7. Estimation and hypothesis testing

Objective

In this chapter, we show how the election of estimators can be represented as a decision problem. Secondly, we consider the problem of hypothesis testing from a Bayesian viewpoint and illustrate the similarities and differences between Bayesian and classical procedures.

Recommended reading

- Kass, R. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Point estimation

Assume that θ is univariate. For Bayesians, the problem of estimation of θ is a decision problem. Associated with an estimator, T, of θ, there is a loss function $L(T, \theta)$ which reflects the difference between the value of T and θ. Then, for an expert, E, with distribution $p(\theta)$, the Bayes estimator of θ is that which minimizes the expected loss

$$E[L(T, \theta)] = \int L(T, \theta) p(\theta) \, d\theta.$$

Definition 18

The *Bayes estimator*, B, of θ is such that

$$E[L(B, \theta)] \leq E[L(T, \theta)] \quad \text{for any } T \neq B.$$

Bayes estimators for some given loss functions

The mean, median and mode of E's distribution for θ can be justified as estimators given certain specific loss functions.

Theorem 29

Suppose that E's density of θ is $p(\theta)$. Then

1. If $L(T,\theta) = (T-\theta)^2$ then B is E's mean, $B = E[\theta]$.

2. If $L(T,\theta) = |T-\theta|$, then B is E's median for θ.

3. If Θ is discrete and $L(T,\theta) = \begin{cases} 0 & \text{if } T = \theta \\ 1 & \text{if } T \neq \theta \end{cases}$ then T is E's modal estimator for θ.

Proof

1.

$$\begin{aligned} E[L(T,\theta)] &= \int (T-\theta)^2 p(\theta)\, d\theta \\ &= \int (T - E[\theta] + E[\theta] - \theta)^2\, p(\theta)\, d\theta \\ &= (T - E[\theta])^2 + V[\theta] \end{aligned}$$

which is minimized when $T = E[\theta] = B$ is the Bayes estimator.

2. Exercise.

3. Exercise.

In the case that Θ is continuous, then if we consider the loss function

$$L(T, \theta) = \begin{cases} 0 & \text{if } |T - \theta| < \epsilon \\ 1 & \text{otherwise} \end{cases},$$

then it is easy to see that T is the centre of the modal interval of width ϵ and letting $\epsilon \to 0$, T approaches the mode.

Interval estimation

A expert's 95% credible interval for a variable is simply an interval for which the expert, E, has a 95% probability We have previously used such intervals in the earlier chapters.

Definition 19

If $p(\theta)$ is E's density for θ, we say that (a, b) is a $100(1-\alpha)$**% credible interval** for θ if

$$P(a \leq \theta \leq b|\mathbf{x}) = \int_a^b p(\theta)\, d\theta = 1 - \alpha.$$

It is clear that in general, E will have (infinitely) many credible intervals for θ. The shortest possible credible interval is called a *highest posterior density* (HPD) interval.

Definition 20
The $100 \times (1-\alpha)\%$ HPD interval is an interval of form

$$C = \{\theta : f(\theta) \geq c(\alpha)\}$$

where $c(\alpha)$ is the largest number such that $P(C) \geq 1-\alpha$.

Example 53
$X|\mu \sim \mathcal{N}(\mu, 1)$. Let $f(\mu) \propto 1$. Therefore, $\mu|\mathbf{x} \sim \mathcal{N}(\bar{x}, 1/n)$ and some 95% posterior credible intervals are

$$(-\infty, \bar{x} + 1.64/\sqrt{n}) \quad \text{or} \quad (\bar{x} - 1.64/\sqrt{n}, \infty) \quad \text{or} \quad (\bar{x} \pm 1.96/\sqrt{n})$$

which is the HPD interval.

We can generalize the definition of a credible interval to multivariate densities $p(\boldsymbol{\theta})$. In this case, we can define a credible region $\mathbf{C}$:

$$P(\boldsymbol{\theta} \in \mathbf{C}) = 1 - \alpha.$$

Hypothesis testing

Assume now that we wish to test the hypothesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus the alternative $H_1 : \boldsymbol{\theta} \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \phi$. In theory, this is straightforward. Given a sample of data, $\mathbf{x}$, we can calculate the posterior probabilities $P(H_0|\mathbf{x})$ and $P(H_1|\mathbf{x})$, and under a suitable loss function, we can decide to accept or reject the null hypothesis H_0.

Example 54

Given the *all or nothing* loss,

$$L(H_0, \theta) = \begin{cases} 0 & \text{if } H_0 \text{ is true} \\ 1 & \text{if } H_1 \text{ is true} \end{cases}$$

we accept H_0 if $P(H_0|\mathbf{x}) > P(H_1|\mathbf{x})$.

For point null and alternative hypotheses or for one tailed tests, then Bayesian and classical solutions are often similar.

