Bayesian inference

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

April 8, 2024

Outline for today

- A genetic example
- Bayes theorem
- Examples
- Priors
- Posterior summaries

Bayes theorem

Bayes theorem for events A, B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Combines marginal probability for A with conditional probability for B given A to obtain conditional probability of $A|B$.

Bayes theorem for random variables X and Y:

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)} \propto f(y|x)f(x)$$

NB: $c = f(y)$ normalizing constant for unnormalized density

$$h(x) = f(y|x)f(x)$$

Example: forensic statistics

Population of n individuals each with bloodtype a or $\neg a$.

Population: $\{x_1, x_2, \ldots, x_n\}$ where $x_i = (i, t_i)$ and t_i is either a or $\neg a$.

Stochastic variables G and B. $G = i$ means ith person guilty. B is bloodtype of guilty person ($G = i \Rightarrow B = t_i$).

Prior distribution for G: $P(G = i) = p_i$. Suppose we know $B = a$. Then

$$P(G = i | B = a) = \frac{P(B = a | G = i) P(G = i)}{P(B = a)}$$

Note $P(B = a | G = l) = 1$ if $t_l = a$ and zero otherwise. Hence if $t_i = a$,

$$P(G = i | B = a) = \frac{p_i}{\sum_{l: t_l = a} p_l}$$

Note $P(B = a) = \sum_{l: t_l = a} p_l$ in general differs from proportion of population with bloodtype a !

The idea of Bayesian inference

Idea: in order to infer an unknown quantity θ we should combine information in the data with *prior information* (e.g. past experience).

Formal approach: unknown parameter θ is regarded as a *random variable*. Prior information expressed using probability density $p(\theta)$ and information in data quantified using likelihood function.

Inference given data obtained via *posterior* distribution (Bayes theorem)

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{f(y)} \propto f(y|\theta)p(\theta) \propto L(\theta)p(\theta)$$

(as usual factors not depending on θ do not matter)

NB: Bayesian inference mimics our daily approaches to handling uncertainty where we implicitly combine sources of data/likelihoods with prior knowledge.

Example: data: child late for dinner. Probability of interest $P($ accident on the way home $|$ child late$)$. Here we use prior probability $P($ accident$)$ as well as "likelihoods" $P(\text{late}|\text{accident})$, $P(\text{late}|\text{not accident}) = q$. If q big we worry less.

Advantage: *enables* the use of prior information when this is available.

Disadvantage: *requires* the use of prior information. This may be hard to obtain or different persons may have different prior opinions.

Example: beta-binomial

Suppose we observe $X \sim b(n, \theta)$. Use beta prior

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

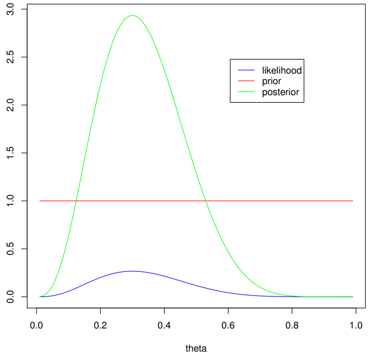
Posterior

$$p(\theta|x) \propto \theta^{x}(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

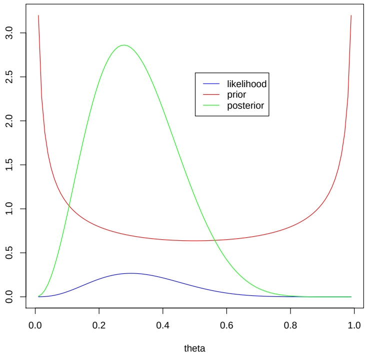
Hence posterior $p(\theta|x)$ is beta-distributed $(\text{Beta}(x+\alpha, n-x+\beta))$ too !

