Bayes Stat

27 December 2019 16:20

(1) Base Rate Paradox:

Suppose $P(B|A_1) \gg P(B|A_2)$; Does that imply $P(A_1|B) > P(A_2|B)$

$\Rightarrow$ NO!! the base matters

From Bayes formula we know:

$$P(B|A_i) = \frac{P(B, A_i)}{P(A_i)} = \frac{P(A_i|B)\cdot P(B)}{P(A_i)}$$

$$P(B|A_k) = \frac{P(A_k|B)\cdot P(B)}{P(A_k)}$$

hence the nominator can be relatively small but if the rate really small $P(B|A_i)$ very high! so the base might influence the conditional prob. more than $P(A_i|B)$!

$$\frac{P(B|A_i)}{P(B|A_k)} = \frac{P(A_i|B)}{P(A_k|B)} \cdot \frac{P(A_k)}{P(A_i)}$$

the base influence the relative size of the cond. prob.

(2) Point Estimation and decision Theory

Given that you have obtained a posterior distribution the question is how to choose the parameter out of it.

For this reason it makes sense to refer back to decision Theory; it is then possible to choose a loss function and choose the parameter as the one minimising the expected loss for the function. For instance choosing:

$$L: \Theta \in \mathbb{R}^n \to \mathbb{R}$$

(i) $L := (d-\theta)^2$ (squared Loss) $\Rightarrow \hat{\theta} = \arg\min_d E((d-\theta)^2) \Rightarrow d_{Bayes} = E(\theta|x)$

(ii) $L := |d-\theta|$ (absolute Loss) $\Rightarrow \hat{\theta} = \arg\min_d E(|d-\theta|) \Rightarrow d_{Bayes} = \frac{1}{2} \int_{-\infty}^{\infty} \pi(\theta)\cdot f(x|\theta)\, d\theta = \int_{\theta}^{\infty} \pi(\theta)\cdot f(x|\theta)\, d\theta$

(iii) $L := \mathbb{1}_{[\theta-\varepsilon,\theta]^c}\cdot(d-\theta)$ (mode Loss) $\Rightarrow \hat{\theta} = \arg\min_d E(|\mathbb{1}_{[\theta,\varepsilon]^c}\cdot(d-\theta)|) \Rightarrow d_{Bayes} = \arg\max_{\theta} \int_{-\infty}^{\infty} \pi(\theta) f(x|\theta)\, d\theta$

We can summarize this approach choosing:

$$\arg\min_d p(\theta,d) \quad ; \quad \text{where} \quad p(\theta,d) = E_\theta(L(\theta,d)|X)$$

$$= \arg\min_d \int_{-\infty}^{\infty} L(\theta,d)\cdot f(\theta|x)\, d\theta$$

Notice that this stays in contrast to frequentist statistics, where we take θ as random and we try to minimize:

$$\hat{\theta} = \arg\min_\theta E_X(R(\theta,d)) = \int_{\mathcal{X}} p(\theta,d)\cdot f(x|\theta)\, dx$$

here we ideally would like to get an estimator minimizing the risk $\forall\ \theta$. This is however not always possible θ but this is not always possible.

To solve this issue we might want to take different approaches:

(i) minimax
(ii) admissibility
(iii) weighted risk.

we showed in Mathstat that the two classes are guaranteed in a vast amount of cases. It is especially possible to prove that in the limit / under each admissible statistic is a Bayes estimator.

Looking at (iii) we have:

$$E(R(\theta,\nu)) = \int_\Theta R(\theta,d)\cdot w(\theta)\, d\nu(\theta)$$

it is then possible to show that this is equivalent to minimize the posterior Loss function described above.

To show this consider:

$$r(\theta,d) = \int_{\mathcal{X}}\int_{\Theta} L(\theta,d)\cdot w(\theta|x)\, d\theta\ f(x|\theta)\, dx$$

$$= \int_{\Theta}\int_{\mathcal{X}} L(\theta,d)\cdot f(x|\theta)\, dx\ \cdot w(\theta)\, d\theta$$

$$= \int_{\Theta} p(\theta,d)\cdot w(\theta)\, d\theta$$

Now given that your bayes decision is the one minimizing the

$d' = \arg\min_d E(p(\theta,d))$, we see

$$E(R(\theta,d)) \geq \int_\Theta p(\theta,d')\, w(\theta)\, d(\theta)$$

and you obtained the previously obtained result with $w(\theta) = \pi(\theta)$; i.e. you minimise the weighted risk where the weights corresponds to the prior distribution.

Bayes Tests

> Because in Bayesian statistics the parameter is random, it becomes possible to speak about the "probability that the null hypothesis is true" or the "probability that θ belongs to some interval". In frequentist statistics, such statements have no meaning, and one has to be very careful if one wants to explain the meaning of a p-value or a confidence interval in words.

To test a given hypothesis against its alternative it is now necessary to create an appropriate Bayes test minimizing the risk.

Defining the risk as the sum of expected Error of first and second type ; i.e.

$$R(\theta,d) = \begin{cases} E_{\theta_0} L(\theta_0, A) & \theta = \theta_0 \\ 1 - E_{\theta_1} L(\theta_0, d) & \theta = \theta_1 \end{cases}$$

Look at fundamental: there we checked at the Bayes test minimizing the expected loss. The test specifies the loss function there; here it is given

looking for the test minimizing the loss

$$\arg\min_{\phi} l(\theta, d) = \phi \int_{\Theta_0} w(\vartheta_0 | x)\, d(\vartheta_0) + (1-\phi) \int_{\Theta_1} w(\vartheta_1 | x)\, d\vartheta_1$$

hence

$$\phi = \begin{cases} 1 & \text{if} \quad w(\vartheta_1 | x) > w(\vartheta_0 | x) \\ q & \text{if} \quad w(\vartheta_1 | x) = w(\vartheta_0 | x) \quad \text{where } q = [0,1] \\ 0 & \text{if} \quad w(\vartheta_1 | x) < w(\vartheta_0 | x) \end{cases}$$

