



The exponential family: conjugate priors

Paul Hewson

Overview: The phrase “exponential family” is a bad choice. It overlaps with other “exponential” functions. However, we are going to meet the “exponential family”. These notes are meant to be slightly interactive, mysterious green dots, squares and boxes appear which you can click on to answer questions and check solutions.

1. The exponential family

For two reasons, we need to mention to exponential family - next week we will use it to let us specify a generalised linear model. This week, we need to say a little more about conjugate priors.

Consider a general p -parameter distribution where in generality $\theta = (\theta_1, \dots, \theta_p)$ (there may be only one parameter θ)

- The support of the distribution must *not* involve the unknown parameter θ .
- The parameters $\eta(\theta)$ are called the “natural parameters”
- The $t(x_i)$ are called natural statistics, and are sufficient (the likelihood for θ depends only on the data through $t(x_i)$)
- Assume there is no linear dependence between $\eta(\theta)$ s or between $t(x_i)$ s.



Back



Some lesser points about the exponential family:

- The family is full if there are as many parameters as statistics
- If there is a non-linear relationship among natural parameters there may be more natural sufficient statistics (a curved exponential family)



Back

◀ Doc

Doc ▶

Where $A(\theta)$, $c(x)$, $\eta(\theta)$ and $t(x_i)$ are known functions for each i , we can rewrite the density in the form:

$$p(x_i|\theta) = c(x_i)A(\theta)e^{\eta(\theta)z t(x_i)} \quad (1)$$

Quiz Consider the Poisson distribution, the density is given here:

$$p(x_i|\lambda) = (1/x_i!) \exp(-\lambda)e^{\log(\lambda)x_i}$$

1. The term $\eta(\theta)$ (the natural parameter) in the exponential family density corresponds to which term in the Poisson:

(a) $1/x_i!$ (b) $\log(\lambda)$ (c) *Constant*

2. The term $c(x_i)$ in the exponential family density corresponds to which term in the Poisson:

(a) $1/x_i!$ (b) $\log(\lambda)$ (c) *Constant*

Given the general expression for the exponential family density in (1), we have the likelihood in this form:

$$L(\theta|y) \propto A(\theta)^n e^{\eta(\theta)^z t(x)}$$

If the prior density can be specified as:

$$p(\theta) \propto A(\theta)^a e^{\eta(\theta)^z b}$$

Then the posterior can be written as:

$$p(\theta|x) \propto A(\theta)^{a+n} e^{\eta(\theta)^z (b+t(x))}$$

- The exponential families are the only distributions which have natural conjugate prior distributions
- With only a few exceptions, the only distributions having a fixed number of sufficient statistics for all n are exponential type

2. The Poisson distribution

We wish to evaluate the posterior distribution of the Poisson rate parameter λ , given some data \mathbf{x} . Recall that the posterior is given by:

$$post(\lambda|\mathbf{x}) \propto p(\lambda)L(\lambda)$$

We will find that we become interested in the Gamma distribution written, expressed in the form of a rate parameter rather than a scale parameter (where rate = 1/scale).

$$f(\lambda) = \frac{r^a}{\Gamma(a)} e^{-\lambda r} \lambda^{(a-1)} \text{ for } \lambda > 0.$$

(When using this parameterisation, $E[\lambda] = a/r$, $Var[\lambda] = a/r^2$ and the mode is given by $(a - 1)/r$ if $a > 1$).

As we have noted, we can ignore constants not involving λ , so we only need to watch:

$$f(\lambda) \propto e^{-\lambda r} \lambda^{(a-1)} \tag{2}$$

instead of the entire *Gamma* density.

2.1. A uniform prior

When we examined Binomial data we considered a uniform prior for π . This makes sense as $\int_0^1 g(\pi) = 1$. Clearly this cannot apply for a Poisson parameter λ as $\int_0^\infty g(\lambda) = \infty$.

For the Poisson distribution, the uniform is an “improper prior”. But inference is still possible provided the posterior is proper.