Example 55

Let $X|\theta \sim N(\theta, 1)$. We wish to test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$. Given the uniform prior, $p(\theta) \propto 1$, we know that $\theta|\mathbf{x} \sim \mathcal{N}\left(\bar{x}, \frac{1}{n}\right)$. Therefore,

$$\begin{aligned} P(H_0|\mathbf{x}) &= P(\theta \leq 0|\mathbf{x}) \\ &= \Phi\left(-\sqrt{n}\bar{x}\right) \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cdf.

This probability is equal to the usual, classical p-value for the test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$.

$$\begin{aligned} P(\bar{X} \geq \bar{x}|H_0) &= P\left(\sqrt{n}\bar{X} \geq \sqrt{n}\bar{x}|H_0\right) \\ &= 1 - \Phi\left(\sqrt{n}\bar{x}\right) = \Phi\left(-\sqrt{n}\bar{x}\right) \end{aligned}$$

The Lindley/Jeffreys paradox

For two-tailed tests, Bayesian and classical results can be very different.

Example 56

Let $X|\theta \sim \mathcal{N}(\theta, 1)$ and suppose that we wish to test the hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta \neq 0$.

Assume first that

$$p_0 = P(H_0) = 0.5 = P(H_1) = p_1$$

with a normal prior distribution,

$$\theta|H_1 \sim \mathcal{N}(0, 1)$$

and suppose that we observe the mean $\bar{x}$ of a sample of n data with likelihood

$$l(\theta|\bar{x}) = \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n}{2}(\theta - \bar{x})^2\right).$$

Then, we can calculate the posterior probability, $\alpha_0 = P(H_0|\mathbf{x})$.

$$\begin{aligned}
\alpha_0 &= P(H_0|\bar{x}) \\
&= \frac{p_0 l(\theta = 0|\bar{x})}{p_0 l(\theta = 0|\bar{x}) + p_1 l(\theta|\bar{x})} \\
&= \frac{p_0 l(\theta = 0|\mathbf{x})}{p_0 l(\theta = 0|\mathbf{x}) + p_1 \int l(\theta|\bar{x}) p(\theta|H_1)\, d\theta} \\
&= \frac{l(\theta = 0|\mathbf{x})}{l(\theta = 0|\bar{x}) + \int l(\theta|\bar{x}) p(\theta|H_1)\, d\theta} \\
&= K \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right)
\end{aligned}$$

for a constant

$$K^{-1} = l(\theta = 0|\bar{x}) + \int l(\theta|\bar{x}) p(\theta|H_1)\, d\theta.$$

We can also evaluate the second term in the denominator.

$$\begin{aligned} l(\theta|\bar{x})p(\theta|H_1) &= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right) \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\theta^2\right) \\ &= \frac{\sqrt{n}}{2\pi} \exp\left(-\frac{1}{2}\left[n(\bar{x}-\theta)^2+\theta^2\right]\right) \\ &= \frac{\sqrt{n}}{2\pi} \exp\left(-\frac{1}{2}\left[(n+1)\left(\theta-\frac{n\bar{x}}{n+1}\right)^2+\frac{n\bar{x}^2}{n+1}\right]\right) \end{aligned}$$

and, integrating, we have,

$$\int l(\theta|\bar{x})p(\theta|H_1)d\theta = \left(\frac{n}{(n+1)2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{n\bar{x}^2}{n+1}\right).$$

Therefore, we have

$$\begin{aligned} K^{-1} &= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right) + \left(\frac{n}{(n+1)2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{n\bar{x}^2}{n+1}\right) \\ &= \left(\frac{n}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{n\bar{x}^2}{2}\right) \left(1 + \frac{1}{\sqrt{n+1}} \exp\left(\frac{-n^2\bar{x}^2}{2(n+1)}\right)\right) \end{aligned}$$

so the posterior probability of H_0 is

$$\alpha_0 = \left\{1 + \frac{1}{\sqrt{n+1}} \exp\left(\frac{-n^2\bar{x}^2}{2(n+1)}\right)\right\}^{-1}.$$

Now suppose that $\bar{x} = 2/\sqrt{n} > 1.96/\sqrt{n}$.

Therefore, for the classical test of H_0 against H_1, with a fixed significance level of 95%, the result is significant and we reject the null hypothesis H_0. However, in this case, the posterior probability is

$$\begin{aligned} \alpha_0 &= \left\{1 + \frac{1}{\sqrt{n+1}} \exp\left(-\frac{2n}{n+1}\right)\right\}^{-1} \\ &\to 1 \quad \text{when } n \to \infty. \end{aligned}$$

Using Bayesian methods, a sample that leads us to reject H_0 using a classical test gives a posterior probability of H_0 that approaches 1 as $n \to \infty$. This is called the Lindley / Jeffreys paradox. See Lindley (1957).

