0.1 The exponential family

The *natural exponential family* with *canonical parameterisation* has p.d.f of the form

$$f(y) = \exp\left\{\frac{y\theta + c(\theta)}{\phi} + d(y,\phi)\right\}.$$

θ is the *canonical* (or *natural*) parameter for the distribution. The parameter ϕ is called the *scale* parameter. Most times we shall consider that this is known, and does not need to be estimated.

0.1.1 Mean and variance

The loglikelihood is

$$l(\theta,\phi;y) = \frac{y\theta + c(\theta)}{\phi} + d(y,\phi).$$

The score function is

$$s(\theta) = \frac{\partial l}{\partial \theta} = \frac{y + c'(\theta)}{\phi}$$

and this has mean zero, so

$$\mathrm{E}[Y] = -c'(\theta).$$

The leading element of the Hessian is

$$\frac{\partial^2 l}{\partial \theta^2} = \frac{c''(\theta)}{\phi}.$$

This does not involve y, so the leading element of the information matrix $\mathcal{I}$ is $-c''(\theta)/\phi$. However this is equal to the variance of $s(\theta)$ which is

$$\mathrm{E}\left[s(\theta)^2\right] = \mathrm{E}\left[\left(\frac{\partial L}{\partial \theta}\right)^2\right]$$

because $\mathrm{E}[s(\theta)] = 0$. So

$$\mathrm{E}\left[\left(\frac{Y - \mathrm{E}[Y]}{\phi}\right)^2\right] = -c''(\theta)/\phi$$

$$\frac{1}{\phi^2}\mathrm{Var}[Y] = -c''(\theta)/\phi$$

$$\mathrm{Var}[Y] = -\phi c''(\theta).$$

We have therefore shown that

$$\mu = \mathrm{E}[Y] = -c'(\theta),$$

and that

$$\sigma^2 = \mathrm{Var}[Y] = -\phi c''(\theta).$$

Note that

$$\mathrm{Var}[Y] = \phi\frac{d\mu}{d\theta} = \phi V(\mu)$$

where $V(\mu)$ is the *variance function*. Here are the variance functions and the *dispersion parameter* ϕ for some standard distributions.

Distribution	Variance function	Dispersion parameter
Normal	1	σ^2
Poisson	μ	1
Binomial	$\mu(1-\mu)$	1
Gamma	μ^2	$1/\alpha$

0.1.2 Moment Generating function

$$\begin{aligned} M_Y(t) &= \int e^{yt} e^{\frac{y\theta + c(\theta)}{\phi} + d(y,\phi)} dy \\ &= \int e^{\frac{y(\theta+t\phi)+c(\theta)}{\phi} + d(y,\phi} dy \\ &= e^{\frac{c(\theta)-c(\theta+t\phi)}{\phi}} \int e^{\frac{y(\theta+t\phi)+c(\theta+t\phi)}{\phi} + d(y,\phi)} dy \\ &= e^{\frac{c(\theta)-c(\theta+t\phi)}{\phi}}. \end{aligned}$$

The cumulant generating function is $\frac{c(\theta)-c(\theta+t\phi}{\phi}$.

0.1.3 Normal distribution

The Normal distribution has

$$\begin{aligned} \ln(f) &= -\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \\ &= \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right) \end{aligned}$$

So we take

$$\begin{aligned} \theta &= \mu, \\ \phi &= \sigma^2, \\ c(\theta) &= -\frac{1}{2}\theta^2, \text{ and} \\ d(y,\phi) &= -\frac{y^2}{2\phi} - \frac{1}{2}\ln(2\pi\phi). \end{aligned}$$

We check the mean

$$\begin{aligned} \mathrm{E}[Y] &= -c'(\theta) \\ &= \theta = \mu \end{aligned}$$

and the variance

$$\begin{aligned} \mathrm{Var}[Y] &= -\phi c''(\theta) \\ &= \phi = \sigma^2. \end{aligned}$$

The cumulant generating function is

$$-\frac{1}{2}\frac{\mu^2}{\sigma^2} + \frac{1}{2}\frac{(\mu+\sigma^2 t)^2}{\sigma^2}$$

which is

$$\mu t + \frac{1}{2}\sigma^2 t^2.$$

0.1.4 The Gamma distribution

$$f(y) = \frac{1}{\Gamma(\nu)} \frac{1}{y} \left(\frac{\nu y}{\mu}\right)^{\nu} e^{-\frac{\nu y}{\mu}}$$

so

$$ln(f) = \nu \left(y \left(-\frac{1}{\mu} \right) - \ln \mu \right) + (\nu - 1) \ln y - \ln \Gamma(\nu) + \nu \ln \nu.$$

