07 - More priors - 02

Master in Foundations of Data Science
Bayesian Statistics and Probabilistic Programming
Fall 2018-2019

Josep Fortiana

Wednesday, November 07, 2018

Universitat de Barcelona

07 - More priors - 02

Pros and cons of conjugate priors

Non-Informative Priors (NIP's)

Jeffreys' NIP's

Entropy

Maximum entropy NIP's

07 - More priors - 02

Pros and cons of conjugate priors

Non-Informative Priors (NIP's)

Jeffreys' NIP's

Entropy

Maximum entropy NIP's

Weighing pros and cons - Misaprehensions

When using a conjugate family, computing the posterior distribution amounts to a parameter update, which:

- Is easy to obtain explicitly in closed form.
- Can often be given an intuitive explanation.

Weighing pros and cons - Misaprehensions

It is not true, in general, that the fact of belonging to a conjugate family entails non-informativeness (or any other specific property) of a given prior.

Weighing pros and cons - Misaprehensions

It may well happen that several NIP's belong to the same conjugate family.

We saw an instance of this phenomenon in the case of a $\mathsf{Ber}(\theta)$ model, where several NIP's (derived by imposing different criteria) belong to the same conjugate family $\mathsf{Beta}(\alpha, \beta)$.

Weighing pros and cons - Misaprehensions

A pitfall to be avoided is that an undue tendency to use a conjugate family can lead to ignore possibly relevant, but possibly incompatible information

Another cautionary remark: priors in a conjugate family are not necessarily the best available choice with respect to robustness* or other desirable property.

**A prior distribution is called robust if a small modification of it has little or no influence on the final inferential result.*

07 - More priors - 02

Pros and cons of conjugate priors

Non-Informative Priors (NIP's)

Jeffreys' NIP's

Entropy

Maximum entropy NIP's

NIP: Subjective, naïve concept

In principle, using a NIP is just a way of formalizing the fact that we have no prior information, previous to data collection.

That is, before data collection all possible parameter values are “equally likely”.

Technical problems

In practice, this simple concept leads to technical problems.

For instance, when the parameter is the probability $\theta \in [0, 1]$ of a binomial experiment, a sensible option is a $\mathsf{Unif}[0, 1]$ prior.

But, what are we to do in cases like θ in a $\mathsf{N}(\theta, \sigma^2)$ model?

Improper priors

The prior density:

$$h(\theta) = 1, \quad -\infty < \theta < +\infty,$$

has not a finite mass (it is not a probability density).

We can work with these *improper prior distributions*, as posterior pdf computation involves a *renormalization*.

Example

$X \sim \mathrm{N}(\theta, \sigma^2)$, with a known σ^2.

$$f(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}.$$

Possible NIP pdf:

$$h(\theta) = 1.$$

Example

Joint pdf:

$$f(x,\theta) = f(x|\theta)\cdot h(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\}\cdot 1.$$

The x marginal (prior predictive pdf):

$$f(x) = \int_{\mathbb{R}} f(x|\theta)\, h(\theta)\, d\theta = 1,$$

still an improper pdf.

Example

From Bayes' rule, the posterior pdf of θ is:

$$h(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{exp}\left\{-\frac{(x-\theta)^2}{2\sigma^2}\right\},$$

i.e., $(\theta|x) \sim \mathrm{N}(x,\sigma^2)$, a perfectly reasonable result.

NIP: towards an objective definition

Desirable properties:

1. A NIP should not prevail over or mask observed data.
2. Its influence should tend to vanish when the observed sample size is large.

07 - More priors - 02

Pros and cons of conjugate priors

Non-Informative Priors (NIP's)

Jeffreys' NIP's

Entropy

Maximum entropy NIP's

Invariance

Reparameterizing $\theta \in \mathbb{R}$ with $\eta = \exp(\theta)$, a bijective transformation, should not change the statistics.

But, if $h(\theta)$ is the prior pdf of θ, the prior pdf of η is:

$$\widetilde{h}(\eta) = h(\log \eta) \cdot \frac{1}{\eta}, \quad 0 < \eta < \infty.$$

Invariance

If $h(\theta) = 1$, $\widetilde{h}(\eta)$ is not constant (and another bijective transformation $\mathbb{R} \to \mathbb{R}_+$ gives another function).

We should seek NIP's whose influence on the posterior pdf does not vary with these reparameterizations.

Also, any invariance property of the likelihood should be matched in the prior pdf.