Comments

The choice of the prior variance of θ is quite important but the example also illustrates that it is not very sensible, from a classical viewpoint, to use fixed significance levels as n increases. (In practice, smaller values of α are virtually always used when n increases so that the power of the test is maintained).

Also, point null hypotheses do not appear to be very sensible from a Bayesian viewpoint. The Lindley / Jeffreys paradox can be avoided by considering an interval $(-\epsilon, \epsilon)$ and considering a prior distribution over this interval, see e.g. Berger and Delampady (1987). The main practical difficulty then lies in the choice of ϵ.

Bayes factors

Good

The original idea for the Bayes factor stems from Good (1958) who attributes it to Turing. The motivation is to find a more objective measure than simple posterior probability.

Motivation

An important problem in many contexts such as regression modeling is that of selecting a model $\mathcal{M}_i$ from a given class $\mathcal{M} = \{\mathcal{M}_i : \ i = 1, \ldots, k\}$. In this case, given the a prior distribution $P(\cdot)$ over the model space, we have

$$P(\mathcal{M}_i|\mathbf{x}) = \frac{P(\mathcal{M}_i)f(\mathbf{x}|\mathcal{M}_i)}{\sum_{j=1}^{k} P(\mathcal{M}_j)f(\mathbf{x}|\mathcal{M}_j)},$$

and given the all or nothing loss function, we should select the most probable model.

However, the prior model probabilities are strongly influential in this choice and, if the class of possible models $\mathcal{M}$ is large , it is often difficult or impossible to precisely specify the model probabilities $P(\mathcal{M}_i)$. Thus, it is necessary to introduce another concept which is less strongly dependent on the prior information.

Consider two hypotheses (or models) H_0 and H_1 and let $p_0 = P(H_0)$, $p_1 = P(H_1)$ and $\alpha_0 = P(H_0|\mathbf{x})$ and $\alpha_1 = P(H_1|\mathbf{x})$. Then Jeffreys (1961) defines the Bayes factor comparing the two hypotheses as follows.

Definition 21

The *Bayes factor* in favour of H_0 is defined to be

$$B_1^0 = \frac{\alpha_0/\alpha_1}{p_0/p_1} = \frac{\alpha_0 p_1}{\alpha_1 p_0}.$$

The Bayes factor is simply the posterior odds in favour of H_0 divided by the prior odds. It tells us about the changes in our relative beliefs about the two models caused by the data. The Bayes factor is almost an objective measure and partially eliminates the influence of the prior distribution in that it is *independent* of p_0 and p_1.

Proof

$$\begin{aligned}
\alpha_0 &= P(H_0|\mathbf{x}) = \frac{p_0 f(\mathbf{x}|H_0)}{p_0 f(\mathbf{x}|H_0) + p_1 f(\mathbf{x}|H_1)} \\
\alpha_1 &= \frac{p_1 f(\mathbf{x}|H_1)}{p_0 f(\mathbf{x}|H_0) + p_1 f(\mathbf{x}|H_1)} \\
\frac{\alpha_0}{\alpha_1} &= \frac{p_0 f(\mathbf{x}|H_0)}{p_1 f(\mathbf{x}|H_1)} \\
B_1^0 &= \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)}
\end{aligned}$$

which is independent of p_0 and p_1. ∎

The Bayes factor is strongly related to the likelihood ratio, often used in classical statistics for model choice. If H_0 and H_1 are point null hypotheses, then the Bayes factor is exactly equal to the likelihood ratio.

Example 57

Suppose that we wish to text $H_0 : \lambda = 6$ versus $H_1 : \lambda = 3$ and that we observe a sample of size n from an exponential density with rate λ,

$$f(x|\lambda) = \lambda e^{-\lambda x}.$$

Then the Bayes factor in favour of H_0 in this case is

$$\begin{aligned} B_1^0 &= \frac{l(\lambda = 6|\mathbf{x})}{l(\lambda = 3|\mathbf{x})} \\ &= \frac{6^n e^{-6n\bar{x}}}{3^n e^{-3n\bar{x}}} \\ &= 2^n e^{-3n\bar{x}}. \end{aligned}$$

For composite hypotheses however, the Bayes factor depends on the prior parameter distributions. For example, in the case that H_0 is composite, then the marginal likelihood is

$$f(\mathbf{x}|H_0) = \int f(\mathbf{x}|\boldsymbol{\theta}, H_0) p(\boldsymbol{\theta}|H_0)\, d\theta$$

which is dependent on the prior parameter distribution, $p(\boldsymbol{\theta}|H_0)$.

Consistency of the Bayes factor

We can see that the Bayes factor always takes values between 0 and infinity. Furthermore, it is obvious that $B_1^0 = \infty$ if $P(H_0|\mathbf{x}) = 1$ and $B_1^0 = 0$ if $P(H_0|\mathbf{x}) = 0$.

