1 Exponential Family

If the probability mass/density function of a probability distribution with R parameters $\boldsymbol{\theta}$ can be written in the form

$$P(x|\boldsymbol{\theta}) = h(x) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(x) - A(\boldsymbol{\theta})\right) \tag{1}$$

where $h(x)$ is any function of x, $\eta(\boldsymbol{\theta})$ is a vector of reparametrisations called the *natural parameters* of P, $\mathbf{T}(x)$ is a vector containing R transformations of x called the *natural sufficient statistics* of P, and $A(\boldsymbol{\theta})$ is a partition function, then P is a member of the exponential family of distributions.

1.1 Maximum likelihood estimation for the exponential family

For any exponential family distribution

$$1 = \int P(x|\boldsymbol{\theta})dx = \int h(x) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(x) - A(\boldsymbol{\theta})\right) dx \tag{2}$$

because it's a probability distribution. Taking the derivative of both sides with respect to the rth natural parameter eventually gives

$$E_P\left[[\mathbf{T}(x)]_r\right] = \frac{\partial A(\boldsymbol{\theta})}{\partial \eta(\boldsymbol{\theta})_r} \tag{3}$$

To estimate the parameters of the distribution using some i.i.d. samples $\{x_1, \ldots, x_n\}$, we maximise the likelihood of these data under the probability distribution by varying the parameters, whatever parameter values maximise the likelihood are our estimates for the parameters. To do this, we take the derivatives of the likelihood with respect to each of the natural parameters and set this derivative to zero. In practice, we use the log-likelihood.

The likelihood takes the form

$$P(\{x_1, \ldots, x_n\}|\boldsymbol{\theta}) = \prod_{i=1}^{n} h(x_i) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(x_i) - A(\boldsymbol{\theta})\right) \tag{4}$$

The log-likelihood takes the form

$$l(X) = \sum_{i=1}^{n} \log h(x_i) + \boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(x_i) - A(\boldsymbol{\theta}) \tag{5}$$

Therefore the derivative takes the form

$$\frac{\partial l(X)}{\partial \eta(\boldsymbol{\theta})_r} = \sum_{i=1}^{n} \left[\mathbf{T}(x_i)\right]_r - \frac{\partial A(\boldsymbol{\theta})}{\partial \eta(\boldsymbol{\theta})_r} \tag{6}$$

Setting this expression equal to 0 gives

$$\sum_{i=1}^{n} \left[\mathbf{T}(x_i)\right]_r = n\frac{\partial A(\boldsymbol{\theta})}{\partial \eta(\boldsymbol{\theta})_r} = nE_P\left[[\mathbf{T}(x)]_r\right] \tag{7}$$

So, if we cannot solve the left hand equations in 7, we can use the equity of the left and right expressions in 7 as requirements in numerically estimating the parameters.

1.2 Bayesian parameter estimation

Every member of the exponential family has a *conjugate prior*. Given some likelihood function, a conjugate prior distribution to that likelihood is one that results in a posterior distribution of the same form as that prior distribution when used in Bayes's Rule. The conjugate prior for a probability distribution in the form of 1 has a has the form

$$P(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) \exp(\boldsymbol{\eta} \cdot \boldsymbol{\chi} - \nu A(\boldsymbol{\eta})) \tag{8}$$

where $\boldsymbol{\eta}$ represent the natural parameters of the likelihood. Then $\boldsymbol{\chi}$ and ν are hyperparameters, $f(\boldsymbol{\chi}, \nu)$ is a normalising function, and $A(\boldsymbol{\eta})$ is the same partition function as in 1. Note that $\nu > 0$ and $\boldsymbol{\chi} \in \mathbb{R}^s$ where s is the dimensionality of $\boldsymbol{\eta}$ (the number of parameters). ν controls the effective number of observations contributed by the prior. $\boldsymbol{\chi}$ controls the total amount that these pseudo-observations contribute to the sufficient statistic over all observations and pseudo-observations. Note also that f is defined exactly by the the other functions.

