Bayesian updating of distributions in the exponential family

1 The likelihood distribution

The probability density function (or probability mass function, in the case of a discrete random variable) of an exponential family distribution is

$$p(x \mid \eta) = g(\eta)h(x)\exp(\eta \cdot T(x)) \tag{1}$$

where x is the random variable, η are the natural parameters, $g(\eta)$ is the normalization factor, $h(x)$ is the base measure, and $T(x)$ is a sufficient statistic. The sufficient statistic fully summarizes the data within the probability density function. For data $X = (x_1, \ldots, x_n)$, the likelihood is

$$\begin{aligned} p(X \mid \eta) &= \prod_{i=1}^{n} g(\eta)h(x)\exp(\eta \cdot T(x)) \\ &= g(\eta)^n \left(\prod_{i=1}^{n} h(x) \right) \exp\left(\eta \cdot \sum_{i=1}^{n} T(x_i) \right) \end{aligned} \tag{2}$$

1.1 Example: Poisson distribution

The Poisson distribution is defined by one parameter λ, which represents the expected value and variance of the distribution. It can be expressed in the form of an exponential family distribution as follows:

$$\eta = \ln \lambda \tag{3}$$

$$h(x) = \frac{1}{x!} \tag{4}$$

$$T(x) = x \tag{5}$$

$$g(\eta) = \exp(-\exp \eta) = \exp(-\lambda) \tag{6}$$

Hence

$$
\begin{aligned}
p(x \mid \eta) &= g(\eta)h(x)\exp(\eta \cdot T(x)) \\
&= \exp(-\lambda)\frac{1}{x!}\exp(x \ln \lambda) \\
&= \frac{\lambda^x}{x!}\exp(-\lambda)
\end{aligned} \tag{7}
$$

yielding the familiar expression for the probability mass function.

1.2 Example: Normal distribution

The normal distribution is defined by two parameters μ and λ, which represent the mean and variance of the distribution, respectively. It can be expressed in the form of an exponential family distribution as follows:

$$
\eta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \tag{8}
$$

$$
h(x) = \frac{1}{\sqrt{2\pi}} \tag{9}
$$

$$
T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \tag{10}
$$

$$
g(\eta) = \exp\left(\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\ln(-2\eta_2)\right) = \exp\left(-\frac{\mu^2}{2\sigma^2} - \ln \sigma\right) \tag{11}
$$

$$
= \sqrt{-2\eta_2}\exp\left(\frac{\eta_1^2}{4\eta_2^2}\right) = \frac{1}{\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) \tag{12}
$$

Hence

$$
\begin{aligned}
p(x \mid \eta) &= g(\eta)h(x)\exp(\eta \cdot T(x)) \\
&= \frac{1}{\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \cdot \begin{bmatrix} x \\ x^2 \end{bmatrix}\right) \\
&= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\right) \\
&= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2}\right) \\
&= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\end{aligned} \tag{13}
$$

yielding the familiar expression for the probability density function.

2 The posterior distribution

Consider the problem of determining the parameters of the distribution given an observation x. From Bayes' theorem, the posterior distribution is the product of the likelihood distribution $p(x \mid \eta)$ and the prior distribution $p(\eta)$, normalized by the probability $p(x)$ of the data:

$$p(\eta \mid x) = \frac{p(x \mid \eta)p(\eta)}{p(x)} = \frac{p(x \mid \eta)p(\eta)}{\int_{\eta'} p(x \mid \eta')p(\eta')\mathrm{d}\eta'} \tag{14}$$

For certain distributions, the posterior can be determined analytically. The conjugate prior gives a closed-form expression for the posterior. All exponential family distributions have conjugate priors, which take the form

$$p(\eta \mid \chi, \nu) = f(\chi, \nu)g(\eta)^{\nu} \exp(\eta \cdot \chi) \tag{15}$$

where $f(\chi, \nu)$ is a normalization constant and χ and ν are hyperparameters. Hyperparameters describe how the parameters of a distribution are themselves distributed. Hence the posterior distribution is

$$\begin{aligned} p(\eta \mid \chi, \nu, X) &\propto p(X \mid \eta)p(\eta \mid \chi, \nu) \\ &= g(\eta)^n \left(\prod_{i=1}^{n} h(x)\right) \exp\left(\eta \cdot \sum_{i=1}^{n} T(x_i)\right) f(\chi, \nu)g(\eta)^{\nu} \exp(\eta \cdot \chi) \\ &= g(\eta)^{\nu+n} \exp\left(\eta \cdot \left(\chi + \sum_{i=1}^{n} T(x_i)\right)\right) f(\chi, \nu) \left(\prod_{i=1}^{n} h(x)\right) \\ &\propto g(\eta)^{\nu+n} \exp\left(\eta \cdot \left(\chi + \sum_{i=1}^{n} T(x_i)\right)\right) \\ &= p\left(\eta \mid \chi + \sum_{i=1}^{n} T(x_i), \nu + n\right) f\left(\chi + \sum_{i=1}^{n}, \nu + n\right)^{-1} \\ &\propto p\left(\eta \mid \chi + \sum_{i=1}^{n} T(x_i), \nu + n\right) \end{aligned} \tag{16}$$

This is the kernel of the prior distribution, hence

$$\begin{aligned} p(\eta \mid \chi, \nu, X) &= p\left(\eta \mid \chi + \sum_{i=1}^{n} T(x_i), \nu + n\right) \\ &= p\left(\eta \mid \chi', \nu'\right) \end{aligned} \tag{17}$$

where χ' and ν' are the posterior (updated) hyperparameters.