Thus, the Bayes factor is consistent so that if H_0 is true, then $B_1^0 \to \infty$ when $n \to \infty$ and if H_1 is true, then $B_1^0 \to 0$ when $n \to \infty$.

Returning to Example 57, it can be seen that if $\lambda = 6$, then when $n \to \infty$, we have

$$B_1^0 \to 2^n e^{-n/2} \to \infty.$$

Bayes factors and scales of evidence

We can interpret the statistic $2 \log B_1^0$ as a Bayesian version of the classical log likelihood statistic. Kass and Raftery (1995) suggest using the following table of values of B_1^0 to represent the evidence against H_1.

$2 \log_{10} B_1^0$	B_1^0	Evidence against H_1.
0 a 2	1 a 3	Hardly worth commenting
2 a 6	3 a 20	Positive
6 a 10	20 a 150	Strong
> 10	> 150	Very strong

Jeffreys (1961) suggests a (similar) alternative scale of evidence.

Relation of the Bayes factor to standard model selection criteria

The Schwarz (1978) criterion or Bayesian information criterion for evaluating a model, $\mathcal{M}$ is

$$BIC = -2 \log l(\hat{\boldsymbol{\theta}}|\mathcal{M}, \mathbf{x}) + d \log n$$

where $\hat{\boldsymbol{\theta}}$ is the MLE and d is the dimension of the parameter space $\boldsymbol{\Theta}$ for $\mathcal{M}$.

Then, it is possible to show that when the sample size, $n \to \infty$, then

$$BIC_0 - BIC_1 \approx -2 \log B_1^0$$

where BIC_i represents the Bayesian information for model i and B_1^0 is the Bayes factor. See Kass and Raftery (1995).

The deviance information criterion (DIC)

The DIC (Spiegelhalter et al 2002) is a Bayesian alternative to the BIC, AIC etc. appropriate for use in hierarchical models, where the Bayes factor is difficult to calculate. For a model $\mathcal{M}$ and sample data, $\mathbf{x}$, the deviance is:

$$D(\mathbf{x}|\mathcal{M}, \boldsymbol{\theta}) = -2 \log l(\boldsymbol{\theta}|\mathcal{M}, \mathbf{x}).$$

The expected deviance, $\bar{D} = E[D|\mathcal{M}, \boldsymbol{\theta}]$, is a measure of lack of fit of the model. The effective number of model parameters is

$$p_D = \bar{D} - D(\mathbf{x}|\mathcal{M}, \bar{\theta})$$

where $\bar{\theta}$ is the posterior mean. Then the deviance information criterion is $DIC = \bar{D} + p_D$.

An advantage of the DIC is that it is easy to calculate when using Gibbs sampling and this criterion is implemented automatically in `Winbugs`.

For more details, see:

`http://en.wikipedia.org/wiki/Deviance_information_criterion`

However, the DIC has some strange properties which make it inappropriate in certain contexts, e.g. it is not guaranteed that p_D is positive. For alternatives, see e.g. Celeux et al (2006).

Calculation of the Bayes factor

In many cases, when non-conjugate prior distributions are used, the marginal likelihoods $f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$ needed for the calculation of the Bayes factor cannot be evaluated analytically. In this case, there are various possibilities:

- Use of an alternative criterion, e.g. BIC or DIC.
- Use of the Laplace approximation. See chapter 8.
- Gibbs sampler or MCMC based approximations of the marginal likelihood. See e.g. Gelfand and Dey (1994), Chib (1995).

Chib's method

Suppose that for a given model, the data are $X|\boldsymbol{\theta} \sim f(\cdot|\boldsymbol{\theta})$. Given data, $\mathbf{x}$, then we wish to estimate the marginal likelihood $f(\mathbf{x}) = \int l(\boldsymbol{\theta}|\mathbf{x})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, where $p(\cdot)$ is the prior for $\boldsymbol{\theta}$.

Assume that $p(\boldsymbol{\theta}|\mathbf{x})$ and therefore $f(\mathbf{x})$, cannot be evaluated analytically, but that we can set up a Gibbs sampler in order to simulate a sample from $p(\boldsymbol{\theta}|\mathbf{x})$.

