Bayesian Statistics

Fabio Sigrist

ETH Zurich, Autumn Semester 2019

Today's topics

- Reference priors
- Expert priors

Reference priors

Reference priors

- A **reference prior** is a prior π for which the distance between the prior π and the posterior $\pi(. \mid x)$ is maximal
  - Idea: if the prior has a small influence on the posterior, the data x have the largest possible impact

- 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 $f(x) = \int_\Theta f(x \mid \theta)\pi(\theta)d\theta$

Kullback-Leibler divergence

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

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

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

Reference prior

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

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

- $I(X,\theta)$ is called the **mutual information** of X and θ
  - Denoting by $\pi(x,\theta)$ the joint density of x and θ, this can also be written as

$$I(X,\theta) = \int_{X\times\Theta} \pi(x,\theta) \log \frac{\pi(x,\theta)}{\pi(\theta) f(x)} dx d\theta$$

Problem

- **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

Remedy: asymptotic solution

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

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

Reference prior: asymptotic solution

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

See blackboard for details

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 $\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2 \mid \theta_1)$

Bernardo's approach for nuisance parameters

- $\pi(\theta_2 \mid \theta_1)$ 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

See blackboard for example

Expert priors

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

Concluding comments on non-informative priors

Concluding comments on non-informative 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
- 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

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