Math 459 Lecture 18

Todd Kuffner

Last Time

The Bayesian linear regression model.

Gibbs sampling for Bayesian linear regression.

`MCMCregress` and the mercury risks for Kuwaiti fishermen.

Today

Bayes factors, model comparison, and measure of evidence

Bayesian hypothesis testing

Next time: approximating the marginal likelihood, computing Bayes factors, model and variable selection

Broad Overview

Three 'old' schools of thought regarding statistical testing:

- Fisher
- Neyman (and Pearson)
- Jeffreys

Example

Suppose we have i.i.d. $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$ with known σ^2, $n = 10$ and we wish to test $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$. Suppose $z = \sqrt{n}\bar{x}/\sigma = 2.3$.

Fisher would give $p-$values $p = 0.021$

Jeffreys would report posterior probabilities of H_0, $Pr(H_0|x_1, \ldots, x_n) = 0.28$, if he put equal prior probabilities on the two hypotheses and used a Cauchy$(0, \sigma)$ prior on H_1

Neyman would prespecify Type I error probability $\alpha = 0.05$ and would report $\alpha = 0.05$ and a Type II error probability β

Fisher's significance testing

Suppose one observes data $X \sim f(x|\theta)$ and wishes to test $H_0 : \theta = \theta_0$

- choose a test statistic $T = t(X)$, large values of which give evidence against H_0
- compute p-value $p = Pr_0(t(X) \geq t(x))$, reject H_0 if this is small

Neyman-Pearson testing

Believe you can only test a point null $H_0 : \theta = \theta_0$ vs. some alternative $H_1 : \theta = \theta_1$.

- reject H_0 if $T \geq c$, where c is a pre-specified critical value
- compute type I and type II error probabilities, $\alpha = Pr_0(\text{reject the null})$, $\beta = Pr_1(\text{fail to reject null})$

Jeffreys approach

Agreed with Neyman that you need an alternative to be able to test the null.

- define Bayes factor $B(x) = f(x|\theta_0)/f(x|\theta_1)$
- reject H_0 if $B(x) \leq 1$
- report objective posterior error probabilities (i.e. the posterior probabilities of the hypotheses), e.g.

$$Pr(H_0|x) = \frac{B(x)}{1 + B(x)}$$

based on assigning equal prior probabilities of 1/2 to the two hypotheses and then applying Bayes theorem

Motivating Bayes Factors

Recall in decision theory (and frequentist testing) the following terminology.

- we have a statistical model $f(x|\theta)$ with $\theta \in \Theta$
- want to test a null hypothesis of the form

$$H_0 : \theta \in \Theta_0$$

  where $\Theta_0 \subset \Theta$, e.g. could be the single point $\{\theta_0\}$ (a point null hypothesis)
- in linear regression, Θ_0 is often a *subspace* of the vector space Θ; then testing is the same as model selection

Example (Robert)

Suppose we have a **logistic regression model**

$$Pr_\alpha(y=1) = 1 - Pr_\alpha(y=0) = \exp(\alpha^t x)/(1+\exp(\alpha^t x)), \quad \alpha, x \in \mathbb{R}^p$$

- could be a model for the probability of winning the lottery, or developing a water buffalo allergy during your lifetime
- the explanatory variables $x = (x_1, \ldots, x_p)$ represent factors you believe can influence this probability
- wish to test whether some coefficient, say α_{i0} corresponding to x_{i0} is zero or not

Neyman-Pearson perspective

Formalize the decision space $\mathcal{D}$ which is restricted to $\{1, 0\}$.

- thus such testing problems can be viewed as inference about the indicator function $I_{\Theta_0}(\theta)$, and thus answers should be in $I_{\Theta_0}(\Theta) = \{0, 1\}$

Frequently we have more information about the support of θ, specifically that $\theta \in \Theta_0 \cup \Theta_1 \neq \Theta$.

- then define the alternative hypothesis *against which* we test H_0 as

$$H_1 : \theta \in \Theta_1$$

In these terms, every test procedure, say φ, is interpreted as an estimator of $I_{\Theta_0}(\theta)$.