So we take

$$\begin{aligned} \theta &= -\frac{1}{\mu}, \\ \phi &= \frac{1}{\nu}, \\ c(\theta) &= -\ln \mu = \ln(-\theta), \text{ and} \\ d(y, \phi) &= -\ln y - \ln \Gamma(\nu) + \nu \ln \nu. \end{aligned}$$

We check the mean

$$\begin{aligned} \mathrm{E}[Y] &= -c'(\theta) \\ &= -1/\theta = \mu \end{aligned}$$

and the variance

$$\begin{aligned} \mathrm{Var}[Y] &= -\phi c''(\theta) \\ &= \phi \frac{1}{\theta^2} = \frac{\mu^2}{\nu}. \end{aligned}$$

The cumulant generating function is

$$K_Y(t) = \frac{\ln(1\mu) + \ln(1/\mu - t/\nu)}{1/\nu} = \frac{1}{\nu} \ln \frac{1}{1 - \frac{\mu t}{\nu}}$$

0.1.5 The Poisson distribution

The probability mass function is

$$f(y) = e^{-\lambda} \lambda^y / y!$$

so

$$ln(f) = \frac{y \ln(\lambda) - \lambda}{1} - \ln y!.$$

So we take

$$\begin{aligned} \theta &= \ln \lambda \\ \phi &= 1 \\ c(\theta) &= -\lambda = -e^{\theta}, \text{ and} \\ d(y, \phi) &= -\ln y!. \end{aligned}$$

We check the mean

$$\begin{aligned} \mathrm{E}[Y] &= -c'(\theta) \\ &= e^{\theta} = \lambda. \end{aligned}$$

and the variance

$$\begin{aligned}\mathrm{Var}\,[Y] &= -\phi c''(\theta)\\ &= -1 \times (-e^{\theta}) = \lambda.\end{aligned}$$

The cumulant generating function is

$$\begin{aligned}K_X(t) &= -e^{\ln\lambda} + e^{\ln\lambda + t}\\ &= \lambda(e^t - 1).\end{aligned}$$

0.1.6 Binomial distribution

The probability mass function is

$$f(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}.$$

so

$$ln(f) = \frac{y\ln\frac{\pi}{1-\pi} + n\ln(1-\pi)}{1} + \ln\binom{n}{y}.$$

So we take

$$\begin{aligned}\theta &= \ln\frac{\pi}{1-\pi}\\ \phi &= 1\\ c(\theta) &= n\ln(1-\pi) = -n\ln(1+e^{\theta}),\text{ and}\\ d(y,\phi) &= \ln\binom{n}{y}.\end{aligned}$$

We check the mean

$$\begin{aligned}\mathrm{E}\,[Y] &= -c'(\theta)\\ &= n\frac{e^{\theta}}{1+e^{\theta}} = n\pi.\end{aligned}$$

and the variance

$$\begin{aligned}\mathrm{Var}\,[Y] &= -\phi c''(\theta)\\ &= n\frac{e^{\theta}}{(1+e^{\theta})^2} = n\pi(1-\pi).\end{aligned}$$

The cumulant generating function is

$$\begin{aligned}K_X(t) &= n\ln(1-\pi) + n\ln(1+e^{\ln\frac{\pi}{1-\pi}+t})\\ &= n\ln(1-\pi) + n\ln(1+\frac{\pi}{1-\pi}e^t)\\ &= n\ln(1-\pi) + n\ln(\frac{1-\pi+\pi e^t}{1-\pi})\\ &= n\ln(1-\pi+\pi e^t).\end{aligned}$$

0.1.7 Application to a sample

The loglikelihood of n *independent* and *identically distributed* observations from the distribution is

$$l(\boldsymbol{\beta}, \phi; y) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta + c(\theta)}{\phi} + d(y_i, \phi) \right\}$$

. With ϕ considered as fixed, it is immediately clear from the factorisation criterion that $\sum_i y_i$ is a sufficient statistic for θ.