In general, the posterior is given by:

$$post(\lambda|\mathbf{x}) \propto p(\lambda)L(\lambda|\mathbf{x})$$

[Back](#)[◀ Doc](#)[Doc ▶](#)

so using $p(\lambda) = 1$ (for $0 < \lambda < \infty$) then we have:

$$\text{post}(\lambda|\mathbf{x}) \propto 1 \times \lambda^{\sum x} e^{-n\lambda}$$

This is the same shape as the likelihood, and is a $\text{Gamma}(a', r')$ distribution but where we have updated the values of a and r :

Quiz When using a Uniform Prior with a Poisson likelihood, the posterior could be considered as a Gamma distribution with the following parameters:

1. The shape a' is given by:

(a) $\sum x$ (b) $\sum x + 1$ (c) n (d) $n\lambda$

2. The rate r' is given by:

(a) $\sum x$ (b) $\sum x + 1$ (c) n (d) $n\lambda$



Back

◀ Doc

Doc ▶

2.2. Jeffrey's prior

Here we would use $p(\lambda) \propto \frac{1}{\sqrt{\lambda}}$ for $0 < \lambda < \infty$. This also is improper!

$$post(\lambda|\mathbf{x}) \propto \frac{1}{\sqrt{\lambda}} \times \lambda^{\sum x} e^{-n\lambda}$$

this becomes

$$post(\lambda|\mathbf{x}) \propto \lambda^{\sum x - \frac{1}{2}} e^{-n\lambda}$$

Quiz When using a Jeffreys Prior $p(\lambda) \propto \frac{1}{\sqrt{\lambda}}$ with a Poisson likelihood, the posterior is a Gamma distribution with the following parameters:

1. The shape a' is given by:

(a) $\sum x + \frac{1}{2}$ (b) $\sum x - \frac{1}{2}$ (c) n (d) $n\lambda$

2. The rate r' is given by:

(a) $\sum x + \frac{1}{2}$ (b) $\sum x - \frac{1}{2}$ (c) n (d) $n\lambda$



Back

◀ Doc

Doc ▶

2.3. Conjugate prior

You may have noticed already that the Gamma density provides a conjugate prior.

Ignoring constants not involving λ , we have an expression for the prior as follows:

$$p(\lambda) \propto \lambda^{a-1} e^{-r\lambda}$$

Our likelihood can be reduced to:

$$L(\lambda|x) \propto \lambda^{\sum x} e^{-n\lambda}$$

We \therefore have an expression for the Posterior distribution as:

$$post(\lambda|x) \propto \lambda^{a-1+\sum x} e^{-(r+n)\lambda}$$

Quiz When using a Conjugate Prior $p(\lambda) \propto \lambda^{r-1} e^{-a\lambda}$ with a Poisson likelihood, the posterior is a Gamma distribution with the following parameters:

1. The shape a' is given by:

(a) $a + \sum x$ (b) $a - 1 + \sum x$ (c) $r + n$ (d) $-(r + n)$

2. The rate r' is given by:

(a) $a + \sum x$ (b) $a - 1 + \sum x$ (c) $r + n$ (d) $-(r + n)$



Back

◀ Doc

Doc ▶

3. Choosing a conjugate prior

Standard results for the Gamma distribution tell us that $E[\lambda] = \frac{a}{r}$ and $Var(\lambda) = \frac{a}{r^2}$

We can calculate an equivalent sample size using this formula:

$$\frac{\lambda_{Prior}}{n_{eq}} = \frac{a}{r^2}$$

Quiz We wish to investigate the Poisson rate parameter for the number of accidents on a stretch of traffic highway over $n = 24$ weeks. We wish the prior mean to be 2 and the prior standard deviation to be 1. We decide that we shall use a value of $a = 4$. What is the effective sample size of our prior?



Back

◀ Doc

Doc ▶

4. Summarising the posterior

For example, with a conjugate prior, the posterior was:

$$\text{post}(\lambda|x) \propto \lambda^{a-1+\sum x} e^{-(r+n)\lambda}$$

which is a $\text{Gamma}(a + \sum x, r + n)$ distribution

Notice it is more convenient to use notation a' and r' to express the posterior distribution in terms of a $\text{Gamma}(a', r')$ distribution, where in this case $a' = a + \sum x$ and $r' = r + n$.

Hence, we can express key summary measures for the Posterior as:

- Mode = $\frac{a'-1}{r'}$
- Posterior Mean $m_p = \frac{a'}{r'}$
- Posterior Variance $\text{Var}_p = \frac{a'}{(r')^2}$

4.1. Posterior mean square error

Strictly speaking, Mean Square Error does not make sense in a Bayesian framework, as our inference is concerned with a posterior distribution rather than a point. However, it is common to compare the performance of Bayesian and Frequentist estimates using this approach.

