Math 459 Lecture 10

Todd Kuffner

Background Concept 1: Prior Independence

Consider $\theta = (\theta_1, \ldots, \theta_d)^T$.

- If parameters *a priori* independent, then

$$p(\theta) = p(\theta_1)p(\theta_2)\cdots p(\theta_d).$$

Does not imply *a posteriori* independence.

Background Concept 2: Fisher Information

Consider $\theta = (\theta_1, \ldots, \theta_d)$. The Fisher information of f, or equivalently, X, is

$$I(\theta) = E_\theta \left[\nabla_\theta \log f(X;\theta) \nabla_\theta \log f(X;\theta)^T \right]$$

which is the covariance of the score function, $\nabla_\theta \log f(X;\theta)$.

Elements of Fisher Information

The (i, j)th element of the Fisher information matrix is given by

$$I_{ij} = E_\theta \left[\left(\frac{\partial}{\partial \theta_i} \log f(X; \theta) \right) \left(\frac{\partial}{\partial \theta_j} \log f(X; \theta) \right) \right]$$

$$\text{(under regularity)} = -E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X; \theta) \right]$$

- Regularity conditions needed to ensure validity of interchanging differentiation and integration.
- Sufficient condition: Lebesgue's dominated convergence theorem, i.e. there exists a function g such that $g(x) \geq \|\nabla_\theta f(x; \theta)\|$ for all θ

Background Concept 3: Orthogonality

θ_i and θ_j are **orthogonal** if $I_{ij} = 0$.

However, prior independence plus orthogonality **does not imply** posterior independence!

For prior independence to imply posterior independence, also need the likelihood to factor:

$$L(\theta) = L(\theta_1)L(\theta_2)\cdots L(\theta_d)$$

Bayesian Orthogonality

Related to normality of the posterior.

Selecting Priors by Formal Rules

Definition

A formal rule for selecting a prior is a prescription for how to specify a prior, which is derived from some chosen principle, and this principle is broadly applicable to a wide variety of settings (i.e. the rule is not specific to the particular sampling model).

Let Ψ be a rule for choosing a prior $p(\theta)$, i.e. Ψ could be something like

- principle of insufficient reason
- maximum entropy
- match posterior intervals and frequentist intervals
- invariance under transformations
- maximize the 'information' provided by the data
- decision theory: 'least favorable priors', unbiased decision rules

Noninformative Priors & Objective Bayes

Prior introduces information into the model.

- Objective Bayesians want priors to have little *influence* on the posterior
- not as easy as it seems!

Historical Approach (Bayes, Laplace): use flat priors

- expresses ignorance
- all values equiprobable

Problem: Not *invariant*

Consider a prior for a variance parameter σ^2, and the reparameterization $\eta = \log \sigma^2$.

uniform for η $p(\eta) \propto 1$ implies

$$p(\sigma^2) \propto \sigma^{-2}$$

uniform for σ^2 $p(\sigma^2) \propto 1$ implies

$$p(\eta) \propto \exp(\eta)$$

Another example

Suppose $X \sim \text{Bin}(n, \theta)$. Want a prior for θ.

- clearly $\theta \in (0, 1)$
- flat prior: uniform $p(\theta) = 1$

Consider reparameterization using log-odds ratio:

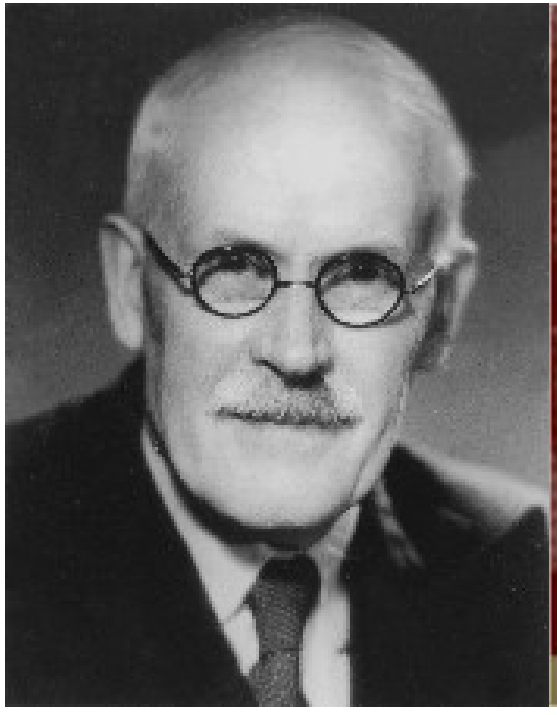
$$\eta = \log \frac{\theta}{1-\theta}$$

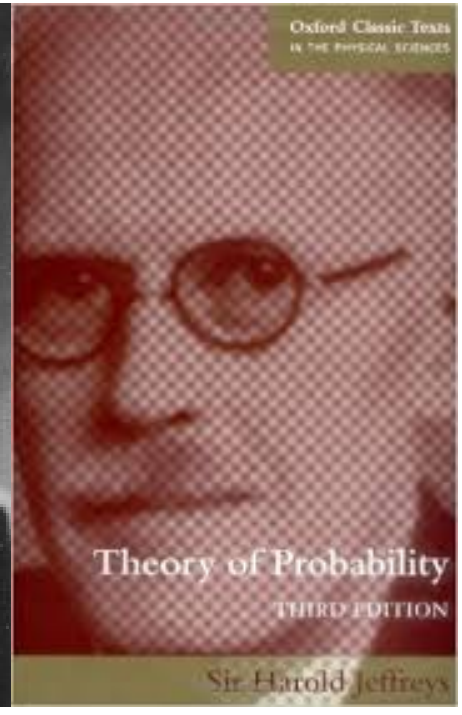
- valid reparameterization–natural for mapping θ to the real line
- but the prior $p(\eta) = 1$ is not flat

In original parameterization, $p(\cdot)$ is noninformative, but it is informative in new parameterization.

$\Rightarrow$ not invariant under one-to-one reparameterizations!

Harold Jeffreys (1891-1989)





Idea: use principle of invariance w.r.t. one-to-one transformations

Invariance Argument

Suppose we have a likelihood and some data.

- should be able to get a prior for θ using the likelihood only
- note: by contrast, a *subjective* Bayesian would first choose a prior, then apply it to the likelihood to derive the posterior

Answer:

$$\pi_J \propto \sqrt{\det I(\theta)}$$

Explanation Part I

Define a new parameter $\eta = h(\theta)$, $h(\cdot)$ one-to-one. For simplicity, assume θ, η are scalar.

- if we calculate $\pi_J(\theta)$ w.r.t. θ, then transform variables, we will get a prior π on η by change-of-variables formula
- if this prior $\pi(\eta)$ is the same as $\pi_J(\eta)$, that would be computed using η from the beginning, then the Jeffreys rule is invariant under one-to-one transformations

Explanation Part II

Apply Jeffreys's rule to η and use chain rule to re-express in terms of θ:

$$\begin{aligned} I(\eta) &= -E\left[\frac{d^2 \log f(X;\eta)}{d\eta^2}\right] \\ &= -E\left[\frac{d^2 \log f(X;\theta)}{d\theta^2}\left(\frac{d\theta}{d\eta}\right)^2 + \frac{d \log f(X;\theta)}{d\theta}\frac{d^2\theta}{d\eta^2}\right] \\ &= -E\left[\frac{d^2 \log f(X;\theta)}{d\theta^2}\right]\left(\frac{d\theta}{d\eta}\right)^2 + E\left[\frac{d \log f(X;\theta)}{d\theta}\right]\frac{d^2\theta}{d\eta^2}. \end{aligned}$$

We have exactly what we want, provided that

$$E\left[\frac{d \log f(X;\theta)}{d\theta}\right] = 0$$

Reminder

For all θ, $\int f(X;\theta)dX = 1$. Assume sufficiently regularity, differentiate w.r.t. θ

$$
\begin{aligned}
0 &= \frac{d}{d\theta}\int f(X;\theta)dX \\
&= \int \frac{df(X;\theta)}{d\theta}\frac{f(X;\theta)}{f(X;\theta)}dX \\
&= \int \left[\frac{df(X;\theta)}{d\theta}\frac{1}{f(X;\theta)}\right] f(X;\theta)dX \\
&= \int \left[\frac{d\log f(X;\theta)}{d\theta}\right] f(X;\theta)dX \\
&= E\left[\frac{d\log f(X;\theta)}{d\theta}\right]
\end{aligned}
$$

Need DCT on distributions.

Final Result

Taking expectation over X with fixed θ is equivalent to taking expectation with η fixed $\Rightarrow$

$$I(\eta) = I(\theta) \left(\frac{d\theta}{d\eta} \right)^2$$

take square root

$$\sqrt{I(\eta)} = \sqrt{I(\theta)} \left| \frac{d\theta}{d\eta} \right|$$

By change-of-variable formula, this shows Jeffreys prior $\pi_J(\theta) = \sqrt{I(\theta)}$ is invariant to a change of variable.

Example 1

Let $X \sim \text{Exp}(\lambda)$, so $f(x;\lambda) = \lambda e^{-\lambda x}$, $x > 0$

- $E(X) = \lambda^{-1}$, $\text{Var}(X) = \lambda^{-2}$, $E(X^k) = (k!)\lambda^{-k}$
- $\log f(x;\lambda) = \log \lambda - \lambda x$

$$
\begin{aligned}
I(\theta) &= E\left[\left(\frac{d \log f(x;\lambda)}{d\lambda}\right)^2\right] = E(\lambda^{-2} + X^2 - 2X\lambda^{-1}) \\
&= -E\left[\frac{d^2 \log f(x;\lambda)}{d\lambda^2}\right] = -E(-\lambda^{-2}) \\
&= \lambda^{-2}
\end{aligned}
$$