- to derive a Bayes estimator, we need a loss function $L(\theta, \varphi)$

Example

Neyman-Pearson proposed the 0–1 loss

$$L(\theta, \varphi) = \begin{cases} 1 & \text{if } \varphi \neq I_{\Theta_0}(\theta), \\ 0 & \text{otherwise} \end{cases}$$

For the 0–1 loss, the Bayesian 'test procedure' (i.e. estimator) is

$$\varphi^{\pi}(x) = \begin{cases} 1 & \text{if } Pr^{\pi}(\theta \in \Theta_0 | x) > Pr^{\pi}(\theta \in \Theta_0^C | x), \\ 0 & \text{otherwise.} \end{cases}$$

Intuitive justification: this estimator choose the hypothesis with largest posterior probability.

Generalization of Example

One could penalize errors made according to differnt weights, depending on whether H_0 is true or false.

- the weighted 0–1 losses defined below are called "$a_0 - a_1$"

$$L(\theta, \varphi) = \begin{cases} 0 & \text{if } \varphi = I_{\Theta_0}(\theta), \\ a_0 & \text{if } \theta \in \Theta_0 \text{ and } \varphi = 0, \\ a_1 & \text{if } \theta \notin \Theta_0 \text{ and } \varphi = 1. \end{cases}$$

Question: what is the Bayesian estimator in this case?

The Bayes estimator associated with prior π is

$$\varphi^{\pi}(x) = \begin{cases} 1 & \text{if } Pr^{\pi}(\theta \in \Theta_0 | x) > \frac{a_1}{a_0 + a_1}, \\ 0 & \text{otherwise.} \end{cases}$$

Sketch of proof: Note that

$$\begin{aligned} L(\pi, \varphi | x) &= \int_{\Theta} L(\theta, \varphi) \pi(\theta | x) d\theta \\ &= a_0 Pr^{\pi}(\theta \in \Theta_0 | x) I_{\{0\}}(\varphi) + a_1 Pr^{\pi}(\theta \notin \Theta_0 | x) I_{\{1\}}(\varphi) \end{aligned}$$

and from this the Bayes estimator follows.

Comment

For the class of losses to which this generalized 0–1 loss belongs, H_0 is rejected when *the posterior probability of* H_0 *is small.*

- the acceptance level is determined by the form of the loss function
- note that φ^π depends only on a_0/a_1
- also, the larger a_0/a_1 is, meaning wrong answers are more important under H_0 relative to H_1, the smaller the posterior probability of H_0 needs to be for H_0 to be accepted

Example

Consider $x \sim \text{Bin}(n, p)$ and $\Theta_0 = [0, 0.5]$.

- with uniform prior $\pi(p) = 1$, the posterior probability of H_0 is

$$Pr^{\pi}(p \leq \frac{1}{2}|x) = \frac{\int_0^{1/2} p^x (1-p)^{n-x} dp}{B(x+1, n-x+1)}$$

$$= \frac{(1/2)^{n+1}}{B(x+1, n-x+1)} \left\{ \frac{1}{x+1} + \frac{n-x}{(x+1)(x+2)} + \cdots + \frac{(n-x)!x!}{(n+1)!} \right\}$$

which can be computed and compared with the acceptance level.

Example

Suppose $x \sim \mathcal{N}(0, \sigma^2)$ and $\theta \sim \mathcal{N}(\mu, \tau^2)$.

- then $\pi(\theta|x)$ is $\mathcal{N}(\mu(x), \omega^2)$ where

$$\mu(x) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2}, \quad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2 + tau^2}$$

To test $H_0 : \theta < 0$, compute

$$Pr^{\pi}(\theta < 0|x) = Pr^{\pi}\left(\frac{\theta - \mu(x)}{\omega} < \frac{-\mu(x)}{\omega}\right) = \Phi(-\mu(x)/\omega).$$

Denoting the $a_1/(a_0 + a_1)$ quantile by z_{a_0,a_1}, i.e. $\Phi(z_{a_0,a_1}) = a_1/(a_0 + a_1)$, then H_0 is accepted when

$$-\mu(x) > z_{a_0,a_1}\omega.$$

Comment

Note that to a Bayesian it is natural to base a decision on the posterior probability that the hypothesis is true.

This can help motivate the Bayes factor, which is really a 1-to-1 transformation of the posterior probability, but it has its own rich history.

The Bayes Factor

The **Bayes factor** is the ratio of the posterior probabilities of the null and alternative hypotheses *over* the ratio of the prior probabilities of the null and alternative hypotheses,

$$B_{01}^{\pi}(x) = \frac{Pr(\theta \in \Theta_0|x)}{Pr(\theta \in \Theta_1|x)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}.$$

This measures how the odds of Θ_0 against Θ_1 change due to the observed data.

