MT4531/5731: (Advanced) Bayesian Inference

Model choice/comparison

Nicolò Margaritella

School of Mathematics and Statistics, University of St Andrews

–

University of St Andrews

Outline

1. Introduction
2. The Bayes Factor
3. Computing Bayes Factors

Outline

1. Introduction

2. The Bayes Factor

3. Computing Bayes Factors

Model choice as choice of hypothesis

- Suppose that we are interested in the parameter $\theta \in \Theta$. Then, our hypotheses are of the form,

$$H_0 : \theta \in \Theta_0; \qquad H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 are disjoint and exhaustive subsets of the parameter space Θ.

$\Rightarrow$ This can translate to a model comparison setting.

- For example, to choose between fitting a constant or a simple linear regression,

$$M_0 : \mathrm{E}(Y_i) = \beta_0; \qquad M_1 : \mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i,$$

one has to choose between the hypotheses,

$$H_0 : \beta_1 = 0; \qquad H_1 : \beta_1 \neq 0.$$

Introduction

- Focusing on model choice, one simple example is the choice between the following 2 linear models for the relation between an outcome y and an explanatory variable z, so that,for n observations, and $i = 1, ..., n$,

  $$\text{Model 0: } y_i = \alpha + \epsilon_i; \qquad \text{Model 1: } y_i = \alpha + \beta z_i + \epsilon_i,$$

  where, given σ, we assume, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
- Thus, we could express the models in the form:

  $$\text{Model 0: } y_i \sim N(\alpha, \sigma^2); \qquad \text{Model 1: } y_i \sim N(\alpha + \beta z_i, \sigma^2).$$

Model comparison for many models (1)

- We use multiple regression to illustrate the idea.
- Assume J explanatory variables, such that $\boldsymbol{z}_i = \{z_{i1}, \ldots, z_{iJ}\}$.
- The "full" model is given by,

$$y_i = \alpha + \sum_{j=1}^{J} \beta_j z_{ij} + \epsilon_i,$$

  where, given σ, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
- Equivalently, given α, σ, and β_j, $j = 1, \ldots, J$,

$$y_i \sim N\left(\alpha + \sum_{j=1}^{J} \beta_j z_{ij}, \sigma^2\right), \text{ independently.}$$

- Sub-models are specified by setting some β_js equal to zero.
- 2^J possible models corresponding to combinations of explanatory variables present. (Intercept is always present.)

Model comparison for many models (2)

- For example, for $J = 3$, a full model is given by,

$$y_i = \alpha + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \epsilon_i,$$

- The $2^3 = 8$ possible models can be compared by considering posterior model probabilities, or pairwise Bayes factors that generate a ranking of the possible models.
- Note that for interpretability, each of the explanatory variables should all be measured on the same scale.
- Typically, explanatory variables and response variable are normalised (by taking the raw values, subtracting their mean and dividing by the sample standard deviation)
- This is important for calculating Bayes factors, and posterior model probabilities.

Outline

1 Introduction

2 The Bayes Factor

3 Computing Bayes Factors

The Bayes factor - Simple example (1)

- We have already seen the concept of the Bayes factor.
- Consider 2 possible statistical models for observations y_i. (A constant mean, or a simple linear regression.)
- Under Model 0 ($M = 0$) the set of model parameters is $\boldsymbol{\theta}_0 = \{\alpha, \sigma^2\}$; alternatively, under Model 1 ($M = 1$), the model parameters are $\boldsymbol{\theta}_1 = \{\alpha, \beta, \sigma^2\}$.
- The Bayes factor for choosing one of the 2 models is defined in the same manner as for different hypotheses by,

$$B_{01} = \frac{p(M=0|\boldsymbol{x})/p(M=1|\boldsymbol{x})}{p(M=0)/p(M=1)}, \text{ or, } \frac{\pi(M=0|\boldsymbol{x})/\pi(M=1|\boldsymbol{x})}{p(M=0)/p(M=1)},$$

where $p(M = m)$ denotes the probability that m is the true model, for $m = 0, 1$.