If equation 4 is our likelihood and equation 8 is our prior, then our posterior takes the form

$$\begin{aligned} P(\boldsymbol{\eta}|\mathbf{X}, \boldsymbol{\chi}, \nu) &= P(X|\boldsymbol{\eta}) P(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) && (9) \\ &= \left(\prod_{i=1}^{n} h(x_i) \exp\left(\boldsymbol{\eta} \cdot \mathbf{T}(x_i) - A(\boldsymbol{\eta})\right) \right) f(\boldsymbol{\chi}, \nu) \exp(\boldsymbol{\eta} \cdot \boldsymbol{\chi} - \nu A(\boldsymbol{\eta})) && (10) \\ &= \left(f(\boldsymbol{\chi}, \nu) \prod_{i=1}^{n} h(x_i) \right) \exp\left(\sum_{i=1}^{n} \boldsymbol{\eta} \cdot \mathbf{T}(x_i) - nA(\boldsymbol{\eta}) + \boldsymbol{\eta} \cdot \boldsymbol{\chi} - \nu A(\boldsymbol{\eta}) \right) && (11) \\ &= \left(f(\boldsymbol{\chi}, \nu) \prod_{i=1}^{n} h(x_i) \right) \exp\left(\boldsymbol{\eta} \cdot \left(\boldsymbol{\chi} + \sum_{i=1}^{n} \mathbf{T}(x_i) \right) - (n + \nu) A(\boldsymbol{\eta}) \right) && (12) \end{aligned}$$

This is the same form as the prior. We can write the posterior as

$$P(\boldsymbol{\eta}|\mathbf{X}, \boldsymbol{\chi}, \nu) = P(\boldsymbol{\eta}|\boldsymbol{\chi} + \sum_{i=1}^{n} \mathbf{T}(x_i), n + \nu) \tag{13}$$

The maximum a priori estimate for the parameters $\boldsymbol{\eta}$ is the mode of this posterior distribution.

2 Normal binomial distribution

2.1 Probability mass function

The probability mass function (p.m.f.) for a random variable X with a 'normal' binomial distribution with probability of success p, and m trials is

$$P(X = k) = \binom{m}{k} p^k (1 - p)^{m-k} \tag{14}$$

that is k successes and $m - k$ failures with a coefficient for the number of different ways we can have k successes and $m - k$ failures.

2.2 Likelihood function

If we take n i.i.d. samples from X consisting of n integers between 0 and m, $\{k_1, \dots, k_n\}$, the probability of these data, aka the likelihood is

$$P(\{k_1, \dots, k_n\}|p) = L(K|p, m) = \prod_{i=1}^{n} \binom{m}{k_i} p^{k_i}(1-p)^{m-k_i} \tag{15}$$

that is the product of all the samples.

2.3 Maximum likelihood estimation

In order to estimate the value of p from the sample $\{k_1, \dots, k_n\}$, we maximise the value of the likelihood function with respect to p. Since $\log$ is an increasing function, maximising the log of the likelihood function is equivalent to maximising the likelihood function. But because likelihood functions tend to take the form of large products, maximising the log of the likelihood is often easier than maximising the likelihood. So we will maximise the log of the likelihood function, a.k.a. the 'log likelihood'.

$$l(p, m) = \log L(K|p, m) = \log \left(\prod_{i=1}^{n} \binom{m}{k_i} p^{k_i}(1-p)^{m-k_i} \right) \tag{16}$$

$$= \sum_{i=1}^{n} \log \binom{m}{k_i} + k_i \log p + (m - k_i) \log(1-p) \tag{17}$$

The derivative with respect to p is

$$\frac{\partial l(p, m)}{\partial p} = \sum_{i=1}^{n} \frac{k_i}{p} - \frac{m - k_i}{1-p} \tag{18}$$

Letting the derivative equal 0 gives

$$\sum_{i=1}^{n} \frac{k_i}{\hat{p}} - \frac{m - k_i}{1-\hat{p}} = 0 \tag{19}$$

$$\implies \sum_{i=1}^{n} \frac{k_i(1-\hat{p}) - \hat{p}(m - k_i)}{\hat{p}(1-\hat{p})} = 0 \tag{20}$$

$$\implies \sum_{i=1}^{n} k_i - k_i\hat{p} - m\hat{p} + k_i\hat{p} = 0 \tag{21}$$

$$\implies \hat{p} = \frac{1}{nm} \sum_{i=1}^{n} k_i \tag{22}$$

As you might expect, our maximum likelihood estimate for p is the total number of successes divided by the total number of trials. It's interesting to note that m is technically

a parameter of the binomial distribution, and if we didn't know m beforehand, we would be unable to estimate p or m using the maximum likelihood method (I think).