Hence given the risk function above we search for a rejection criteria

$$\pi(\Theta_0 | x) > c \Rightarrow \text{accept } H_0 \quad \text{else: reject } H_0$$

that minimizes the risk; i.e. a_1 := type I loss a_2 := type II loss

Then if $\phi = 1$ reject H_0; with expected loss $a_1 \cdot \pi(\Theta_0 | x)$

and when $\phi = 0$; i.e. accept H_0; expected loss $a_2 (1 - \pi(\Theta_0 | x))$

(a_2: loss from accepting H_0 when true; $(1-\pi(\Theta_0|x))$: prob H_A)

Then we reject H_0 if

$$a_1 \cdot \pi(\Theta_0 | x) > a_2 (1 - \pi(\Theta_0 | x))$$

$$\pi(\Theta_0 | x) > \frac{a_2}{a_1 + a_2}$$

□

Instead of having the acceptance according to the $\pi(\Theta_0 | x)$ statistic and the corresponding critical value; it is also possible to look at the so called Bayes factor looking how much likely is the posterior probability of observing the Null compared to the alternative:

$$\frac{\pi(\Theta_0 | x)}{\pi(\Theta_0)} = \frac{f(x | \Theta_0) \cdot \pi(\Theta_0)}{\pi(\Theta_0) \cdot f(x)}$$

$$\frac{\pi(\Theta_1 | x)}{\pi(\Theta_1)} = \frac{f(x | \Theta_1) \cdot \pi(\Theta_1)}{\pi(\Theta_1) \cdot f(x)}$$

this is why in case of independent $\Theta_0 \wedge \Theta_1$ and simple hypothesis $H_0: \theta = \theta_0$ vs. $H_A: \theta \neq \theta_0$ this simplifies to the ratio of likelihoods and to the NP-test

It is then possible to state that if the above ratio is $[1/3, 1]$ there is mild evidence for H_0; if it is $[0{,}01, 1/3]$ there is substantial evidence against the Null and if $< 0{,}01$ there is strong evidence against the Null.

In many applications the null hypothesis consists of a subset of Lebesgue measure zero, typically a lower dimensional subset of Θ, e.g. in the case of $\mathcal{N}(\mu, \sigma^2)$-observations $\Theta_0 = \{(\mu, \sigma^2); \mu = \mu_0\}$. In this case, if we choose a prior which has a density w.r. to the Lebesgue measure, the prior and the posterior give zero probability to the null hypothesis. Hence there would be no need to collect data as data cannot change the prior belief in Θ_0.

In such situations one should therefore choose a prior which assigns to Θ_0 a probability strictly between 0 and 1. This can be achieved by a mixture

this then will insure that probability at the null $\neq 0$.

$$\pi(d\theta) = \rho_0 \pi_0(d\theta) + (1 - \rho_0)\pi_1(\theta) d\theta$$

where π_0 is a distribution which is concentrated on Θ_0 and ρ_0 is the prior probability of Θ_0. Because π_0 cannot have a density, we use here the general notation of measure theory. With such a prior, the posterior probability of Θ_0 is

$$\pi(\Theta_0 \mid x) = \frac{\rho_0 \int_{\Theta_0} f(x \mid \theta) \pi_0(d\theta)}{\rho_0 \int_{\Theta_0} f(x \mid \theta) \pi_0(d\theta) + (1 - \rho_0) \int_{\Theta} f(x \mid \theta) \pi_1(\theta) d\theta}.$$

Note that whether testing in the case of such a point null hypothesis is reasonable in the first place or not is currently controversially debated.

In frequentist statistics, the p-value is sometimes taken as a measure of evidence against the null hypothesis. It is defined as the smallest significance level for which the null hypothesis is still rejected. Although conceptually this is not the same as the posterior probability of the null hypothesis, it would be nice if these two measures were close at least in the case where the null and the alternative are a priori equally likely. Let us consider the case where $\Theta_0 = \{\theta_0\}$. Then for $\rho_0 = \frac{1}{2}$

$$\pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \int_\Theta f(x \mid \theta)\pi_1(\theta)d\theta}$$

which depends on the chosen prior for the alternative, but there is the trivial lower bound

$$\inf_{\pi_1} \pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \sup_\theta f(x \mid \theta)}.$$

i.e. when the prior is heavily biased towards the alternative the p-value is way too conservative.

Numerical comparisons show that in many situations this lower bound is substantially larger than the p-value, see Table 1.1. This means that the p-value overestimates the evidence against the null even when the prior is heavily biased towards the alternative. If we assume θ to be scalar and one restricts π_1 to the class $\mathcal{S}$ of symmetric unimodal densities, then one can show that

because $\rho_0 = 1/2$ assumed above.

$$\inf_{\pi_1 \in \mathcal{S}} \pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \sup_c \frac{1}{2c} \int_{\theta_0 - c}^{\theta_0 + c} f(x \mid \theta)d\theta}.$$

A Bayesian confidence set with level $1-\alpha$ is called a $(1-\alpha)$-*credible set*. It is a subset $C_x \subset \Theta$ (depending on x) such that

probability of C_x given parameter and data

$$P(\theta \in C_x \mid X = x) = \pi(C_x \mid x) \geq 1 - \alpha.$$

Among the many $(1-\alpha)$-credible sets, the one minimizing the volume (Lebesgue measure) is particularly attractive. It is obtained by taking C_x as a level set of the posterior, $C_x = L_{k_\alpha}$, where

$$L_k = \{\theta; \pi(\theta \mid x) \geq k\}, \quad k_\alpha = \sup\{k; \pi(L_k \mid x) \geq 1 - \alpha\}.$$

It is thus called a highest posterior density (HPD) credible set. That it minimizes the volume can be seen as follows. For simplicity, we assume that $\pi(L_{k_\alpha} \mid x) = 1 - \alpha$. If C is another $(1-\alpha)$-credible set, then by the definition of the level set

are we assuming we have such sets.

$$\begin{aligned} 0 \overset{\text{because of the supremum characteristic above in the definition of } k_\alpha}{\geq} & \pi(L_{k_\alpha} \mid x) - \pi(C \mid x) = \int_{L_{k_\alpha} \cap C^c} \pi(\theta \mid x)d\theta - \int_{L_{k_\alpha}^c \cap C} \pi(\theta \mid x)d\theta \\ \geq & \; k_\alpha \left(|L_{k_\alpha} \cap C^c| - |L_{k_\alpha}^c \cap C| \right) \end{aligned}$$

where $|C|$ denotes the volume (Lebesgue measure) of a set C. In high dimensions, the computation of L_{k_α} can be difficult.