The assume for simplicity that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and that the conditional distributions $p(\boldsymbol{\theta}_1|\mathbf{x}, \boldsymbol{\theta}_2)$ and $p(\boldsymbol{\theta}_2|\mathbf{x}, \boldsymbol{\theta}_1)$ are available. Chib's method for estimating the marginal likelihood proceeds as follows:

First note that via Bayes theorem, for any $\boldsymbol{\theta}$, we have

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\mathbf{x})}.$$

Therefore,

$$\begin{aligned}\log f(\mathbf{x}) &= \log f(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{x}) \\ &= \log l(\boldsymbol{\theta}|\mathbf{x}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}_2|\mathbf{x}, \boldsymbol{\theta}_1) - \log p(\boldsymbol{\theta}_1|\mathbf{x})\end{aligned}$$

Now, assume that we run a Gibbs sampler for T iterations. Then we can fix some high posterior density point $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$ such as the estimated posterior mode or posterior mean. Then all of the terms composing the log marginal likelihood can be estimated directly from the Gibbs sampler output, e.g.

$$\log p(\tilde{\boldsymbol{\theta}}_1|\mathbf{x}) \approx \frac{1}{T}\sum_{t=1}^{T} \log p\left(\tilde{\boldsymbol{\theta}}_1|\mathbf{x}, \boldsymbol{\theta}_2^{(t)}\right).$$

If $\boldsymbol{\theta}$ is higher (k) dimensional then the algorithm may be operated in the same way but now writing

$$\log p(\tilde{\boldsymbol{\theta}}|\mathbf{x}) = \log p(\tilde{\boldsymbol{\theta}}_1|\mathbf{x}) + \log p(\tilde{\boldsymbol{\theta}}_2|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1) + \ldots + \log p(\tilde{\boldsymbol{\theta}}_k|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{k-1}).$$

In this case, $p(\tilde{\boldsymbol{\theta}}_1|\mathbf{x})$ can be estimated directly from the Gibbs output as earlier. In order to estimate $p(\tilde{\boldsymbol{\theta}}_2|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1)$, we can run the Gibbs sampler for a further T iterations but holding $\boldsymbol{\theta}_1 = \tilde{\boldsymbol{\theta}}_1$ fixed when we have

$$p(\tilde{\boldsymbol{\theta}}_2|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1) \approx \frac{1}{N} \sum_{t=1}^{T} p\left(\tilde{\boldsymbol{\theta}}_2|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}\right).$$

In order to estimate $p(\tilde{\boldsymbol{\theta}}_3|\mathbf{x}, \tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$, the algorithm is run again but with $\boldsymbol{\theta}_1 = \tilde{\boldsymbol{\theta}}_1$ and $\boldsymbol{\theta}_2 = \tilde{\boldsymbol{\theta}}_2$ fixed and so on. Thus, in order to estimate all the terms, we need to run the Gibbs sampler a total of k times.

In general, this algorithm will be efficient if the point $\tilde{\boldsymbol{\theta}}$ is chosen to have sufficiently high mass, although in theory, it will work for any choice of $\tilde{\boldsymbol{\theta}}$. The disadvantage of the algorithm is the extra execution time needed to run the Gibbs sampler various times.

Chib and Jeliazkov (2001,2005) provide extensions of this algorithm to more complex Markov chain Monte Carlo samplers.

Problems and generalizations of the Bayes factor

If we use improper priors for the model parameters, then in general, the Bayes factor is not defined because

$$B_1^0 = \frac{l(H_0|\mathbf{x})}{l(H_1|\mathbf{x})} = \frac{\int p(\boldsymbol{\theta}_0|H_0) f(\mathbf{x}|\boldsymbol{\theta}_0)\, d\boldsymbol{\theta}_0}{\int p(\boldsymbol{\theta}_1|H_1) f(\mathbf{x}|\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}$$

depends on the undefined constants of the prior distributions $p(\cdot|H_0)$ and $p(\cdot|H_1)$.

In such cases, we need to modify the Bayes factor. Various possibilities have been considered.

Intrinsic Bayes factors

Berger

This approach was developed by Berger and Perrichi (1996).

The idea is to divide the sample into two parts; $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ where $\mathbf{y}$ is thought of as training data.

Then we can define the partial Bayes factor in favour of model H_0 based on the data $\mathbf{z}$, after observing $\mathbf{y}$ as

$$B(\mathbf{z}|\mathbf{y}) = \frac{\int f(\boldsymbol{\theta}_0|\mathbf{y}) f(\mathbf{z}|\boldsymbol{\theta}_0)\, d\boldsymbol{\theta}_i}{\int f(\boldsymbol{\theta}_1|\mathbf{y}) f(\mathbf{z}|\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1}$$

which exists whenever the distributions $f(\boldsymbol{\theta}_i|\mathbf{y})$ are proper for $i = 0, 1$, even though the priors are improper.

A problem with the partial Bayes factor is that this depends on the arbitrary partition of $\mathbf{x} = (\mathbf{y}, \mathbf{z})$. One possibility is to define a measure based on averaging over all of the possible sets $\mathbf{y}$ of the least dimension that gives a proper Bayes factor.