The number of trials m is also a parameter of the binomial distribution. In order to find an estimate for m using maximum likelihood estimation, we take the derivative of $l(p, m)$ with respect to m.

$$\frac{\partial l(p,m)}{\partial m} = \sum_{i=1}^{n} \frac{\partial \log \binom{m}{k_i}}{\partial m} + \log(1-p) \tag{23}$$

Using

$$\frac{\partial \log \binom{m}{k_i}}{\partial m} = H_m - H_{m-k_i} \tag{24}$$

where H_n is the nth Harmonic number [1], and letting the derivative equal 0 gives

$$\sum_{i=1}^{n} H_m - H_{m-k_i} + \log(1-p) = 0 \tag{25}$$

$$\tag{26}$$

Theoretically we could sub in our expression for $\hat{p}$ and solve for m (??? is this true ???). But I don't know how to do this right now.

2.4 Bayesian parameter estimation

If we consider m to be known and fixed, then the binomial distribution is an exponential family member.

$$P(k|p) = \binom{m}{k} p^k (1-p)^{m-k} \tag{27}$$

$$= \binom{m}{k} \exp\left(k \log p + (m-k)\log(1-p)\right) \tag{28}$$

$$= \binom{m}{k} \exp\left(k \log \frac{p}{1-p} + m \log(1-p)\right) \tag{29}$$

this takes the form of equation 1 where $h(x) = \binom{m}{k}$, $\boldsymbol{\eta}(\boldsymbol{\theta}) = \log p/(1-p)$, $\mathbf{T}(k) = k$, and $A(\boldsymbol{\theta}) = -m \log 1 - p$.

This means the conjugate prior for the binomial distribution takes the form

$$P(p|\chi, \nu) = f(\chi, \nu) \exp\left(\chi \log \frac{p}{1-p} - \nu m \log(1-p)\right) \tag{30}$$

or in natural parameter form

$$P(\eta|\chi, \nu) = f(\chi, \nu) \exp\left(\chi\eta + \nu m \log(1+e^{\eta})\right) \tag{31}$$

[1] https://en.wikipedia.org/wiki/Harmonic_number

which gives a posterior of the form

$$P(\eta|\mathbf{X},\chi,\nu) = \left(\prod_{i=1}^{n}\binom{m}{k_i}\exp\left(\eta k_i + m\log(1+e^{\eta})\right)\right) f(\chi,\nu)\exp\left(\chi\eta + \nu m\log(1+e^{\eta})\right) \tag{32}$$

$$= \left(f(\chi,\nu)\prod_{i=1}^{n}\binom{m}{k_i}\right)\exp\left(\eta(\chi+\sum_{i=1}^{n}k_i) + m\log(1+e^{\eta})(n+\nu)\right) \tag{33}$$

$$\propto \exp\left(\eta(\chi+\sum_{i=1}^{n}k_i) + m\log(1+e^{\eta})(n+\nu)\right) \tag{34}$$

In order to estimate p for our distribution, we pick values for χ and ν, we find $\hat{\eta} = \arg\max_{\eta} P(\eta|\mathbf{X},\chi,\nu)$, then $\hat{p} = e^{\hat{\eta}}/(1+e^{\hat{\eta}})$.

Now we know how to do maximum likelihood estimation, we move onto a more complicated distribution.

3 Conway-Maxwell binomial distribution

The Conway-Maxwell binomial distribution is similar to the binomial distribution in that we can think of it as a sum of Bernoulli trials. But for the Conway-Maxwell binomial distribution, the Bernoulli variables are associated with each other. Not dependent, not correlated, but associated.

3.1 Probability mass function

The p.m.f. of a random variable X with the Conway-Maxwell binomial distribution with probability of success p, dispersion parameter ν, and number of trials m is

$$P(X=k) = \frac{1}{S(p,\nu,m)}\binom{m}{k}^{\nu} p^k(1-p)^{m-k} \tag{35}$$

where

$$S(p,\nu,m) = \sum_{i=0}^{m}\binom{m}{i}^{\nu} p^i(1-p)^{m-i} \tag{36}$$

is a normalising function.

The use of the dispersion parameter ν enables the Conway-Maxwell binomial distribution to 'over-disperse' or 'under-disperse' relative to a binomial distribution i.e., have greater or lesser variance.