Bayes Asymptotics

Bayesian asymptotics says – again under regularity conditions – that for any smooth prior which is strictly positive in a neighborhood of θ_0

$$\theta \mid (x_1, \ldots, x_n) \overset{\text{approx}}{\sim} \mathcal{N}\left(\hat{\theta}_n, \frac{1}{n} I(\hat{\theta}_n)^{-1}\right).$$

Therefore the influence of the prior disappears asymptotically and the posterior is concentrated in a $\sqrt{1/n}$ neighborhood of the MLE. There is a nice symmetry in the asymptotic statements, but note again the difference in what is considered fixed and what is random in the two approaches.

This is known as the Bernstein von Mises Theorem and it is hence possible to see that the influence of the prior disappears asymptotically and the posterior is concentrated in an $1/\sqrt{n}$ neighborhood of the MLE $\hat{\theta}_n$

Likelihood Principle

Chapter 2

The choice of a prior is a point which has lead to an intensive debate and which is often considered to be the weak point of the Bayesian approach. We discuss three approaches. The first one chooses prior distributions such that the posterior can be easily computed. The second one tries to determine a prior which contains as little information as possible. The third one tries to choose a prior based on the opinion of one or several experts.

Definition 2.1. *A parametric family $\mathcal{P}_\Xi = \{\pi_\xi(\theta); \xi \in \Xi\}$, $\Xi \subset \mathbb{R}^q$, of prior densities is called* conjugate *for the model $\{f(x \mid \theta); \theta \in \Theta\}$ if, for any $\pi \in \mathcal{P}_\Xi$ and any x, $\pi(\theta \mid x)$ is again in $\mathcal{P}_\Xi$.*

prior — posterior

Written out, this means that to any $\xi \in \Xi$ and any x there must be a $\xi' = \xi'(\xi, x)$ such that

$$\pi_\xi(\theta) f(x \mid \theta) \propto \pi_{\xi'}(\theta).$$

Computing the posterior amounts then to computing $\xi'(\xi, x)$?

It is obvious that $\mathcal{P}_\Xi$ is conjugate if the following two conditions are satisfied

1. To any x there is a $\xi(x) \in \Xi$ such that $f(x \mid \theta) \propto \pi_{\xi(x)}(\theta)$.
2. To any pair $\xi_1, \xi_2 \in \Xi$ there is a $\xi_3 \in \Xi$ such that $\pi_{\xi_1}(\theta)\pi_{\xi_2}(\theta) \propto \pi_{\xi_3}(\theta)$.

A class of conjugate priors $\mathcal{P}_\Xi$ remains conjugate under repeated sampling i.e. it is also conjugate for the model where $X_1, \ldots, X_n$ are i.i.d., $X_i \sim f(x \mid \theta)dx$ and n is arbitrary because for instance

$$\pi(\theta \mid x_1, x_2) \propto \pi(\theta \mid x_1) f(x_2 \mid \theta).$$

It follows that if:

If $\mathcal{P}_\Xi$ is conjugate for $f(x \mid \theta)$, for arbitrary, but fixed ξ_0, we can write

$$\prod_{i=1}^{n} f(x_i \mid \theta) = \frac{\pi_{\xi_n(x_1,\ldots,x_n)}(\theta)}{\pi_{\xi_0}(\theta)} f_n(x_1, \ldots, x_n)$$

where

- f_n is the prior predictive density of $X_1, \ldots, X_n$
- ξ_n maps n-tupels of observed values to Ξ

- ξ_n is a sufficient statistic whose dimension is independent of n

Proof:

- One can show that if the set $\{x; f(x \mid \theta) > 0\}$ does not depend on θ, exponential families are the only class of distributions which allow for sufficient statistics whose dimension is independent of n

- An **exponential family** has densities of the following form

$$f(x \mid \theta) = \exp(c_1(\theta)T_1(x) + \ldots c_q(\theta)T_q(x) + d(\theta))h(x)$$

The conjugate family consists then of densities

$$\pi_\xi(\theta) \propto \exp(c_1(\theta)\xi_1 + \ldots c_q(\theta)\xi_q + d(\theta)\xi_{q+1})$$

On the Hyperparameter issue

Conjugate priors have again parameters, usually called *hyperparameters*, which have to be chosen (In the general formula the hyperparameters are called ξ_i, in Table 2.1 different

symbols are used). Hence using a conjugate prior does not answer the question "which prior ?". As typically there are more hyperparameters than parameters, choosing a hyperparameter seems even more difficult than choosing a parameter value. Note however that usually one of the hyperparameters can be regarded as a hypothetical sample size of the prior: In the general case, it is the parameter ξ_{q+1}, in the binomial case it is $\alpha + \beta$, in the Poisson case it is γ, in the normal case it is n_0 for μ and γ for τ. So this parameter can be determined by asking how much we want to rely on the prior. The other parameters usually are related to a location parameter of the prior which helps to choose its value. For instance in the case of a Beta distribution, the mean is $\alpha/(\alpha + \beta)$.

