

BAYESIAN LEARNING - LECTURE 2

Mattias Villani

**Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University**

LECTURE OVERVIEW

- ▶ The Poisson model
- ▶ Conjugate priors
- ▶ Prior elicitation - how to come up with a prior.
- ▶ Non-informative priors

POISSON MODEL

► **Model:**

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Pois}(\theta)$$

► **Likelihood** from iid Poisson sample $y = (y_1, \dots, y_n)$

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

► **Prior:**

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

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

POISSON MODEL, CONT.

► *Posterior*

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

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

► Prior-to-Posterior mapping:

$$\text{Model: } y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Pois}(\theta)$$

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

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

POISSON EXAMPLE - NUMBER OF BOMB HITS IN LONDON

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

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

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

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

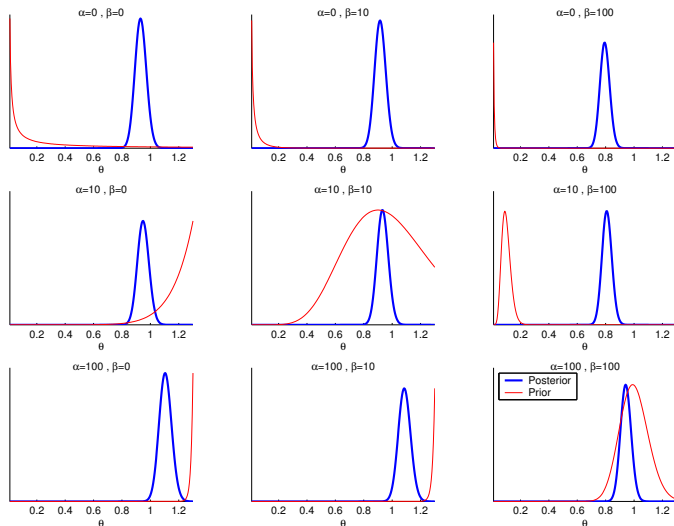
and

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

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

POISSON BOMB HITS IN LONDON

Analysis of bomb hits in regions of London – Poisson model with Gamma prior



POISSON EXAMPLE - POSTERIOR PROBABILITY INTERVALS

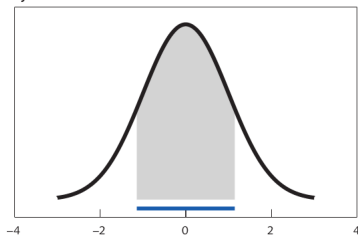
- ▶ **Bayesian 95% interval**: the probability that the unknown parameter θ lies in the interval is 0.95. What a relief!
- ▶ Approximate 95% **credible interval** for θ (for small α and β):

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

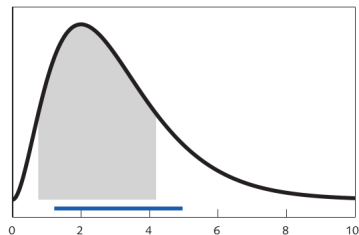
- ▶ An exact 95% equal-tail interval is $[0.8550; 1.0125]$ (assuming $\alpha = \beta = 0$)
- ▶ **Highest Posterior Density (HPD)** interval contains the θ values with highest pdf.
- ▶ An exact Highest Posterior Density (HPD) interval is $[0.8525; 1.0144]$. Obtained numerically, assuming $\alpha = \beta = 0$.

ILLUSTRATION OF DIFFERENT INTERVAL TYPES

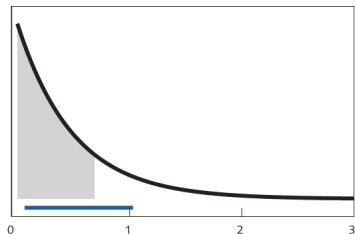
Symmetrical distribution



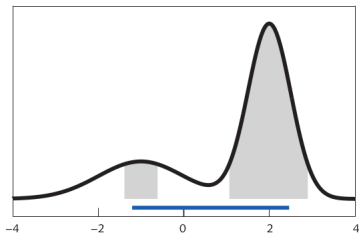
Skewed distribution



Skewed monotonous distribution



Bimodal distribution



CONJUGATE PRIORS

- ▶ Normal likelihood: Normal prior \rightarrow Normal posterior. (posterior belongs to the same distribution family as prior)
- ▶ Bernoulli likelihood: Beta prior \rightarrow Beta posterior.
- ▶ Poisson likelihood: Gamma prior \rightarrow Gamma posterior.
- ▶ **Conjugate priors**: A prior is conjugate to a model (likelihood) if the prior and posterior belong to the same distributional family.
- ▶ Formal definition: 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 $p(y|\theta) \in \mathcal{F}$.