The Bayes factor - Simple example (2)

- The prior is a prior probability $p(M = m)$, $m = 0, 1$, and the prior distribution $p(\boldsymbol{\theta}_m|M = m)$ conditional on each model.
- Given a data vector $\boldsymbol{x}$, the posterior probability of model m is

$$\pi(M = m|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|M = m)p(M = m)}{f(\boldsymbol{x})} \propto f(\boldsymbol{x}|M = m)p(M = m).$$

- $f(\boldsymbol{x}|M = m)$ is calculated as,

$$\int f(\boldsymbol{x}, \boldsymbol{\theta}_m|M = m)d\boldsymbol{\theta}_m \int f(\boldsymbol{x}|M = m, \boldsymbol{\theta}_m)p(\boldsymbol{\theta}_m|M = m)d\boldsymbol{\theta}_m$$

The constant of proportionality $f(\boldsymbol{x})$ is,

$$f(\boldsymbol{x}) = [f(\boldsymbol{x}|M = 0)p(M = 0) + f(\boldsymbol{x}|M = 1)p(M = 1)]^{-1}.$$

The Bayes factor - General case (1)

- The posterior distribution over parameter and model space is given by,

$$\begin{aligned}\pi(\boldsymbol{\theta}_m, M=m|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\boldsymbol{\theta}_m, M=m)p(\boldsymbol{\theta}_m|M=m)p(M=m)}{f(\boldsymbol{x})} \\ &\propto f(\boldsymbol{x}|\boldsymbol{\theta}_m, M=m)p(\boldsymbol{\theta}_m|M=m)p(M=m).\end{aligned} \tag{1}$$

- The posterior model probability of model m is then simply the marginal posterior distribution:

$$\begin{aligned}\pi(M=m|\boldsymbol{x}) &= \int \pi(\boldsymbol{\theta}_m, M=m|\boldsymbol{x})d\boldsymbol{\theta}_m \\ &= \frac{p(M=m)}{f(\boldsymbol{x})}\int f(\boldsymbol{x}|\boldsymbol{\theta}_m, M=m)p(\boldsymbol{\theta}_m|M=m)d\boldsymbol{\theta}_m \\ &= \frac{p(M=m)}{f(\boldsymbol{x})}\int f(\boldsymbol{x},\boldsymbol{\theta}_m|M=m)d\boldsymbol{\theta}_m = \frac{p(M=m)}{f(\boldsymbol{x})}f(\boldsymbol{x}|M=m).\end{aligned}$$

The Bayes factor - General case (2)

- Therefore,

$$\pi(M = m|\boldsymbol{x}) \quad \propto \quad f(\boldsymbol{x}|M = m)p(M = m), \qquad (2)$$

where the likelihood of the data given the model is given by,

$$f(\boldsymbol{x}|M = m) = \int f(\boldsymbol{x}|\boldsymbol{\theta}_m, M = m)p(\boldsymbol{\theta}_m|M = m)d\boldsymbol{\theta}_m. \qquad (3)$$

- The constant of proportionality is,

$$1/f(\boldsymbol{x}) = \left[\sum_m f(\boldsymbol{x}|M = m)p(M = m)\right]^{-1}.$$

The Bayes factor

- Calculating posterior model probabilities (and hence Bayes factors), by performing the above integration, provides a *quantitative* discrimination between competing models.
- However, in general the necessary integration will be analytically intractable....... (problem number 1!)
- In addition, often we may not want to compare only a limited number of possible models, but may have many competing models, e.g. 2^J for some large J considering a multiple regression. (problem number 2!).

Important Notes

- Caution is required when using a proper prior with a large variance.
- For example, when comparing nested (linear) models, adopting, say, a $N(0, 10^6)$ prior for the additional parameters may give different results compared to adopting, say, a $N(0, 10^{10})$ prior;
- The Bayes Factor is consistent. If the true model belongs to the set of fitted models, the Bayes Factor will select the true model with probability one as $n \to \infty$. (The same is true for the BIC, a frequentist criterion that is an approximation of the Bayes factor).
- In the next lecture we will discuss the Deviance Information Criterion (DIC). It is designed to find the best model in terms of predictive performance, in the same spirit as the frequentist AIC.
- The DIC is not consistent. However, this is not considered to be a problem by some investigators, arguing that the true model is rarely among the fitted models, as nature is too complex to allow for that.

