Bayesian Ingredients

Hedibert Freitas Lopes
Professor of Statistics
hedibert.org

Insper

Outline

Example *i*. Normal model and normal prior

Let us now consider a simple measurement error model with normal prior for the unobserved measurement.

$$\begin{aligned} X|\theta &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\theta_0, \tau_0^2) \end{aligned}$$

with σ^2, θ_0 and τ_0^2 known for now. It is easy to show that the posterior of θ given $X = x$ is also normal.

More precisely, $(\theta|X = x) \sim N(\theta_1, \tau_1^2)$ where

$$\begin{aligned} \theta_1 &= w\theta_0 + (1-w)x \\ {\tau_1}^{-2} &= {\tau_0}^{-2} + \sigma^{-2} \\ w &= {\tau_0}^{-2}/({\tau_0}^{-2} + \sigma^{-2}) \end{aligned}$$

w measures the relative information contained in the prior distribution with respect to the total information (prior plus likelihood).

Example from Box & Tiao (1973)

Prior A: Physicist A (large experience): $\theta \sim N(900, (20)^2)$

Prior B: Physicist B (not so experienced): $\theta \sim N(800, (80)^2)$.

Model: $(X|\theta) \sim N(\theta, (40)^2)$.

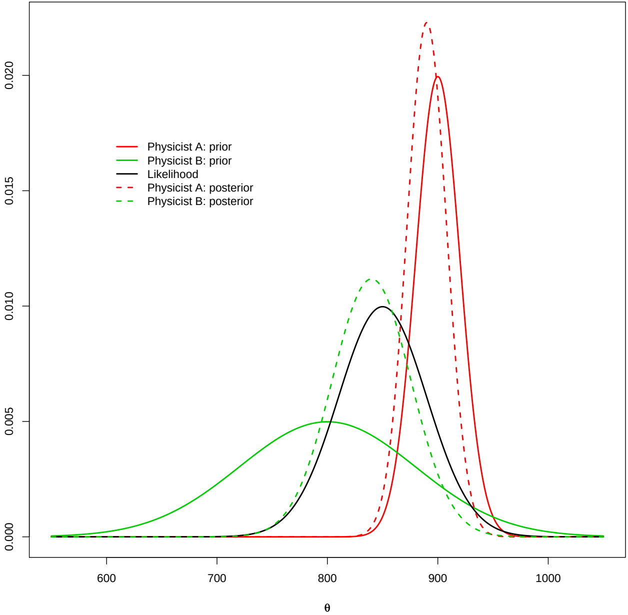
Observation: $X = 850$

$$\begin{aligned} (\theta|X = 850, H_A) &\sim N(890, (17.9)^2) \\ (\theta|X = 850, H_B) &\sim N(840, (35.7)^2) \end{aligned}$$

Information (precision)

Physicist A: from 0.002500 to 0.003120 (an increase of 25%)

Physicist B: from 0.000156 to 0.000781 (an increase of 400%)


Physicist A: prior
Physicist B: prior
Likelihood
Physicist A: posterior
Physicist B: posterior
0.000
0.005
0.010
0.015
0.020
600
700
800
900
1000
θ

Two additional examples

Observations $\boldsymbol{x}$, parameters $\boldsymbol{\theta}$ and history H.

Likelihood functions/models - $p(\boldsymbol{x}|\boldsymbol{\theta}, H)$

$$\begin{aligned} \text{Example ii} &: \quad x_i|\theta, H \sim N(\theta z_i; \sigma^2) \qquad i = 1, \ldots, n \\ \text{Example iii} &: \quad x_t|\theta, H \sim N(0; e^{\theta_t}) \qquad t = 1, \ldots, T \end{aligned}$$

Prior distributions - $p(\boldsymbol{\theta}|H)$

$$\begin{aligned} \text{Example ii} &: \quad \theta|H \sim N(\theta_0, \tau_0^2) \\ \text{Example iii} &: \quad \theta_t|H \sim N(\alpha + \beta\theta_{t-1}, \sigma^2) \qquad t = 1, \ldots, T \end{aligned}$$

Assume, for now, that $(z_1, \ldots, z_n, \sigma^2, \nu_0, \theta_0, \tau_0^2)$ and $(\alpha, \beta, \sigma^2, \theta_0)$ are known and belong to H.