When $\nu > 1$ the distribution is under-dispersed relative to a binomial distribution. In the limit that $\nu \to \infty$ all the mass accumulates at $m/2$ for even m, and at $\lfloor m/2 \rfloor$ and $\lceil m/2 \rceil$ for odd m. This corresponds to negatively associated Bernoulli variables. For $\nu < 1$, the distribution is over-dispersed relative to a binomial distribution. In the case where $\nu \to -\infty$ all the mass is distributed at 0 and m. This is the extreme case of

positive association, where all the Bernoulli variables have the value. So, ν measures the strength of the negative or positive association between the Bernoulli variables that make up the Conway-Maxwell binomial distribution. Note that if $\nu = 1$ we have the 'normal' binomial distribution described in section 2.

3.2 Likelihood function

If we take n i.i.d. samples from a random variable X with the Conway-Maxwell binomial distribution with parameters p, ν, and m this gives us n integers between 0 and m, $\{k_1, \dots, k_n\}$. These probability or 'likelihood' of these data is

$$P(\{k_1, \dots, k_n\}|p, \nu, m) = L(K|p, \nu, m) = \prod_{i=1}^{n} \frac{\binom{m}{k_i}^{\nu} p^{k_i}(1-p)^{m-k_i}}{S(p, \nu, m)} \tag{37}$$

$$= \frac{\prod_{i=1}^{n} \binom{m}{k_i}^{\nu} p^{k_i}(1-p)^{m-k_i}}{S(p, \nu, m)^n} \tag{38}$$

Again, just the product of the individual probabilities of each sample.

3.2.1 Exponential family form

The likelihood function for a single data point sampled from the Conway-Maxwell binomial distributed variable can be rewritten as

$$P(k|p, \nu, m) = \frac{\binom{m}{k}^{\nu} p^k (1-p)^{m-k}}{S(p, \nu, m)} \tag{39}$$

$$= \frac{m!^{\nu}}{S(p, \nu, m)} \frac{(1-p)^m}{k!^{\nu}(m-k)!^{\nu}} \left(\frac{p}{1-p}\right)^k \tag{40}$$

$$= \exp(k \log \frac{p}{1-p} - \nu \log(k!(m-k)!) \tag{41}$$

$$+ \nu \log(m!) + m \log(1-p) - \log S(p, \nu, m)) \tag{42}$$

If we let $\pi = \log \frac{p}{1-p}$, then we have

$$P(k|p, \nu, m) = \exp\left(k\pi - \nu \log(k!(m-k)!) - m \log(1+e^{\pi}) + \nu \log m! - \log S(\pi, \nu, m)\right) \tag{43}$$

$$= h(k) \exp\left(\boldsymbol{\theta} \cdot \mathbf{T}(x) - A(\boldsymbol{\theta})\right) \tag{44}$$

where $\boldsymbol{\theta} = (\pi, -\nu)$, $h(k) = 1$, $A(\boldsymbol{\theta}) = m \log(1 + e^{\pi}) - \nu \log m! + \log S(\pi, \nu, m)$ and $\mathbf{T}(x) = (k, \log(k!(m-k)!)$. When we consider m to be fixed and known, $\boldsymbol{\theta}$ are the natural parameters of the distribution, and $\mathbf{T}(x)$ are the natural sufficient statistics. After the parameter transformation

$$S(\pi, \nu, m) = \sum_{j=0}^{m} \binom{m}{j}^{\nu} \frac{e^{\pi j}}{(1+e^{\pi})^m} \tag{45}$$

3.3 Bayesian parameter estimation

The conjugate prior for the Conway-Maxwell binomial distribution takes the form

$$P(p, \nu|\boldsymbol{\chi}, c) = f(\boldsymbol{\chi}, c) \exp\left((p, \nu) \cdot \boldsymbol{\chi} - cA(p, \nu)\right) \tag{46}$$

or in natural parameter form

$$P(\pi, \nu|\boldsymbol{\chi}, c) = f(\boldsymbol{\chi}, c) \exp\left((\pi, \nu) \cdot \boldsymbol{\chi} - cA(\pi, \nu)\right) \tag{47}$$

where we use c to represent the pseudocount hyperparameter instead of ν, as in section 1.2. This is to avoid confusion with the dispersion parameter of the Conway-Maxwell binomial distribution, which we will still represent using ν.

