Bayesian Learning - Lecture 2

Mattias Villani

Division of Statistics and Machine Learning
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:
$$x_1, ..., x_n | \theta, \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2).$$
- Prior:
$$p(\theta) \propto c$$
- Likelihood
$$\begin{aligned} p(x_1, ..., x_n | \theta, \sigma^2) &= \Pi_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(x_i - \theta)^2\right] \\ &\propto \exp\left[-\frac{1}{2(\sigma^2/n)}(\theta - \bar{x})^2\right]. \end{aligned}$$
- Posterior
$$\theta | x_1, ..., x_n \sim N(\bar{x}, \sigma^2/n)$$

Normal with known variance - normal prior

- Prior

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

- Posterior

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

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.

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

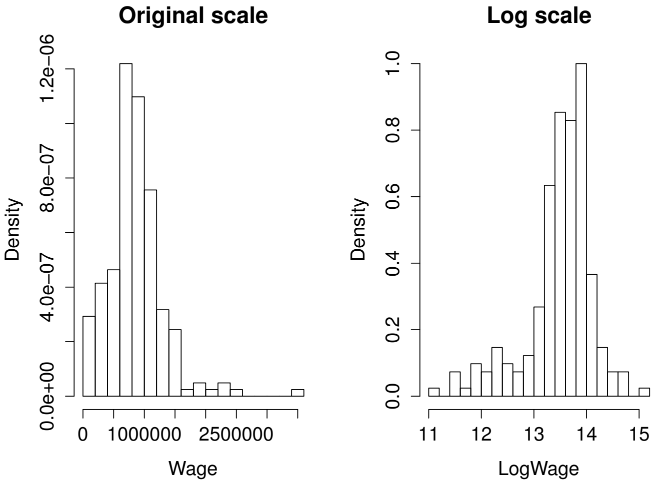
$$\text{Posterior precision} = \text{Data precision} + \text{Prior precision}$$

$$\text{Posterior mean} =$$

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

Canadian wages data

- Data on wages for 205 Canadian workers.



Canadian wages

- Model

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

- Prior

$$\theta \sim N(\mu_0, \tau_0^2)\ \mu_0 = 12 \text{ and } \tau_0 = 10$$

- Posterior

$$\theta|x_1, ..., x_n \sim N\left(\mu_n, \tau_n^2\right),$$

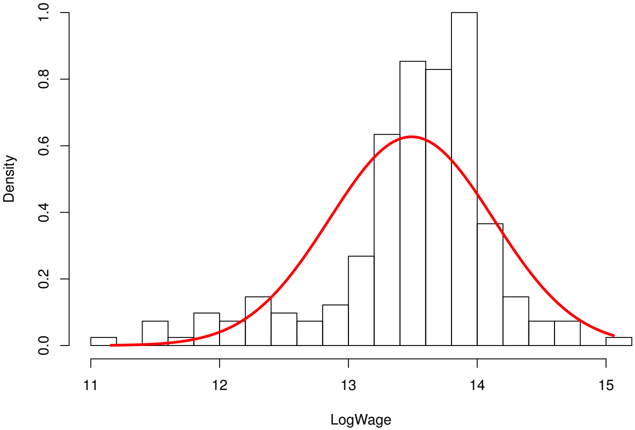
  where $\mu_n = w\bar{x} + (1 - w)\mu_0$.
- For the Canadian wage data:

$$w = \frac{\sigma^{-2}n}{\sigma^{-2}n + \tau_0^{-2}} = \frac{2.5 \cdot 205}{2.5 \cdot 205 + 1/100} = 0.99998.$$

$$\mu_n = w\bar{x} + (1 - w)\mu_0 = 0.99998 \cdot 13.48988 + (1 - 0.99998) \cdot 12 = 13.48$$

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

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**: $p(\theta) = c \cdot p(y_1, ..., y_n|\theta)$ 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