0.2 Exponential Family: General case

The exponential family with canonical parameters are distributions with density functions or probability mass functions of the form

$$f(\boldsymbol{y}) = \exp\left[\sum_{i=1}^{s} \theta_i T_i(\boldsymbol{y}) + c(\boldsymbol{\theta}) + d(y) \right].$$

Here we don't explicitly put in the scale parameter that we considered before as a known parameter, but it may be absorbed into the functions $d(y), \theta_i$ so this family includes the previous one. The support of the density function or probability mass function is assumed not to depend on the parameters θ_i. To make for simplicity we assume that there is no linear dependence in the set of T_is or in the set of θ_is, and also that there is an s-dimensional rectangle of values θ_i for which the density or probability mass function is properly defined.

0.2.1 sufficiency

The factorisation criterion gives immediately that $(T_1(\boldsymbol{y}), \ldots, T_s(\boldsymbol{y}))$ is a sufficient statistic for $(\theta_1, \ldots, \theta_s)$.

Essentially, the only nice regular cases of distributions with simple sufficient statistics are those in the exponential family.

Note that $\boldsymbol{y}$ can be a vector of values in the definition, so that the joint density for a random sample of observations from a natural exponential family with canonical parameters is a member of the general exponential family.

0.2.2 Mean and Covariance

The loglikelihood is

$$l(\boldsymbol{\theta}; \boldsymbol{T}(\boldsymbol{y})) = \boldsymbol{T}(\boldsymbol{y})^T\boldsymbol{\theta} + c(\boldsymbol{\theta}) + d(\boldsymbol{y}).$$

The score function is

$$s(\boldsymbol{\theta}) = \frac{\partial l}{\partial \boldsymbol{\theta}} = \boldsymbol{T}(\boldsymbol{y}) + c'(\boldsymbol{\theta})$$

and this has mean zero, so

$$\mathrm{E}[\boldsymbol{T}(y)] = -c'(\boldsymbol{\theta}).$$

The Hessian is

$$\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = c''(\theta).$$

This does not involve y, so the information matrix $\mathcal{I}$ is $-c''(\boldsymbol{\theta})$. However this is equal to the variance-covariance matrix of $s(\boldsymbol{\theta})$ which is

$$\mathrm{E}\left[s(\boldsymbol{\theta})s(\boldsymbol{\theta})^T\right] = \mathrm{E}\left[\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)^T\right]$$

because $\mathrm{E}[s(\boldsymbol{\theta})] = 0$. So

$$\mathrm{E}\left[\left(\boldsymbol{T}(\boldsymbol{y}) - \mathrm{E}[\boldsymbol{T}(\boldsymbol{y})]\right)\left(\boldsymbol{T}(\boldsymbol{y}) - \mathrm{E}[\boldsymbol{T}(\boldsymbol{y})]\right)^T\right] = -c''(\boldsymbol{\theta}).$$

We have therefore shown that

$$\mu = \mathrm{E}[\boldsymbol{T}(\boldsymbol{y})] = -c'(\boldsymbol{\theta}),$$

and that

$$\mathrm{Var}[\boldsymbol{T}(\boldsymbol{y})] = -c''(\boldsymbol{\theta}).$$

0.2.3 Joint Cumulant Generating Function

It is easy to show that the joint cumulant generating function of $T_1(\boldsymbol{y}), T_2(\boldsymbol{y}), \ldots, T_k(\boldsymbol{y})$ is

$$c(\theta_1, \theta_2, \ldots, \theta_k) - c(t_1 + \theta_1, t_2 + \theta_2, \ldots, t_k + \theta_k)$$

0.2.4 Multinomial Distribution

For the multinomial distribution we have probability mass function

$$f(\boldsymbol{y}) = \frac{n!}{y_1! y_2! \ldots y_k!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_k^{y_k}$$

where the $\{y_i\}$ are non-negative integers with $\sum_i y_i = n$, and the $\{\pi_i\}$ are probabilities satisfying $\sum_i \pi_i = 1$. The probability mass function is

$$\begin{aligned} f(\boldsymbol{y}) &= \exp[y_1 \ln \pi_1 + y_2 \ln \pi_2 + \ldots y_k \ln \pi_k + \ln \frac{n!}{y_1! y_2! \ldots y_k!}] \\ &= \exp[y_1 \ln \pi_1 + y_2 \ln \pi_2 + \ldots (n - \sum_i y_i) \ln \pi_k + \ln \frac{n!}{y_1! y_2! \ldots y_k!}] \\ &= \exp[y_1 \ln(\pi_1/\pi_k) + y_2 \ln(\pi_2/\pi_k) + \cdots + y_{k-1} \ln(\pi_{k-1}/\pi_k) + n \ln \pi_k + \ln \frac{n!}{y_1! y_2! \ldots y_k!}]. \end{aligned}$$