2.2 Non-informative priors

The search for a non-informative prior is motivated by the wish to reduce the subjective element in a Bayesian analysis. A first attempt defines a uniform prior on Θ to be non-informative: This has however two drawbacks. First, the uniform distribution on Θ is a probability distribution only if Θ has finite volume. Second, the uniform distribution is not invariant under different ways to parametrize the same family of distribution. Assume we use $\tau = g(\theta)$ as our new parameter where g is invertible and smooth, i.e. $\theta = g^{-1}(\tau)$. If θ has density π, then by the change-of-variables formula, $\tau = g(\theta)$ has the density

$$\lambda(\tau) = \pi(g^{-1}(\tau))|\det Dg(g^{-1}(\tau))|^{-1},$$

check this link for under. formula

https://stats.stackexchange.com/questions/239588/derivation-of-change-of-variables-of-a-probability-density-function

where Dg is the matrix whose (ij)-th entry is $\partial g_i/\partial\theta_j$, the so-called Jacobi matrix. Hence, if π is constant and g is not linear, then λ is not constant.

2.2.1 Improper priors

We call a (σ-finite) measure π on Θ which ~~is not a probability measure~~, i.e.,

$$\int_\Theta \pi(\theta)d\theta = \infty,$$

an improper prior.

So despite your prior is improper, as long as the normalizing function will be a proper posterior will be a proper defined pdf due to the constant

If π has infinite total mass, $\pi(\theta)f(x \mid \theta)$ can have both finite or infinite total mass, depending on the likelihood. If the total mass is finite, then we have by formal analogy the posterior density

$$\pi(\theta \mid x) = \frac{\pi(\theta)f(x \mid \theta)}{\int \pi(\theta')f(x \mid \theta')d\theta'}$$

and we can construct Bayesian point estimates, tests and credible intervals as before. Typically, this can be justified by approximating the improper prior by a sequence of proper priors π_k and showing that the associated sequence of posteriors $\pi_k(\theta \mid x)$ converges to the above expression. However, even if this convergence holds, paradoxes can occur. Moreover, in complicated models it is not always easy to check whether $\pi(\theta)f(x \mid \theta)$ has finite total mass.

this is how you explicitly talk about

kind: Extended Bayes

2.2.2 Jeffreys prior

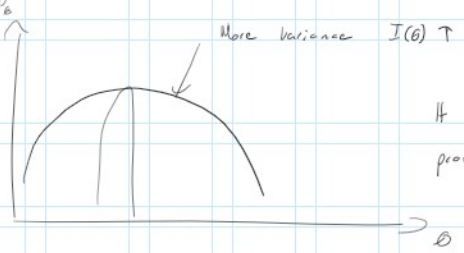
Jeffreys proposed to take

$$\pi(\theta) \propto \det(I(\theta))^{1/2}$$

to check this how much a relative change the volume of a matrix.

Heuristically this means

$\log P_\theta$



$I(\theta)\uparrow \Rightarrow$ higher stability of the estimator.

$J_\theta = E(\zeta_\theta^T S)$

It follows intuitively and can be proved that

$$\sqrt{N}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, I(\theta)^{-1})$$

Hence taking

$\pi(\theta) \propto \det(I(\theta)^{1/2})$ we set a higher prior where the volume of the $I(\theta)$ is higher and hence the estimation is less affected by the choice of θ and a lower prior where the parameter θ is most effective, leaving the posterior decision mostly to the data.

On top of the above mentioned important property of the prior as being non-informative we have the second nice property that the Jeffrey's prior is equivariant by reparametrization of the prior dist.

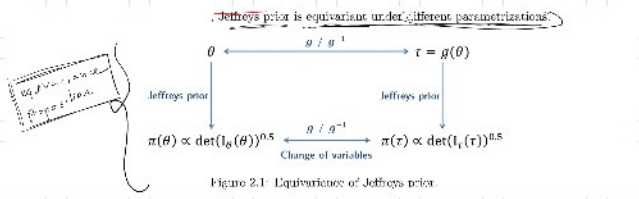


Figure 2.1: Equivariance of Jeffreys prior

It is possible to prove the above mentioned equivariance of the Jeffery's prior by using the change of variable formula; i.e.

$\theta \sim \pi(\cdot)$ $\quad \tau \sim g(\theta)$. $\quad \theta \sim g^{-1}(\tau)$ } 1:1 map.

then

$F_\pi(Y \leq \tau) = F_\pi(g(\theta) \leq \tau)$ } CDF

$= F_\pi(\theta \leq g^{-1}(\tau))$

$F_\tau(\tau) = F_\pi(g^{-1}(\tau))$

taking the first derivative

$$P_\tau(y) = P_\pi(\theta) \cdot \left| \frac{1}{g'(g^{-1}(\tau))} \right|$$ considering the two possible cases

In higher dimension this translates to

$$P_\tau(y) = P_\pi(\theta) \cdot |\det D \cdot g^{-1}(\tau)|$$

$P_T(y) = P_\theta(\theta) \cdot |\det D g^{-1}(\tau)|$

The argument in the general multiparameter case $\tau = g(\theta)$, where g is one-to-one and differentiable, goes as follows. Denoting the Jacobi matrices of g and g^{-1} by Dg and Dg^{-1}, respectively, the chain rule implies

$$\frac{\partial}{\partial \tau} \log f(x \mid g^{-1}(\tau)) = (Dg^{-1}(\tau))^T \frac{\partial}{\partial \theta} \log f(x \mid g^{-1}(\tau)).$$

Hence, the Fisher information with respect to τ is → change of variable

$$I_\tau(\tau) = (Dg^{-1}(\tau))^T I_\theta(g^{-1}(\tau)) Dg^{-1}(\tau).$$

as $I_\theta = E(s_\theta^T s_\theta) = E\left(\frac{\partial}{\partial \theta} \log p_\theta^T \cdot \frac{\partial}{\partial \theta} \log p_\theta\right)$

Moreover, $Dg^{-1}(\tau)$ is the inverse matrix of $Dg(g^{-1}(\tau))$ and therefore Jeffreys prior for τ is proportional to

$$\det(I_\tau(\tau))^{1/2} = |\det Dg(g^{-1}(\tau))|^{-1} \det(I_\theta(g^{-1}(\tau)))^{-1/2},$$

in accordance with the above result for the transformation of densities.