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

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

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

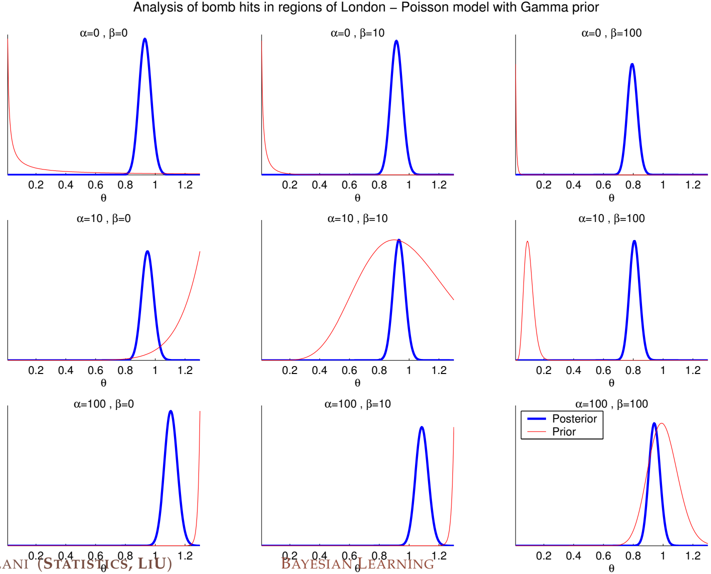
$$E(\theta|y) = \frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n} \approx \bar{y} \approx 0.9323,$$

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 bomb hits in London



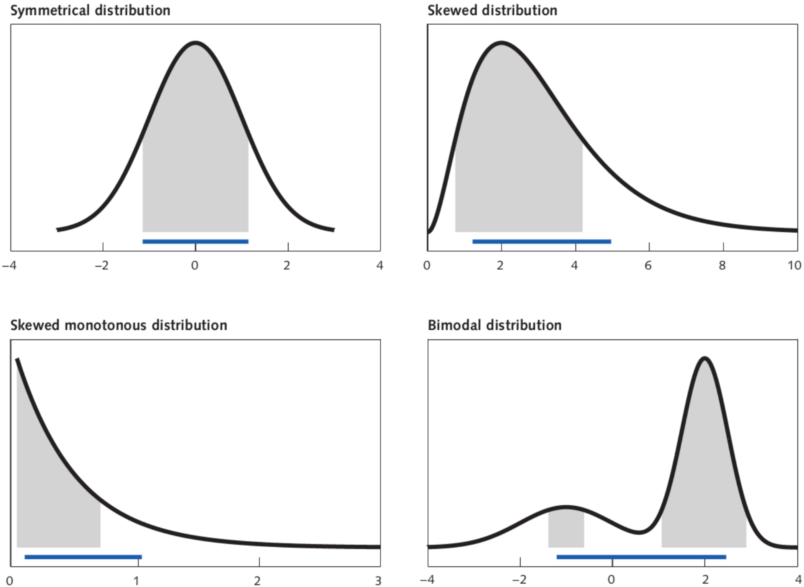
Poisson example - posterior probability intervals

- 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



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 (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 his elicitated prior. If he does not agree with these consequences, iterate the above steps until he is happy.

Prior elicitation - AR(p) example

- Autoregressive process or order p

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

- Informative prior on the unconditional mean: $\mu \sim N(\mu_0, \tau_0^2)$. 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 $\phi_i, i = 1, ..., p$ are independent a priori, and $\phi_i \sim N(\mu_i, \psi_i)$
- Prior on $\phi = (\phi_1, ..., \phi_p)$ centered on persistent AR(1) process: $\mu_1 = 0.8, \mu_2 = ... = \mu_p = 0$
- Prior variance of the ϕ_i decay towards zeros: $Var(\phi_i) = \frac{c}{i^\lambda}$, 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.

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, cont.

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

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

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

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

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

where I_θ is the Fisher information.

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?

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.