Example: Translation invariant priors

If the likelihood $f(x|\theta)$ belongs to a *location family*, where θ is the *location* parameter:

$$f(x|\theta) = f_0(x - \theta).$$

Such family is invariant under translations: for any given a, the distribution of $\widetilde{x} = x - a$ belongs to the same family,

$$f(\widetilde{x}|\widetilde{\theta}), \quad \text{on } \widetilde{\theta} = \theta - a$$

Example: Translation invariant priors

Imposing translation invariance on the prior θ pdf,

$$h(\theta) = h(\theta - a), \quad \text{for any } a.$$

Thus, a translation invariant prior is:

$$h(\theta) = \text{ a constant value.}$$

Jeffrey's prior definition

The *Jeffreys' prior* for θ is defined from the Fisher information $I(\theta)$:

$$h(\theta) \propto [I(\theta)]^{1/2},$$

normalizing, if possible, to obtain a pdf. When normalization is not possible, $h(\theta)$ is an improper prior.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society of London A186: 453–461.

Fisher information

In a model with likelihood $f(x|\theta)$, with $\theta \in \Theta \subset \mathbb{R}$, the *Fisher information* is:

$$I(\theta) = \mathrm{E}_\theta \left(\frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2,$$

provided some regularity conditions hold.

Intuitive justification for Jeffreys' prior

Fisher's information quantifies *the information on* θ *provided by the observed* x (the larger $I(\theta)$, the better is the maximum possible precision of an unbiased estimator of θ).

The idea is that, *a priori*, θ values for which $I(\theta)$ is larger are more likely.

Invariance of Jeffreys' prior

Assume a reparameterization by g, a bijective and differentiable function.

The Fisher information for $\psi = g(\theta)$ is:

$$I(g(\theta)) = \frac{I(\theta)}{g'(\theta)^2},$$

Invariance of Jeffreys' prior

Jeffreys' NIP for $\psi = g(\theta)$ is the correct one, corresponding to the variable change in the pdf.

Indeed, the new prior pdf for ψ is:

$$\widetilde{h}(\psi) = \sqrt{I(\psi)} = \frac{\sqrt{I(\theta)}}{|g'(\theta)|} = h(\theta)\,\left|\frac{d\theta}{d\psi}\right|.$$

Example 1: binomial likelihood

$x \sim \mathrm{B}(n, \theta)$.

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

$$\log f = \log \binom{n}{x} + x\, log(\theta) + (n-x)\, \log(1-\theta),$$

$$\frac{\partial \log f}{\partial \theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x - n\,\theta}{\theta\,(1-\theta)}.$$

Fisher information for the binomial model

$$\mathrm{E}_\theta \left(\frac{\partial \log f}{\partial \theta} \right)^2 = \frac{\mathrm{E}_\theta (x - n\,\theta)^2}{\theta^2\,(1-\theta)^2}.$$

Since $\mathrm{E}_\theta(x) = n\,\theta,$

$$\mathrm{E}_\theta (x - n\,\theta)^2 = \mathrm{E}_\theta (x - \mathrm{E}_\theta(x))^2 = \mathrm{var}_\theta(x) = n\,\theta\,(1-\theta),$$

thus, $I(\theta) = \dfrac{n}{\theta\,(1-\theta)}.$

Jeffreys' prior for the binomial model

Jeffreys' prior pdf is:

$$h(\theta) = A\,\theta^{-1/2}\,(1-\theta)^{-1/2},$$

where the normalization constant A is finite, giving a (proper) prior pdf:

$$\theta \sim \mathrm{Beta}(\frac{1}{2}, \frac{1}{2}).$$

Example 2: Negative Binomial

Let $X \sim \mathrm{NegBin}(a, \theta)$, $a > 0$, $\theta \in [0, 1]$, with pmf:

$$f(x|a, \theta) = \frac{\Gamma(a + x)}{\Gamma(a)\, x!}\, \theta^a\, (1 - \theta)^x, \quad x \in \mathbb{N} \cup \{0\}.$$

$$\mathrm{E}(X) = a\, \frac{1 - \theta}{\theta}, \quad I(\theta) = \frac{a}{\theta^2\, (1 - \theta)}.$$

Jeffreys' prior has, in this case, a pdf proportional to:

$$\frac{1}{\theta\, (1 - \theta)^{1/2}},$$

an improper distribution.

Example 3: Location family

$$f(x|\theta) = f_0(x-\theta), \quad \theta \in \mathbb{R}.$$

Since

$$I(\theta) = \mathrm{E}_\theta \left(\frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 = \int_{\mathbb{R}} \left(\frac{f_0'(x)}{f_0(x)} \right)^2 f_0(x)\, dx,$$

does not depend on θ, Jeffreys' prior has a constant density.

When f's support is unbounded, Jeffreys' prior is an improper density.

07 - More priors - 02

Pros and cons of conjugate priors

Non-Informative Priors (NIP's)

Jeffreys' NIP's

Entropy

Maximum entropy NIP's

Mathematical theory of information

Seminal article:

Claude Shannon (1948)

A Mathematical Theory of Communication,

Bell Systems Technical Journal.