Outline

1 Introduction

2 The Bayes Factor

3 Computing Bayes Factors

Simple Monte Carlo approach (1)

- For each model m, we need to evaluate,

$$f(\boldsymbol{x}|M=m) = \int f(\boldsymbol{x}|\boldsymbol{\theta}_m, M=m)p(\boldsymbol{\theta}_m|M=m)d\boldsymbol{\theta}_m.$$

- We express the above equation in the form,

$$f(\boldsymbol{x}|M=m) = \mathbb{E}_p[f(\boldsymbol{x}|\boldsymbol{\theta}_m, M=m)],$$

- We estimate this expectation by drawing K observations from the prior distribution of the parameters in model m, denoted by $\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^K$ and then use the Monte Carlo estimate,

$$\widehat{f}(\boldsymbol{x}|M=m) = \frac{1}{K}\sum_{k=1}^{K} f(\boldsymbol{x}|\boldsymbol{\theta}^k, M=m).$$

Simple Monte Carlo approach (2)

- An estimate of the posterior probability of model m is,

$$\hat{\pi}(M = m|\boldsymbol{x}) \propto p(M = m)\widehat{f}(\boldsymbol{x}|M = m).$$

- This estimate converges to the posterior model probability (up to proportionality), as $K \to \infty$.
- We repeat this process for each model $m = 1, ..., m^*$, and renormalise the estimates to obtain an estimate of the corresponding posterior model probabilities,

$$\hat{\pi}(M = m|\boldsymbol{x}) = \frac{p(M = m)\widehat{f}(\boldsymbol{x}|M = m)}{\sum_{q=1}^{m^*} p(M = q)\widehat{f}(\boldsymbol{x}|M = q)}.$$

Example

- Consider 3 possible models, $m = 1, 2, 3$. Then,

$$f(\boldsymbol{x}|M = m) = \int f(\boldsymbol{x}|\boldsymbol{\theta}_m, M = m)p(\boldsymbol{\theta}_m|M = m)d\boldsymbol{\theta}_m$$
$$= E_{prior}(f(\boldsymbol{x}|\boldsymbol{\theta}_m, M = m)).$$

- After obtaining estimates $\widehat{f}(\boldsymbol{x}|M = 1)$, $\widehat{f}(\boldsymbol{x}|M = 2)$, $\widehat{f}(\boldsymbol{x}|M = 3)$, consider that,

$$\pi(M = m|\boldsymbol{x}) = \frac{p(M = m)}{f(\boldsymbol{x})} f(\boldsymbol{x}|M = m),$$

- As $\pi(M = 1|\boldsymbol{x}) + \pi(M = 2|\boldsymbol{x}) + \pi(M = 3|\boldsymbol{x}) = 1$,

$$\widehat{f}(\boldsymbol{x}) = p(M = 1)\widehat{f}(\boldsymbol{x}|M = 1) + p(M = 2)\widehat{f}(\boldsymbol{x}|M = 2) + p(M = 3)\widehat{f}(\boldsymbol{x}|M = 3).$$

- Now, for $m = 1, 2, 3$,

$$\widehat{\pi}(M = m|\boldsymbol{x}) = \frac{p(M = m)\widehat{f}(\boldsymbol{x}|M = m)}{p(M = 1)\widehat{f}(\boldsymbol{x}|M = 1) + \ldots + p(M = 3)\widehat{f}(\boldsymbol{x}|M = 3)}.$$

Low efficiency

- However, the above estimates of the likelihood $f(\boldsymbol{x}|M = m)$ are generally inefficient and very unstable, as a result of the parameters being drawn from the prior distribution.
- (For a computational example see the relevant question in Tutorial 8.)
- The likelihood values evaluated at the parameter values sampled from the prior distribution are often very small, as the prior, in general, is far more dispersed over the parameter space than the likelihood.
- Thus, the expectation is heavily dominated by only few sampled values, even for large values of K, resulting in a very large variance, and the instability of the estimate.
- A more stable approach makes use of a sampling technique called *Importance Sampling*. This will be discussed later in the course.

Alternative approaches

- Monte Carlo estimates for model comparison are the easiest to program and conceptualise. However, they are often very inefficient and do not always converge within a feasible number of iterations.
- One alternative approach for calculating posterior model probabilities is the (RJ) reversible jump MCMC. The basic idea is to construct a Markov chain with stationary distribution equal to the joint posterior distribution over both parameter and model space. See the lecture notes for more details (not examinable).
- Another approach, pertinent to linear modelling, is to use variable/covariate selection, with methods based on the ‘Spike and Slab’ and ‘Horseshoe’ priors (not discussed in this course).
- Finally, another approach is to use a model comparison criterion such as the DIC or WAIC, similarly to the AIC or BIC in classical statistics. This is what we will discuss in the next lecture.