

# BAYESIAN LEARNING - LECTURE 2

Mattias Villani

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

# LECTURE OVERVIEW

- ▶ The Normal model
- ▶ The Poisson model
- ▶ Conjugate priors
- ▶ Non-informative priors

# NORMAL DATA WITH KNOWN VARIANCE - UNIFORM PRIOR

- ▶ Model:

$$x_1, \dots, x_n | \theta, \sigma^2 \stackrel{iid}{\sim} N(\theta, \sigma^2).$$

- ▶ Prior:

$$p(\theta) \propto c$$

- ▶ Likelihood (see Technical Appendix A):

$$\begin{aligned} p(x_1, \dots, x_n | \theta, \sigma^2) &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp \left[ -\frac{1}{2\sigma^2} (x_i - \theta)^2 \right] \\ &\propto \exp \left[ -\frac{1}{2(\sigma^2/n)} (\theta - \bar{x})^2 \right]. \end{aligned}$$

- ▶ Posterior

$$\theta | x_1, \dots, x_n \sim N(\bar{x}, \sigma^2/n)$$

# NORMAL WITH KNOWN VARIANCE - NORMAL PRIOR

- Prior

$$\theta \sim N(\mu_0, \tau_0^2)$$

- Posterior (see Technical Appendix A)

$$\begin{aligned} p(\theta | x_1, \dots, x_n) &\propto p(x_1, \dots, x_n | \theta, \sigma^2) p(\theta) \\ &\propto N(\theta | \mu_n, \tau_n^2), \end{aligned}$$

where

$$\frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2},$$

$$\mu_n = w\bar{x} + (1 - w)\mu_0,$$

and

$$w = \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

## NORMAL WITH KNOWN VARIANCE - NORMAL PRIOR, CONT.

$$\theta \sim N(\mu_0, \tau_0^2) \xrightarrow{x_1, \dots, x_n} \theta|x \sim N(\mu_n, \tau_n^2).$$

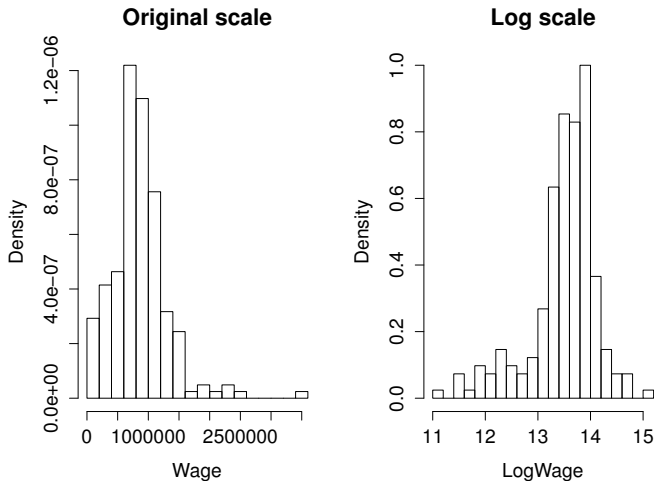
Posterior precision = Data precision + Prior precision

Posterior mean =

$$\frac{\text{Data precision}}{\text{Posterior precision}}(\text{Data mean}) + \frac{\text{Prior precision}}{\text{Posterior precision}}(\text{Prior mean})$$

# CANADIAN WAGES DATA

- Data on wages for 205 Canadian workers.



# CANADIAN WAGES

- ▶ Model

$$X_1, \dots, X_n | \theta \sim N(\theta, \sigma^2), \sigma^2 = 0.4$$

- ▶ Prior

$$\theta \sim N(\mu, \tau^2) \quad \mu = 12 \text{ and } \tau = 10$$

- ▶ Posterior

$$\theta | x_1, \dots, x_n \sim N \left[ \tilde{\theta}, (\sigma^{-2}n + \tau^{-2})^{-1} \right],$$

where  $\tilde{\theta} = w\bar{x} + (1 - w)\mu$ .

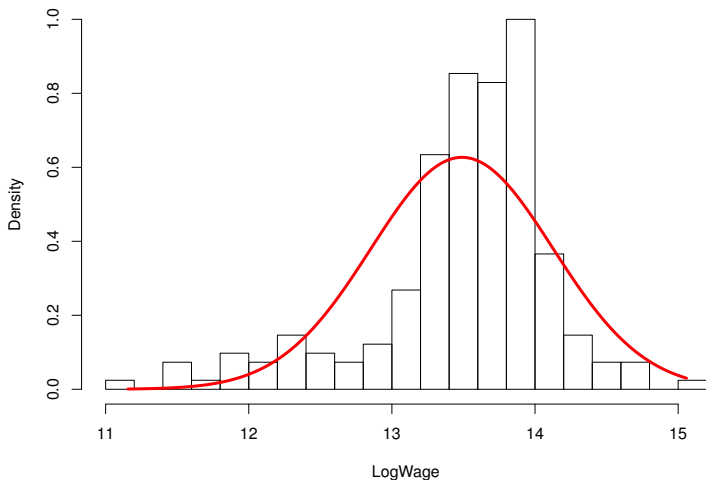
- ▶ For the Canadian wage data:

$$w = \frac{\sigma^{-2}n}{\sigma^{-2}n + \tau^{-2}} = \frac{2.5 \cdot 205}{2.5 \cdot 205 + 1/100} = 0.99998.$$

$$\tilde{\theta} = w\bar{x} + (1 - w)\mu = 0.99998 \cdot 13.48988 + (1 - 0.99998) \cdot 12 = 13.4898$$

$$(\sigma^{-2}n + \tau^{-2})^{-1} = (2.5 \cdot 205 + 1/100)^{-1} = 0.00195$$

# CANADIAN WAGES DATA - MODEL FIT





# 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.
- ▶ **Conjugate priors:** A prior is conjugate to a model (likelihood) if the prior and posterior belong to the same distributional family.
- ▶ *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  $p(y|\theta) \in \mathcal{F}$ .