Taking the square and the determinant of it

For a scalar parameter, Jeffreys prior is usually a good choice, although it violates the likelihood principle because the Fisher information contains an integral over X. However, for vector parameters, it can have undesirable features.

For instance we showed in class a Bayes Estimator that does not satisfy the consistency property.

$\Rightarrow \pi((\mu, \sigma)) \neq \pi(\mu)\pi(\sigma)$

product of univariate Jeffery's priors not equal to multivariate Jeffery's prior.

Important property of Jeffreys Prior

Reference Prior

It has two aspects:

(1) A new justification for Jeffreys prior

(2) Distinction between parameters of interest and nuisance parameters

We begin with the former and call a prior π non-informative if the difference between the prior π and the posterior $\pi(. \mid x)$ is maximized in some distance. This seems reasonable because if the data x have the largest possible impact, the impact of the prior is minimal.

- There are **two issues**:
  1. Choice of distance
  2. Dependence on data
- Bernardos proposal:
  1. Use Kullback-Leibler divergence
  2. Integrate over the data according to the prior predictive distribution $f(x) = \int_\Theta f(x \mid \theta)\pi(\theta)d\theta$

Result

The **Kullback-Leibler divergence** between two densities f and g is defined as

$$K(f,g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

weighting function

if equal $\log(1) = 0$

- This is not a true distance since in general $K(f,g) \neq K(g,f)$
- It satisfies $K(f,g) \geq 0$ and $K(f,g) = 0$ iff $f(x) = g(x)$ for almost all x

Bernardos idea is to choose π such that is maximizes the expected Kullback-Leibler divergence:

integration over data.

$$\begin{aligned} I(X,\theta) &= \int_X f(x) \underbrace{\int_\Theta \pi(\theta \mid x) \log \frac{\pi(\theta \mid x)}{\pi(\theta)} d\theta}_{KL(\pi(\theta \mid x), \pi(\theta))} dx \\ &= \int_\Theta \int_X \underbrace{\pi(\theta) f(x \mid \theta)}_{=\pi(x,\theta)} \log \frac{\pi(\theta) f(x \mid \theta)}{\pi(\theta) f(x)} dx d\theta \end{aligned}$$

$\pi(\theta \mid x)$

Also a Kullback Leibler

This is called mutual information in information theory.

Issues

- **Maximizing $I(X,\theta)$ is** often **unfeasible**
  - Finding the maximizer of $I(X,\theta)$ is complicated and there is in general no closed form solution
  - The resulting distribution $\pi(\theta)$ typically has a finite support which is a very undesirable property for a prior that is thought to be non-informative

Solution:

Use an possibly infinite sample of data and calculate the mutual information based on it.

- Assume n i.i.d. observations $X_1, \ldots, X_n$ with density $f(x \mid \theta)$
- Denote the corresponding mutual information by $I((X_1, \ldots, X_n), \theta)$
- Let n go to infinity and choose $\pi(\theta)$ that maximizes

$$I_\infty(\pi) = \lim_{n \to \infty} I((X_1, \ldots, X_n), \theta)$$

so that support vanishes.

Then

- Still a problem: $I_\infty(\pi)$ is usually infinite
- Remedy: appropriately standardize the mutual information

Given the two:

- We obtain the following approximation for the standardized mutual information

$$I((X_1, \ldots, X_n), \theta) - \frac{p}{2} \log\left(\frac{n}{2\pi e}\right) \approx \int_\Theta \pi(\theta) \log \frac{\det I(\theta)^{1/2}}{\pi(\theta)} d\theta$$

- This is maximal for $\pi(\theta) = c^{-1} \det I(\theta)^{1/2}$. We thus have again Jeffreys prior in the limit

normalizing constant. Jeffrey's Prior.

Bernardo's approach for nuisance parameters

Often, the parameter $\theta = (\theta_1, \theta_2)$ can be decomposed in **parameters of interest θ_1 and nuisance parameters θ_2**.

- Nuisance parameters are parameters which we are not of primary interest when doing statistical inference (e.g. scale / variance parameters).

Bernardo's approach:

1. Condition on θ_1 and find Jeffreys prior for $\pi(\theta_2 \mid \theta_1)$
2. Calculate
$$f^*(x \mid \theta_1) = \int_{\Theta_2} f(x \mid \theta) \pi(\theta_2 \mid \theta_1) d\theta_2$$
and find Jeffreys prior for $f^*(x \mid \theta_1)$
3. Set $\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2 \mid \theta_1)$

Jeffreys prior obtained above.

- $\pi(\theta_2 \mid \theta_1)$ needs to be a proper prior in order that $f^*(x \mid \theta_1)$ is a probability density
- Workaround in this case:
  - Construct a sequence of compact subsets $\Theta_1^1 \subseteq \Theta_1^2 \subseteq \cdots \subseteq \Theta_1$ and determine corresponding reference priors for θ_1
  - Obtain $\pi(\theta_1)$ as the limit of this sequence

2.3 Expert priors

- Idea: elicit a prior from one or several experts
- Challenge: expert judgement is subject to various kinds of heuristics and biases. The size of unwanted effects depends strongly on how questions are phrased
- Procedure for a univariate prior:
  1. Elicit a number of summary statistics (e.g., the median and the quartiles or the 33% and 67% quantiles)
  2. Fit a distribution which takes these summaries into account

2.4 Concluding Remarks on Priors

- Non-informative priors are difficult to implement in complex models with many parameters $\Rightarrow$ some subjective choices are often unavoidable
- If there is enough data, any reasonable choice leads to similar conclusions because the likelihood tends to dominate

} Bernstein Von-Mises Theorem.

