Conjugate Priors, Non-informative Prior, Jeffreys prior and Hierarchical Bayesian models

September 12, 2018

Outline

Conjugate Priors

Non-informative priors

Jeffreys Noninformative Prior

Hierarchical Bayesian Models

Conjugate Priors

Conjugate Priors

- Consider a class of probability distribution P. For every prior $\pi(\theta) \in P$, if the posterior distribution $\pi(\theta|x)$ belongs to P and the likelihood $f(x|\theta)$ to a family F, then the P class is **conjugate** for F.

- Conjugate priors are analytically tractable. Finding the posterior reduces to an updating of the corresponding parameters of the prior.

- Consider a coin flipping example:

  - Let θ the probability that the coin will draw heads
  - Prior $\theta \sim \mathscr{Be}(a,b)$
  - Data: the coin flipped n times with n_H of those were heads (binomial)
  - Posterior:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_0^1 f(x|\theta)\pi(\theta)d\theta} = \frac{\theta^{a+n_H-1}(1-\theta)^{b+n-n_H-1}}{beta(a+n_H, b+n-n_T)} = \mathscr{Be}(a+n_H, b+n-n_H)$$

- The role of conjugate priors is generally to provide a first approximation to the adequate prior distribution which should be followed by a robustness analysis.

Consider the likelihood $f(x|\theta)$: Binomial distribution.

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

Shape depends on $\theta^x(1-\theta)^{1-x}$ and $\theta \in [0,1]$.

Prior: Beta distribution (Beta(a,b)):

$$Beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \tag{2}$$

Expression for posterior in terms of likelihood and prior:

$$\begin{align} posterior &\propto likelihood \times prior \tag{3}\\ &\propto \theta^x(1-\theta)^{n-x} \times \theta^{a-1}(1-\theta)^{b-1} \tag{4}\\ &\propto \theta^{(a+x)-1}(1-\theta)^{(b+n+x)-1} \tag{5}\end{align}$$

$$\boxed{posterior = \text{Beta}(a+x, b+n+x)} \tag{6}$$

Standard Exponential Families

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $N(\theta,\sigma^2)$	Normal $\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu+\tau^2 x),\rho\sigma^2\tau^2)$ $\rho^{-1}=\sigma^2+\tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{P}(\theta)\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(v,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+v,\beta+x)$
Binomial $\mathcal{B}(\mathrm{n},\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+x,\beta+n-x)$
Negative Binomial $\mathcal{N}eg(\mathrm{m},\theta)$	Beta $\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+m,\beta+x)$
Multinomial $\mathcal{M}_k(\theta_1,\ldots,\theta_k)$	Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$	$\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$
Normal $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}a(\alpha,\beta)$	$\mathcal{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$

Consider a random variable x described by an exponential distribution with parameter λ:

$$x \sim p(x;\lambda) = \lambda e^{\lambda x}. \tag{7}$$

We are uncertain about the value of λ and can choose to model this uncertainty by defining a Gamma distribution over it:

$$\lambda \sim \mathsf{Gamma}(\alpha, \beta), \tag{8}$$

where the Gamma distribution is the conjugate prior for the exponential distribution.

Conjugate prior: If the posterior distributions $p(\lambda|x)$ are in the same probability distribution family as the prior probability distribution $p(\lambda)$, the prior is called a conjugate prior for the likelihood function. Therefore, the posterior can be expressed as:

$$p(\lambda|x) \propto \mathsf{Gamma}(\alpha^*, \beta^*) \tag{9}$$

Obtain an analytic form of the posterior distribution of Eq. (9)

Let us consider the data $X = x_1, x_2, ..., x_n$. The posterior distribution $p(\lambda|X)$ is given by:

$$p(\lambda|X) = \frac{p(X|\lambda)p(\lambda)}{\int p(X|\lambda)p(\lambda)} \tag{10}$$

$$\propto p(X|\lambda)p(\lambda) \tag{11}$$

$$\propto \lambda \exp\{\lambda \sum_{i=1}^{N} x_i\}\mathsf{Gamma}(\alpha, \beta) \tag{12}$$

$$\propto \lambda^n e^{\lambda \sum_{i=1}^{N} x_i} \lambda^{\alpha-1} e^{-\beta\lambda} \tag{13}$$

$$\propto e^{-\lambda(\sum_{i=1}^{N} x_i - \beta)} \lambda^{n+\alpha-1} \tag{14}$$

$$\boxed{p(\lambda|X) \propto \mathsf{Gamma}(\alpha + n, \sum_{i=1}^{N} x_i + \beta)} \tag{15}$$

(a) Derive the maximum likelihood estimate (MLE) (λ_{MLE}) of Eq. (7)

(b) Obtain an analytic form of the posterior distribution of Eq. (9) and
Derive the maximum a posteriori estimator (MAP) λ_{MAP} as a function of α, β.