$\pi_J(\lambda) \propto \sqrt{I(\lambda)} = \lambda^{-1}$

Example 2

Suppose $X \sim \text{Bin}(n, \theta)$ (with n fixed). Then

$$f(x;\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

- also
$$\frac{d^2 \log f(x;\theta)}{d\theta^2} = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2},$$
- then
$$I(\theta) = \frac{n}{\theta(1-\theta)},$$

Therefore $\pi_J(\theta) \propto [\theta(1-\theta)]^{-1/2}$ which is a proper Beta$(1/2, 1/2)$ density

Example 3

Consider $X \sim \mathcal{N}(\mu, \sigma^2)$ and let $\theta = (\mu, \sigma^2)^T$. We have

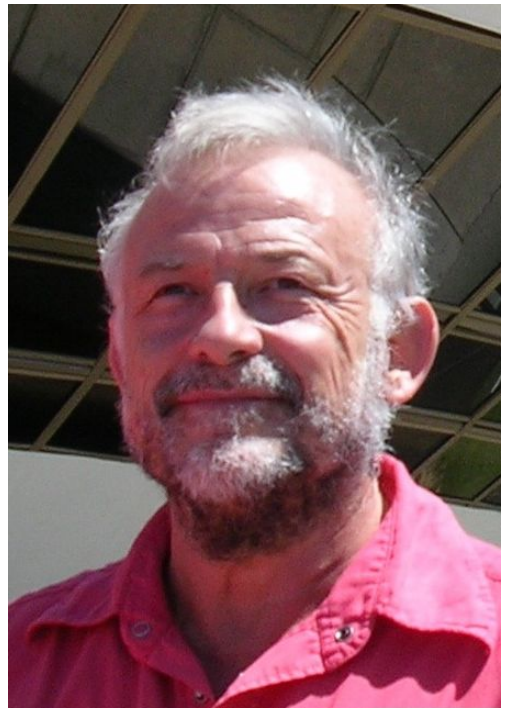
$$I(\theta) = E_\theta \begin{pmatrix} \sigma^{-2} & 2(X-\mu)\sigma^{-3} \\ 2(X-\mu)\sigma^{-3} & 3(\mu - X)^2\sigma^{-4} - \sigma^{-2} \end{pmatrix}$$
$$= \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-2} \end{pmatrix}$$

Thus

$$\pi_J(\theta) = \sqrt{\det I(\theta)} \propto \sigma^{-2}$$

Note: if μ and σ are assumed to be *a priori* independent, the corresponding prior would be σ^{-1} (can you show this?)

Reference Priors



Jose M. Bernardo

- formalize what is meant by 'uninformative' prior:
- it is a function that maximizes some measure of *distance* or **divergence** between the posterior and prior, as data are observed
- e.g. Hellinger distance, K-L divergence
- allows data to have maximum effect on the posterior

Preliminary Comments

Equivalent to Jeffreys prior in scalar (one-dimensional) parameter models.

- Question: how to maximimize divergence between posterior and prior before seeing the data?
- Reference prior proposal: take the **expectation** of the divergence, given a sampling model
- notice this sounds a bit frequentist: base inference on 'imagined'/'hypothetical' (not yet observed) data
- But once prior is chosen, proceed as usual with Bayesian inference—in frequentist inference you continue to imagine repeated sampling even after observing the actual data.

Nuisance Parameters

The reference prior approach distinguishes between parameters of interest and nuisance parameters.

- suppose $X \sim f(X;\theta)$ where $\theta = (\theta_1, \theta_2)$, and θ_1 is the interest parameter
- to find reference prior, first define $\pi(\theta_2|\theta_1)$ as the Jeffreys prior associated with $f(x;\theta)$ when θ_1 is fixed
- then derive the marginal distribution

$$\tilde{f}(x;\theta_1) = \int f(x;\theta_1,\theta_2)\pi(\theta_2|\theta_1)d\theta_2$$

- then compute the Jeffreys prior $\pi(\theta_1)$ associated with $\tilde{f}(x;\theta_1)$
- Principle: eliminate the nuisance parameter by using a Jeffreys prior where the parameter of interest remains fixed

K-L and Reference Priors

Let $T \equiv T(X)$ be a sufficient statistic for data X from likelihood $f(x;\theta)$.

- the K-L divergence between the posterior and prior is

$$\int p(\theta|t) \log \frac{p(\theta|t)}{\pi(\theta)} d\theta$$

- reference prior is $\pi(\cdot)$ that maximizes the expected value of this as $n \to \infty$
- the expectation of the divergence is taken under the marginal distribution of the sufficient statistic T (or of the data)

Probability Matching Priors

Consider a scalar θ. Let $C_x \equiv C_x(\pi(\theta), X)$ be a $1-\alpha$ posterior credible set, i.e.

$$\Pr_{\theta|X}(\theta \in C_x) = 1 - \alpha.$$

A probability matching prior $\pi_M(\theta)$ is chosen so that

$$\Pr_\theta(\theta \in C_x(\pi(\theta), X)) = 1 - \alpha + O(n^{-(j+1)/2})$$

- $\Pr_\theta$ is frequentist probability under repeated sampling of X
- $j = 0$ is true for all smooth priors