- In any practical application, one should
  - check that, at least marginally, the **prior is approximately constant in a highest probability density credible set**
  - or do a **sensitivity analysis** by varying the prior

} this you good if ex-post, so that you will have to run the analysis first.

} i.e. where the mass will be concentrated the prior should have a constant effect without influencing this too much.

Connection between regularization and prior

- If the **number of parameters is large compared to the number of observations, the prior often matters**. This seems unavoidable
- In that situation, frequentist statistics often uses **regularization methods** which usually have a Bayesian interpretation

Connection between regularization and prior

- If the **number of parameters is large compared to the number of observations, the prior often matters**. This seems unavoidable
- In that situation, frequentist statistics often uses **regularization methods** which usually have a Bayesian interpretation
- For instance, if we use penalized maximum likelihood estimation

$$\widehat{\theta} = \arg\max(\log f(x \mid \theta) + P(\theta))$$

the penalty $P(\theta)$ can usually be interpreted as the log of a prior density

Chapter 3

Hierarchical Bayes models

- In hierarchical models, the prior $\pi(\theta \mid \xi)$ for θ depends on other parameters ξ, called **hyperparameters**, which also have a prior distribution $\pi(\xi)$
- This leads to a triple of random variables (ξ, θ, x) with joint density

$$\pi(\xi)\pi(\theta \mid \xi)f(x \mid \theta)$$

Proof:
Idea is that you can start from the usual equation.

$$\pi(\xi,\theta \mid x) = \frac{\pi(\theta,\xi,x)}{\text{marginal}(x)} = \frac{\pi(x \mid \xi,\theta)\cdot f(\theta,\xi)}{\text{marginal}}$$

$$= \frac{\pi(x \mid \xi,\theta)\cdot f(\theta \mid \xi)\cdot \pi(\xi)}{\text{marginal}}$$

given that conditioning on θ no further information added by ξ

$$= \frac{\pi(x \mid \theta)\cdot f(\theta \mid \xi)\cdot \pi(\xi)}{\text{marginal}}$$

The basic approach of Bayesian statistics remains unchanged: Once we have specified the three factors of the joint distribution of the triple (ξ, θ, x), we compute the posterior, that is the conditional distribution of the unobserved variables (ξ, θ) given the observed variables by the rules of probability, and then we base our conclusions on this posterior.

Often, the primary interest is in the original parameter θ and then we need the marginal posterior $\pi(\theta \mid x)$. There are two ways to compute it.

- Method 1

In the first approach, we begin by computing the marginal prior

$$\pi(\theta) = \int \pi(\theta \mid \xi)\pi(\xi)d\xi$$

$\xi = \pi(\theta) = \int \pi(\theta,\xi)\,d\xi$ by def. } follows hierarchical structure broken down piece by piece.

and then use Bayes formula

$$\pi(\theta \mid x) \propto \pi(\theta)f(x \mid \theta).$$

This shows in particular that the introduction of hyperparameters is equivalent to a special choice of a prior for θ.

- Method 2

There are however situations where the approach based on the following formula is computationally easier: (for proportionality)

$$\pi(\theta \mid x) = \int \pi(\theta \mid x,\xi)\pi(\xi \mid x)d\xi \propto \int \pi(\theta \mid x,\xi)\pi(\xi)f(x \mid \xi)d\xi,$$

. Further, $\pi(\xi \mid x)$ can be calculated either by marginalizing the joint posterior over θ

$$\pi(\xi \mid x) = \int \pi(\theta,\xi \mid x)d\theta$$

or

or by using

$$\pi(\theta \mid x) = \int \pi(\theta,\xi \mid x)\,d\xi = \int \pi(\xi \mid x)\cdot\pi(\theta \mid \xi,x)\,d\xi$$

$$\rightarrow \quad \pi(\xi \mid x) = \frac{\pi(\theta,\xi \mid x)}{\pi(\theta \mid \xi,x)}$$

Similarly, $f(x \mid \xi)$ can be obtained by marginalizing the joint distribution of x and θ given

$$f(x \mid \xi) = \int \pi(x,\theta \mid \xi)d\theta = \int f(x \mid \theta)\pi(\theta \mid \xi)d\theta.$$

3.1 Empirical Bayes

- Recall that the marginal posterior can be computed as

$$\pi(\theta \mid x) \propto \int \pi(\theta \mid x, \xi) f(x \mid \xi) \pi(\xi) d\xi$$

- Instead of approximating this integral, the **empirical Bayes method** uses

$$\pi(\theta \mid x) \approx \pi(\theta \mid x, \widehat{\xi}(x)) = \frac{f(x|\theta)\pi(\theta \widehat{\xi}(x))}{f(x \mid \widehat{\xi}(x))}$$

where

$$\widehat{\xi}(x) = \arg\max_{\xi} f(x \mid \xi)$$

and

$$f(x \quad \xi) = \int f(x|\theta)\pi(\theta \xi) d\theta$$

- $\widehat{\xi}(x)$ is the marginal maximum likelihood estimator of the hyperparameter

- Instead of taking a weighted average, one takes the value with maximal weight (assuming that $\pi(\xi)$ is flat around $\widehat{\xi}(x)$)

This MAP estimator of fundamental stat.

Advantage

- This method avoids not only the computation of the integral, but also the choice of a hyperprior $\pi(\xi)$

Drawbacks

- The data x is used twice:
  1. First, to select the prior $\pi(\theta \mid \widehat{\xi}(x))$
  2. Then, to compute the posterior according to Bayes formula
  - This is somewhat undesirable from a conceptual point of view
- In general, the uncertainty is underestimated

Important for the oral !!

Conclusion

- Due to this, Bayesians often avoid empirical Bayes methods
- From a pragmatic point of view, empirical Bayes methods can be useful and have good frequentist properties

3.3 Linear Model

This chapter addresses the linear model in the bayesian frameworks.
We will deal with two estimations of the linear model:

(1) Parameter Estimation without variable selection := Bayesian Linear Model.
(2) Model selection via Bayesian Modeling := Bayesian Variable Selection.