The likelihood is given by

$$p(x|\lambda) = \prod_{i=1}^{n} \lambda \exp\{\lambda x_i\} \tag{16}$$

The log likelihood:

Now set the derivative w.r.t. λ to 0:

Therefore,

$$\boxed{\lambda_{MLE} = \frac{1}{\bar{x}}}. \text{ Where, } \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{17}$$

The posterior is given by:

$$p(\lambda|X) \propto e^{-\lambda(\sum_{i=1}^{N} x_i - \beta)} \lambda^{n+\alpha-1} \tag{18}$$

The log posterior:

set the derivative to 0:

$$\boxed{\lambda_{MAP} = \frac{n+1}{\sum_{i=1}^{N} x_i + \beta}} \tag{19}$$

Generate $N = 20$ samples drawn from an exponential distribution with parameter $\lambda = 0.2$. Fix $\beta = 100$ and vary α over the range $(1, 40)$ using a step-size of 1.

Compute the corresponding MLE and MAP estimates for λ.

For each α, compute the mean squared error [1] of both estimates compared against the true value and then plot the mean squared error as a function of α.

Now, fix $\alpha = 30$, $\beta = 100$ and vary N over the range (1, 500) using a step-size of 1. Plot the mean squared error for each N of the corresponding estimates and explain under what conditions is the MAP estimator better.

[1] Mean square error (MSE) is defined as : MSE $= \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$. Where Y_i is the true value and $\hat{Y}_i$ is the estimated value

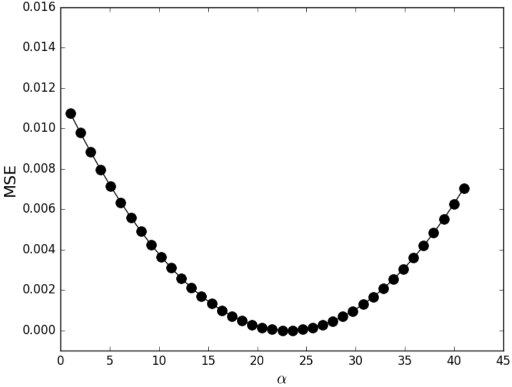

Figure 1: Mean square error of both estimates compared against the true value for varying α

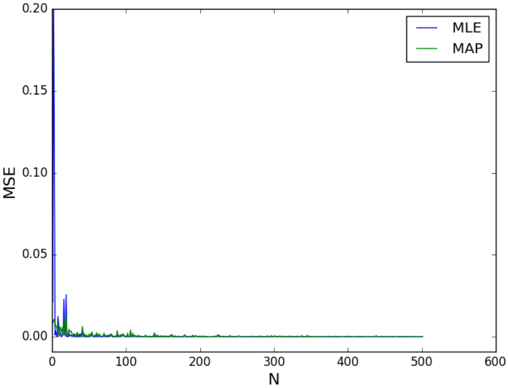

Figure 2: Mean square error of both estimates compared against the true value for varying α

Maximum Entropy Priors

- If nothing is known about a distribution except that it belongs to a certain class, then the distribution with the largest entropy should be chosen as the default.[a]
- The entropy is defined as
  - discrete case $\mathbb{H}(\pi) = -\sum_k \pi(\theta_k)\log\left(\pi(\theta_k)\right)$
- When some statistics (moments) of the prior distribution are known,

$$\mathbb{E}_\pi\left[g_k(\theta)\right] = w_k,\ k = 1,..,K$$

the maximum entropy distribution is of the form:

$$\pi(\theta_i) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_i)\right)}{\sum_j \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta_j)\right)},\ \lambda_k = Lagrange\ multipliers$$

- However, the constraints may not be compatible, e.g. $\mathbb{E}\left(\theta^2\right) \geq \mathbb{E}^2(\theta)$.

[a] C. P. Robert, *The Bayesian Choice*, *Springer, 2nd edition,* chapter 3 (full text available)

Maximum Entropy Priors

- For the continuous case, we define the entropy as the Kullback-Leibler divergence between π and some invariant non-informative prior for the problem π_0, i.e.

$$\mathbb{H}(\pi) = -\int \pi_0(\theta) \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta$$

- As for the discrete case, the maximum entropy distribution is of the form:

$$\pi(\theta) = \frac{\exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta)\right) \pi_0(\theta)}{\int \exp\left(\sum_{k=1}^{K} \lambda_k g_k(\theta)\right) \pi_0(\theta) d\theta}, \lambda_k = Lagrange\ multipliers$$

- The selection of π_0 is not obvious or easy.