2.1 Example: Poisson distribution

The conjugate prior of the Poisson distribution is the Gamma distribution, with prior hyperparameters α and β:

$$\mathrm{Gamma}(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \tag{18}$$

The interpretation of these is that there are α occurrences of an event in β intervals. The posterior hyperparameters are

$$\alpha' = \alpha + \sum_{i=1}^{n} x_i \tag{19}$$

$$\beta' = \beta + n \tag{20}$$

2.2 Example: Normal distribution

The conjugate prior of the normal distribution is the normal-gamma distribution, with prior hyperparameters μ_0, ν, α, and β:

$$\mathrm{NG}(x, \tau \mid \mu_0, \nu, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi}} \tau^{\alpha - \frac{1}{2}} \exp\left(-\tau \frac{2\beta + \nu(x-\mu)^2}{2}\right) \tag{21}$$

The interpretation of these is that mean was estimated from ν observations with sample mean μ_0 and the variance was estimated from 2α observations with a sum of squared deviations 2β. The posterior hyperparameters are

$$\mu_0' = \frac{\nu\mu_0 + n\bar{x}}{\nu + n} \tag{22}$$

$$\nu' = \nu + n \tag{23}$$

$$\alpha' = \alpha + \frac{n}{2} \tag{24}$$

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\nu n(\bar{x} - \mu_0)^2}{2(\nu + n)} \tag{25}$$

where $\bar{x}$ is the sample mean.

3 The predictive distribution

The probability density function of the model distribution can be expressed in terms of the hyperparameters by marginalizing over the parameters:

$$
\begin{aligned}
p(x \mid \chi, \nu) &= \int_\eta p(x \mid \eta) p(\eta, \chi, \nu) \mathrm{d}\eta \\
&= \int_\eta g(\eta) h(x) \exp(\eta \cdot T(x)) f(\chi, \nu) g(\eta)^\nu \exp(\eta \cdot \chi) \mathrm{d}\eta \\
&= h(x) f(\chi, \nu) \int_\eta g(\eta)^{\nu+1} \exp(\eta \cdot (\chi + T(x))) \mathrm{d}\eta \\
&= h(x) f(\chi, \nu) \int_\eta \frac{p(\eta \mid \chi + T(x), \nu + 1)}{f(\chi + T(x), \nu + 1)} \mathrm{d}\eta \\
&= \frac{h(x) f(\chi, \nu)}{f(\chi + T(x), \nu + 1)} \int_\eta p(\eta \mid \chi + T(x), \nu + 1) \mathrm{d}\eta \\
&= \frac{h(x) f(\chi, \nu)}{f(\chi + T(x), \nu + 1)}
\end{aligned} \tag{26}
$$

This is the predictive distribution of observing a new data point x given the data observed so far, with the parameters marginalized out.

3.1 Example: Poisson distribution

The predictive distribution of the Poisson distribution is given by

$$
p(x \mid \alpha, \beta) = \mathrm{NB}\left(x \mid \alpha', \frac{1}{1 + \beta'}\right) \tag{27}
$$

where primed variables indicate the posterior values of the hyperparameters and $\mathrm{NB}(x \mid r, p)$ is the function of a negative binomial distribution with r failures and a probability p of success in each trial:

$$
\mathrm{NB}(x \mid r, p) = \binom{x + r - 1}{x} (1 - p)^r p^x \tag{28}
$$

3.2 Example: Normal distribution

The predictive distribution of the normal distribution is given by

$$
p(x \mid \mu, \eta, \alpha, \beta) = t_{2\alpha'}\left(x \mid \mu', \frac{\beta'(\nu' + 1)}{\alpha' \nu'}\right) \tag{29}
$$

where $t_\nu(x \mid \mu, \sigma)$ refers to Student's t-distribution with n degrees of freedom, centered at μ and scaled by σ:

$$
t_\nu(x \mid \mu, \sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma \Gamma(\frac{\nu}{2}) \sqrt{\nu \pi}} \left(1 + \frac{1}{\nu} \left(\frac{x - \mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}} \tag{30}
$$

Note that σ in this equation does not correspond to a standard deviation. It simply sets the overall scaling of the distribution.