Turning the Bayesian crank

Posterior (Bayes' Theorem):

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}, H) &= \frac{p(\boldsymbol{\theta}, \boldsymbol{x}|H)}{p(\boldsymbol{x}|H)} \\ &= \frac{p(\boldsymbol{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H)}{p(\boldsymbol{x}|H)} \\ &\propto p(\boldsymbol{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H) \end{aligned}$$

Prior predictive distribution:

$$p(\boldsymbol{x}|H) = \int_{\Theta} p(\boldsymbol{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H)\, d\boldsymbol{\theta} = E_{\boldsymbol{\theta}}[p(\boldsymbol{x}|\boldsymbol{\theta}, H)]$$

Posterior predictive distribution

Let $\boldsymbol{y}$ be a new set of observations conditionally independent of $\boldsymbol{x}$ given $\boldsymbol{\theta}$. Then,

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x},H) &= \int_\Theta p(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x},H)d\boldsymbol{\theta} = \int_\Theta p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x},H)p(\boldsymbol{\theta}|\boldsymbol{x},H)d\boldsymbol{\theta} \\ &= \int_\Theta p(\boldsymbol{y}|\boldsymbol{\theta},H)p(\boldsymbol{\theta}|\boldsymbol{x},H)d\boldsymbol{\theta} = E_{\boldsymbol{\theta}|\boldsymbol{x}}\left[p(\boldsymbol{y}|\boldsymbol{\theta},H)\right] \end{aligned}$$

Note 1: In general, but not always (time series, for example) $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent given $\boldsymbol{\theta}$.

Note 2: It might be more useful to concentrate on prediction rather than on estimation since the former is *verifiable*. In other words, $\boldsymbol{x}$ and $\boldsymbol{y}$ can be observed; not $\boldsymbol{\theta}$.

Sequential Bayes

Experimental result: $\boldsymbol{x}_1 \sim p_1(\boldsymbol{x}_1 \mid \boldsymbol{\theta})$

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}_1) \propto l_1(\boldsymbol{\theta}; x_1) p(\boldsymbol{\theta})$$

Experimental result: $\boldsymbol{x}_2 \sim p_2(\boldsymbol{x}_2 \mid \boldsymbol{\theta})$

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}_2, \boldsymbol{x}_1) &\propto l_2(\boldsymbol{\theta}; \boldsymbol{x}_2) p(\boldsymbol{\theta} \mid \boldsymbol{x}_1) \\ &\propto l_2(\boldsymbol{\theta}; \boldsymbol{x}_2) l_1(\boldsymbol{\theta}; \boldsymbol{x}_1) p(\boldsymbol{\theta}) \end{aligned}$$

Experimental results: $\boldsymbol{x}_i \sim p_i(\boldsymbol{x}_i \mid \boldsymbol{\theta})$, for $i = 3, \cdots, n$

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}_n, \cdots, \boldsymbol{x}_1) &\propto l_n(\boldsymbol{\theta}; \boldsymbol{x}_n) p(\boldsymbol{\theta} \mid \boldsymbol{x}_{n-1}, \cdots, \boldsymbol{x}_1) \\ &\propto \left[\prod_{i=1}^{n} l_i(\boldsymbol{\theta}; \boldsymbol{x}_i) \right] p(\boldsymbol{\theta}) \end{aligned}$$

Example ii. Simple linear regression

Model, prior and posterior:

$$\begin{aligned} x_i|\theta, H &\sim N(\theta z_i; \sigma^2) \qquad i = 1, \ldots, n \\ \theta|H &\sim N(\theta_0, \tau_0^2) \\ \theta|\boldsymbol{x}, H &\sim N(\theta_1, \tau_1^2) \end{aligned}$$

where

$$\tau_1^{-2} = \tau_0^{-2} + \boldsymbol{z}'\boldsymbol{z}/\sigma^2 \quad \text{and} \quad \theta_1 = \tau_1^2 \left(\theta_0 \tau_0^{-2} + \boldsymbol{z}'\boldsymbol{x}/\sigma^2\right)$$

Note 1: As n increases, $\tau_1 \to \sigma^2(\boldsymbol{z}'\boldsymbol{z})^{-1}$ and $\theta_1 \to (\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{x}$.

Note 2: The same applies when $\tau_0^{-2} \to 0$, i.e. with 'little' prior knowledge about θ.

Example iii. SV model

Model, prior and posterior:

$$\begin{aligned} x_t|\theta_t, H &\sim N(0; e^{\theta_t}) \qquad t = 1, \ldots, T \\ \theta_t|H &\sim N(\alpha + \beta\theta_{t-1}, \sigma^2) \qquad t = 1, \ldots, T \\ p(\boldsymbol{\theta}|\mathbf{x}, H) &\propto \prod_{t=1}^{T} e^{-\theta_t/2} \exp\left\{-\frac{1}{2} x_t e^{-\theta_t}\right\} \\ &\times \prod_{t=1}^{T} \exp\left\{-\frac{1}{2\sigma^2}(\theta_t - \alpha - \beta\theta_{t-1})^2\right\} \end{aligned}$$

Unfortunately, closed form solutions are rare.