Plots of prior, likelihood and posterior when $X = 3$ and $n = 10$ with different choices of (α, β):

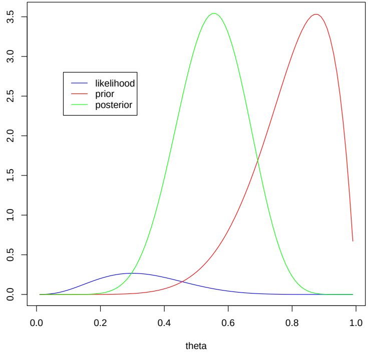
(1,1) (uniform/flat)



(0.5,0.5) (symmetric)

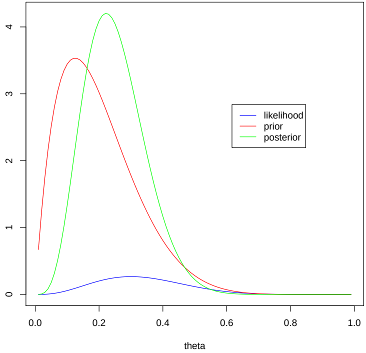


(8,2)



likelihood
prior
posterior
theta
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

(2,8)



likelihood
prior
posterior
theta
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4

Conjugate prior distributions

Beta distribution is an example of a prior which is conjugate for the binomial likelihood: posterior distribution is beta too !

Other examples:

- Gamma is conjugate for Poisson
- normal/scaled inverse χ^2 conjugate for linear normal model

Conjugate priors only available in simple situations.

Poisson-Gamma

Suppose $Y_1, \ldots, Y_n | \lambda$ independent Poisson with mean λ and we choose $\Gamma(\alpha, \beta)$ prior for λ.

Posterior:

$$p(\lambda|y) \propto \lambda^{y.} \exp(-n\lambda)\lambda^{\alpha-1} \exp(-\lambda/\beta) = \lambda^{y.+\alpha-1} \exp(-\lambda/[\beta/(1+n\beta)])$$

Hence posterior for λ is $\Gamma(y. + \alpha, \beta/(1 + n\beta))$.

Expressions for posterior means and variances for binomial-beta and Poisson-gamma can be found in Chapter 6 in M & T.

Linear normal model

$$Y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I).$$

Priors: $\beta|\sigma^2 \sim N(0, \phi I)$ and $\sigma^2 \sim S\chi^{-2}(f)$.

We already know from our treatment of linear mixed models that

$$\beta|\sigma^2, y \sim N\left((\frac{\sigma^2}{\phi} I + X^\mathsf{T} X)^{-1} X^\mathsf{T} Y, \sigma^2 (\frac{\sigma^2}{\phi} I + X^\mathsf{T} X)^{-1} \right) \quad (1)$$

Note this converges to proper limit $N(\hat{\beta}, \sigma^2 (X^\mathsf{T} X)^{-1})$ when $\phi \to \infty$. Note *formal* similarity with frequentist result for MLE $\hat{\beta}$.

We can also show that $\sigma^2|y$ is scaled χ^{-2}, see next slides.

With

$$p(\beta, \sigma^2) \propto (\sigma^2)^{-\frac{f}{2}-1} \exp(-S/(2\sigma^2))$$

and using Pythagoras

$$\|y - X\beta\|^2 = \|y - X\hat{\beta}\|^2 + \|X\hat{\beta} - X\beta\|^2$$

we obtain

$$\begin{aligned} p(\beta, \sigma^2|y) \propto & (\sigma^2)^{-n/2} \mathrm{e}^{-\frac{1}{2\sigma^2}\|y-X\beta\|^2} (\sigma^2)^{-\frac{f}{2}-1} \mathrm{e}^{-\frac{S}{2\sigma^2}} \\ = & \mathrm{e}^{-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^\mathsf{T} X^\mathsf{T} X(\beta-\hat{\beta})} (\sigma^2)^{-\frac{f+n}{2}-1} \mathrm{e}^{-\frac{S+RSS}{2\sigma^2}} \end{aligned}$$

where $RSS = \|y - X\hat{\beta}\|^2$ is the sum of squared residuals.

From this we (again) obtain $\beta|\sigma^2, y \sim N(\hat{\beta}, \sigma^2(X^\mathsf{T} X)^{-1})$

Further,

$$
\begin{aligned}
p(\sigma^2|y) &\propto \int \mathrm{e}^{-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^\mathsf{T} X^\mathsf{T} X(\beta-\hat{\beta})} (\sigma^2)^{-\frac{f+n}{2}-1} \mathrm{e}^{-\frac{S+RSS}{2\sigma^2}} \mathrm{d}\beta \\
&= (2\pi)^{p/2} (\sigma^2)^{p/2} |X^\mathsf{T} X|^{-1/2} (\sigma^2)^{-\frac{f+n}{2}-1} \mathrm{e}^{-\frac{S+RSS}{2\sigma^2}} \\
&\propto (\sigma^2)^{-\frac{f+n-p}{2}-1} \mathrm{e}^{-\frac{S+RSS}{2\sigma^2}}
\end{aligned}
$$

Hence, $\sigma^2|y \sim (RSS + S)\chi^{-2}(f + n - p)$.