Summary: Conjugate Priors

PROS.:

Simple to handle.

Conjugate priors are analytically tractable. Finding the posterior reduces to an updating of the corresponding parameters of the prior.

CONS.:

Not applicable to all likelihood functions.

Not flexible, cannot account for constraints e.g. $\theta > 0$

...

Non-informative priors

Non-informative priors

Non-informative priors: If there is a small amount of prior information on the parameters of interest, the hyper-parameters can be set at values to reflect this, leading to non-informative, vague, flat

Non-informative priors can be improper: $p(\sigma) \propto \frac{1}{\sigma}$.

Also, improper priors can have proper posterior distribution: Normal liklehood $+\ p(\sigma)$

Noninformative Priors

- A second difficulty arises from the transformation behavior of a probability density under a nonlinear change of variables.

- If a function $h(\lambda)$ is constant, and we change variables to $\lambda = \eta^2$, then $h(\eta) = h(\eta^2)$ will also be constant. However, if we choose the density $p_\lambda(\lambda)$ to be constant, then the density of η will be given by

$$p_\eta(\eta) = p_\lambda(\lambda)\left|\frac{d\lambda}{d\eta}\right| = p_\lambda(\eta^2)2\eta \propto \eta$$

  and so the density over η will not be constant.

- This issue does not arise when we use maximum likelihood, because the likelihood function $p(x|\lambda)$ is a simple function of λ and so we are free to use any convenient parameterization.

- If, however, we are to choose a prior distribution that is constant, we must take care to use an appropriate representation for the parameters.

Jeffreys Noninformative Prior

Jeffrey's Noninformative Priors

- Jeffrey's proposes a more intrinsic approach which avoids the need to take the invariance structure into account.

- Given a likelihood $f(x \mid \theta)$, Jeffrey's noninformative prior distributions are based on Fisher information, given by

$$I(\theta) = \mathbb{E}_{X|\theta}\left(\frac{\partial \log f(X \mid \theta)}{\partial \theta}\frac{\partial \log f(X \mid \theta)^T}{\partial \theta}\right) = -\mathbb{E}_{X|\theta}\left(\frac{\partial^2 \log f(X \mid \theta)}{\partial \theta^2}\right)$$

the corresponding prior distribution is

$$\pi(\theta) \propto \underbrace{\left|I(\theta)\right|}_{\text{Determinant of } I}{}^{1/2}$$

Sir Harold Jeffreys
(1891–1989)

Jeffrey's Noninformative Priors

- Jeffreys Invariance Principle:
  - Any rule for defining the prior distribution on θ should lead to an equivalent result when using a transformed parameterization
  - Let $\phi = h(\theta)$ and h be an invertible function with inverse function $\theta = g(\phi)$, then

$$\pi(\phi) = \pi(g(\phi))\left|\frac{dg(\phi)}{d\phi}\right| = \pi(\theta)\left|\frac{d\theta}{d\phi}\right|$$

  - Jeffreys noninformative priors $\pi(\phi) \propto \left|I(\phi)\right|^{1/2}$ satisfy this invariant reparameterization requirement.

$$I(\phi) = -\mathbb{E}_{X|\phi}\left(\frac{\partial^2 \log f(X|\phi)}{\partial \phi^2}\right) = -\mathbb{E}_{X|\theta}\left(\frac{\partial^2 \log f(X|\phi)}{\partial \theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right) = I(\theta)\left|\frac{d\theta}{d\phi}\right|^2$$

Consider a random variable x described by a Poisson distribution:

$$x \sim p(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}. \tag{20}$$

Determine the Jeffreys prior π^J for θ. Is the scale invariant prior $\pi_0(\theta) = \frac{1}{\theta}$ preferable to π^J? Why?