- How to compute $E(\theta_{43}|\mathbf{x}, H)$ or $V(\theta_{11}|\mathbf{x}, H)$?
- How to obtain a 95% credible region for $(\theta_{35}, \theta_{36}|\mathbf{x}, H)$?
- How to sample from $p(\theta|\mathbf{x}, H)$?
- How to compute $p(\mathbf{x}|H)$ or $p(x_{T+1}, \ldots, x_{T+k}|\mathbf{x}, H)$?

Example iv. Multiple linear regression

The standard Bayesian approach to multiple linear regression is

$$(y|X,\beta,\sigma^2) \sim N(X\beta,\sigma^2 I_n)$$

where $y=(y_1,\ldots,y_n)$, $X=(x_1,\ldots,x_n)'$ is the $(n\times q)$, design matrix and $q=p+1$.

The prior distribution of (β,σ^2) is $NIG(b_0,B_0,n_0,S_0)$, i.e.

$$\begin{aligned}\beta|\sigma^2 &\sim N(b_0,\sigma^2 B_0)\\ \sigma^2 &\sim IG(n_0/2,n_0S_0/2)\end{aligned}$$

for known hyperparameters b_0, B_0, n_0 and S_0.

Example iv. Conditionals and marginals

It is easy to show that (β, σ^2) is $NIG(b_1, B_1, n_1, S_1)$, i.e.

$$\begin{aligned} (\beta|\sigma^2, y, X) &\sim N(b_1, \sigma^2 B_1) \\ (\sigma^2|y, X) &\sim IG(n_1/2, n_1 S_1/2) \end{aligned}$$

where

$$\begin{aligned} B_1^{-1} &= B_0^{-1} + X'X \\ B_1^{-1} b_1 &= B_0^{-1} b_0 + X'y \\ n_1 &= n_0 + n \\ n_1 S_1 &= n_0 S_0 + (y - Xb_1)'y + (b_0 - b_1)' B_0^{-1} b_0. \end{aligned}$$

It is also easy to derive the full conditional distributions, i.e.

$$\begin{aligned} (\beta|y, X) &\sim t_{n_1}(b_1, S_1 B_1) \\ (\sigma^2|\beta, y, X) &\sim IG(n_1/2, n_1 S_{11}/2) \end{aligned}$$

where

$$n_1 S_{11} = n_0 S_0 + (y - X\beta)'(y - X\beta).$$

Example iv. Ordinary least squares

It is well known that

$$
\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \\
\hat{\sigma}^2 &= \frac{S_e}{n-q} = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n-q}
\end{aligned}
$$

are the OLS estimates of β and σ^2, respectively.

The conditional and unconditional sampling distributions of $\hat{\beta}$ are

$$
\begin{aligned}
(\hat{\beta}|\sigma^2, y, X) &\sim N(\beta, \sigma^2(X'X)^{-1}) \\
(\hat{\beta}|y, X) &\sim t_{n-q}(\beta, S_e)
\end{aligned}
$$

respectively, with

$$
(\hat{\sigma}^2|\sigma^2) \sim IG\left((n-q)/2, ((n-q)\sigma^2/2\right).
$$

Example iv. Sufficient statistics

Recall $(y_t|x_t,\beta,\sigma^2) \sim N(x_t'\beta,\sigma^2)$ for $t=1,\ldots,n$, with prior $\beta|\sigma^2 \sim N(b_0,\sigma^2 B_0)$ and $\sigma^2 \sim IG(n_0/2, n_0S_0/2)$.

Then, for $y^t=(y_1,\ldots,y_t)$ and $X^t=(x_1,\ldots,x_t)'$, it follows:

$$\begin{aligned}(\beta|\sigma^2,y^t,X^t) &\sim N(b_t,\sigma^2 B_t)\\ (\sigma^2|y^t,X^t) &\sim IG(n_t/2, n_tS_t/2)\end{aligned}$$

where $n_t = n_{t-1}+1$, $B_t^{-1} = B_{t-1}^{-1} + x_t x_t'$, $B_t^{-1}b_t = B_{t-1}^{-1}b_{t-1} + y_t x_t$ and $n_tS_t = n_{t-1}S_{t-1} + (y_t - b_t'x_t)y_t + (b_{t-1}-b_t)'B_{t-1}^{-1}b_{t-1}$.

The only ingredients needed are: $x_t x_t'$, $y_t x_t$ and y_t^2.

These recursions will play an important role later on when deriving **sequential Monte Carlo** methods for conditionally Gaussian dynamic linear models, like many stochastic volatility models.