Hence provided $RSS > 0$ and $n - p > 0$, posterior also proper with the improper prior $p(\beta, \sigma^2) \propto 1/\sigma^2$ (i.e. $S = f = 0$).

Results with improper prior for β and σ^2

With $R = (\beta - \hat{\beta})/\sigma$ we obtain $R|\sigma^2, y \sim N(0, (X^\mathsf{T}X)^{-1})$. Thus R and σ^2 are conditionally independent given y.

With $s^2 = RSS/(n-p)$ and $p(\beta, \sigma^2) \propto 1/\sigma^2$:

$$\frac{\beta - \hat{\beta}}{\sqrt{s^2}} = R\frac{\sigma}{s} \quad \text{and} \quad R\frac{\sigma}{s}|y \sim N(0, (X^\mathsf{T}X)^{-1})\sqrt{(n-p)\chi^{-2}(n-p)}$$

The product of independent $N(0, (X^\mathsf{T}X)^{-1})$ and $\sqrt{(n-p)\chi^{-2}(n-p)}$ gives a p-dimensional t distribution with $n-p$ degrees of freedom. Thus

$$\frac{\beta - \hat{\beta}}{\sqrt{s^2}}|y \sim t(p, (X^\mathsf{T}X)^{-1}, n-p)$$

With v_i the ith diagonal element of $(X^\mathsf{T}X)^{-1}$ we obtain

$$\frac{\beta_i - \hat{\beta}_i}{\sqrt{v_i s^2}}|y \sim t(n-p)$$

Note again *formal* similarity with frequentist t-statistic !

Improper priors

Priors

$$p(\beta) \propto 1, \quad \beta \in \mathbb{R}^p$$

and

$$p(\sigma^2) \propto 1/\sigma^2, \quad \sigma^2 > 0$$

are improper (do not integrate to one).

In case of normal likelihood posterior is nevertheless proper (limiting cases of normal and χ^{-2} priors).

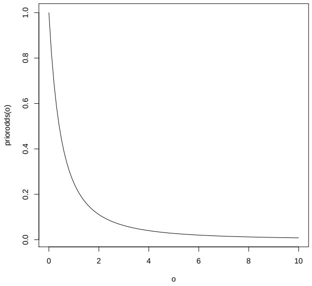
Reason for using improper prior: a) may seem more objective (but this is not really true, see next slide for a cautionary example) b) avoids choosing parameters like ϕ, S, f in the normal example.

Danger: in complex models it may be hard to check that a posterior is proper when improper priors are used.

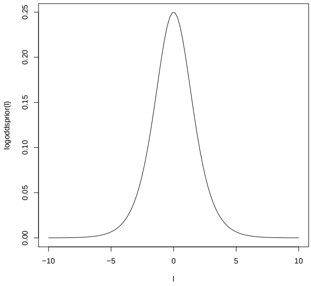
'Non-informative' and priors

Consider flat prior for $\theta \in [0, 1]$. Priors for odds and log odds not flat !:

odds



log odds



Hence whether a prior is non-informative depends on scale.

Rule of thumb: use non-informative priors on the scale that we wish to draw inference for.

Priors for odds and log odds obtained using transformation theorem:

Suppose $X \sim f_X$ and $Y = h(X)$ for differentiable and injective function h. Then density of Y is

$$f_Y(y) = \frac{1}{|\mathrm{d}y/\mathrm{d}x|} f_X(x) \quad \text{where } x = h^{-1}(y)$$

Also valid in the multivariate case. Then $|\cdot|$ is determinant and

$$\frac{\mathrm{d}y}{\mathrm{d}x} = [\frac{\mathrm{d}y_i}{\mathrm{d}x_j}]_{ij}$$

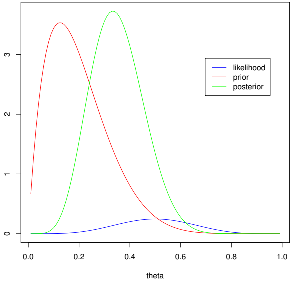
is Jacobian matrix of partial derivatives.