Information of an event A

$$\boxed{I(A) = -\log_2 \mathrm{P}(A)}$$

$\log_2$ is the base 2 logarithm

Remarks

$$y = \log_2 x \iff x = 2^y \iff y = \frac{\log x}{\log 2}$$

$$\log_2 1 = 0, \quad \log_2 2 = 1, \quad \log_2 8 = 3, \quad \log_2 \tfrac{1}{2} = -1.$$

Since P lies between 0 and 1, its $\log_2$ is negative.

The minus sign $-$ in the definition makes the result a positive quantity.

Information as a measure of “surprise”

Intuitively:

The more unlikely an event,

the larger is the quantity of information

provided by knowing it actually happened.

Examples

- Tossing a coin: $I(\text{heads}) = 1$.
- Throwing a dice: $I(\text{six}) = 2.58$.
- Tossing 50 coins: $I(\text{45 or more heads}) = 28.82$.

Information and communication

Message: a sequence of symbols (letters, bytes).

Predictability: degree of surprise caused by the next letter, given we know the preceding ones.

Redundancy and data compression

Lossless compression methods increase information density, keeping its total.

Redundancy decreases. Predictable letters are omitted.

Tradeoff: harder or impossible to correct transmission errors.

Example

The sequence:

AACAACAACABSABSABSAACABSAASAACAACAACABSABSABSAACABS

can be compacted with the table:

AAC	ABS	AAS
1	2	3

plus the list:

1 1 1 2 2 2 1 2 3 1 1 1 2 2 2 1 2

Definition of entropy of a finite, discrete probability

The *entropy of a finite, discrete pmf,*

$$\boldsymbol{p} = (p_1, \dots, p_m),$$

is the weighted average (expectation) of the m informations:

$$H(\boldsymbol{p}) = -\sum_{j=1}^{m} p_j \log_2 p_j$$

Entropy for a partition of a set

A *partition* of a set T of n elements in m classes:

$$T = C_1 \sqcup \cdots \sqcup C_m$$

gives the vector $\boldsymbol{n}$ of absolute frequencies:

$$\boldsymbol{n} = (n_1, \dots, n_m), \quad \text{where } n_j = \sharp(C_j),$$

$$n = \sum_{j=1}^{m} n_j.$$

Entropy for a partition of a set

The vector $\boldsymbol{f} = \boldsymbol{n}/n$ of relative frequencies, $f_j = n_j/n$,

$$\boldsymbol{f} = (f_1, \ldots, f_m),$$

is a probability.

The *entropy of the partition* is:

$$H(\boldsymbol{f}) = -\sum_{j=1}^{m} f_j \log_2 f_j$$

Example: entropy of a binary partition

Partitioning a set into two classes:

$$T = C_1 \sqcup C_2, \quad \boldsymbol{f} = (f, 1 - f),$$

The resulting entropy is:

$$H(\boldsymbol{f}) = -f\ \log_2 f - (1 - f)\ \log_2(1 - f).$$

Maximum entropy

How does entropy vary depending on the manner of locating elements into the classes?

Maximum entropy

How does entropy vary depending on the manner of locating elements into the classes?

Is there a maximum value?

Maximum entropy

How does entropy vary depending on the manner of locating elements into the classes?

Is there a maximum value?

If so, for which partition this maximum is attained?

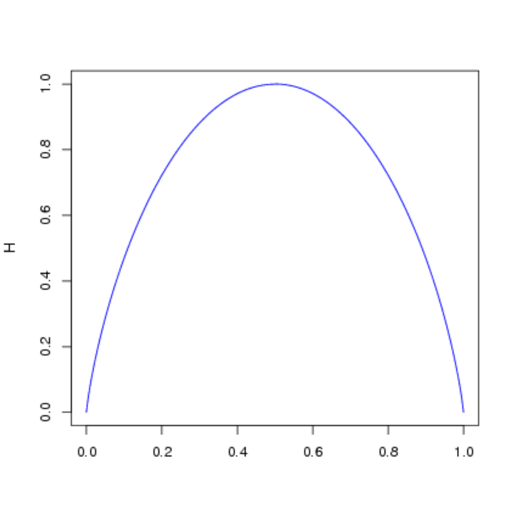
Entropy of a binary partition as a function of f

Differentiate and equate to 0:

$$\begin{aligned} H(f) &= -f\ \log_2 f - (1-f)\ \log_2(1-f) \\ &= \frac{1}{\log 2}\left(-f\ \log f - (1-f)\ \log(1-f)\right). \\ H'(f) &= \frac{1}{\log 2}\left(-\log f + \log(1-f)\right) = 0. \end{aligned}$$

Hence, $H(0.5) = 1$ for $f = 0.5$ is a maximum.

