sample size**: $\theta \sim \mathrm{Beta}(\epsilon, \epsilon)$ with $\epsilon \downarrow 0$.
  Posterior $\rightarrow \mathrm{Beta}(s, f)$.
  Issue: posterior is improper if $s = 0$ or $f = 0$.

Invariant prior

- **Observed information**

$$J_{\mathbf{x}}(\hat{\theta}) = -\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}}$$

- **Fisher information**

$$I(\theta) = E_{\mathbf{x}|\theta}\left(J_{\mathbf{x}}(\theta)\right)$$

- **Jeffreys' rule** to construct prior

$$p(\theta) = I(\theta)^{1/2}.$$

- **Invariance** under 1:1 parameter transformation $\phi = g(\theta)$.
Example: $\phi = \log \frac{\theta}{1-\theta}$.
    - Specify $p_\theta(\theta)$ directly
    - Specify $p_\phi(\phi)$ and then obtain $p_\theta(\theta) = p_\phi(g^{-1}(\theta)) \left| \frac{dg^{-1}(\theta)}{d\theta} \right|$.

Jeffreys' prior for Bernoulli sampling

$$x_1, ..., x_n|\theta \overset{iid}{\sim} Bern(\theta).$$

$$\ln p(\mathbf{x}|\theta) = s \ln \theta + f \ln(1-\theta)$$

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

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

$$I(\theta) = \frac{E_{\mathbf{x}|\theta}(s)}{\theta^2} + \frac{E_{\mathbf{x}|\theta}(f)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

Thus, the Jeffreys' prior is

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

Jeffreys' prior for negative binomial sampling

- Jeffreys' prior:

$$n|\theta \overset{iid}{\sim} NegBin(s, \theta).$$

$$\ln p(\mathbf{x}|\theta) = \ln \binom{n-1}{s-1} + s \ln \theta + f \ln(1-\theta)$$

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

$$I(\theta) = \frac{s}{\theta^2} + \frac{E_{n|\theta}(n-s)}{(1-\theta)^2} = \frac{s}{\theta^2} + \frac{s/\theta - s}{(1-\theta)^2} = \frac{s}{\theta^2(1-\theta)}$$

- Thus, the Jeffreys' prior is

$$p(\theta) = |I(\theta)|^{1/2} \propto \theta^{-1}(1-\theta)^{-1/2} \propto Beta(\theta|0, 1/2).$$

- Jeffreys' prior is **improper**, but the posterior is proper: $\theta|n \sim \mathrm{Beta}(s, f + 1/2)$ which is proper since $s \geq 1$.
- Jeffreys' prior **violates the likelihood principle** because $I(\theta)$ is sampling-based.