Starting with (1) the setting is as usual:
Given a linear Model:

$$Y = \underbrace{\alpha 1}_{\text{single common intercept}} + X\beta + \varepsilon$$

you are interested in getting

$$\pi(\alpha, \beta, \sigma^2 \mid y)$$

To get it you can leverage the common Bayes Formulae:

$$\pi(\alpha, \beta, \sigma^2 \mid y) \propto f(y \mid \alpha, \beta, \sigma^2) \cdot \pi(\alpha, \beta, \sigma^2)$$

$$\pi(\alpha, \beta, \sigma^2) = \pi(\alpha, \beta \mid \sigma^2) \cdot \pi(\sigma^2) = \pi(\alpha \mid \beta, \sigma^2) \quad \pi(\beta \mid \sigma^2) \cdot \pi(\sigma^2)$$

Assuming that the intercept is independent of the variance of the error term and the β:

$$\pi(\alpha, \beta, \sigma^2) = \overset{(1)}{\pi}(\alpha) \cdot \overset{(2)}{\pi}(\sigma^2) \cdot \overset{(3)}{\pi}(\beta \mid \sigma^2)$$

It follows that to fully describe the model we need a prior for (1) - (3) above.

(1) - (2) Jeffreys prior $\quad \pi(\alpha) \propto 1$, $\pi(\sigma^2) \propto 1/\sigma^2$

(3) G-prior: $\quad \pi(\beta \mid \sigma^2) \sim N(\beta^0, g\sigma^2 (X^T X)^{-1})$

Given the three it is possible to solve above and obtain

We obtain the following marginal and conditional posteriors:

- $\beta_\gamma \mid y, \sigma^2 \sim \mathcal{N}\left(\frac{g}{g-1}\hat{\beta}_\gamma + \frac{1}{g+1}\beta_\gamma^0, \frac{g\sigma^2}{g-1}(X_\gamma^T X_\gamma)^{-1}\right)$
- $\alpha \mid y, \sigma^2 \sim N(\bar{y}, \frac{\sigma^2}{n})$
- $\sigma^2 \mid y \sim \text{Gamma}\left(\frac{n-1}{2}, \frac{g^{-1}+g^{-1}(\hat{\beta}_\gamma - \beta_\gamma^0)^T X_\gamma^T X_\gamma (\hat{\beta}_\gamma - \beta_\gamma^0)}{2}\right)$

- $g > 0$ is a hyperparameter which can be interpreted as a measure of the amount of information available in the prior relative to the data
- The g-prior arises as the posterior from a flat prior and a response vector $y = X_\gamma \beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ if the same design matrix X_γ is employed, and with an inverse gamma prior on g^{-1}.
- We cannot use a flat prior if we want to do model selection later because it is would leave posterior probabilities of different models underspec.

Notes

Turning to (2): Bayesian Variable Selection

We expand the model above by considering:

$$Y = \alpha 1 + X\beta_\gamma + \varepsilon$$

where we have

$\alpha, \beta_\gamma, \sigma^2$: the usual parameters previously seen

γ : A dummy parameter expressing the variables that should be included in the model.

it is now clear that β_γ should reflect them; thus the subindex.

Here we are interested in

$$\pi(\gamma \mid y).$$

less on other parameters

It is then possible to write the above through Bayesian Probability:

$$\pi(\gamma \mid y) = \frac{f(y \mid \gamma) \cdot \pi(\gamma)}{f(y)} = \frac{f(y \mid \gamma) \cdot \pi(\gamma)}{\sum_{\gamma'} f(y \mid \gamma') \cdot \pi(\gamma')} \quad (1)$$

Where we can express the PDF $f(y)$ as:

$$f(y) = \iiint f(y \mid \beta_\gamma, \alpha, \sigma^2) \cdot \pi(\beta_\gamma, \alpha, \sigma^2) \, d\beta_\gamma \, d\alpha \, d\sigma^2$$

↳ (idea $f(y) = P(x, y)\,dx = f(y \mid x) \cdot f(x)\,dx$)

same idea

From (1) we see that we can take an improper prior for α and σ^2 as it would cancel in the ratio; but not for β_γ as the different γ' will imply different terms that will not cancel leading to an improper posterior.

For the *g*-prior*, $f(y \mid \gamma)$ can be computed in closed form

$$f(y \mid \gamma) \propto \frac{(1+g)^{(n-1-|\gamma|)/2}}{(1+g(1-R^2_\gamma))^{(n-1)/2}}$$

} it is possible to prove this building on top of the posterior dist. obtained for the parameters in the previous section.

where

- $R^2_\gamma = 1 - \frac{s^2_\gamma}{s^2_0}$,
  and where $s^2_0 = (y - \bar{y}\mathbf{1})^T(y - \bar{y}\mathbf{1})$ is the sum of squared errors in the null model $\gamma = 0$
- "$\propto$" means up to factors which contain neither γ nor g

Given a sensible choice for $\pi(\gamma)$ the ratio above is then fully specified and it is possible to analyze the posterior distribution of the Bayesian Model selection.

- The simplest choice is the **uniform prior $\pi(\gamma) = 2^{-p}$** for all γ. I.e., each explanatory variable is included with probability $\frac{1}{2}$, independently of the other. For large p, this is however **informative for the size of the model** because with high prior probability $|\gamma| \approx \frac{p}{2}$.

→ Expectation bernoulli

- A **uniform prior for $|\gamma|$** is obtained by assuming that each explanatory variable is included with probability r where r is unknown and uniform on $(0, 1)$

} Solution

Bayes factor for model selection

- The posterior model probabilities $\pi(\gamma \mid y)$ depend on the prior $\pi(\gamma)$

} Natural by-product of Bayes formula, as above.