PRIOR ELICITATION

- ▶ The prior should be determined (elicited) by an **expert**. Typically, expert \neq statistician.
- ▶ Elicit the prior on a **quantity that she knows well** (maybe log odds $\ln \frac{\theta}{1-\theta}$ when the model is $Bern(\theta)$). The statistician can always compute the implied prior on other quantities after the elicitation.
- ▶ Elicit the prior by asking the expert probabilistic questions:
 - ▶ $E(\theta) = ?$
 - ▶ $SD(\theta) = ?$
 - ▶ $Pr(\theta < c) = ?$
 - ▶ $Pr(y > c) = ?$
- ▶ **Show the expert some consequences** of her elicited prior. If she does not agree with these consequences, iterate the above steps until she is happy.
- ▶ **Beware of psychological effects**, such as anchoring.

PRIOR ELICITATION - AR(P) EXAMPLE

- ▶ **Autoregressive process** of order p

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

- ▶ Informative prior on the unconditional mean: $\mu \sim N(\mu_0, \tau_0^2)$. Usually, μ_0 and τ_0^2 can be specified accurately.
- ▶ “Noninformative” prior on σ^2 : $p(\sigma^2) \propto 1/\sigma^2$
- ▶ Assume for simplicity that all $\phi_i, i = 1, \dots, p$ are independent a priori, and $\phi_i \sim N(\mu_i, \psi_i)$
- ▶ Prior on $\phi = (\phi_1, \dots, \phi_p)$ centered on persistent AR(1) process: $\mu_1 = 0.8, \mu_2 = \dots = \mu_p = 0$
- ▶ Prior variance of the ϕ_i decay towards zeros: $\text{Var}(\phi_i) = \frac{c}{i^\lambda}$, so that “longer” lags are more likely to be zero a priori. λ is a parameter that can be used to determine the rate of decay.

DIFFERENT TYPES OF PRIOR INFORMATION

- ▶ Real **expert information**. Combo of previous studies and experience.
- ▶ Vague prior information, or even **noninformative priors**.
- ▶ **Reporting priors**. Easy to understand the information they contain.
- ▶ **Smoothness priors**. Regularization. Shrinkage. Big thing in modern statistics/machine learning.

NON-INFORMATIVE PRIORS

- ▶ ... do not exist!
- ▶ ... may be **improper** and still lead to proper posterior
- ▶ **Regularization priors.**
- ▶ Ideal: Present the posterior distributions for all possible priors.
- ▶ Practical communication - **Reference priors.**
- ▶ Model the prior in terms of a few **hyperparameters.**

NON-INFORMATIVE PRIORS

- ▶ **Subjective consensus**: when extreme priors give essentially the same posterior.

$$p(\theta|y) \rightarrow N\left(\hat{\theta}, J_{\hat{\theta}, \mathbf{x}}^{-1}\right) \text{ for all } p(\theta) \text{ as } n \rightarrow \infty,$$

where $J_{\theta, \mathbf{x}}$ is the **observed information**

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

- ▶ A common non-informative prior is **Jeffreys' prior**

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

where I_{θ} is the **Fisher information**

$$I_{\theta} = E_{\mathbf{x}|\theta} (J_{\theta, \mathbf{x}})$$

JEFFREYS' PRIOR FOR BERNOULLI TRIAL DATA

$$x_1, \dots, x_n | \theta \stackrel{iid}{\sim} \text{Bern}(\theta).$$

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

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

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

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

Thus, the Jeffreys' prior is

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

PROPERTIES OF JEFFREYS PRIOR

- ▶ **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.