Some Considerations about Choosing a Prior Distribution

[Based on Gutierrez-Peña, E. and Muliere, P.(2004) Paper]

Maria D. Lomeli-Garcia

Gatsby Computational Neuroscience Unit

September 20, 2012

What prior should we choose?

Example:
Let $X_1, \ldots, X_n$ be a random sample such that
$X_i \sim \text{Bernoulli}(X \mid \theta)$.

What prior should we choose for this likelihood model?

A "natural" choice of a prior for θ is $\text{Beta}(\alpha_0, \beta_0)$, i.e. a conjugate prior.

Why?

What prior should we choose?

Example:
Let $X_1, \ldots, X_n$ be a random sample such that
$X_i \sim \text{Bernoulli}(X \mid \theta)$.

What prior should we choose for this likelihood model?

A "natural" choice of a prior for θ is Beta(α_0, β_0), i.e. a conjugate prior.

Why?

As MacEachern would say,

"... prior distributions are not priors of belief, but priors of convenience."

What if we want to be as *minimally informative* as possible about our prior knowledge?

We could use Jeffrey's prior:

$$p(\theta) \propto [I(\theta)]^{\frac{1}{2}}$$

Jeffrey's prior asymptotically maximizes the mutual information between the distribution of the sample and that of the parameter (Clarke and Barron, 1994).

As MacEachern would say,

"... prior distributions are not priors of belief, but priors of convenience."

What if we want to be as *minimally informative* as possible about our prior knowledge?

We could use Jeffrey's prior:

$$p(\theta) \propto [I(\theta)]^{\frac{1}{2}}$$

Jeffrey's prior asymptotically maximizes the mutual information between the distribution of the sample and that of the parameter (Clarke and Barron, 1994).

What does it mean to use a conjugate prior in terms of mutual information between the parameter and the sample?

Main result: The prior that minimizes the mutual information between the sample and the parameter is natural conjugate when the model belongs to an exponential family. So

..."conjugate priors should probably not be used for representing prior knowledge without substantive justification".

What does it mean to use a conjugate prior in terms of mutual information between the parameter and the sample?

Main result: The prior that minimizes the mutual information between the sample and the parameter is natural conjugate when the model belongs to an exponential family. So

..."conjugate priors should probably not be used for representing prior knowledge without substantive justification".

1. **Kullback-Leibler Divergence**

$$D_{KL}(p||q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) \eta(dx)$$

2. **Mutual Information**

$$\begin{aligned} I(X,Y) &= \int \int p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \eta_1(dx)\eta_2(dy) \\ &= D_{KL}(p(x,y)||p(x)p(y)) \end{aligned}$$

Amount of information that X gives about Y and viceversa, measure of dependence between X and Y.

1. **Capacity of the Channel**

$$C = \overset{\text{Sup}}{p(x)} \; I(X, Y)$$

Information can be sent through the channel with arbitrary low probability of error at any rate less than C. X is the source and $Y \mid X$ is the channel. (Data Transmission)

2. **Rate Distortion Function**

$$R(I) = \overset{\text{Inf}}{p(y|x) \in \mathcal{P}} \; I(X, Y)$$

where $\mathcal{P}$ is the class of conditional distributions which satisfy the following constraint:

$$\int \int p(x)p(y \mid x)L(x,y)\eta_1(dx)\eta_2(dy) \leq I$$

and $L(x,y)$ is a given distortion (loss) function. $p(y \mid x)$ is a description of how X is represented by Y, it should be represented by as few bits as possible.(Data Compression)

Exponential Families

Let X be a random variable with probability mass function (or probability density) given by

$$p(x \mid \theta) = b_1(x) \exp\left[x'\theta - M(\theta)\right]$$

where

$$M(\theta) = \log \int b_1(x) \exp\left[x'\theta\right] \eta(dx)$$

then X is a member of a regular[1] exponential family of distributions and θ is the natural parameter.