- One can avoid this if one uses the Bayes factor which is independent of the prior:

$$B(\gamma, \gamma') = \frac{\pi(\gamma \mid y)\,\pi(\gamma')}{\pi(\gamma' \mid y)\,\pi(\gamma)} = \frac{f(y \mid \gamma)}{f(y \mid \gamma')}$$

by def; check the notes above to see why.

$$= \underbrace{(1+g)^{(|\gamma'|-|\gamma|)/2}}_{\text{"Complexity penalty"}} \underbrace{\left(\frac{1+g(1-R^2_{\gamma'})}{1+g(1-R^2_\gamma)}\right)^{(n-1)/2}}_{\text{"Goodness of fit"}}$$

(decreases with $|\gamma|$) (increases with R^2_γ)

- The Bayes factor for comparing γ with the null model is

$$B(\gamma, 0) = \frac{(1+g)^{(n-|\gamma|-1)/2}}{(1+g(1-R^2_\gamma))^{(n-1)/2}}$$

Bayesian model averaging

Alternative to Model Selection

- For predicting a new observation y_{n+1} for a given vector x_{n+1} of explanatory variables, **Bayesian model averaging** is an alternative to model selection.
- Bayesian model averaging works by
  1. making predictions under each model,
  2. averaging all predictions according to the posterior probability of each model.
- For instance, the prediction of the mean of y_{n+1} is (for known g) given by

$$\mathbb{E}(y_{n+1} \mid y) = \bar{y} + \frac{g}{g+1} \sum_\gamma x^T_{n+1,\gamma} \hat{\beta}_\gamma \pi(\gamma \mid y).$$

$$E(y_{n+1} \mid y) = E(E(y_{n+1} \mid y, \gamma) \mid y)$$
$$= E(E(\alpha + x^T_{n+1}\beta + \varepsilon \mid y, \gamma))$$
$$= \hat{\alpha} + \frac{g}{g+1} \sum_\gamma x^T_{n+1} \hat{\beta}_\gamma \pi(\gamma \mid y)$$

($\hat{\alpha} + x^T_{n+1}\hat{\beta}_\gamma$ = $\hat{y}_{n+1}$)

Choosing g

- Bayes factors (and also posterior distributions) depend on the choice of g.
- As ***g* tends to infinity**, the prior becomes non-informative. However, as $g \to \infty$, $B(\gamma, 0) \to 0$ for any $\gamma \neq 0$. I.e., we always choose the null model (**"Bartlett's paradox"**)
- Choosing any **fixed value for *g*** also leads to problems: if g is fixed and $R^2_\gamma \to 1$ then $B(\gamma, 0) \to (1+g)^{(n-1-\gamma)/2}$ which is finite although one would expect that that this goes to infinity (**"information paradox"**)

} Recall that a higher R^2_γ means a better fit, so higher evidence for H_γ

A solution in this case consists in treating g as unknown and leverage:

A solution in this case consists in treating g as unknown and leverage :

(i) empirical Bayes

- In an **empirical Bayes approach**, we can determine $\widehat{g}$ either separately for each model γ or globally for all models together
- **Separately**:

$$\widehat{g} = \arg\max\left((n-1-|\gamma|)\log(1+g) - (n-1)\log(1+g(1-R^2_\gamma))\right)$$
$$= \max\left(\frac{(n-1-|\gamma|)R^2_\gamma}{|\gamma|(1-R^2_\gamma)} - 1, 0\right)$$

  The above ratio is the standard F-test statistics for the null hypothesis $\beta_\gamma = 0$

- **Globally**:

$$\widehat{g} = \arg\max \sum_\gamma \pi(\gamma) f(y \mid \gamma)$$

  This has to be computed numerically

- In both cases, one can show that the information paradox does not occur any more.
- The empirical Bayes approaches do have model selection consistency except if the true model is the null model

(ii) hierachical Models

Unknown g: fully Bayesian

- In a **fully Bayesian approach**, one can both avoid the above paradoxes and have model selection consistency for all true models
- It is desirable to have a prior $\pi(g)$ such that

$$f(y \mid \gamma) \propto \int \frac{(1+g)^{(n-1-|\gamma|)/2}}{(1+g(1-R^2_\gamma))^{(n-1)/2}} \pi(g) dg$$

  can be computed easily

- In order to avoid the information paradox, it is sufficient to have

$$\int (1+g)^{(n-1-|\gamma|)/2} \pi(g) dg = \infty \quad (|\gamma| \leq p)$$

Hierarchical Bayes and Empirical Bayes:

1. Hierarchical Bayes Method 1 (sequential)

Normal prior ; normal likelihood => no known dist.

↳ possible to see that if sample and prior mean differ

=> mode is closer to the sample mean.

↳ more influence from the data

Also using approach 2 you would have high values for the hyper prior τ^2 and convergence to prior mean

Using Empirical Bayes

=> if prior conflicts with the data we choose a wider prior; given more weight to the data.

↳ interesting here is to see how when

$$|\bar{x} - \mu| \leq \frac{\sigma}{\sqrt{n}}$$

we set posterior variance to 0 this shows how in fact empirical bayes underestimates the risk

Anova : empirical bayes shrinkes the individual treatment effects towards the global mean.

Hierarchical Methods:

The inner integral can be computed in closed form. It gives a normal density with mean equal to a convex combination of all group means $\bar{y}_1, \ldots, \bar{y}_J$. The weights in the convex combination and the variance depend on τ^2, and the integration over τ^2 has to be done by numerical integration or using one of the approximate methods that we will discuss in Chapter 4.

Poisson Model

Emp. Bayes: shrinks the individual experience towards the average experience of all contracts

Hierarchical Bayes:

Proof Reversibility => Invariance

Start assuming a $B \in \mathbb{R}^p$

then transition matrix

$$\Pi(A) = \int_A \pi(x)\, dx = \int_A \pi(x) \cdot P(x, \mathbb{R}^p)$$

transition matrix

invariance

$$= \int_{\mathbb{R}^p} \pi(x) \cdot P(x, A)$$