So, given a suitable point estimate, we can express the Posterior Mean Square for our estimator as:

$$\begin{aligned} PMS(\lambda) &= \int_0^{\infty} (\hat{\lambda} - \lambda)^2 post(\lambda|x) d\lambda \\ &= \int_0^{\infty} (\hat{\lambda} - m_p + m_p - \lambda)^2 post(\lambda|x) d\lambda \\ &= Var(\lambda|x) + 0 + (m_p - \hat{\lambda})^2 \end{aligned}$$

If we use the maximum likelihood estimator for the Poisson parameter λ :

$$\lambda_f = \frac{\sum x_i}{n}$$

we know this is unbiased. Accordingly, the mean square equals its variance:

$$MS(\lambda_f) = \frac{\lambda}{n}$$



Back

◀ Doc

Doc ▶

When we use a $Gamma(a, r)$ prior, the bias will be:

$$\begin{aligned} Bias(\lambda_B) &= E(\hat{\lambda} - \lambda) \\ &= E\left(\frac{a + \sum x_i}{r + n}\right) - \lambda \\ &= \frac{a - r\lambda}{r + n} \end{aligned}$$

The corresponding variance will be:

$$\begin{aligned} Var(\lambda_B) &= \left(\frac{1}{r + n}\right)^2 \sum var(x_i) \\ &= \frac{n\lambda}{(r + n)^2} \end{aligned}$$

Remember, $MSE = Bias^2 + Variance$. Maybe you can see circumstances in which the MSE of the Bayesian estimator could be lower than the Frequentist estimator.

Quiz We are investigating accidents on a stretch of road. We believe the rate parameter will be around $\lambda = 3$. We collect data for $n = 6$ weeks. We decide to perform analysis using Frequentist methods (based on the Maximum Likelihood) and Bayesian methods, using a conjugate Gamma Prior with $a = 2$ and $r = 1$

The frequentist estimate for λ has mean square error:

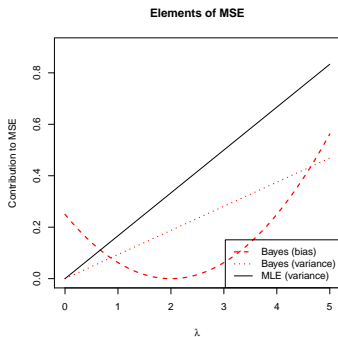
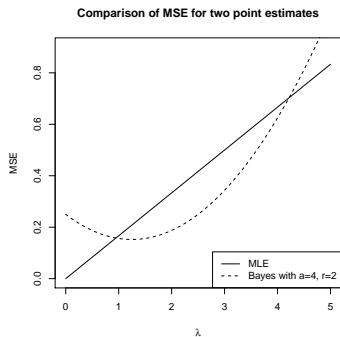
$$MSE(\hat{\lambda}_F) = \frac{\lambda}{6} \quad (3)$$

For the Bayesian estimate, the mean square error is:

$$MSE(\hat{\lambda}_B) = \left(\frac{4 - 2\lambda}{2 + 6} \right)^2 + \left(\frac{6\lambda}{(2 + 6)^2} \right)$$

1. Assume $\lambda = 3$, find the MSE for the Frequentist estimator:
2. Assume $\lambda = 3$, find the MSE for the Bayesian estimator:

We can sketch the mean square error as a function of λ for both the maximum likelihood estimate and the Bayesian estimate (with a $\text{Gamma}(2, 1)$ prior).



Note that for a certain range, a point estimate derived from the (Bayesian) posterior distribution has lower mean square error than a point estimate derived from the maximum likelihood estimate.

Solutions to Quizzes

Solution to Quiz: Firstly we have $r = \frac{\text{mean}}{\text{variance}} = \frac{2}{1} = 2$. Now we can solve $a = \frac{\text{mean}^2}{\text{variance}} = \frac{2^2}{1} = 4$. The effective sample size is therefore found by solving $\frac{2}{\lambda} = \frac{4}{2^2}$ which gives 2.

We might feel that a prior with an “Equivalent sample size” of 2 should not unduly influence a posterior which is based on a likelihood with $n = 24$ data points.

Click on that green button to return to the quiz →



Solution to Quiz: The Maximum Likelihood Estimate is unbiased, hence the only contribution to the MSE comes from the Variance. From equation 3 we have:

$$MSE(\hat{\lambda}_F) = \frac{3}{6} = 0.5$$



Solution to Quiz: $MSE(\hat{\lambda}_B) = \left(\frac{4-2 \times 3}{2+6}\right)^2 + \left(\frac{6 \times 3}{(2+6)^2}\right) = 0.34375$