3.4 Maximum likelihood estimation

Since the natural parameters of the Conway-Maxwell binomial distribution are $\left(\log \frac{p}{1-p}, -\nu\right)$ and the natural sufficient statistics are $(\sum_{i=1}^{n} k_i, \sum_{i=1}^{n} \log(k!(m-k)!))$, the values of p and ν satisfy

$$E_{P(p,\nu)}[k] = \frac{\partial A(\boldsymbol{\theta})}{\partial \pi} \tag{48}$$

$$E_{P(p,\nu)}[\log(k!(m-k)!)] = \frac{\partial A(\boldsymbol{\theta})}{\partial \nu} \tag{49}$$

at the point of maximum likelihood.

Once again, we maximise the log-likelihood function, which is

$$l(p, \nu, m) = \log L(K|p, \nu, m) = \log \frac{\prod_{i=1}^{n} \binom{m}{k_i}^{\nu} p^{k_i}(1-p)^{m-k_i}}{S(p, \nu, m)^n} \tag{50}$$

$$= -n \log S(p, \nu, m) + \sum_{i=1}^{n} \log \binom{m}{k_i}^{\nu} + k_i \log p + (m - k_i) \log(1-p) \tag{51}$$

Maximising will involve calculating the partial derivative of $\log S(p, \nu, m)$ with respect to the parameters.

$$\frac{\partial \log S(p, \nu, m)}{\partial p} = \frac{1}{S(p, \nu, m)} \frac{\partial S(p, \nu, m)}{\partial p} \tag{52}$$

For the partial derivative without the $\log$, we have

$$\frac{\partial S(p, \nu, m)}{\partial p} = \frac{\partial \sum_{i=0}^{m} \binom{m}{i}^{\nu} p^i (1-p)^{m-i}}{\partial p} \tag{53}$$

$$= \sum_{i=0}^{m} \binom{m}{i}^{\nu} \left[ip^{i-1}(1-p)^{m-i} - (m-i)p^i(1-p)^{m-i-1}\right] \tag{54}$$

which gives

$$\frac{\partial \log S(p,\nu,m)}{\partial p} = \sum_{i=0}^{m} \frac{\binom{m}{i}^{\nu}\left[ip^{i-1}(1-p)^{m-i} - (m-i)p^{i}(1-p)^{m-i-1}\right]}{\binom{m}{i}^{\nu}p^{i}(1-p)^{m-i}} \tag{55}$$

$$= \sum_{i=0}^{m} ip^{-1} - (m-i)(1-p)^{-1} \tag{56}$$

$$= \sum_{i=0}^{m} \frac{i}{p} - \frac{m-i}{1-p} \tag{57}$$

$$= \sum_{i=0}^{m} \frac{i(1-p) - (m-i)p}{p(1-p)} \tag{58}$$

$$= \sum_{i=0}^{m} \frac{i-mp}{p(1-p)} \tag{59}$$

$$= \frac{m(m+1) - 2m(m+1)p}{2p(1-p)} \tag{60}$$

$$= \frac{m(m+1)(1-2p)}{2p(1-p)} \tag{61}$$

using $\sum_{i=0}^{m} i = m(m+1)/2$.

So the partial derivative of the log-likelihood function with respect to p is

$$\frac{\partial l(p,\nu,m)}{\partial p} = -n\frac{\partial \log S(p,\nu,m)}{\partial p} + \sum_{i=1}^{n} \frac{k_i}{p} - \frac{m-k_i}{1-p} \tag{62}$$

$$= -n\frac{\partial \log S(p,\nu,m)}{\partial p} + \sum_{i=1}^{n} \frac{k_i - mp}{p(1-p)} \tag{63}$$

$$= \frac{-nm(m+1)(1-2p)}{2p(1-p)} + \sum_{i=1}^{n} \frac{k_i - mp}{p(1-p)} \tag{64}$$

$$= \frac{-nm(m+1)(1-2p) - 2nmp + 2\sum_{i=1}^{n} k_i}{2p(1-p)} \tag{65}$$

$$= \frac{-nm\left[(m+1)(1-2p) + 2p\right] + 2\sum_{i=1}^{n} k_i}{2p(1-p)} \tag{66}$$

$$= \frac{-nm\left[m - 2mp + 1 - 2p + 2p\right] + 2\sum_{i=1}^{n} k_i}{2p(1-p)} \tag{67}$$

$$= \frac{-nm\left[m(1-2p) + 1\right] + 2\sum_{i=1}^{n} k_i}{2p(1-p)} \tag{68}$$

References

[1] Joseph B. Kadane, *Sums of possibly associated Bernoulli variables: The Conway-Maxwell binomial distribution*. Bayesian Analysis 11, 403 - 420, (2016)