Entropy of a binary partition as a function of f



Differential entropy

Substituting pdf's for pmf's and integral for summation we obtain a definition of a *differential entropy,* which shares some properties of entropy.

For a (univariate) r.v. with pdf f:

$$h[f] = -\int_{\mathbb{R}} \log f(x)\, f(x)\, dx.$$

Differential entropy for a $\mathrm{Unif}(0, b)$

When f is the uniform (rectangular) pdf on $[0, b]$,

$$f(x) = \frac{1}{b}, \quad 0 < x < b,$$

$$h[f] = -\int_0^b \frac{1}{b} \log \frac{1}{b} \, dx = \log b.$$

Differential entropy is NOT non-negative

- If $b > 1$, $h[f] > 0$ can be made arbitrarily large,
- If $b = 1$, $h[f] = 0$,
- If $b < 1$, $h[f] < 0$; its absolute value can be made arbitrarily large.

Differential entropy is NOT limit of discrete entropies

Partition $[0, 1]$ into n intervals, each of length $\dfrac{1}{n}$.

Assign a $\dfrac{1}{n}$ probability mass to a point in each interval.

This discrete uniform distribution, U_n, has entropy:

$$H_n = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n.$$

Differential entropy is NOT limit of discrete entropies

When $n \to \infty$, also $H_n \to \infty$.

But

$$\{U_n\} \xrightarrow[n\to\infty]{\mathscr{L}} \mathrm{Unif}[0,1],$$

with differential entropy 0.

07 - More priors - 02

Pros and cons of conjugate priors

Non-Informative Priors (NIP's)

Jeffreys' NIP's

Entropy

Maximum entropy NIP's

Edwin T. Jaynes (1922 - 1998)

Proposed the *maximum entropy principle* to find NIP's.

Rationale behind the maximum entropy principle

A prior pdf should collect any available information on the parameters, but nothing else.

In Jaynes' words:

> *The goal of avoiding unjustified conclusions brings us to ask ourselves whether there exists a reasonable measure of the degree of uniformity of a probability distribution, which could be maximized, subjected to constraints representing the available information.*

Maximum entropy discrete distribution

Given a discrete r.v., with m values:

$$\theta = (\theta_1, \ldots, \theta_m),$$

and vector of probabilities:

$$\boldsymbol{p} = (p_1, \ldots, p_m),$$

the entropy is:

$$H(\boldsymbol{p}) = -\sum_{j=1}^{m} p_j \log p_j .$$

Maximum entropy discrete distribution

To locate the maximum, construct the Lagrangian:

$$\Phi(\boldsymbol{p}) = -\sum_{j=1}^{m} p_j \log p_j - \lambda \left(\sum_{j=1}^{m} p_j - 1 \right),$$

where λ is a Lagrange multiplier.

Differentiate Φ and equate to zero.

Maximum entropy discrete distribution

$$\frac{\partial \Phi}{\partial p_j} = -\log p_j - 1 - \lambda = 0, \quad 1 \leq j \leq m,$$

hence all the p_j's are equal to $1/m$.

Maximum attained for the uniform pmf on

$$\theta = (\theta_1, \dots, \theta_m).$$

Maximum entropy discrete distribution, with constraints

Assume the expectation $\overline{\theta} = \sum_{j=1}^{m} p_j \, \theta_j.$ is known.

The maximum is obtained from the Lagrangian:

$$\begin{aligned}\Phi(\boldsymbol{p}) = -\sum_{j=1}^{m} p_j \, \log p_j - \lambda_0 \, \left(\sum_{j=1}^{m} p_j - 1\right) \\ + \lambda_1 \, \left(\sum_{j=1}^{m} p_j \, \theta_j - \overline{\theta}\right),\end{aligned}$$

with two Lagrange multipliers.

Maximum entropy discrete distribution, with constraints

Differentiating and equating to 0 as above,

$$p_i = \frac{\exp(\lambda_1 \theta_i)}{\sum_{j=1}^{m} \exp(\lambda_1 \theta_j)}, \quad 1 \leq i \leq m,$$

This is the *Gibbs distribution*.

For $\lambda_1 = 0$ we recover the discrete uniform distribution.

Maximum differential entropy

Differential entropy, for unconstrained pdf's is unbounded.

Problem: find a continuous pdf f such that:

- $Supp(f) = \mathbb{R}$.
- $H[f]$ is a maximum.
- The mean and variance of f are finite and equal to the given quantities μ and σ^2, respectively.

Entropy of the gaussian law

The Gaussian distribution $\mathsf{N}(\mu, \sigma^2)$ is the solution.

Similarly, the maximum entropy continuous pdf on a finite interval $[a, b]$ is the $\mathsf{Unif}(a, b)$.

And the maximum entropy continuous pdf on $[0, \infty)$, with a given expectation $1/\lambda$, is the $\mathsf{Exp}(\lambda)$.