- ▶ **Natural conjugate prior:**  $p(\theta) = c \cdot p(y_1, \dots, y_n|\theta)$  for some constant  $c$ , i.e. the prior is of the same functional form as the likelihood.

# POISSON MODEL

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

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

so that the sum of counts  $\sum_{i=1}^n y_i$  is a sufficient statistic for  $\theta$ .

- *Natural conjugate prior for Poisson parameter  $\theta$*

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

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

## POISSON MODEL, CONT.

- *Posterior for Poisson parameter  $\theta$ .* Multiplying the poisson likelihood and the Gamma prior gives the 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.

- In summary

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

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

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  $= \bar{y} = 537/576 \approx 0.9323$ .

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

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

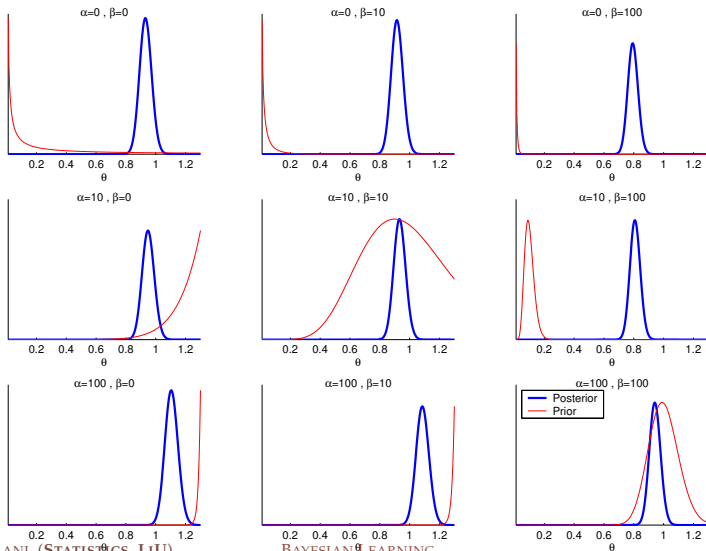
and

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

if  $\alpha$  and  $\beta$  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  $\theta$  lies in the interval is 0.95. What a relief!
- ▶ Approximate 95% credible interval for  $\theta$  (for small  $\alpha$  and  $\beta$ ):

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

- ▶ 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

SEE SEPARATE FIGURE!

# PRIOR ELICITATION

- ▶ The prior should be determined (elicited) by an expert. Typically, expert  $\neq$  statistician.
- ▶ Elicit the prior on a quantity that he knows well (he may prefer the log odds  $\ln \frac{\theta}{1-\theta}$  over the success probability  $\theta$  in a Bernoulli experiment). The statistician can always compute the implied prior on other quantities after the elicitation.
- ▶ Elicit the prior by asking the expert probabilistic questions:  $E(\theta) = ?$ ,  $SD(\theta) = ?$  or  $Pr(\theta < c) = ?$  or even  $Pr(y > c) = ?$ .
- ▶ Show the expert some consequences of his elicited prior. If he does not agree with these consequences, iterate the above steps until he is happy.



## 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,  $\mu_0$  and  $\tau_0^2$  can be specified accurately.
- ▶ “Noninformative” prior on  $\sigma^2$ :  $p(\sigma^2) \propto 1/\sigma^2$
- ▶ 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  $\phi_i$  decay towards zeros:  $\text{Var}(\phi_i) = \frac{c}{i^\lambda}$ , so that “longer” lags are more likely to be zero a priori.  $\lambda$  is a parameter that can be used to determine the rate of decay.

# 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.
- ▶ Subjective consensus: when extreme priors give essentially the same posterior. This will happen, given enough data as

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

where  $J_{\hat{\theta}, \mathbf{x}}$  is the (observed) information (matrix).

# JEFFREYS' PRIOR

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

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

where  $I_{\theta}$  is the Fisher information.

## JEFFREYS' PRIOR FOR BERNOULLI TRIAL DATA

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

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

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

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

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

Thus, the Jeffreys' prior is

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

# JEFFREYS' PRIOR BINOMIAL VS NEGATIVE BINOMIAL SAMPLING

- ▶ Bernoulli experiment: Perform  $n$  independent trials with success probability  $\theta$  and count the number of successes. Here

$$y|\theta \sim \text{Bin}(\theta)$$

- ▶ Inverse Bernoulli experiment: Perform independent trials with success probability  $\theta$  until you have observed  $y$  successes. Here

$$y|\theta \sim \text{NegBin}(\theta)$$

- ▶ 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  $\theta$  in both cases, but that Jeffreys prior is not the same in both models. Is this reasonable?

# PROPERTIES OF JEFFREYS PRIOR

- ▶ Invariant to 1:1 transformations of  $\theta$ . 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.