Large data sets

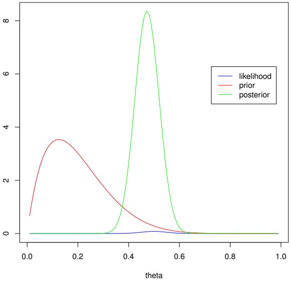
With large datasets, posterior results less sensitive to choice of prior (likelihood dominates).

Example beta-binomial with $x = 5, n = 10$ and $x = 50, n = 100$ (in both cases MLE is 0.5):

$L(0.5)/L(0.1) = 165.4$



$L(0.5)/L(0.1) = 1.53e22$!!



Note: likelihoods look small compared to prior and posterior because not normalized to integrate to one !

Summarizing the posterior

For a vector $(\theta_1, \ldots, \theta_n)$ posterior summaries are often computed for the components separately.

Hence for θ_i we may compute posterior mean or median and express posterior uncertainty in terms of posterior variance (not so useful if posterior far from normal).

Posterior 95 % credibility interval: interval $[l, u]$ (depending on *data*) such that $P(\theta_i \in [l, u]|y) = 95\%$. Often a *central* interval is used: $P(\theta_i < u|y) = P(\theta_i > l|y) = 0.025$.

95% Highest posterior density (HPD) region : H chosen so that $P(\theta \in H|y) = 0.95$ and $p(\theta|y) > p(\tilde{\theta}|y)$ whenever θ inside H and $\tilde{\theta}$ outside.

More sophisticated possibilities: e.g. posterior probability that $\theta_1 > \theta_2$ or look at ranks for components of θ (e.g. which treatment is best ?).

Confidence intervals versus posterior intervals

95% confidence interval: random interval which in 95% of future hypothetical repetitions of the experiment would contain the (fixed) unknown parameter θ (frequentist interpretation).

95% posterior interval: Given the data y the posterior interval is fixed while θ is random. The 95% probability associated with the posterior interval is the probability that θ is in the interval given the data. No reference to hypothetical repetitions of experiment.

Exercises

1. Consider m *iid* binomial observations $X_i \sim b(n_i, \theta)$ where θ is the common probability parameter. Compute the posterior distribution of θ when a beta prior is used for θ.
2. Suppose $y|\lambda$ is Poisson(λ) and λ is $\Gamma(\alpha, \beta)$. Show that y marginally has a negative binomial distribution.
3. Compute the prior for p when $\text{logit}(p) = \log(p/(1-p)) = \beta$ and the prior for β is $N(0, \tau^2)$. What happens if $\tau^2 \to \infty$ (try to plot the prior for large τ^2) ?
4. Consider the linear normal model $Y_i \sim N(\beta, \sigma^2)$ (i.e. the design matrix X is a column of 1's) and use the prior $p(\beta, \sigma^2) \propto 1/\sigma^2$.
    4.1 Compute a 95% posterior credibility interval for β.
    4.2 Compare with the frequentist 95% confidence interval. What are the interpretations of the two intervals and how do the interpretations differ ?

5. Suppose observations $4, 6, 6, 7, 3, 5, 3, 11, 10, 5$ are observations of *iid* Poisson random variables with mean λ. Use a Gamma prior with mean 6 and variance 10. Compute the posterior mean, variance, and 95% central posterior interval for λ.
6. Verify (1) using results from prediction lecture (slide Prediction in linear mixed model).

A few results needed for the exercises

The density of $\Gamma(\alpha, \beta)$ with shape α and scale β is

$$f(x; \alpha, \beta) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta), \quad x > 0$$

where $\Gamma(\cdot)$ is the gamma function. Mean and variance are $\alpha\beta$ and $\alpha\beta^2$. If β is interpreted as the rate (inverse scale) then

$$f(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \quad x > 0$$

The density of a negative binomial distribution with parameters α and β is

$$f(y) = \frac{\Gamma(y + \alpha)}{y!\Gamma(\alpha)} \left(\frac{1}{1+\beta}\right)^{\alpha} \left(\frac{\beta}{1+\beta}\right)^{y} \quad y = 0, 1, 2, \ldots$$