So this is in the general exponential family with

$$\begin{aligned} T_i(\boldsymbol{y}) &= y_i \text{ for } i = 1, 2 \ldots, (k-1) \\ \theta_i &= \ln(\pi_i/\pi_k) \\ c(\boldsymbol{\theta}) &= n \ln \pi_k = n \ln(1/(1 + \sum_1^{k-1} \exp \theta_i) \\ d(\boldsymbol{y}) &= \ln \frac{n!}{y_1! y_2! \ldots y_k!}. \end{aligned}$$

We use in the above $\pi_k = \sum_{i=1}^{k-1} \pi_i$.

A sufficient statistic for $\pi_1, \pi_2, \ldots, \pi_{k-1}$ is $y_1, y_2, \ldots, y_{k-1}$.

The mean for $T_1(\boldsymbol{y})$ is

$$\begin{aligned}
-\frac{\partial}{\partial\theta_1}c\left(\boldsymbol{\theta}\right) &= \frac{\partial}{\partial\theta_1}n\ln(1/(1+\sum_1^{k-1}\exp\theta_i) \\
&= n\frac{\exp\theta_1}{1+\sum_1^{k-1}\exp\theta_i} \\
&= n\frac{\frac{\pi_1}{\pi_k}}{1+\sum_1^{k-1}\frac{\pi_i}{\pi_k}} \\
&= n\pi_1.
\end{aligned}$$

The variances and covariances follow on further differentiation.

0.2.5 Normal Distribution

With both mean and variance unknown, the normal distribution has density function

$$\begin{aligned}
f(y) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right] \\
&= \exp\left[y\frac{\mu}{\sigma^2}+y^2\frac{-1}{2\sigma^2}-\frac{\mu^2}{2\sigma^2}-\ln(\sqrt{2\pi}\sigma)\right].
\end{aligned}$$

So we could take

$$\begin{aligned}
T_1(y) &= y \\
T_2(y) &= y^2 \\
\theta_1 &= \frac{\mu}{\sigma^2} \\
\theta_2 &= \frac{-1}{2\sigma^2} \\
c(\boldsymbol{\theta}) &= -\frac{\mu^2}{2\sigma^2}-\ln\sigma = \frac{\theta_1^2}{4\theta_2}+\frac{1}{2}\ln(-\theta_2).
\end{aligned}$$

The means are

$$\begin{aligned}
E(y) &= -\frac{\partial}{\partial\theta_1}c(\boldsymbol{\theta}) = -\frac{2\theta_1}{4\theta_2} = \mu \\
E(y^2) &= -\frac{\partial}{\partial\theta_2}c(\boldsymbol{\theta}) \\
&= \frac{\theta_1^2}{4\theta_2^2}-\frac{1}{2\theta_2} \\
&= \mu^2+\sigma^2.
\end{aligned}$$

0.2.6 Minimum Variance Unbiased Estimators

In the simple case when there is just one parameter, the exponential family has density or probability mass function

$$f(\boldsymbol{y}) = \exp\left[\theta T(\boldsymbol{y}) + c(\theta) + d(\boldsymbol{y})\right].$$

The score function is

$$s(\theta) = T(\boldsymbol{y}) + c'(\theta)$$

and so the Fisher information for θ, which is the variance of the score function is

$$I(\theta) = \mathrm{Var}T(\boldsymbol{y}).$$

The Cramér-Rao lower bound on the variance of an estimator $S(\boldsymbol{y})$ which is unbiased for $-c'(\theta)$ is

$$\mathrm{Var}S \geq \frac{(-c'')^2}{I(\theta)}.$$

Since the Fisher information for θ can also be written as $-c''$, we have

$$\mathrm{Var}S \geq \frac{I(\theta)^2}{I(\theta)} = I(\theta) = \mathrm{Var}T.$$

So, $T(\boldsymbol{y})$ is a minimum variance unbiased estimator of $-c'(\theta)$. A little more work will show that only an unbiased estimator of $-c'(\theta)$ can achieve the Cramér-Rao lower bound. For other functions of θ, the CR bound can't be used to show a minimum variance property.