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \tag{21}$$

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(x|\theta)\right] \tag{22}$$

$$= \frac{\theta}{\theta^2} = \frac{1}{\theta} \tag{23}$$

Therefore the Jeffreys prior is given by:

$$\boxed{\pi^J(\theta) \propto \theta^{-\frac{1}{2}}} \tag{24}$$

Find the maximum entropy prior for θ for the reference measure π^J subject to the constraints $\mathbb{E}^\pi[\theta] = 1$, $Var^\pi[\theta] = 1$.
Considering the reference measure as $\pi_{ref} = \pi^J(\theta) \propto \frac{1}{\sqrt{\theta}}$
The maximum entropy prior under the constraints that the prior mean and variance of θ are both 1:

$$\hat{\pi} = \frac{\pi_{ref}(\theta) \exp(\sum_{k=1}^{K} \lambda_k g_k(\theta))}{\int \pi_{ref}(\theta) \exp(\sum_{k=1}^{K} \lambda_k g_k(\theta))} \tag{25}$$

In this problem,

$$\boxed{\hat{\pi} \propto \theta^{-\frac{1}{2}} \exp(\lambda_1 \theta + \lambda_2 (\theta - 1)^2)} \tag{26}$$

Hierarchical Bayesian Models

Hierarchical Bayes

- A key requirement for computing the posterior $p(\theta|\mathcal{D})$ is the specification of a prior $p(\theta|\eta)$, where η are the hyper-parameters.
- What if we don't know how to set η?
- In some cases, we can use uninformative priors as discussed earlier.
- A more Bayesian approach is to put a prior on our priors! *In terms of graphical models (showing explicitly dependence relations)*, we can represent the situation as follows:

$$\eta \to \theta \to \mathcal{D}$$

- This is an example of a hierarchical Bayesian model, also called a multi-level model, since there are multiple levels of unknown quantities.

Suppose that a group of scientists conduct a number of experiments to investigate the development of tumors in rats. For a given experiment j, let y_j denote the number of rats in the experiment that are observed to develop a tumor. n_j is the total number of rats in experiment j. θ_j describes the probability that a given rat within experiment j develops a tumor. $J = 71$ experiments are conducted.

One can model y_j using a Binomial distribution:

$$y_j = \mathsf{Bin}(n_j, \theta_j) \tag{27}$$

We are uncertain about the value of θ_j and can choose to model this uncertainty by defining a beta distribution over it:

$$\theta_j \sim \mathsf{Beta}(\alpha, \beta), \tag{28}$$

where α and β are hyperparameters.

One can write down the joint posterior distribution of the parameters as

$$p(\theta, \alpha, \beta | y) \propto p(\alpha, \beta) p(\theta | \alpha, \beta) p(y | \theta, \alpha, \beta). \tag{29}$$

Suppose we define the following noninformative hyperprior:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}. \tag{30}$$

One can show that this is equivalent to the following density over transformed variables:

$$p\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha + \beta) \right) \propto \alpha\beta(\alpha + \beta)^{-5/2} \tag{31}$$

Obtain an analytic forms of (a) the posterior distribution of Eq. (29) and (b) the marginal posterior distribution over α and β: $p(\alpha, \beta|y)$.
The data from the experiments j = 1,2,3,4,...,71. are assumed to follow independent binomial distribution. The joint posterior distribution of all parameters is given by:

$$p(\theta, \alpha, \beta|y) \propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta) \tag{32}$$

$$\boxed{p(\theta, \alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha + y_i - 1}(1 - \theta_j)^{\beta + n_j - y_j - 1}} \tag{33}$$

Given (α, β) the components of θ have independent posterior densities that are of the form $\theta_j^A(1-\theta_j)^B$- that is, beta densities and the joint density is expressed as

$$p(\theta|\alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1} \tag{34}$$

We can determine the marginal posterior distribution of (α, β) $[p(\alpha, \beta|y)]$ by using eq. 32, eq. 34 and the hint provided. [Hint: For example, the marginal posterior distribution of ϕ can be computed algebraically using the conditional probability formula, $p(\phi|y) = \frac{p(\theta,\phi|y)}{p(\theta|\phi,y)}$. Where θ is the parameter and y is the fixed data.]

Plot the marginal posterior density $p(\alpha, \beta|y)$ as a function of the transformed variables $\log \frac{\alpha}{\beta}$ and $\log(\alpha + \beta) \in [(-1.3, -2.3); (1, 5)]$. In the above plot of the marginal posterior distribution of the hyperparameters, under the transformation, is approximately symmetric about the mode, approximately (-1.79,2.8). Obtain the corresponding value of (α, β).

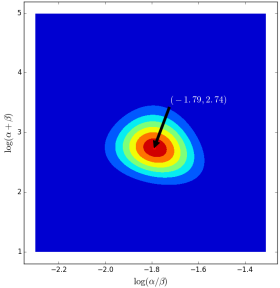

Figure 3: Marginal posterior density $p(\alpha, \beta|y)$ as a function of the transformed variables $\log \frac{\alpha}{\beta}$ and $\log(\alpha + \beta) \in [(-1.3, -2.3); (1, 5)]$.