One possibility is to define an arithmetic Bayes factor

$$B_A = \frac{1}{L}\sum_{l=1}^{L} B(\mathbf{z}(l)|\mathbf{y}(l)),$$

where the index l runs over all sets $\mathbf{y}(l)$ of minimum dimension. This has the disadvantage in the model selection context that if we define B^i_j as the Bayes factor in favour of model i as against model j, then $B^i_{Aj} \neq 1/B^j_{Ai}$ so that if we wish to use this Bayes factor in this context, then it is necessary to have a predefined ordering of the models under consideration.

An alternative which does not require such an ordering constraint is the geometric Bayes factor

$$B_G = \left\{\prod_{l=1}^{L} B(\mathbf{z}(l)|\mathbf{y}(l))\right\}^{1/L}.$$

Example 58

Recall Example 55. Suppose that we wish to calculate the Bayes factor in favour of H_0 as against H_1. Earlier, we assumed a uniform prior for θ, i.e.

$$p(\theta|H_0) \propto 1 \text{ and } p(\theta|H_1) \propto 1.$$

Obviously, the Bayes factor is not defined but if we observe a single datum x_l, we have

$$\theta|x_l \sim \mathcal{N}(x_l, 1)$$

which is proper, and

$$\begin{aligned} p(\theta|H_0, x_l) &= \frac{p(\theta|x_l)}{P(\theta \leq 0|x_l)} = \frac{p(\theta|x_l)}{\Phi(-x_l)} \\ p(\theta|H_1, x_l) &= \frac{p(\theta|x_l)}{1 - \Phi(-x_l)} \end{aligned}$$

Therefore, defining $\mathbf{z}(l) = (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_n)$, we are able to calculate the intrinsic Bayes factor.

$$
\begin{aligned}
B(\mathbf{z}(l)|x_l) &= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x_l)^2} \left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{1}{2}\sum_{i\neq l}(x_i-\theta)^2\right) \, d\theta}{\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x_l)^2} \left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{1}{2}\sum_{i\neq l}(x_i-\theta)^2\right) \, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\int_{-\infty}^{0} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \, d\theta}{\int_{0}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\int_{-\infty}^{0} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right) \, d\theta}{\int_{0}^{\infty} \exp\left(-\frac{n}{2}(\theta-\bar{x})^2\right) \, d\theta} \\
&= \frac{1-\Phi(-x_l)}{\Phi(-x_l)} \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)} \quad \text{so the intrinsic, geometric Bayes factor is} \\
B^0_{G1} &= \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)} \left(\prod_{l=1}^{n} \frac{1-\Phi(-x_l)}{\Phi(-x_l)}\right)^{\frac{1}{n}}.
\end{aligned}
$$

Fractional Bayes factors

O' Hagan

This approach stems from O' Hagan (1995).

We define $B_F(b) = \frac{g_0(\mathbf{x}|b)}{g_1(\mathbf{x}|b)}$ where

$$g_i(\mathbf{x}|b) = \int \frac{p(\boldsymbol{\theta}_i|H_i) f(\mathbf{x}|\boldsymbol{\theta}_i)}{\int p(\boldsymbol{\theta}_i|H_i) \left\{f(\mathbf{x}|\boldsymbol{\theta}_i)\right\}^b \, d\boldsymbol{\theta}_i} \, d\boldsymbol{\theta}_i$$

for $i = 0, 1$, where b may be interpreted as the proportion of the data chosen for the training sample. We might elect the minimum value of b possible although larger values will produce more robust results.

Example 59

In Example 55, let $b = 1/n$. Therefore

$$g_0(\mathbf{x}|b) = \frac{\int_{-\infty}^{0} \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\, d\theta}{\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2n}\sum_{i=1}^{n}(x_i-\theta)^2\right)\, d\theta}$$

$$= \frac{\left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{(n-1)s^2}{2}\right) \Phi\left(-\sqrt{n}\bar{x}\right)}{\exp\left(-\frac{(n-1)s^2}{2n}\right) \Phi\left(-\bar{x}\right)} \quad \text{and similarly,}$$

$$g_1(\mathbf{x}|b) = \frac{\left(\frac{1}{2\pi}\right)^{\frac{n-1}{2}} \exp\left(-\frac{(n-1)s^2}{2}\right) \left(1-\Phi\left(-\sqrt{n}\bar{x}\right)\right)}{\exp\left(-\frac{(n-1)s^2}{2n}\right) \left(1-\Phi\left(-\bar{x}\right)\right)} \quad \text{and therefore}$$

$$B_F\left(\frac{1}{n}\right) = \frac{\Phi\left(-\sqrt{n}\bar{x}\right)}{1-\Phi\left(-\sqrt{n}\bar{x}\right)} \left(\frac{1-\Phi\left(-\bar{x}\right)}{\Phi\left(-\bar{x}\right)}\right).$$

Fractional and intrinsic Bayes factors do not have all of the properties of simple Bayes factors. See O' Hagan (1997). Also, many variants are available in particular for intrinsic Bayes factors.