- simple interpretation is to compare B_{01} to 1 to assess the evidence against each model; in practice the comparison scale depends on the loss function

Explanation from First Principles

Probability of model M given data D, using Bayes's formula:

$$Pr(M|D) = \frac{Pr(D|M) \cdot Pr(M)}{Pr(D)}$$

- to compare two models

$$\begin{aligned}
\text{Posterior model odds} &= \frac{Pr(M_1|D)}{Pr(M_2|D)} \\
&= \frac{Pr(D|M_1) \cdot Pr(M_1)/Pr(D)}{Pr(D|M_2) \cdot Pr(M_2)/Pr(D)} \\
&= \left(\frac{Pr(M_1)}{Pr(M_2)}\right)\left(\frac{Pr(D|M_1)}{Pr(D|M_2)}\right) \\
&= \text{prior model odds} \times \text{Bayes factor}
\end{aligned}$$

Connection with Likelihood Ratios

When $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, the Bayes factor is just the usual likelihood ratio

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}$$

Bayes factors are like Bayesian likelihood ratios because, if π_0 is a prior distribution under H_0 and π_1 is a prior distribution under H_1, then

$$B_{01}^{\pi}(x) = \frac{\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(x|\theta_1)\pi_1(\theta)d\theta} = \frac{m_0(x)}{m_1(x)},$$

the **ratio of marginal likelihoods**.

Another Viewpoint of the Connection

Let $\hat{\theta}$ and $\hat{\theta}_1$ be the MLE on the sets Θ_0 and Θ_1, respectively.

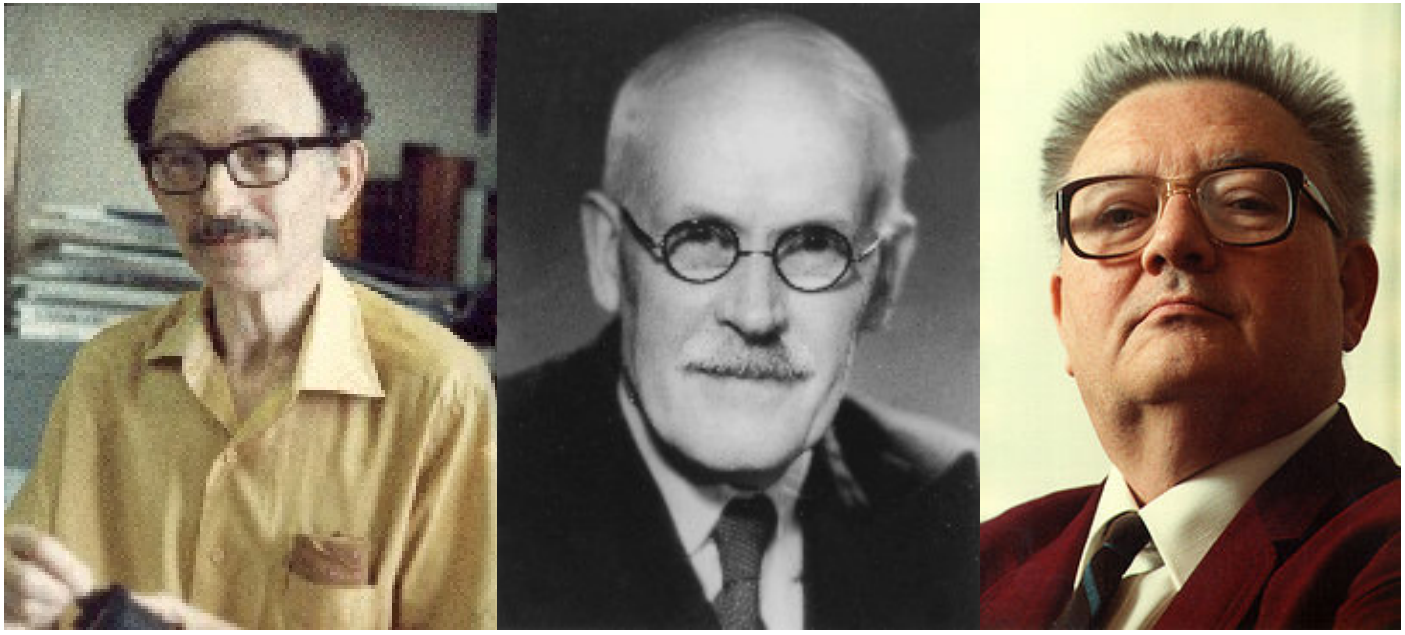
- the likelihood ratio

$$\frac{f(x|\hat{\theta}_0)}{f(x|\hat{\theta}_1)} = \frac{\sup_{\Theta_0} f(x|\theta)}{\sup_{\Theta_1} f(x|\theta)}$$

  which appears as a particular case of $B^{\pi}_{01}(x)$ when π_0 and π_1 are Dirac masses at $\hat{\theta}_0$ and $\hat{\theta}_1$.