[1] Subject to regularity conditions

Mean Parameter

If $T = \frac{1}{n}\sum X_i$ and $\mu = \frac{d}{d\theta}M(\theta)$ is the mean parameter, then

$$p(t \mid \mu, n) = B(tn, n)\exp\left[n\left(t\theta(\mu) - M_\theta(\mu)\right)\right]$$

The corresponding conjugate prior is

$$p(\mu \mid t_0, n_0) = H(t_0 n_0, n_0)\exp\left[n_0\left(t_0\theta(\mu) - M_\theta(\mu)\right)\right]V(\mu)^{-1}$$

where

$$V(\mu) = \frac{d^2}{d\theta^2}M(\theta(\mu))$$

A loss function between members of the exponential family is

$$L(t, \mu) = D_{KL}\left(p(. \mid \tilde{\mu}) || p(\mid \mu)\right) = M_\theta(t) - M_\theta(\mu) + (\theta(t) - \theta(\mu))t$$

Main result

We wish to minimize the mutual information between T and μ over a class of posterior distributions defined by an expected distortion constraint, keeping $m(t)$ (source distribution) fixed, for a given loss function $L(t, \mu)$, i.e.

$$\mathcal{L} = \int \int w(\mu) w(\mu \mid t) \log \left(\frac{w(\mu \mid t)}{w(\mu)} \right) dt d\mu$$

$$+\lambda \int \int m(t) w(\mu \mid t) L(t, \mu) dt d\mu + \int \int g(t) w(\mu \mid t) d\mu dt$$

Solution to the minimization problem

Setting $\frac{d}{dw(\mu|t)}\mathcal{L} = 0$, the solution is

$$w(\mu \mid t) \quad = \quad \frac{w^*(\mu)e^{-\lambda L(t,\mu)}}{\int w^*(u)e^{-\lambda L(t,u)}du}$$

and, if $w^*(\mu) > 0$

$$\frac{\int m(t)e^{-\lambda L(t,\mu)}dt}{\int w^*(u)e^{-\lambda L(t,u)}du} = 1 \ \forall \ \mu \in \mathcal{M}$$

Finally, λ is determined by

$$\int \int m(t)w^*_\lambda(\mu \mid t)L(t,\mu)d\mu dt = I$$

Solution to the minimization problem in the Exponential Family Case

$$w^*(\mu \mid t) = \frac{w^*(\mu)e^{[-\lambda(t\theta(\mu)-M_\theta(\mu))]}}{\int w^*(u)e^{[-\lambda(t\theta(u)-M_\theta(u))]}du}$$

where $w^*(\mu)$ is determined by

$$1 = \int \frac{p(t)e^{[-\lambda(t\theta(\mu)-M_\theta(\mu))]}}{\int w^*(u)e^{[-\lambda(t\theta(u)-M_\theta(u))]}du}dt$$

Setting $w^*(\mu) = H(\delta\gamma, \gamma)e^{[-\gamma(\delta\theta(\mu)-M_\theta(\mu))]}V(\mu)^{-1}$ then

$$1 = \int \frac{H(\delta\gamma + \lambda t, \gamma + \lambda)H(n_0 t_0, n_0)B(nt, n)}{H(n_0 t_0 + nt, n_0 + n)H(\delta\gamma, \gamma)B(\lambda t, \lambda)}p(t \mid \mu, \lambda)dt$$

Conclusions

- The prior which is closest to the posterior, in expected Kullback-Leibler sense, is the natural conjugate.
- Such priors attain the rate distortion function lower bound and thus, can be regarded as *maximally informative* and the corresponding likelihoods as *mainimally informative*.

Conclusions

- Conjugate priors should be used in situations where the researcher wishes to minimize the weight of the observations on inferences about the parameters.
- Conjugate priors and Jeffrey's prior can be regarded as extremes of the problem of elliciting a prior distribution.
- It is tempting to regard as "conjugate" any prior minimizing the mutual information between the parameter and the sample in more general settings.

References

- Gutierrez-Peña, E. and Muliere, P."Conjugate Priors Represent Strong Pre-Experimental Assumptions", Journal of Scandinavian Statistics, Vol. 31, Num. 2, (Jun 2004), pp. 235–246.