Also there are a number of alternative approaches for Bayesian model comparison. See e.g. Wassermann (2000)

http://www.stat.cmu.edu/cmu-stats/tr/tr666/tr666.html

Application II continued: choosing between half-normal and half-t models

In Application II, we considered fitting a half-normal model ($\mathcal{M}_0$) to athletes body fat data and we suggested that a half-t distribution ($\mathcal{M}_1$) might be a reasonable alternative.

Analogous to the half-normal model, we write the half-t model as $X = \xi + \frac{1}{\sqrt{\tau}}|T|$ where $T|d \sim \mathcal{T}_d$ is a Student's t random variable. To simplify the inference, we introduce a latent variable θ such that

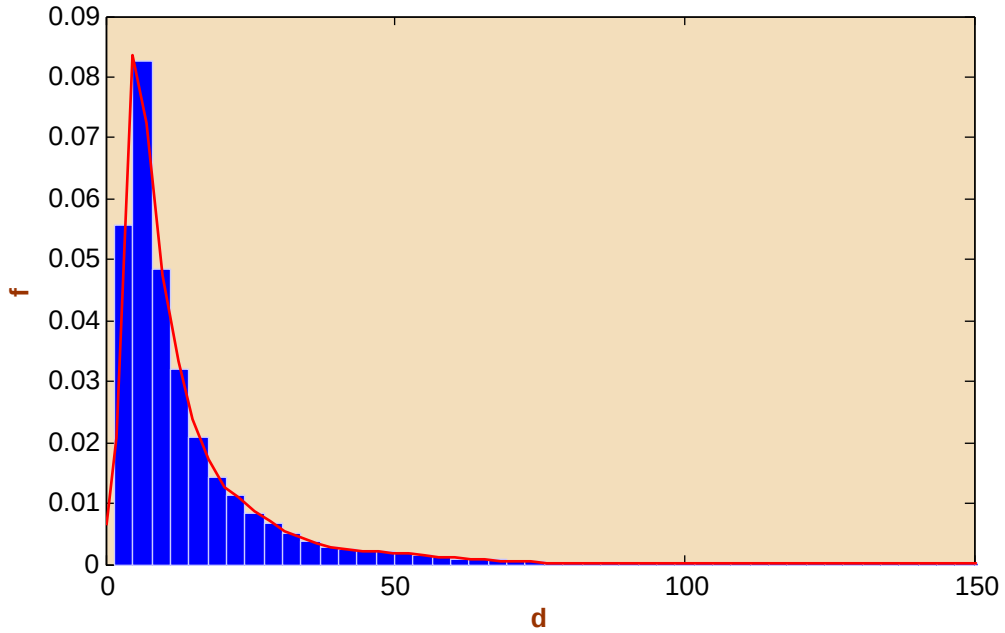
$$\begin{aligned} \theta|d &\sim \mathcal{G}\left(\frac{d}{2}, \frac{d}{2}\right) \quad \text{when} \\ X|\theta, \xi, \tau, \mathcal{M}_1 &\sim \mathcal{HN}\left(\xi, \frac{1}{\tau\theta}\right) \quad \text{has a half-normal distribution.} \end{aligned}$$

Assume that we define improper prior distributions $p(\xi, \tau) \propto \frac{1}{\tau}$ as for the half-normal model and a proper, exponential prior for d, say $d \sim \mathcal{E}(\kappa)$. Then, given a sample, $\mathbf{x}$, the conditional posteriors are:

$$
\begin{aligned}
\tau|\mathbf{x}, d, \xi, \boldsymbol{\theta} &\sim \mathcal{G}\left(\frac{n}{2}, \frac{\sum_{i=1}^{n} \theta_i (x_i - \xi_i)^2}{2}\right) \\
\xi|\mathbf{x}, d, \tau, \boldsymbol{\theta} &\sim \mathcal{TN}\left(\frac{\sum_{i=1}^{n} \theta_i x_i}{\sum_{i=1}^{n} \theta_i}, \frac{1}{\tau \sum_{i=1}^{n} \theta_i}\right) \quad \text{truncated so that } \xi < \min\{\mathbf{x}\} \\
\theta_i|\mathbf{x}, d, \xi, \tau &\sim \mathcal{G}\left(\frac{d+1}{2}, \frac{d + \tau(x_i - \xi)^2}{2}\right) \\
p(d|\mathbf{x}, \xi, \tau, \boldsymbol{\theta}) &\propto e^{-\kappa d} \frac{(d/2)^{\frac{nd}{2}}}{\Gamma(d/2)^n} \prod_{i=1}^{n} \theta_i^{\frac{d}{2}} \exp\left(-\frac{d}{2} \sum_{i=1}^{n} \theta_i\right)
\end{aligned}
$$

The only density that has a slightly complicated form is that of d and the joint posterior can be sampled using a Gibbs algorithm with a *Metropolis* step for d.