To a Bayesian, however, this doesn't mean the likelihood ratio is valid as an inferential tool, because the priors π_0 and π_1 both depend on x.

Good and Jeffreys and the Bayes factor



- I.J. Good (1916–2009): weight of evidence (log of Bayes factor)
- Sir Harold Jeffreys (1891–1989): advocated BFs
- Edwin Jaynes (1922–1998): WashU physics professor, *Probability Theory: The Logic of Science*

Strength of Evidence

Let K be the Bayes factor for H_0 relative to H_1.

Jeffreys scale

K	**Strength of evidence for H_0**
< 1	negative (supports H_1)
$(1, \sqrt{10})$	weak
$(\sqrt{10}, 10)$	substantial
$(10, 10\sqrt{10})$	strong
$(10\sqrt{10}, 100)$	very strong
> 100	decisive

Kass & Raftery (1995 survey)

$2 \log K$	K	**Strength of evidence for H_0**
$(0, 2)$	$(1, 3)$	weak
$(2, 6)$	$(3, 20)$	positive
$(6, 10)$	$(20, 150)$	strong
> 10	> 150	very strong

Useful Departure in Terminology

Moving away from the hypothesis testing framework, Bayes factors provide a means of comparing different models, which could be of completely different types.

- given data $y = (y_1, \ldots, y_n)^T$ i.i.d. with true density $p_0(y)$
- given two candidate models M_k, M_l, where each model M_α, where α is in a countable index set A, prescribes

$$M_\alpha = \{\mathcal{F}_\alpha, \Pi_\alpha, \lambda_\alpha\}$$

  where $\mathcal{F}_\alpha$ is a set (or family) of densities, Π_α is a prior distribution, which is a probability measure on $\mathcal{F}_\alpha$ and λ_α is a probability measure on A
- e.g. the usual parametric setup has

$$\mathcal{F} = \{p_\theta(y), \theta \in \Theta \subset \mathbb{R}^p, \pi(\theta)\}$$

- usually we compare two models with $\lambda_\alpha = 1/\text{card}(A) = 1/2$ $\forall \alpha \in A$, so these cancel out

Definition (again)

The Bayes factor for comparing M_k to M_l is

$$BF_{kl} = \frac{m(y|M_k)}{m(y|M_l)}$$

- observe that the crucial thing for computing Bayes factors is to be able to compute the **marginal likelihood** in each model
- recall this is a messy integral; until the late 1980s, we had trouble computing this, or even finding a decent estimate, in complex models (Laplace approximation)
- mid-1990s was the real breakthrough (MCMC estimation of marginal likelihood)

Bayes factor consistency

The BF for comparing model k to model l, BF_{kl} is **consistent** if both

1. $BF_{kl} \to \infty$ in probability if the true density (p_0) is in model M_k
2. $BF_{kl} \to 0$ in probability if the true density is in model M_l

Equivalently

1. $\log BF_{kl} \to \infty$ in probability if $p_0 \in M_k$
2. $\log BF_{kl} \to -\infty$ in probability if $p_0 \in M_l$

Improper Priors

Using improper priors can cause problems for Bayes factors (e.g. Lindley's paradox).

Several proposals to modify Bayes factors to incorporate improper priors.

- posterior Bayes factor (Aitkin, early 1990s)
- fractional Bayes factors (O'Hagan, 1995)
- objective Bayes factors using intrinsic priors (Berger & Pericchi, Bayarri, etc., late 1990s-present)

Other measures of evidence

- posterior likelihood ratio (PLR); Dempster
- Ockham's factor; Jaynes and others
- relative belief ratios calibrated by strength; Evans (2015 book)
- Aitkin: p-value derived from profile likelihood
- Shafer et al. (2011): test martingales
- Smith & Ferrari (2014): extending PLR for composite hypotheses