Example iv. Predictive

The predictive density can be seen as the *marginal likelihood*, i.e.

$$p(y|X) = \int p(y|X,\beta,\sigma^2)p(\beta|\sigma^2)p(\sigma^2)d\beta d\sigma^2$$

or, by Bayes' theorem, as the *normalizing constant*, i.e.

$$p(y|X) = \frac{p(y|X,\beta,\sigma^2)p(\beta|\sigma^2)p(\sigma^2)}{p(\beta|\sigma^2,y,X)p(\sigma^2|y,X)}$$

which is valid for all (β, σ^2).

Closed form solution is available for the multiple normal linear regression:

$$(y|X) \sim t_{n_0}(Xb_0, S_0(I_n + XB_0X')).$$

Bayesian model criticism

Suppose that the competing models can be enumerated and are represented by the set $M = \{M_1, M_2, \ldots\}$, and that the *true model* is in M (Bernardo and Smith, 1994).

The posterior model probability of model M_j is given by

$$Pr(M_j|y) \propto f(y|M_j)Pr(M_j)$$

where

$$f(y|M_j) = \int f(y|\theta_j, M_j)p(\theta_j|M_j)d\theta_j$$

is the prior predictive density of model M_j and $Pr(M_j)$ is the prior model probability of model M_j.

Posterior odds

The posterior odds of model M_j relative to M_k is given by

$$\underbrace{\frac{Pr(M_j|y)}{Pr(M_k|y)}}_{\text{posterior odds}} = \underbrace{\frac{Pr(M_j)}{Pr(M_k)}}_{\text{prior odds}} \times \underbrace{\frac{f(y|M_j)}{f(y|M_k)}}_{\text{Bayes factor}} .$$

The Bayes factor can be viewed as the weighted likelihood ratio of M_j to M_k.

The main difficulty is the computation of the marginal likelihood or normalizing constant $f(y|M_j)$.

Therefore, the posterior model probability for model j can be obtained from

$$\frac{1}{Pr(M_j|y)} = \sum_{M_k \in M} B_{kj} \frac{Pr(M_k)}{Pr(M_j)}.$$

Bayes factor

Jeffreys (1961) recommends the use of the following rule of thumb to decide between models j and k:

$\log_{10} B_{jk}$	B_{jk}	Evidence against k
0.0 to 0.5	1.0 to 3.2	Not worth more than a bare mention
0.5 to 1.0	3.2 to 10	Substantial
1.0 to 2.0	10 to 100	Strong
> 2	> 100	Decisive

Kass and Raftery (1995) argue that "it can be useful to consider twice the natural logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood ratio test statistics". Their slight modification is:

$2 \log_e B_{jk}$	B_{jk}	Evidence against k
0.0 to 2.0	1.0 to 3.0	Not worth more than a bare mention
2.0 to 6.0	3.0 to 20	Substantial
6.0 to 10.0	20 to 150	Strong
> 10	> 150	Decisive

Bayesian Model Averaging

See Hoeting, Madigan, Raftery and Volinsky (1999), *Statistical Science*, 14, 382-401.

Let $\mathcal{M}$ denote the set that indexes all entertained models.

Assume that Δ is an outcome of interest, such as the future value y_{t+k}, or an elasticity well defined across models, etc. The posterior distribution for Δ is

$$p(\Delta|y) = \sum_{m \in \mathcal{M}} p(\Delta|m, y) Pr(m|y)$$

for data y and posterior model probability

$$Pr(m|y) = \frac{p(y|m)Pr(m)}{p(y)}$$

where $Pr(m)$ is the prior probability model.

Posterior predictive criterion

Gelfand and Ghosh (1998) introduced a posterior predictive criterion that, under squared error loss, favors the model M_j which minimizes

$$D_j^G = P_j^G + G_j^G$$

where

$$P_j^G = \sum_{t=1}^{n} V(\tilde{y}_t|y, M_j)$$

$$G_j^G = \sum_{t=1}^{n} [y_t - E(\tilde{y}_t|y, M_j)]^2$$

and $(\tilde{y}_1, \ldots, \tilde{y}_n)$ are predictions/replicates of y.

The first term, P_j, is a penalty term for model complexity.

The second term, G_j, accounts for goodness of fit.

Deviance Information Criterion

See Spiegelhalter, Best, Carlin and van der Linde (2002), *JRSS-B*, 64, 583-616.

If $\theta^* = E(\theta|y)$ and $D(\theta) = -2\log p(y|\theta)$ is the deviance, then the DIC generalizes the AIC

$$\begin{aligned} DIC &= \bar{D} + p_D \\ &= \text{goodness of fit} + \text{model complexity} \end{aligned}$$

where $\bar{D} = E_{\theta|y}(D(\theta))$ and $p_D = \bar{D} - D(\theta^*)$.

The p_D is the *effective number of parameters*.

Small values of DIC suggests a better-fitting model.