The histogram shows an estimate of the posterior of d when the half-t model was fitted to the athletes data with a prior distribution $d \sim \mathcal{E}(1/20)$.



The distribution is quite long tailed although the mode of d is below 10. We saw earlier that the fit of the half-t model looks better than the half-normal fit, but how can we calculate a Bayes factor?

Calculation of the Bayes factor

We saw earlier that, in general if improper prior distributions are used, then the Bayes factor is undefined. However, in this case, the usual Bayes factor construction:

$$
\begin{aligned}
B_1^0 &= \frac{f(\mathbf{x}|\mathcal{M}_0)}{f(\mathbf{x}|\mathcal{M}_1)} \\
&= \frac{\int f(\mathbf{x}|\xi,\tau,\mathcal{M}_0)p(\xi,\tau|\mathcal{M}_0)\,d\xi d\tau}{\int f(\mathbf{x}|d,\xi,\tau,\mathcal{M}_1)p(\xi,\tau|\mathcal{M}_1)p(d|\mathcal{M}_1)\,d\xi d\tau, dd} \\
&= \frac{\int_{-\infty}^{\min\{\mathbf{x}\}}\int_0^\infty f(\mathbf{x}|\xi,\tau,\mathcal{M}_0)\frac{1}{\tau}\,d\xi d\tau}{\int_0^\infty\int_{-\infty}^{\min\{\mathbf{x}\}}\int_0^\infty\int f(\mathbf{x}|d,\xi,\tau,\mathcal{M}_1)\frac{1}{\tau}p(d|\mathcal{M}_1)\,d\xi d\tau dd}
\end{aligned}
$$

can be shown to produce a well-calibrated, intrinsic Bayes factor because the support and improper priors for ξ, τ are the same under $both$ models. See Cano et al (2004).

The numerator of this formula can be calculated explicitly for the half-normal model. Thus:

$$\int_{-\infty}^{\min\{\mathbf{x}\}} \int_{0}^{\infty} f(\mathbf{x}|\xi,\tau,\mathcal{M}_0)\frac{1}{\tau}\,d\xi d\tau = \frac{2\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{n}}\left(\frac{2}{\sqrt{(n-1)\pi}s}\right)^{n-1}\Phi_{n-1}\left(\frac{\min\{\mathbf{x}\}-\bar{x}}{s/\sqrt{n}}\right).$$

The denominator can be calculated using Chib's approach. We can write

$$\begin{aligned}\log f(\mathbf{x}|\mathcal{M}_1) &= \log p(\tilde{\xi},\tilde{\tau}) + \log p(\tilde{d}) + \log l(\tilde{\xi},\tilde{\tau},\tilde{d}|\mathbf{x}) \\ &\quad - \log p(\tilde{\xi}|\mathbf{x}) - \log p(\tilde{\tau}|\mathbf{x},\tilde{\xi}) - \log p(\tilde{d}|\mathbf{x},\tilde{\xi},\tilde{\tau})\end{aligned}$$

and the integrating constant of the density in the final term in this expression can be evaluated by using standard, one dimensional, numerical integration.

Using this approach, the Bayes factor in favour of the half-t model is $B_0^1 = 4$ which suggests positive evidence in favour of this model. For more details, see Wiper et al (2008).

References

Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, **2**, 217–352.

Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.

Cano, J.A., Kessler, M. and Moreno, E. (2004). On intrinsic priors for nonnested models. *Test*, **13**, 445–463.

Celeux, G., Forbes, F., Robert, C.P. and Titterington, D.M. (2006). Deviance Information Criteria for Missing Data Models (with discussion). *Bayesian Analysis*, **1**, 651–706.

Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis Hastings output. *Journal of the American Statistical Association*, **96**, 270–281.

Chib, S. and Jeliazkov, I. (2005). Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation. *Statistica Neerlandica*, **59**, 30–44.

Gelfand, A.E., and Dey, D.K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations, *Journal of the Royal Statistical Society, Series B*, **56**, 501–514.

Good, I.J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, **53**, 799–813.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.), Oxford: University Press.

Kass, R. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Lindley, D.V. (1957). A Statistical Paradox. *Biometrika*, **44**, 187-192.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 99–138.

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factors. *Test*, **6**, 101–118.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Spiegelhalter, D.J., Best, N.G., Carlin B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**, 583-640.

Wassermann, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, **44**, 92–107.

Wiper, M.P., Girón, F.J. and Pewsey, A. (2008). Objective Bayesian inference for the half-normal and half-t distributions. *Communications in Statistics: Theory and Methods*, **37**, 3165–3185.